Lecture Notes in Electrical Engineering 425

Kuinam J. Kim
Nikolai Joukov
Editors

Mobile and Wireless Technologies 2017

ICMWT 2017

Springer

Lecture Notes in Electrical Engineering

Volume 425

About this Series

"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

LNEE publishes authored monographs and contributed volumes which present cutting edge research information as well as new perspectives on classical fields, while maintaining Springer's high standards of academic excellence. Also considered for publication are lecture materials, proceedings, and other related materials of exceptionally high quality and interest. The subject matter should be original and timely, reporting the latest research and developments in all areas of electrical 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.

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

Kuinam J. Kim · Nikolai Joukov
Editors

Mobile and Wireless Technologies 2017

ICMWT 2017

Springer

Editors
Kuinam J. Kim
Convegence Security Department
Kyonggi University
Seongnam-si, Kyonggi-do
Korea (Republic of)

Nikolai Joukov
modelizeIT Inc., CEO and NYU
Stony Brook, NY
USA

ISSN 1876-1100 ISSN 1876-1119 (electronic)
Lecture Notes in Electrical Engineering
ISBN 978-981-10-5280-4 ISBN 978-981-10-5281-1 (eBook)
DOI 10.1007/978-981-10-5281-1

Library of Congress Control Number: 2017943073

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is 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 Mobile and Wireless Technology, iCatse International Symposium on Software Networking, iCatse International Symposium on Electrical Engineering and ITAIWAN workshop, which were held in Kuala Lumpur, Malaysia, during June 26–29, 2017.

The conferences received over 200 paper submissions from various countries. After a rigorous peer-review process, 77 full-length articles were accepted for presentation at the conference. This corresponds to an acceptance rate was very low and is intended for maintaining the high standards of the conference proceedings.

The conferences provide an excellent forum for sharing knowledge and results in mobile, wireless, software networking and electrical engineering technology. The aim of the conferences is to provide a platform to the researchers and practitioners from both academia and industry to meet and share the cutting-edge developments 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 related fields.

On behalf of the Organizing Committee, we would like to thank Springer for publishing the proceedings of the conferences. 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 volume depends mainly on the expertise and dedication of its 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 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's conference.

Kuinam J. Kim
Nikolai Joukov

ICMWT

General Co-chairs

Dato' Ahmad Mujahid Ahmad Zaidi	National Defence University of Malaysia, Malaysia
Hyeun Cheol Kim	NamSeoul University, Republic of Korea
Norazman bin Mohamad Nor	National Defence University of Malaysia, Malaysia

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
Xiaoxia Huang	University of Science and Technology Beijing, China
Naruemon Wattanapongsakorn	King Mongkut's University of Technology Thonburi, Thailand

Publicity Chairs

Dmitri A. Gusev	Purdue University, USA
Chan Shiau Wei	Universiti Tun Hussein Onn Malaysia, Malaysia
Azizi bin Miskon	National Defence University of Malaysia, Malaysia
Naruemon Wattanapongsakorn	King Mongkut's University of Technology Thonburi, Thailand

Suresh Thanakodi	National Defence University of Malaysia, Malaysia
Hongseok Jeon	ETRI, Republic of Korea
Tomas Cerny	Czech Technical University, Czech Republic
Dan (Dong-Seong) Kim	University of Canterbury, New Zealand

Financial Chair

Hara Paul Kim	Institute of Creative Advanced Technologies, Science and Engineering, Republic of Korea

Publication Chair

Nakhoon Baek	Kyungpook National University, Republic of Korea

Program Chair

Kuinam J. Kim	Kyonggi University, Republic of Korea

Organizers and Supporters

Institute of Creative Advanced Technologies, Science and Engineering (iCatse)
Lecture Notes in Electrical Engineering, Springer
Korean Industry Security Forum (KISF)
Korea Information Assurance Society (KIAS)
National Defence University of Malaysia, Malaysia
River Publishers, the Netherlands
Korea Institute of Science and Technology Information (KISTI)
Electronics and Telecommunications Research Institute (ETRI)

ICSN

General Co-chairs

HyeunCheol Kim	NamSeoul University, Republic of Korea
Nikolai Joukov	New York University and modelizeIT Inc., USA
Dato' Ahmad Mujahid Ahmad Zaidi	National Defence University of Malaysia, Malaysia

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, Republic of Korea
Naruemon Wattanapongsakorn	King Mongkut's University of Technology, Thailand
Xiaoxia Huang	University of Science and Technology Beijing, China
Dato' Ahmad Mujahid Ahmad Zaidi	National Defence University of Malaysia, Malaysia
junkyun Choi	KAIST, Republic of Korea

Publicity Chairs

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

Minki Noh	KISTI, Republic of Korea
Hongseok Jeon	ETRI, Republic of Korea

Financial Chair

Kyoungho Choi	iCatse, Republic of Korea

Publication Chair

Kiwook Sohn	ETRI, Republic of Korea

Program Chairs

Kuinam J. Kim	Kyonggi University, Republic of Korea
Byung Yun Lee	ETRI, Republic of Korea

Organizers and Supporters

Institute of Creative Advanced Technologies, Science and Engineering (iCatse)
Chinese Management Science Society (CMSS)
Korean Industry Security Forum (KISF)
Korea Information Assurance Society (KIAS)
River Publishers, the Netherlands
KISTI
ETRI
Czech Technical University, Czech Republic
National Defence University of Malaysia, Malaysia

ISEE

General Co-chairs

Masrah Azrifah Azmi Murad	University Putra Malaysia, Malaysia
HyeunCheol Kim	NamSeoul 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
Xiaoxia Huang	University of Science and Technology Beijing, China
Naruemon Wattanapongsakorn	King Mongkut's University of Technology Thonburi, Thailand

Publicity Chairs

Nakhoon Baek	Kyungpook National University, Republic of Korea
Dan (Dong-Seong) Kim	University of Canterbury, New Zealand

Financial Chair

Hara Paul Kim	Institute of Creative Advanced Technologies, Science and Engineering, Republic of Korea

Publication Chairs

Donghwi Lee	University of Colorado, USA
Suresh Thanakodi	National Defence University of Malaysia, Malaysia

Program Chair

Kuinam J. Kim	Institute of Creative Advanced Technologies, Science and Engineering, Republic of Korea

Organizers and Supporters

The Korean Federation of Science and Technology Societies (KOFST)
Institute of Creative Advanced Technologies, Science and Engineering (iCatse)
iCatse Electrical Engineering Society
Korea Information Assurance Society (KIAS)
National Defence University of Malaysia, Malaysia
River Publishers, the Netherlands
KISTI, ETRI

ITaiwan

Organizing Committee

Workshop Chair

Lin-huang Chang	National Taichung University, Taiwan

Workshop Co-chair

Jiun-Jian Liaw	ChaoYang University, Taiwan

Program Committee

Lin-huang Chang	National Taichung University, Taiwan
J.C. Chen	National Chiao Tung University, Taiwan
Sheng-Tzong Cheng	National Cheng Kung University, Taiwan
Hung-Chi Chu	ChaoYang University, Taiwan
Jenq-Neng Hwang	University of Washington, USA
T.H. Lee	National Taichung University, Taiwan
Victor C.M. Leung	University of British Columbia, Canada
Jiun-Jian Liaw	ChaoYang University, Taiwan
Ai-Chun Pang	National Taiwan University, Taiwan
Frode Eika Sandnes	Oslo University College, Norway
S.C. Wang	ChaoYang University, Taiwan
K.M. Yap	Sunway University, Malaysia
Sue-Chen Hsueh	ChaoYang University, Taiwan
Chin-Feng Lee	ChaoYang University, Taiwan
Yu-lung Lo	ChaoYang University, Taiwan
Chuan-Pin Lu	Meiho University, Taiwan
Chu-Hui Lee	ChaoYang University, Taiwan

Contents

Mobile and Wireless Technology

Software Networking

Electrical Engineering

Technology and Applications of IoT with Wireless Advanced Networking

Mobile and Wireless Technology

A Lightweight Mutual Authentication Protocol for the IoT

Mohamed Tahar Hammi[1(✉)], Erwan Livolant[2], Patrick Bellot[1],
Ahmed Serhrouchni[1], and Pascale Minet[3]

[1] LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, 75013 Paris, France
{mohamed-tahar.hammi,bellot,ahmed.serhrouchni}@telecom-paristech.fr
[2] AFNet, 30 Rue de Miromesnil, 75589 Paris, France
erwan.livolant@afnet.fr
[3] Inria-Paris, EVA Team, 2 Rue Simone IFF, 75589 Paris Cedex 12, France
pascale.minet@inria.fr

Abstract. The Internet of Things enables the interconnection of smart physical and virtual objects, managed by highly developed technologies. WSN, is an essential part of this paradigm. The WSN uses smart, autonomous and usually limited capacity devices in order to sense and monitor industrial environments. However, if no authentication mechanism is deployed, this system can be accessible, used and controlled by non-authorized users. In this paper, we propose a robust WSN mutual authentication protocol. A real implementation of the protocol was realized on *OCARI*, one of the most interesting Wireless Sensor Network technologies. All nodes wanting to access the network should be authenticated at the *MAC* sub-layer of *OCARI*. This protocol is especially designed to be implemented on devices with low storage and computing capacities.

Keywords: Security · Mutual authentication · WSN · IoT · OCARI · MAC Sub-layer · OTP · Industrial environment

1 Introduction

According to [11], more than 50 billions of devices will be connected in 2020. This huge infrastructure of devices, which is managed by highly developed technologies, is called *Internet of Things (IoT)*. The latter provides advanced services, and brings economical and societal benefits. This is the reason why thousands of workers and researchers of both industry and scientific community are interested in this area.

Wireless Sensor Network (WSN), is a part of the *IoT* domain. A *WSN* is a network composed of clusters of devices that are equipped with (1) sensors to gather data about the environmental conditions, and/or (2) actuators to interact with the real world. Each cluster is managed by a specific device called *Personal Area Network Coordinator (CPAN)*. Devices are generally characterized by the use of a small computation and memory capacity, low bit rate, low

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_1

power consumption, and small packet size. The data produced by each device is transmitted via multiple hops to the *CPAN*, which can use them, or forward them to another network. Usually *WSN* technologies are based on IEEE 802.15.4 physical layer (*PHY*) [9] that provides a good foundation for building ad-hoc mesh networks.

Optimization of Communication for Ad hoc Reliable Industrial networks (OCARI) [5] is a promising WSN. It is characterized by its optimized energy consumption, its time-constrained communication at the *MAC* sub-layer, and its support of pedestrian mobility [1]. However, it needs to be secured against the different threats, especially those that concern confidentiality, data integrity, and entities authentication.

In order to secure *OCARI* specification, our work aims to create a robust security protocol, which ensures a mutual authentication between the device and the *CPAN* at the *MAC* sub-layer, to protect the system against malicious intruders. The proposed protocol also provides a secure algorithm for the exchange of symmetric keys (used to ensure the data integrity). A real implementation with C language, deployed on the OCARI platform, were realized in this work. In this paper, the authentication and data integrity services are our primary concern. We do not consider the confidentiality service.

The rest of this paper is organized as follows. Section 2 presents the related *WSN* and *IoT* technologies and their authentication mechanisms. Section 3 describes our proposed approach and its implementation. Section 4 details a real tests that we have made. Then we provide an evaluation of our authentication protocol. Finally, our conclusions and future work are drawn in Sect. 5.

2 Related Work

IoT and networks in general, represent the working environment of hackers. Everyday, companies and individuals are victims of different kinds of attacks: Denial/Distributed Denial of Service (Dos/DDos) attacks, usurpation of identity, intrusions, data theft, etc. In return, several researches have been realized in order to secure and protect the Information Technology (IT) systems.

In [8], we proposed an authentication protocol based on pre-shared keys. It provides only the authentication of a device during its association to a cluster. Although this solution is lightweight and fast, the authentication of the CPAN is missing in this work. In addition it does not use a good mechanism for the generated keys exchange (this keys are required for the data authentication once the association is realized).

An interesting work described in [3] intends to secure the IoT devices. They propose an authentication mechanism based on shared key between constrained (Cd) and unconstrained (Ud) devices. Both sides use the same security policy, and no gateway is required. For an easier understanding, authors give an example using an IPsec-based security association (see Fig. 1). First, the concerned entities agree on the security policy. Then, exchange the keying material. After, authenticate each other. And finally, create a secure channel. This mechanism is

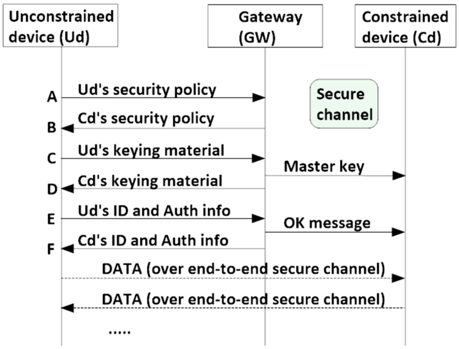


Fig. 1. Association and secure channel establishment solution

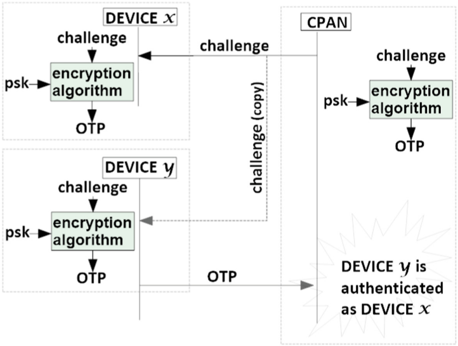


Fig. 2. An asynchronous authentication operation

based on IPsec, which is known by its robustness. However, it has some weaknesses: (1) In C (Fig. 1), the GW, receives the Ud's keying material, and without authenticating Ud, generates and sends a master key to the constrained device. In this case if Ud is malicious, it can send a big number of C messages to the gateway, which, therefore, sends master keys to the constrained device. Knowing that the reception of messages consumes a lot of energy, as a result, this increases the energy consumption of the constrained device. (2) In D, without authenticating the gateway, Ud generates it's own master key. Thus, if a malicious GW generates a Dos/DDos attack (by sending a lot of data about the keying material, or a lot of D messages), this can stop the operation of Ud.

Authors in [4], propose an authentication mechanism for the IEEE 802.1x technologies, based on Extensible Authentication Protocol (EAP). In order to ensure a secure communication between two entities, first, they exchange their identities (without any proof). Then, an authentication server (Remote Authentication Dial-In User Service server) is involved to check the authenticity of the entities, using algorithms and mechanisms as MD5 or TLS protocol. This solution is flexible, based on standardized algorithms, and can be deployed on different systems. However it can not be deployed without a trusted third party (authentication server). Furthermore, it requires the exchange of a very big number of messages (10 messages). Thus, the execution time and the energy consumption of this approach is very high.

One can note that works seen above does not sufficiently address the problem of the energy and time constraints, and can have some weaknesses. In this paper, we propose an approach that provides an energy efficient and optimal mutual authentication method, as well as a key exchange algorithm, designed for WSN systems.

3 Proposed Approach

We implemented our authentication mechanism in the MAC sub-layer, in order to provide more transparency and interoperability in the highest layers (network and application).

3.1 Algorithms

To ensure a mutual authentication in the MAC sub-layer association step, we designed a method based on an **asynchronous** One Time Password (OTP) using a Challenge/Response mechanism. We opted for the asynchronous mode, because generally WSNs does not support an *absolute time* of sufficient precision, which can be exploited in the synchronous mode. In addition, this mode compared to the **synchronous** one (see RFC4226 [10]), does not require any prior approval between the communicating entities, which represents a great advantage, and this allows more flexibility for the system. The OTP by definition is a password, valid for only one transaction, used for proving the identity of an entity. In other words, even if it is transmitted without any encryption, a malicious user cannot exploit it to authenticate itself. Figure 2 shows a possible architecture using an asynchronous authentication operation in the association step.

All the devices of the same cluster have the same secret pre-shared key *psk*. If a device x wants to join the cluster, it computes an OTP using a received random number (*challenge*) and the *psk*), then sends it to the CPAN. However, any device y which can catch the transmitted *challenge* can generate the OTP (because it has the same *psk*), and use it to authenticate itself on behalf of device x. In other words, y steals the identity of x and gets all the authentication/encryption keys, that normally should be secret and shared only between x and the CPAN. Therefore, the data integrity and confidentiality are no longer ensured.

3.2 Preparation of Nodes

Our protocol allows a mutual strong authentication, and solves the problem of the internal identity usurpation due to the "personalization" of the keys. This operation is described in the Fig. 3. The trusted authority, which is generally the provider, should setup in out-of-band channel, what we call "key_{mother}" into the CPAN, and derives from it a personalized "$key_{daughter}$" attributed to each legitimate device belonging to the same cluster. The derivation of this keys is based on the function $f(key_{mother},\, UI)$, where UI (Unique Identifier) is the 8 bytes device's IEEE address. This function is defined as below:

$$\begin{cases} key_{daughter} = f(key_{mother},\, UI) \\ f(key_{mother},\, UI) = hash\_func(key_{mother},\, UI) \end{cases} \tag{1}$$

Where: $hash\_func$ is an irreversible function that generates a strong key, and which protects the key_{mother} against deductive attacks. Once the $key_{daughter}$ is created and set into the device, the latter becomes able to be associated to the cluster.

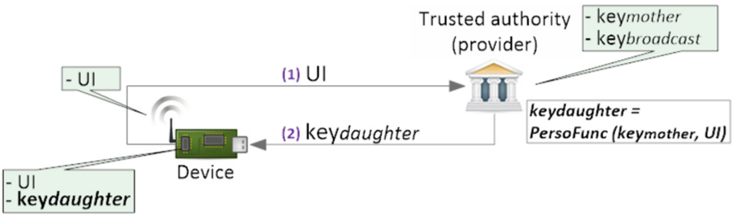


Fig. 3. The personalization of keys

3.3 Association Step

Figure 4 illustrates our mutual authentication protocol in the MAC sub-layer association process. During the association step, the personalized $key_{daughter}$ of a device D can be generated only by the CPAN that has the appropriate key_{mother}. Thus this $key_{daughter}$ is known only by D and by the CPAN, which ensures the protection of the $key_{daughter}$. To start the association operation, the device sends an "association request" message to the CPAN, if needed the request passes through one or several device(s), called relay(s), to reach the intended destination. Receiving the request, the CPAN checks if the UI of the device is blacklisted or not. (1) If it is the case, the association request is directly rejected and no processing is done. This method prevents the system from reserving uselessly the memory and doing additional processing, thus this protects the CPAN from some kind of DoS attacks. (2) Otherwise, it answers by an "authentication request" containing a *challenge*. With the received *challenge*, its own $key_{daughter}$ and using an *encryption algorithm*, the device computes "$otp1$" and sends it through an "authentication response" message. For the *encryption algorithm* we opted for the *HMAC-Based One-Time Password Algorithm* described in the RFC4226 [10] which is proposed for the **synchronous** OTP mode, and which basically uses an increasing counter value. After modifications, we compute $otp1$ using the function below (Eq. 2):

$$\begin{cases} HOTP(key,\ challenge) = \\ Truncate(HMAC - SHA256(key,\ challenge)) \end{cases} \tag{2}$$

Receiving the device's response, the CPAN generates the appropriate $key_{daughter}$ using the same personalization function (described above in (Eq. 1), and computes "$otp1\prime$". Then compares the two otps, if they match then the device is authenticated, otherwise the association operation fails. If the same device is rejected consecutively MAX_ASSOC_REQ times, then it is blacklisted. The fact that there is only the legitimate device that can have the $key_{daughter}$ which is used to compute $otp1$, represents a proof of for its identity. Once the latter is authenticated, the CPAN generates an authentication key called "key_{auth}" using the standard Pseudo Random Function (PRF) defined in the SSL/TLS specification [6]. This key will be used to securely exchange the broadcast key "$key_{broadcast}$", and then to authenticate the exchanged data after the association

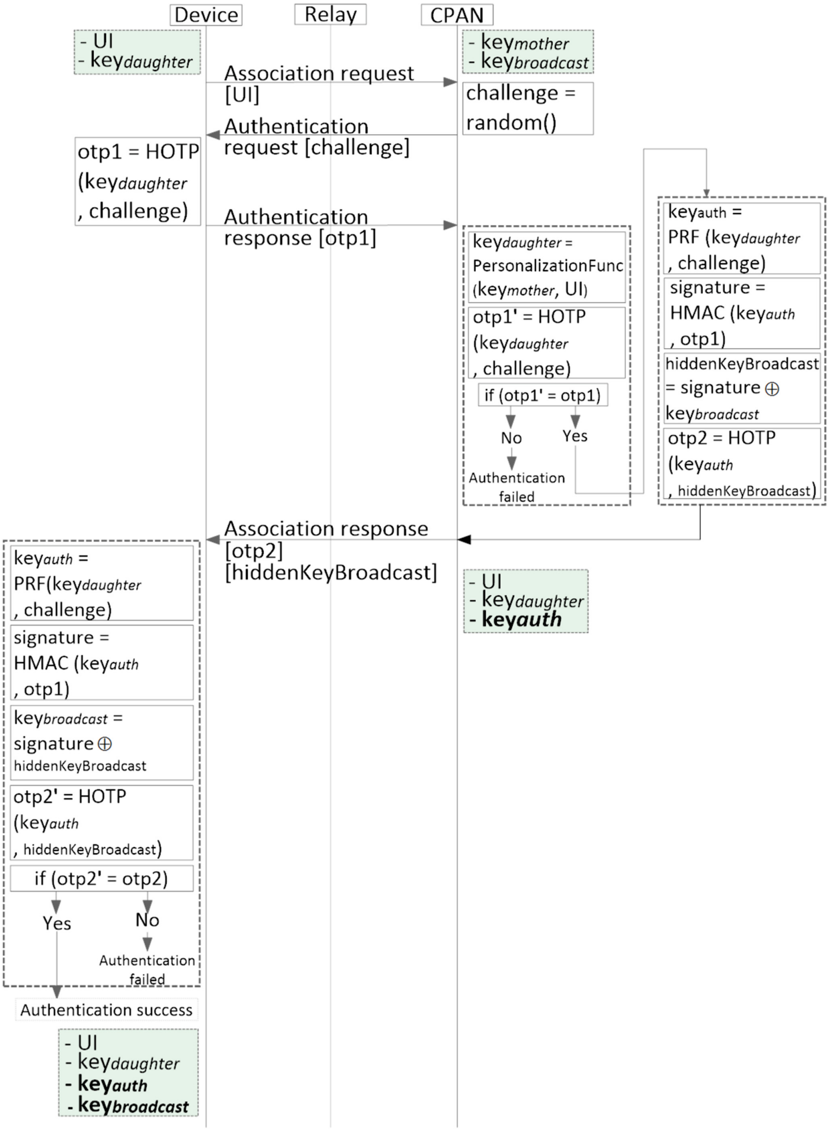

Device
Relay
CPAN
- UI
- keydaughter
- keymother
- keybroadcast
Association request [UI]
challenge = random()
Authentication request [challenge]
otp1 = HOTP (keydaughter , challenge)
Authentication response [otp1]
keydaughter = PersonalizationFunc (keymother, UI)
otp1' = HOTP (keydaughter , challenge)
if (otp1' = otp1)
No
Yes
Authentication failed
keyauth = PRF (keydaughter , challenge)
signature = HMAC (keyauth , otp1)
hiddenKeyBroadcast = signature ⊕ keybroadcast
otp2 = HOTP (keyauth , hiddenKeyBroadcast)
Association response [otp2] [hiddenKeyBroadcast]
- UI
- keydaughter
- keyauth
keyauth = PRF(keydaughter , challenge)
signature = HMAC (keyauth , otp1)
keybroadcast = signature ⊕ hiddenKeyBroadcast
otp2' = HOTP (keyauth , hiddenKeyBroadcast)
if (otp2' = otp2)
Yes
No
Authentication failed
Authentication success
- UI
- keydaughter
- keyauth
- keybroadcast

Fig. 4. Our proposed mutual authentication protocol

step in the unicast mode. The $key_{broadcast}$ is used to ensure the integrity of the broadcasted messages (after the association step). To share this key with the authenticated device, the CPAN should hide it into a "*hiddenKeyBroadcast*" value using the function below (function 3):

$$\begin{cases} hiddenKeyBroadcast = signature \oplus key_{broadcast} \\ where : signature = HMAC - SHA256(key_{auth}, otp1) \end{cases} \tag{3}$$

We created this function in order to securely share the $key_{broadcast}$ (Fig. 5 shows the visible information for the different kind of users). (1) If an external attacker intercepts all the exchanged information ($challenge$, $otp1$, $hiddenKeyBroadcast$, and $otp2$ (explained below)), it can not get any secret information (keys), because it does not have the couple ($key_{daughter}$, $key_{broadcast}$) nor (key_{auth}, $key_{broadcast}$). (2) For an internal attacker which has the $key_{broadcast}$ in addition to all the exchanged information, it cannot also get the keys of other devices. That is to say, when an internal attacker attempts to get the key_{auth} of another device, it computes the xor ($\oplus$) between the $key_{broadcast}$ and the $hiddenKeyBroadcast$ in order to obtain the $signature$, and because the latter is generated by an irreversible function (HMAC), even using $otp1$, the attacker cannot get the key_{auth}. $otp2$ is computed by the CPAN for hitting two targets with one shot. Firstly to ensure the integrity of the $hiddenKeyBroadcast$, and secondly to authenticate itself. To be generated, $otp2$ needs a secret (key), and a unique challenge. For this reason the CPAN uses key_{auth} as a secret, and exploits the $hiddenKeyBroadcast$ as challenge. The latter is unique, because it is based on a unique signature, that is based on unique otp ($otp1$). Then $otp2$ is sent accompanied by the $hiddenKeyBroadcast$ through an "association response" message. Finally, when the device receives the message, it computes also key_{auth} and $signature$ using the same inputs applied by the CPAN. Then to retrieve the $key_{broadcast}$ it computes the xor between $signature$ and $hiddenKeyBroadcast$ (see following function 4):

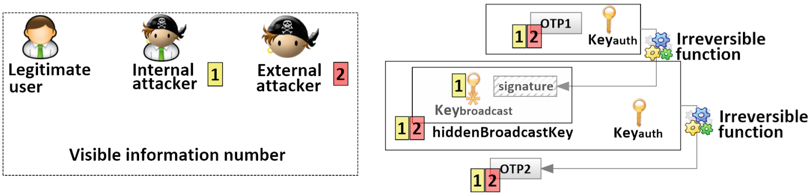


Fig. 5. HiddenBroadCastKey mechanism

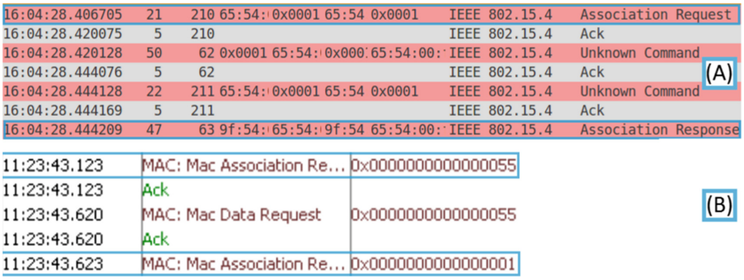
$$key_{broadcast} = signature \oplus hiddenKeyBroadcast \tag{4}$$

The device gets a $key_{broadcast}$ which needs to be verified (check for its integrity). That is why it computes $otp2\prime$ based on the received $hiddenKeyBroadcast$ and key_{auth}, then the device compares the two otps, if they match, then this means that the $hiddenKeyBroadcast$ is correct, thus the $key_{broadcast}$ is correct and the CPAN is authenticated. Otherwise if the retrieved $otp2$ or the $hiddenKeyBroadcast$, or both of them are wrong or modified during their transmission, then $otp2$ and $otp2\prime$ will not match, hence the $key_{broadcast}$ is not accepted, the CPAN is not be authenticated, and the association operation stops. The fact that the device receives from the CPAN a correct $otp2$, validates the identity of the CPAN. Because $otp2$ is computed by key_{auth} which is derived from $key_{daughter}$. Accordingly, the protocol ensures a mutual authentication and a secure exchange of keys. Once the DEVICE is associated to the network, it will use the obtained keys (key_{auth} and $key_{broadcast}$) for the authentication of the exchanged messages under an authenticated channel. In fact, exchanged packets in the authenticated channel should be signed. Which means that each packet, which is made of a header and a body, should add a signature in order to ensure its integrity. The signature is an HMAC of 16 bytes of the header and the body computed using the key_{auth} (unicast mode) or the $key_{broadcast}$ (broadcast mode).

4 Test and Evaluation

For testing our authentication mechanism, we implemented our protocol in the OCARI stack software [1]. OCARI technology is based on the IEEE 802.15.4 physical layer, which is adapted to harsh environment such as power plants and factories. This layer ensures a good signal transmission, that is resilient to radio interferences. Unlike the IEEE 802.15.4 physical layer, the IEEE.802.15.4 MAC layer was replaced by "MaCARI", that is designed taking into consideration two different factors that are "determinism" and "energy optimization". As explained in [1], a deterministic MAC layer should guarantee an access to the medium for each node, every certain period of time. While an energy-efficient MAC layer has to make all the nodes sleep as much as possible. MaCARI uses different access methods to the medium. It uses CSMA/CA for control messages, combined with TDMA for data messages.

Our code is developed with C language and the hardware used is Dresden Elektronik deRFsam3 23M10-R3. It has 48 kb of RAM, 256 kb of ROM and a Cortex M-3 Processor. There is no specific hardware for the authentication mechanism. For our experiment, we created an architecture composed of 2 devices and one CPAN. First, we personalized the devices by flashing into them their associated keys ($key_{daughter}$) as explained in Sect. 3.2. Using a Zolertia z1 sniffer (hardware) and Wireshark [7], we can follow the association and the authenticated channel establishment operations. Figure 6 shows the exchanged frames during the association step. The two "Unknown Command" represents respectively: the "Authentication request message" and the "Authentication response message". For OCARI, at the beginning, devices at 1 hop from the CPAN sends directly an association request to it. Then device at 2 hops sends a request to the CPAN by means of the authenticated ones, that play the role of router. Then we measured the delays of the association operation with (auth) and without (none) the authentication procedure, and compared this results with studies and evaluations of a 1 hop device analyze, realized on an implementation of the Zigbee protocol (WSN technology) [2], using a similar hardware equipment. As shown in Fig. 7, the average delay of the association operation is increased from 0,5243 ms without authentication to 34,45 ms with authentication for the DEVICE at 1 hop from the CPAN. The association time needed by the DEVICE at 2 hops from the CPAN is also increased from 34,5076 ms without authentication to 45,876ms with authentication. This increase is mainly due to the exchange of additional messages of the authentication protocol. Indeed, the number of messages needed by the 2 hop DEVICE for its association without authentication is 4 which is the same as the number of messages for the 1 hop DEVICE



```
16:04:28.406705  21   210 65:54:  0x0001 65:54 0x0001    IEEE 802.15.4   Association Request
16:04:28.420075   5   210                                IEEE 802.15.4   Ack
16:04:28.420128  50    62 0x0001 65:54:  0x000: 65:54:00:  IEEE 802.15.4   Unknown Command
16:04:28.444076   5    62                                IEEE 802.15.4   Ack
16:04:28.444128  22   211 65:54:  0x0001 65:54 0x0001    IEEE 802.15.4   Unknown Command
16:04:28.444169   5   211                                IEEE 802.15.4   Ack
16:04:28.444209  47    63 9f:54:  65:54:  9f:54 65:54:00:  IEEE 802.15.4   Association Response
```

```
11:23:43.123   MAC: Mac Association Re...   0x0000000000000055
11:23:43.123   Ack
11:23:43.620   MAC: Mac Data Request        0x0000000000000055
11:23:43.620   Ack
11:23:43.623   MAC: Mac Association Re...   0x0000000000000001
```

Fig. 6. A Wireshark capture of (A) OCARI and (B) Zigbee frames during a node association step

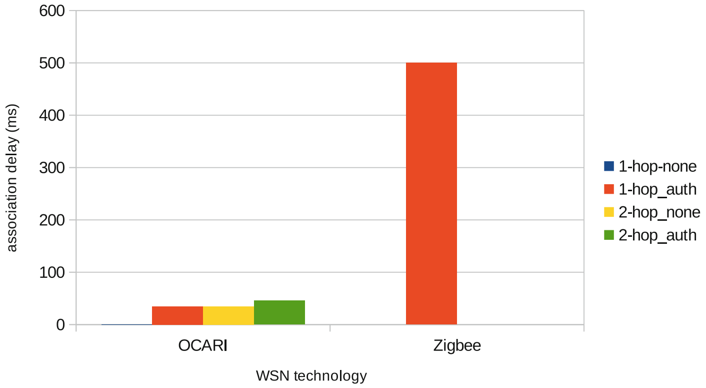


Fig. 7. Association delays comparison with and without authentication

association with authentication. The difference between both is only 0,0576ms (34,5076 - 34,45). Hence, the increase in time caused by the authentication is due to these additional messages and in a very lesser extent to an increase in processing time.

In a similar way as in OCARI, the execution time of the secured Zigbee association can be computed by the subtraction between the Association response message and the Association request message timestamps. The delay of 1 hop with authentication association is equal to 500 ms (11h:23m:43.623s - 11h:23m:43.123s). It is true that this big difference between association delays of OCARI and Zigbee is due also to the nature of the two technologies, but still, the used security protocol plays a main role.

With the real implementation, we proved that our solution is robust, ensures a good mutual authentication and a secure key exchange mechanism.

5 Conclusion and Future Work

Our approach is based on a lightweight, robust, and energy efficient mechanism that allows to solve the problem of the WSN mutual authentication at the MAC sub-layer. This solution provides a protection against “replay attacks”, because the exchanged OTPs are based on random numbers, therefore, they are valid only for one transaction. Using the blacklisting mechanism we can secure our systems against “some DoS” attacks. Finally it is flexible and does not decrease the scalability of the system, and can be deployed in different WSNs technologies, while keeping the same level of robustness.

In our future work we aim to ensure the confidentiality of the transmitted messages exchanged after the MAC sub-layer association and authentication procedure. And thus we will have a secure system which ensures the “Confidentiality”, “Integrity, and “Authentication” services.

References

1. Agha KA, Bertin MH, Dang T, Guitton A, Minet P, Val T, Viollet JB (2009) Which wireless technology for industrial wireless sensor networks? The development of OCARI technology. IEEE Trans Ind Electron 56(10):4266–4278
2. Atmel (2013): Zigbee pro pack et analysis with sniffer, September 2013. http://www.atmel.com/Images/Atmel-32210-ZigBee-PRO-Packet-Analysis-with-Sniffer_AP-Note_AT02597.pdf
3. Bonetto R, Bui N, Lakkundi V, Olivereau A, Serbanati A, Rossi M (2012) Secure communication for smart IoT objects: Protocol stacks, use cases and practical examples. In: 2012 IEEE international symposium on a world of wireless, mobile and multimedia networks (WoWMoM), pp 1–7, June 2012
4. Chen JC, Wang YP (2005) Extensible authentication protocol (EAP) and IEEE 8021.x: tutorial and empirical experience. IEEE Commun Mag 43(12):supl.26–supl.32
5. Dang T, Devic C (2008) OCARI: Optimization of communication for Ad hoc reliable industrial networks. In: 6th IEEE international conference on industrial informatics, INDIN 2008. IEEE, pp 688–693
6. Dierks T (2008) The transport layer security (TLS) protocol version 1.2
7. Foundation TW (2016) Wireshark 2.0.3 and 1.12.11 Released (22 April 2016). https://www.wireshark.org/news/20160422.html. Accessed 28 June 2016
8. Hammi MT, Livolant E, Bellot P, Serhrouchni A, Minet P (2016) MAC sublayer node authentication in OCARI. In: 2016 international conference on performance evaluation and modeling in wired and wireless networks (PEMWN), pp 1–6, November 2016
9. IEEE (2011): IEEE Standard for Local and metropolitan area networks-Part 15.4: Low-Rate Wireless Personal Area Networks (LR-WPANs). IEEE Std 802.15.4-2011 (Revision of IEEE Std 802.15.4-2006), September 2011
10. M’Raihi D, Bellare M, Hoornaert F, Naccache D, Ranen O (2005) HOTP: An HMAC-based one-time password algorithm. IETF, RFC 4226, December 2005
11. online statistics portal, and one of the world’s most successful statistics databases (2016). https://www.statista.com

Improving Consumers' Value and Satisfaction Towards Recreational Facilities Through Apps

Norfaradilla Wahid(✉), Fathin Najwa Masnan, Hanayanti Hafit, and Shahreen Kasim

Department of Web Technology, Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Malaysia
{faradila,hanayanti,shahreen}@uthm.edu.my, fathinnajwamasnan@gmail.com

Abstract. Consumers' value and satisfaction is an important issue that must be address by any service provider as well as product makers. As society rapidly developed, the demand to fulfill people's satisfaction is increasing in all life-centered activities. For example, people are putting demands for quality outdoor activities within their limited quantity of time which constrained by their work requisition, etc. In this paper, we show that consumer's satisfaction can be improved by up to 40% as compared to the result from the pre-survey by the help of technological support, i.e., mobile apps. We support the study by introducing ZoomApp, i.e., a navigational apps for Malacca Zoo to see its effectiveness for the visitors in accessing the facilities as well as every important points in the zoo.

Keywords: Customer · Satisfaction · Apps · ZoomApp

1 Introduction

As society developed and the quality of living improved, consumers lifestyles began to change, and the need for quality outdoor activities activities began to increase. Accordingly, tourism industry is facing challenges to provide quality assurance to the society, in a way that all the facilities are accessible to the users. This is in line with the drastic changes occurred in the work demand, life and holidays especially in Malaysia. In overall, consumers had limited time for recreational activities over the quantity and quality of recreational choices offered in their surroundings. Therefore, the urge to fulfill the value and satisfaction of consumers in each visit to any recreational spot is highly demanded.

According to statistics from the tourism authority [2], over 80% of the survey indicated higher interest in domestic holidays, especially entertainment type activities, and interest in shorter stays (i.e. two days instead of three days). More people began to choose short-term, fixed-point, and recreational holidays over long overseas excursions. In view of this trend, parks and zoos are among the attractions to the local consumers as it normally can fit both the adults as well as the kids. Surveys have also been done to evaluate the value and satisfaction of

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_2

the consumers in recreational parks. One of the survey [2] indicates that accessibility to facilities are among the important point to satisfy the visitors as they would not want any hassle and difficulties through out the visit.

Based on a study that had been done in [4], the visitors categorized zoo as a recreation area, along with its suitability and its facilities for visitors. In Malaysia, a study is carried out in [1] in Ayer Keroh Forest Recreation Area to determine and evaluate the visitors satisfaction towards the facilities and services provided. It shows that four predictors involves in recreation valuation, i.e., amount spend for the facilities and services, perception on sign facilities, participate in recreation activities and income.

In this era of technology, equipping consumers with mobile apps gives advantages to them especially in getting fast and accurate information to get better access to a specific service. [5–7] are examples of available apps which serve different purposes to the society.

In this paper, we show that consumers' value and satisfaction can be improved by gearing apps to support the visibility of facilities and every important points inside the visiting place to the consumers. We support the study by introducing an apps named ZoomApp, i.e., an apps for Malacca Zoo navigation.

The rest of the paper is organised as follows: Sect. 2 will discuss the related work including equivalent apps available in the market and its comparisons with the proposed apps will be discussed. Section 3 discusses the analysis and design of the apps as well as its implementations. In Sect. 4, we will show the result of the testing phase. Lastly, Sect. 5 will conclude the paper.

2 Related Work

Traditionally, zoo visitors will be given a printed version of zoo map upon buying entrance tickets. Then, they can explore the place and do sightseeing. However, any zoo in the world may come in different sizes, i.e., large and small size of zoo. Larger size of zoo normally contain bigger number of animals as well as more complicated pathways. Consequently, the facilities are being distributed within the zoo area with farther distance and may involve tricky confusing route.

Based on the studies that have been done, three equivalent navigational applications are chosen as the benchmark for the proposed application, i.e., the ZoomApp. All the applications will be analysed and compared all in different aspects in order to determined their strength and weaknesses. The apps are Wroclaw Zoo App, Southwick's Zoo Map and Ragunan Zoo App.

1. *Zoo Wroclaw Map.* Zoo Wroclaw Map [1] is an android based location detector application for Wroclaw Zoo. It provides map guides in for the zoo visitors. Using the application, users not only able to view maps of the zoo, but users can view the listing of all events available, view all the paths inside the zoo and the application comes with notification to the users.
2. *Southwicks Zoo Map.* Southwick's Zoo Map [2] is an android based location detector application for Southwick's Zoo in United Kingdom. It offers a lots of

facilities to the zoo visitors. The application is built with a few modules, i.e., map module, zoo information module, event schedule module and emergency module.

3. *Raguan Zoo Apps.* Raguan Zoo Apps [3] too is built for android platform. Similarly, the apps comes with a few interesting modules including map module, zoo information module, animal collection and trivia module as well as it can gives notification whenever appropriate. Users can also jump into its social networking module if interested. The apps shows a cheerful and colorful graphics for users suitable with the animal garden theme.

Table 1 show the similarity and differences of the apps. In the table, we show the comparison of the features in terms of their design and the modules offered by the apps. Based on the analysis of the existing apps, we would like to offer similar features to serve our proposed application based on user requirements.

Table 1. Comparisons between equivalent systems with the proposed apps.

Specification	Wroclaw Zoo App [5]	Southwick Zoo Map [6]	Ragunan Zoo App [7]	ZoomApp
Platform	Android	Android	Android	Mobile
User scope	User	User	User	User and administration
Secure login	No	No	No	Yes
Maps searching	No	No	Yes	Yes
Live traffic	Yes	No	No	Yes
Notification	Yes	No	Yes	Yes

3 ZoomApp: A Navigational Apps for Malacca Zoo

The proposed application is a mobile based application which is developed for the navigation of places in Malacca Zoo. The purpose of developing the apps is to help visitors navigating the zoo. The application offers the zoo map with the help of Global Positioning System (GPS). The apps is expected to reduce the use of paper based map which often been used for visitors where it sometimes cause difficulties to read it.

The proposed application can be accessed by normal users, i.e., zoo visitors as well as the apps administrator. For full access to the apps, it needs to be connected to the Internet and the GPS. However, the navigation can be conducted within the zoo range only. This is done by restricting the apps navigational system where users have to log in to the application if it is within the specified longitude and latitude and auto-reconnect after a specific hours. Using the apps, users can do searching on the animals and all the facilities around the zoo and also can view information of the animals as well as to set notification on a specific event.

Additionally, administrator of the application will be able to update all the data of the application whenever necessary. This is including animal information, facilities information as well as any events occurred in Malacca Zoo. All these information then will be stored into database.

3.1 Analysis and Design

In order to occupy the flow of the project, we conduct the project following the AADIE development model [3]. Based on the model, we do the development following the five phases, i.e., analysis, design, development, evaluation, and implementation. The model allows these phases to be repeated in order to produce good output as it focus to achieve all objectives in each phase of the development process.

Analysis and design is one of important phases in the development of ZoomApp. It is to ensure that each features offered in the apps will occupy all the requirements and issues raised by users. In analysis phase, all the defined problems will become benchmark in developing a good quality system or application. Research about equivalent application is also done in the analysis phase. The analysis is important to support system maintenance in the future. Accordingly, we comes with the application interface design as a guideline for the development phase.

In Fig. 1, we shows the data flow of the ZoomApp application where it involves to users, i.e., the administrator and zoo visitors.

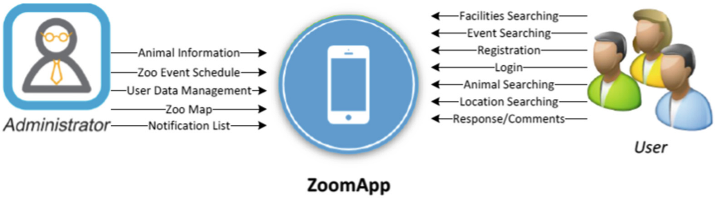
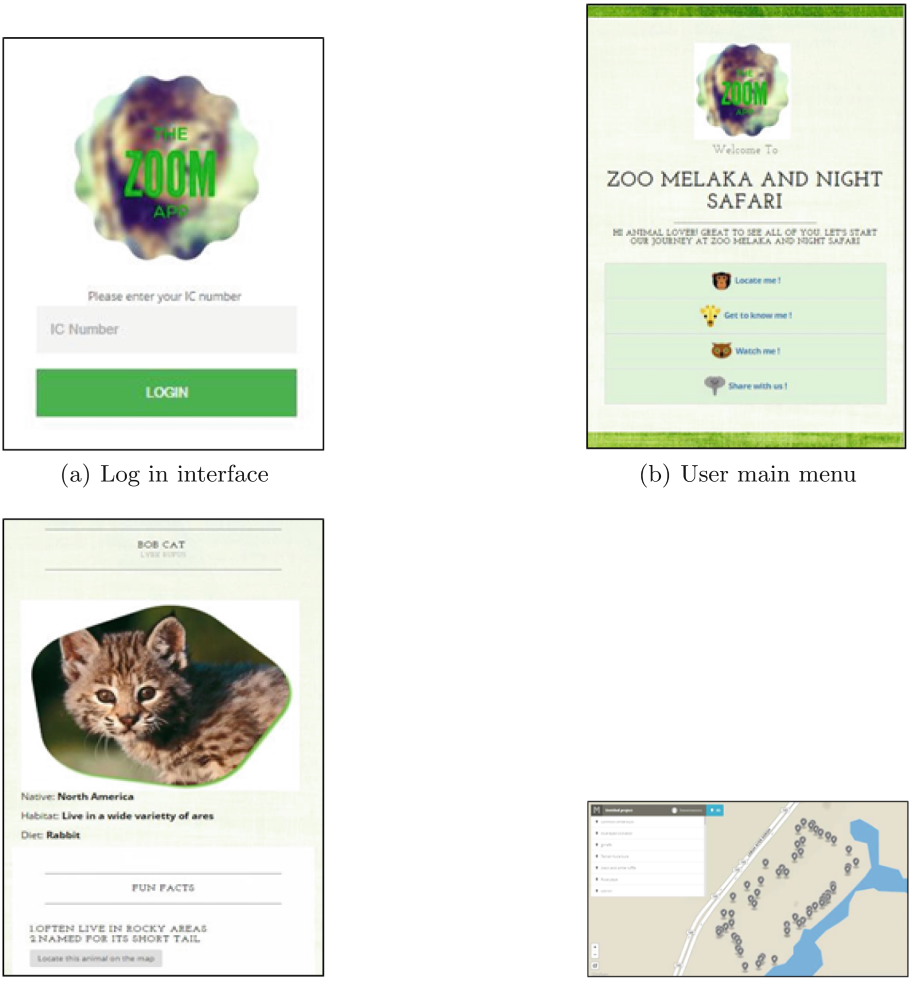


Fig. 1. ZoomApp Overview

3.2 Implementation

Application design is very important in supporting the development phase of application development life cycle. The designed interface needs to be simple, user friendly and easy to be used. Based on the design, we then come out with the real apps during the development phase. Figure 2(a)–(d) shows a few important interfaces for the application.

Figure 2(a) shows the ZoomApp login interface. Users need to key in their identification number in order to log on into the application. Identification number is used upon login as each entrance ticket to the zoo will require identification number to determine the fee rate of the ticket based on the visitors nationality. Therefore, identification number will be registered by the administrator during the ticket purchasing.


THE
ZOOM
APP
Please enter your IC number
IC Number
LOGIN
THE
ZOOM
APP
Welcome To
ZOO MELAKA AND NIGHT SAFARI
HI ANIMAL LOVER! GREAT TO SEE ALL OF YOU. LET'S START OUR JOURNEY AT ZOO MELAKA AND NIGHT SAFARI
Locate me !
Get to know me !
Watch me !
Share with us !
BOB CAT
LYNX RUFUS
Native: North America
Habitat: Live in a wide varietty of ares
Diet: Rabbit
FUN FACTS
1.OFTEN LIVE IN ROCKY AREAS
2.NAMED FOR ITS SHORT TAIL
Locate this animal on the map

(a) Log in interface

(b) User main menu

(c) Get to know me! interface

(d) Locate me! interface

Fig. 2. ZoomApp interfaces

Besides that, Fig. 2(b) shows the main menu of the apps. There are four options on the menu, i.e., Locate me!, Get to know me!, Watch me! and Share with us!. All these options are to identify location of animals and facilities around the zoo, to get further information about animals, to see the available events in the zoo and to give comments about the apps or about the zoo.

Figure 2(c) shows the Get to know me! menu option. In the interface, a collection of all animals in the zoo will be shown. Once user pick the animal, an interface will be shown where brief information about the animal will be displayed. On top of that, it also has a Locate animal button which will display the location of animal displayed on a map. Figure 2(d) is Locate me! menu page.

In this interface, an interactive Malacca Zoo is shown shown. Based on the map, user will be able to navigate all locations in the zoo in much easier way (Fig. 3).

```
<div class="embed-responsive embed-responsive-16by9">
        <iframe class="embed-responsive-item" width='100%'
         height='500px' frameBorder='0'
         src='https://a.tiles.mapbox.com/v4/
         fawamasnan.04jopd4b/attribution,zoompan,zoomwheel,
         geocoder,share.html?access_token=
         pk.eyJ1IjoiZmF3YW1hc25lI6eWNhanZ1In0.
         WiL8exyaaiuXnmqj4ZarLg' height="800" width="1000"
         frameborder="0">
        </iframe>
</div>
```

Fig. 3. Sectional code for applying map with GPS

4 Result

Testing is an important part of any software development. In this study, the testing phase is done in two parts, i.e., (i) the usability of the apps, (ii) the significant of the apps towards visitors satisfaction. The testing was conducted among 50 visitors in Malacca Zoo aged between 7 years old to 55 years old.

The developed apps consist of five main functions, i.e., login, maps view, animal show and information, notifications and comments. Testing on all these functions in the apps shows that more than 60% of the respondents agree that the apps has achieve good quality in terms of its layout, system design and user friendliness (refer to Fig. 4(a)). However, 16.67% think that the apps is user-unfriendly and inconvenient to be used by the visitors. In terms of its usability, average of 80% of the respondents are satisfied with all the functions provided by the apps. This is as shown in Fig. 4(b).

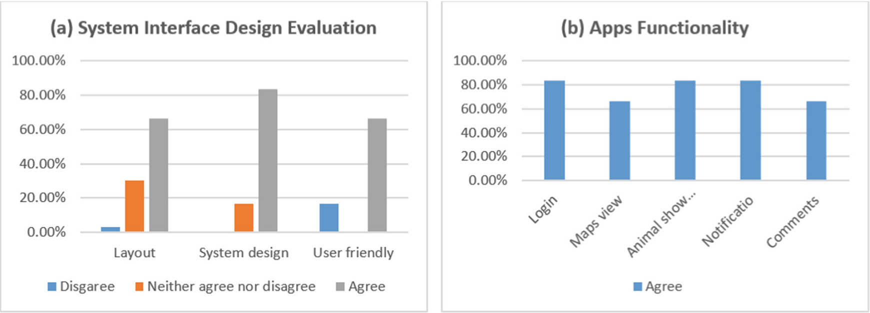


Fig. 4. Visitors response on apps design and functionality

The second part of the evaluation is done to evaluate the effectiveness of the apps towards accessing zoo facilities. Figure 5 shows that at least 82% of the respondents satisfy with the developed apps in facilitating their visit in Malacca Zoo. On top of that 10% of the respondents chose to stand on neither decisions and another 8% was unhappy with the apps. These two categorizes are mostly involved by visitors age 45 and above. Locally, older visitors tend too pick these categories because they are naturally not really exposed with computer system or apps in their daily life (Fig. 6).

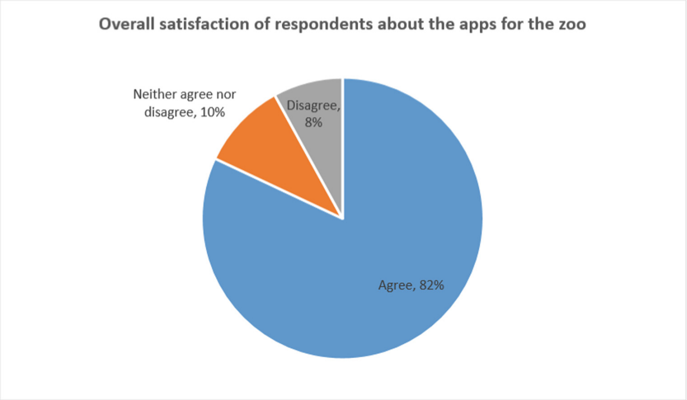


Fig. 5. Overall satisfaction of the respondents about the usability of apps for the Malacca Zoo

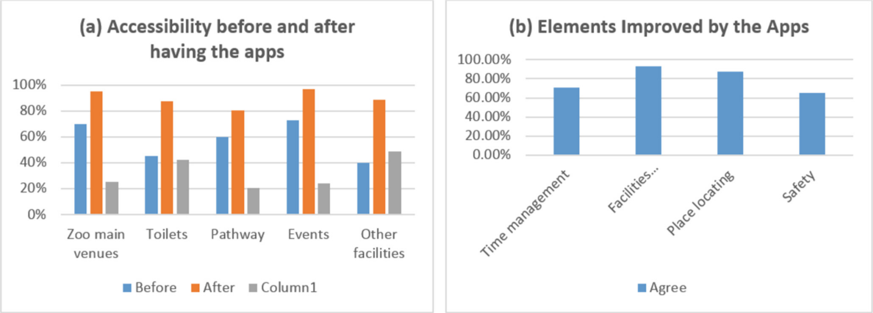


Fig. 6. The effectiveness of the apps in accessing the Zoo

A survey also have been conducted to evaluate the effectiveness of the apps among respondents in accessing zoo amenities. Based on the five things that have been listed, all have shown a significant improvement on the average of accessibility judgement. The noted improvements are 25%, 42.33%, 20.5%, 24% and 48.74% each for accessibility of zoo main venues, toilets, pathway, events and other facilities (Refer to Fig. 5(a)). The highlight of the study is on the increasing percentage of facilities of the zoo and supported by the other elements. Finally, the survey also had evaluated peoples opinion on what items are improved from the proposed apps. More than 60% agree that by using the apps, they can improve their time management, facilities accessibility, place locating and safety. The top percentage can be seen on the second item while the lowest is on the safety of visitors as the apps does not provide any specific module for visitors safety.

5 Conclusion and Future Work

There is growing awareness of the importance of providing citizens with outstanding service and value. To do so, organizations need to have in place strategies which enable them to be knowledge-driven, be responsive, have engaged and skilled staff, and have processes in place to be continually improving.

In this paper, we have shown that consumers' satisfaction of a service can be increased by the help of a specific tool in accessing all the facilities and items being served at the organization. As a case study, we have chosen Malacca Zoo to become part of the research. We developed a navigational application to help users to navigate around the zoo with ease. The application will become the tool to check the location of the intended facilities, animals or even shows at the zoo.

Zoom App is a mobile platform application and it has shown that it has achieve its objectives to occupy the requirements by the users. A few suggestion for future improvements are as follows:

1. To prepare user guideline to improve the usability of the application.
2. To come out with social networking module to allow users to interact with other visitors during the visit.
3. To have quiz module which can be answered by the visitors especially the kids to improve their knowledge.
4. To have user profile page where it can track down numbers of visits to the zoo.

Based on the post survey, it shows that facilities visibilities to the consumers have been improved as compared to the time where paper based map is used. In average, 82% of the zoo visitors agreed that the tool satisfy them in many aspect of the requirements and in overall has improved their experience while at a recreational place.

Acknowledgments. This research was partially supported by Gates IT Solution Sdn Bhd and Research Grant Vot. U540.

References

1. Rahman AA (2007) Consumers satisfaction towards recreational facilities and services in Ayer Keroh Forest Recreational Area, Melaka. Masters Dissertation, UITM Malaysia
2. Armstrong R, de Vaal N, Reynolds J, Taylor J, Wilson J, Boggs M, McLeod B, Ketcheson L (2012) Measuring customer value and satisfaction for parks and recreation: a manual. Parks and Recreation Ontario, USA
3. Morrison GR, Ross SM, Kalman HK, Kemp JE (2016) Designing effective instruction, 8th edn. Wiley, New York
4. Karanikola P, Tampakis S, Tsantopoulos G, Digbasani C (2014) The public zoo as recreation and environmental education area: visitor's perceptions and management implications. WSEAS Trans. Environ. Dev. 10:81–91
5. Sii (2015) Zoo Wroclaw Map, Google Play
6. Androutsopoulos I, Paliouras G, Michelakis (2015) Southwick's Zoo, Google Play
7. Ooredoo I (2015) Ragunan Zoo, Google Play
8. Branch RM (2010) Instuctional design: the ADDIE approach. Springer, New York

N-LibSys: Library System Using NFC Technology

See Pui Mun, Mohd Heikal Husin(✉), Manmeet Mahinderjit Singh,
and Nurul Hashimah Ahamed Hassain Malim

School of Computer Sciences, Universiti Sains Malaysia, Pulau Pinang, Malaysia
spmun.ucom12@student.usm.my,
{heikal,manmeet,nurulhashimah}@usm.my

Abstract. This paper examines the weaknesses of an existing library system which is the Online Public Access Catalog (OPAC) and proposes an enhanced platform that overcomes the weaknesses of the previous system. The existing OPAC platform acts as a search engine for reading materials in the library is commonly complicated and not user friendly. The Near Field Communication (NFC) technology is introduced in this project as part of a mobile application. NFC is a set of protocols on portable devices which perform wireless and contactless short range, two-way communication between compatible devices. Transmission of data between devices is completed over radio waves and magnetic induction with passive NFC devices. As mobile devices, especially smartphones are more commonly utilized by users all around the world including university students, the proposed system would be able to enhance the performance of a library to be more convenient, secure, and more user friendly. Indirectly, the development of this mobile application can improve the visitation rates of students to the library by simplifying the searching, tracking reading materials and borrowing activities.

Keywords: Mobile application · NFC technology · Library system · Android-based application · Smart university

1 Introduction

Today, the Internet is widely used in different areas as it provides access to vast amount of information resources. However, the library still remains as a reliable source for students to obtain accurate and relevant information especially in regards to their university courses. Presently, most libraries are using the OPAC (Online Public Access Catalog) system, which is an online database or bibliography of collected reading materials made available for users to search in the library environment [1].

The OPAC system supports a number of features such as book searching within a library and also allows a user to identify similar books or other sources that may be categorized under similar groupings. It is commonly known as the gateway to a library's collection [2]. But, it is a common theme among existing OPAC implementations that the system is complicated with numerous features built into it [2, 3]. The features that are made available to users via OPAC often complicates the process of searching relevant sources as the platform is often case sensitive and

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_3

utilizes an inefficient searching method [1]. But before we examine further the issues and propose our solution, we would have to explore additional relevant information. As such, this paper is divided into a few sections. Section 2 examines the existing background work. This is then followed by the uniqueness of our proposed system and how the platform works. Section 5 highlights the overall proposed system and this is then followed by the discussion on the evaluation of the system. Finally, the paper concludes with our conclusion on the proposed mobile application.

2 Background and Related Work

Universiti Sains Malaysia (USM) is currently utilizing Web Online Public Access Catalogue (OPAC) system within their libraries. The system has a wide usage coverage by most libraries around the world. The bibliographic records are gathered from various sources by downloading the information in a file directly or indirectly using the in-built platform features. The imported bibliographic data can be modified whenever necessary based on the local practice in the form of main or added entries [4]. Besides that, the records are stored in the MARC (machine-learning readable) format which is enriched with small numbers of controlled subject descriptors which represent the subject content of the item as well as their classification number [5]. When the platform is accessed via a mobile device, the overall system is not user friendly because it is originally a web-based application. Hence, the graphical interface and the features are not suited for a small screen. There is also the lack of an efficient book borrowing functionality for mobile devices [4].

In early 1980s, the first library OPAC was introduced and the usability study of the system was carried out consistently from time to time. A new version of the OPAC system was implemented in 1999 by Chisman, Diller, and Walbridge at the Washington State University Library where several OPAC problems was identified via a usability study and were then rectified [1]. The problems were mainly related to subject indexes and the general understanding of the participants on the usability of OPAC's features [5]. Another issue of the system was discovered by Novotny in 2004 for the Pennsylvania State University Library OPAC, where they determined the system usability by conducting a protocol analysis. Five structured tasks were completed by the participants by using the OPAC catalog and their feedback and understanding about the system were recorded. Their research highlighted that the OPAC system should not necessary be designed to function similarly to internet search engines as this generated problem in the overall users' understandability and navigation of the system [5]. The study shows that there are several essential features which are not offered by OPAC system such as a quick search function and book cover display facilities.

As part of our proposed improvement, we are introducing the NFC technology to the library platform. NFC stands for "Near Field Communication" is a set of standards which enable wireless short range communication between compatible devices [6]. NFC originally evolved from the RFID (Radio Frequency Identification) technology, which is sometimes used interchangeably. Compared to RFID, NFC allows two-way communication and limits the range of communication to only within 4 in. or 10 cm. This limited range allows NFC to be more secure as the transmission is nearly impossible to be intercepted. The technology is now a standard feature in most smartphones, which allows peer-to-peer communications by transmitting data over radio waves. There are two types of devices that could use the NFC standard which are categorized into two: (1) passive and (2) active devices. Active NFC devices are commonly found in smartphones, whereas passive devices include tags and other small transmitters [6].

For the proposed system, the NFC tags will be used to store and transmit relevant information to NFC enabled devices such as smartphones when needed. The NFC tags are small stickers with small unpowered NFC chip embedded into it. NFC tags can be programmed to stored information of books such as ISBN code, book title and author. It has limited storage memory at around 8 KB and literally uses power from the NFC device that reads them due to the magnetic induction to store the information [6]. Figure 1 below shows the example of NFC tags or NFC stickers. The mobile application will work closely with the OPAC database to retrieve relevant information during the book borrowing process such as book ID, ISBN, book's title and student's information. Users can borrow books by scanning the NFC tags at the book using mobile devices embedded with NFC technology. All the related information about a book could be stored in the tags.

Fig. 1. NFC tags

Currently, the Hanno library in Japan is utilizing the NFC technology by applying the tags to the bookshelves, which allow users to gain information on the books by using their smartphones [7]. Besides that, the users can review and reserve books from the library by using NFC technology. But, their current implementation does not fully integrate the NFC as the main method to fulfil the book borrowing process. In saying that, their implementation is an example of a good practice to implement NFC technology in library activities [7, 8]. Moreover, the NFC technology is utilized to simplify certain phases of borrowing activities for users of the Hanno library.

3 Uniqueness of the Proposed Solution

As mentioned earlier, the OPAC system is utilized in different libraries across the world where one of the places is the Universiti Sains Malaysia library. As such, our proposed system is focusing on the improving the students' usage of the library. Essentially, the proposed NFC mobile library application provides several benefits where students can view books' information and availability in a more effective manner. Per se, we are proposing an NFC-enabled library system aptly called N-LibSys. As an initial phase, an

NFC-enabled device such as a smartphone can be used in place of a student matric card for book borrowing. This allows a one-touch borrowing system for the library which simplifies the borrowing activity for users upon exiting the library. Besides that, the application increases the user friendliness of the library system with a decreased in the OPAC usage complexity and improving the efficiency of book tracking processes in the library. The reading material's relevant information is recorded in the library database via the NFC tags and the application transactions is secured with the NFC short-range data transmission [9].

The differences between the OPAC system and the NFC mobile library application is highlighted in Table 1. The OPAC system is an online bibliography or catalog and it is a web-based application with a complicated user interface design which leads to difficult navigation issues for the user [2]. Besides that, it is time consuming and creates security issue as the system allows navigation without any login authentication from the user. The existing borrowing process in the library which utilizes the OPAC system requires users to present their matric card upon borrowing book in the library and the process is physically carried out by the librarian. Thus, there is a need for more manpower in the library. Unlike the OPAC system, the NFC mobile library application is designed with an overall streamlined user interface which is more user-friendly [8]. A login authentication is required from the user before gaining access to the application which makes the system more secure. The authentication is linked to the university's student registration system which validates whether a student is enrolled to the university. Moreover, the mobile application allows the book searching and borrowing process in the library to be completed quicker which saves time as compared to the existing OPAC system. The book borrowing process can be completed without physical contact with a librarian by using the NFC mobile library application. It is more convenient for the users and less manpower is needed as the process is digitized.

Table 1. Comparison between existing OPAC system and our proposed system

Existing OPAC system	N-LibSys
Complicated user interface design	Simplified user interface design (user friendly)
No login authentication required (Security issue)	Login authentication required
Time consuming	Saves time (in the book searching and borrowing process)
Web-based application	Mobile based application
Online bibliography/catalog	Innovative and driven technology used (NFC)
More manpower needed (librarian)	Less manpower needed (book borrowing process digitalized)
Matric card needed in book borrowing processing, carried out physically with librarian	NFC devices allow the book borrowing process to be completed without physical contact with librarian (more convenient for users)

4 How Does the System Works

4.1 Existing OPAC System

The OPAC platform is a standalone online database or bibliography of collected reading materials made available for users to search in the library. It allows students to search for books and other resources in the library. However, the OPAC system is complex and it is hard for users to navigate it without any extensive guidelines. This occurs because the system is not well designed where the navigation controls, menus, and aesthetics of the interface is not consistent throughout the system [1]. Therefore, it would be difficult to use OPAC system if the users are unfamiliar with it.

Once the relevant materials are located by the students, they would have to present their matric card to the librarians at the counter to borrow the books. Thus, the students' matric cards are essential for completing the book borrowing process. Besides that, it might consume much time especially when there is a long queue of students to borrow books in the library.

4.2 N-LibSys

The NFC-enabled mobile library management application provides additional improvements to the existing OPAC library system. It is capable of providing book searching services to the users anywhere and anytime with the availability of Internet connection via their mobile devices. Besides that, users with NFC-enabled device can view additional book details with the NFC tags on the books as the details are recorded in the tags based on the available database information.

Users can also borrow books from the library by scanning the NFC tags equipped on the books. The mobile application captures the book details and places it into a temporary book borrowing list. Moreover, the users can delete selected books whenever needed before the finalized book borrowing process is confirmed. This is more convenience for users to perform the book borrowing process. The finalized book borrowing actions performed by the users via the mobile application will be recorded in database system upon their exit from the library. However, these actions can only be done by registered users in the library system. The users can access to the system once the authentication and verification process are completed. This is to enhance the security of the system. In addition, the mobile application is capable to support multiple users via the available network connectivity.

In addition, the security in the library is improved as the users can only leave the library after they scan their book borrowing information with NFC readers. By utilizing the current security gates and the mobile application, users need to touch their mobile devices to a NFC reader at the exit gate to confirm their borrowing transaction into the library database. This procedure reduces the waiting times as the security guards are not required to manually check the books upon exit. The final borrowing information will be recorded and updated to the related library databases. The overall process is interrelated from when the users scan the books they required via the NFC tags until they are exiting from the library. The system enables the book information to be captured by the users' mobile device and the NFC reader to ensure the data accuracy. Figure 2 shows the complete book borrowing process from book searching until their exit from the library.

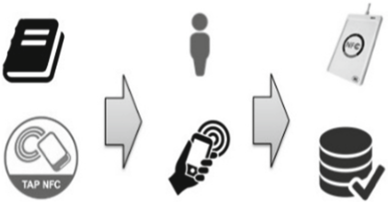


Fig. 2. The borrowing process with the N-LibSys

The mobile library system proposed is only applicable for devices which embeds the NFC technology. Hence, users who do not own a NFC enabled mobile devices will not be able to use the application completely. However, with the fast pace of technology development nowadays, it is possible that all mobile application will be equipped with NFC technology in the future since it is one of the innovative and emerging technologies [8, 9].

5 Proposed System Architecture

The mobile system would be developed based on the Android platform as there are higher numbers of users utilizing such platform globally [9–11]. The Android platform is also used with the integration of SQL as the database management system for the application. The web system which is utilized by the library administrator would be developed based on the HTML and CSS. Figure 3 shows the overall system architecture of the NFC mobile library application. The librarian or administrator manages the students and books record through a website system. The system allows the

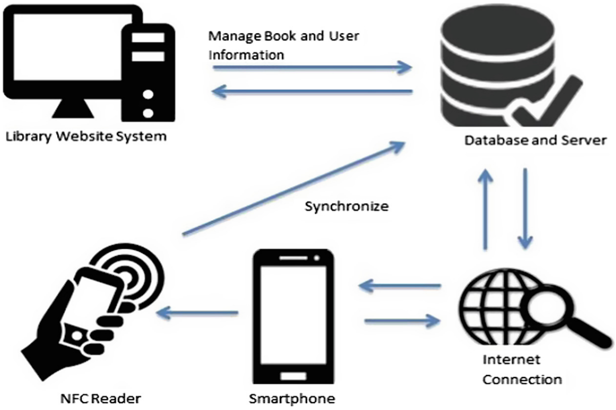


Fig. 3. Overall system architecture for N-LibSys

administrator to create new records, update or delete existing information. The information updated will be stored in the system library database. Besides that, the members' and books' record can also be viewed through the website system.

As highlighted earlier, an internet connection is required for the users to access the mobile application for further functionality such additional information on a book. This is due to the information being accessed directly from the library databases. The user's account login is authenticated by the library database before the user can access the main page of the application. From the main page, the user can view their profile information, search for books available and find the books' location in the library. They can also borrow books by using the NFC technology and view their transaction history via the mobile application.

When a user scans the NFC tags into the application, the books are updated in the library database as 'On Hold'. This notifies other users on the overall availability of the book when it is searched. When the user would like to complete their borrowing process, they would only need to proceed to the library exit which is equipped with a NFC reader. They then scan their mobile devices which has recorded their temporary transaction. This action then finalizes the book borrowing process and the availability status of the related book borrowed is updated in the library database. Figure 4 depicts the user interactions with the proposed application.

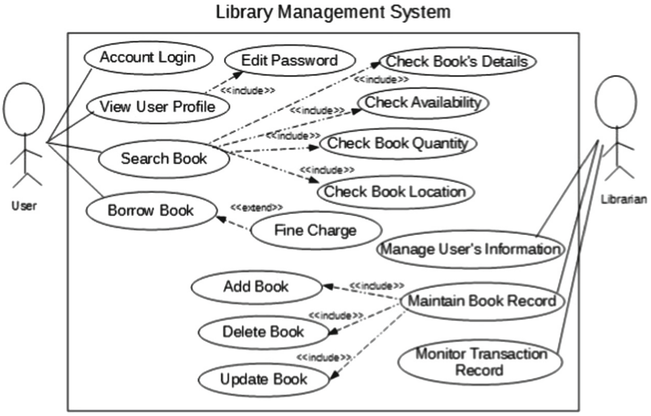


Fig. 4. Use case diagram for the N-LibSys with the user and administrator

6 Initial User Evaluation

The project's features and functionality are clearly identified by completing a series of testing that are unit, integration, system and user acceptance testing. The performance, reliability, usability, etc. of the mobile application are evaluated. There are several advantages, disadvantages, strengths and limitations of the NFC library mobile application.

As part of our testing for the proposed system, we conducted an initial user testing among 30 participants which also included students as they represented the main end users for this system. From the initial users' perspective, the NFC library mobile application is designed with simple user interface which allows easy navigation from one activity to another. They also found the application to be convenient, time saving and user friendly. It can be used as long as network connectivity is available. Besides, it utilizes innovative technology that is NFC technology, hardware such as NFC reader and tags which are relatively cheap and affordable. The books in the library can be borrowed without having to approach the librarian at the counter with matric number. The mobile application increases the usability and functionality of mobile devices. In administrator perspective, they keep track on the record of members, books and transaction from time to time. Besides, the system web application provides ability to manage information such as create, edit, update and delete functions which work closely with the database system of the library.

Nevertheless, there are also disadvantages of the application. It reduces face-to-face inter-action between users and the librarian as the users are borrowing books by using NFC-enabled mobile device but not through the library counter. The limitation of the mobile application is that the NFC technology is not embedded in every mobile device nowadays. However, the technology development is at a fast pace which gives the opportunities that every mobile device will be equipped with NFC technology in the future. Besides, users are not concern and aware of the usage of NFC technology as it is still a relatively new technology nowadays.

7 Conclusion

In conclusion, this project was developed to improve the existing OPAC library system. It allows users to search for books' details and complete the book borrowing process via phone devices which are embedded with NFC technology. Besides that, the location of the books could be accessed and this allows the user to locate for the books required. Users are accessible to the mobile application once their accounts are registered and verified.

Essentially, the system is proposed in order to provide better functionality, usability and convenience to the users. The overall system design provides good navigability and promotes a level of user friendliness via a simplified interface. As most users are equipped with smartphones, the project could provide a relatively convenient and time saving process for users to complete their relevant library activities. The utilization of the NFC technology in the mobile application allows for a faster method to update the library database via a single tap in the library system during the borrowing process. The library mobile application also encourages users to utilize the library facilities more frequently and may promote a healthy reading habit among users.

As part of the future work, the security of the system will be further enhanced by using an encrypted password for the login authentication of the system. The users will be allowed to borrow more than one book for the same book. This is because as part of the limitation of the system, books with same ID or ISBN can only be allowed to borrow in quantities of one. Besides, a student's matric card can be used to pay for fine charge in the future because the card could be used as a multifunction card such as credit top up. For example, the matric card can act like a ‘Touch n Go’ card which is a prepaid electronic cash card for relevant fine charges in the library.

References

1. Villen-Rueda L, Senso JA, de Moya-Anegón F (2007) The use of OPAC in a large academic library: a transactional log analysis study of subject searching. J Acad Librarianship 33(3): 327–337
2. Arshad A, Shafique F (2014) What do users prefer, card catalogue or OPAC? A study of Punjab university library. Electron Libr 32(3):286–295
3. Yesmin S, Ahmed SZ (2016) Preference of Bangladesh university students for searching the library catalogue: OPAC or discovery tool? Electron Libr 34(4):683–695
4. Rajendiran P, Parihar YS, Deshpande AU (2007) Automated bibliographic record capturing from web OPAC and online bibliographic database for library cataloguing in LibSys. Ann Libr Inf Stud 54:140–145
5. Thanuskodi S (2012) Use of online public access catalogue at Annamalai university library. Int J Inf Sci 2:70–74
6. Riekki J, Sanchez I, Pyykkonen M (2012) NFC-based user interfaces. In: 2012 4th International workshop on near field communication, pp 3–9
7. Clark M (2013) Japanese library adds NFC tags. http://www.nfcworld.com/2013/07/16/325011/japanese-library-adds-nfc-tags/. Accessed 16 July 2013
8. Rana S (2013) Japan’s first NFC equipped library lets you rate, review and comment on books with a simple tap of the phone. http://newlaunches.com/archives/japans-first-nfc-equipped-library-lets-you-rate-review-and-comment-on-books-with-a-simple-tap-of-the-phone.php. Accessed 03 July 2013
9. Pachica A, Bernardo J (2016) A middleware application framework for academic institution services utilizing near field communication (NFC) technology. Int J Appl Eng Res (IJAER) 11:7997–8004
10. Liao P, Shieh J (2016) The development of library mobile book-finding system based on NFC. In: 2015 Proceedings of the 4th international congress on advanced applied informatics, pp 148–153
11. Curran K, Millar A, Garvey CM (2012) Near field communication. Int J Electr Comput Eng (IJECE) 2:371–382

The BDD Navigation Tracking Systems Using the Beacon

Sunggyun Jang and Inwhee Joe(✉)

Department of Computer and Software, Hanyang University,
222, Wangsimni-ro, Seongdong-gu, Seoul, Republic of Korea
mrjang28@gmail.com, iwjoe@hanyang.ac.kr

Abstract. There are many guidance systems for people using tracking systems where in railway, airports, trade fairs, shopping centers, office, tourism, and industry. The beacon with Bluetooth Low Energy is a wireless personal area network technology and can be used many applications to calculate the exact location. However, from an accuracy perspective, a big problem with this approach is the Received Signal Strength Indication (RSSI) of weak signal caused by having an obstacle between the BLE modules such as beacons, smartphone, tablet PC. This paper proposes a Beacon Detection and Direction (BDD) algorithm that can track and guide in indoor position more advanced than previous server-based or client-based application. It not depend on only RSSI because of the big problem which is weak signal by hindrance. To calculate accuracy position, the proposed algorithm is combined a server-based approach and a client-based application. It has three tables and beacons that are divided three types are deployed inside location at regular intervals. Using the proposed algorithm we try to test the navigation tracking and the results show that BDD provides an accurate and effective route.

Keywords: IPS · Beacon · Navigation tracking systems · BDD

1 Introduction

Indoor Positioning Systems (IPS) has been widely used in marketing, information, automation, payment, and tracking. The tracking systems consists of router guide, people position tracking, and things position tracking. Indoor navigation are normally used as a client-based application or a server-based application. There are number of techniques for client-based indoor positioning systems such as Wi-Fi, BLE, VLC and Ultra-wideband. Most of BLE services base RSSI signal to detect distinct between the beacon and the smartphone. This idea is simple and useful. However, a problem with this approach is that the RSSI of weak signal has a very big problem if having an obstacle between the beacon and the smartphone. As the result, many studies have proved that it is not enough to calculate the accuracy location using the beacon with RSSI only.

This paper is proposed with an aim to provide the navigation tracking systems using the beacon for the first time visitors in the subway. Specifically, it proposes a Beacon Detection and Direction algorithm with server-based and client-based that can guide which the subway line should I ride for visitor destination in subway station. This paper

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_4

focus on Hanyang University (HYU) Station on the Seoul metro subway line 2 in order to test the navigation tracking systems.

This paper is organized as follows: In Sect. 2, the navigation tracking systems is explained. In Sect. 3, the proposed architecture is explained. In Sect. 4, the proposed method is evaluated by using beacons and we conclude in Sect. 5.

2 Navigation Tracking Systems

2.1 The Subway Guidance System

The Seoul station is center in South Korea and has many gateways than other stations in Seoul, the capital of South Korea. Table 1 shows the number of regions visited in South Korea.

Table 1. The number of regions visited in South Korea

Division	2015	2014	2013
Seoul	78.7	80.4	80.9
Jeju Island	18.3	18.0	16.7
Gyeonggi	13.3	13.0	17.9
Incheon	10.3	8.0	11.7
Gangwon	6.8	5.0	7.8

([a]Source: International Visitor Survey, 2016.2, Policy Statistics Evaluation Office of Korea Culture & Tourism Institute, Multiple responses, unit: %)

First time visitors have questions that where do I go to Gangnam station in Seoul station and which do I get the subway line because of very complex structure that the Seoul station has two train lines, three subway lines and fifteen gateways.

Despite the friendly guide or explanation by crew or people in station, there is a limit to explain detail for the first visitors. Currently there are many the subway guidance systems, but it is not enough to navigate route because of inexact location of the user.

2.2 Beacon Packet and RSSI Analysis

The beacon profile can be used to detect the beacon's location and include advertising data packets that consist of four parts: 1 byte preamble, 4 bytes access address, 39 bytes advertising channel PDU, and 3 bytes CRC [1]. In the data (up to 31 bytes), the proximity UUID is a unique value, major and minor value is beacon location information. The TX power is the signal strength that measure at 1 m far from beacon [1].

This paper tried to analyze RSSI to calculate accuracy distance between beacon and smartphone. For analysis, it can compare with two methods each distance 1, 3, 5, and 8 m; (1) Nothing obstacle between the beacon and the smartphone (2) Obstacle between the beacon and the smartphone. Table 2 illustrates the compare value in two methods, the start 1 (min RSSI value) and end 1 (max RSSI value)have nothing an obstacle between the beacon and the smartphone; the start 2 and end 2 take an obstacle between the beacon and the smartphone.

Table 2. RSSI measure between the beacon and the smartphone

	Start 1	End 1	Start 2	End 2
1 m	−66 dbm	−77 dbm	−76 dbm	−84 dbm
3 m	−74 dbm	−83 dbm	−83 dbm	−92 dbm
5 m	−82 dbm	−91 dbm	−86 dbm	−91 dbm
8 m	−81 dbm	−93 dbm	−85 dbm	−91 dbm

As a result, the first 1 and the end 1 have constant range. It denotes that the beacon and the smartphone distance is can be accuracy gauge. On the other, first 2 and end 2 deviate from the first method and range interval is narrow than the first method. It denotes that the beacon and the smartphone distance is precarious. Also the common issue in 8 m is unstable because far distance is inversely proportioned to precise position. It guesses that if there are a number of barricade, the RSSI value cannot be predict.

3 Design and Implementation

3.1 Beacon Detection and Direction Software Architecture

In the proposed system, called BDD (Beacon Detection and Direction) Architecture, main issue are detection and direction for first visitor in order to track and guide. This proposed systems can be divided into three main parts: (1) Beacon (2) Apps (3) Server. Beacon is deployed in the subway station, and apps is installed on visitor's smartphone. Server is ready to guide visitors in the subway station. Figure 1 is Beacon Detection and Direction Software architecture for navigation tracking systems.

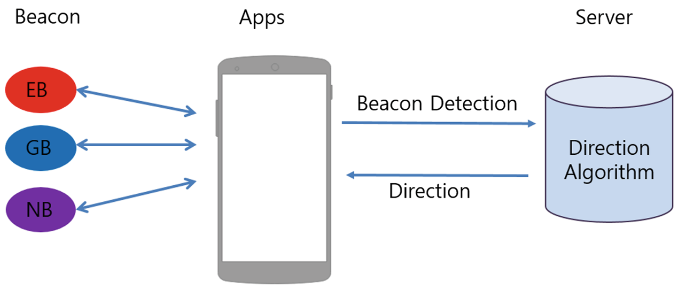


Fig. 1. Beacon Detection and Direction Software Architecture. The direction can be used in the direction algorithm which calculates the correctly direction (1) Beacon can advertiser the beacon packet from beacon to smartphone. (2) The apps could get the beacon packets from the each type of beacon. (3) When collecting 10 beacon packets from each beacon, apps could request a query of detection to server. (4) The server can calculate the direction using three tables, after then response the direction to the apps. It must be enable functions, Wi-Fi and Bluetooth on smartphone.

This proposed systems is possible to detect people using the beacons. However, it is not easy to figure out directions such as north, east, west, and south. For disposition of the beacons, it makes three tables; the first table is separated types of beacons, second table is the station map, and the last table is the location of the beacons. The implementation of beacon devices can be divided into three main modes: (1) Exit Beacon (EB) (2) Gate Beacon (GB) (3) Normal Beacon (NB). All beacons have the major and the minor value in the BLE packet. In the station located each beacon has states of next path to track and detect visits, and that states support movement estimation in the station on Fig. 2.

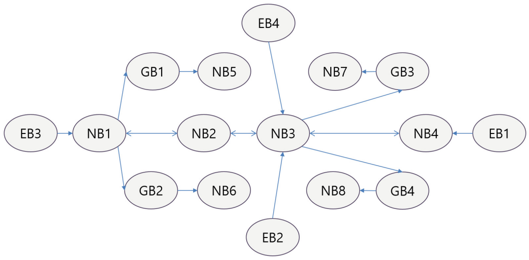


Fig. 2. The beacon state map. A visit stands up in front of exit 3 and the person would like to go the city hall. This case scenario is EB3 -> NB1 -> GB1 -> NB5, this way is best short path. Another way, it is EB3 -> NB1 -> NB2 -> NB3 -> GB3 -> NB7 but not best path.

The application apps support three main functions (1) select source and destination (2) Bluetooth scan (3) display direction. When a visit having the smartphone and an installed application walk in the station, a visit selects source and destination, and the application scan each beacon and get information as major, minor, and tx power. Afterward, the application could request direction information to server, and response direction to application. Final, the application could show direction to user.

The server side has a direction algorithm to calculate navigate direction for first visits in subway station. At the start of the cycle, the visit decides destination and select the destination on apps in front of the subway exit. Afterwards, visit starts walking in the subway station and apps can detect a visit. On detecting, apps requests visit's direction to server. The direction algorithm checks for beacon type to confirm visit location. Afterwards, the algorithm compares with disposition beacon table and calculates the direction. The result of the task is then response the direction to apps.

3.2 Disposition of the Beacons

How many do you need the beacons in HYU station? This paper suggested how to deploy the beacons in station; (a) If there are four exits, you need four exit beacons (b) If there are four gates, you have four gate beacons (c) Then normal beacon is double of number of exit beacons. The HYU has four exits and four gates. Therefore, the station has four EBs, four GBs, and eight NBs. Beacons is located at each floor and each position that EB is in front of EXIT, GB is nearby GATE, and NB is between GATE and EXIT on Fig. 3.

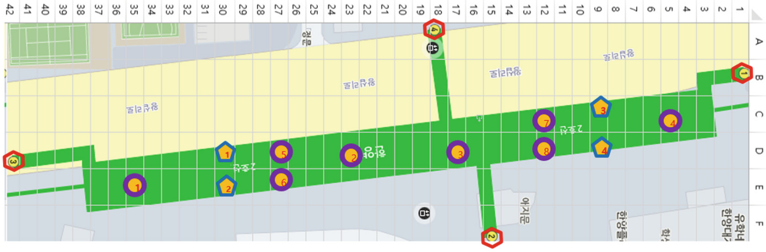


Fig. 3. All beacons arranged at the HYU. We have tried setting the zoom level to 14 using the NAVER, Corporation is Korea' Internet Company, static map for the Station Map Table (SMT). The HYU station has 6 columns and 42 rows due to have width 45 pixel and height 22 pixel. There is a one exit beacon in third floor, there are eight beacons (4 gates and 4 normals) in second floor, and there are seven beacons (3 exits and 4 normals) in first floor. This is all beacons arranged at the HYU station within three floors.

4 Performance and Evaluation

In this section, we mainly evaluate the effectiveness of tracking system via testing. For the implementation of the design, we used set of components, 16 beacons (NRF 51822), a smartphone (Galaxy S6 edge), AWS server, MySQL 5.5.40, R data mining 3.2.0. To test, there are two hypothetical scenarios (1) Source is HYU and destination is City hall station (2) Source is HYU and destination is Gangnam station. We use a station map of the experiment scene. The station map as resource is utilized by the R for data mining. The R language is possible to insert a background image as google map for plot. The station map image of HYU could apply a plot. In the figure, X-axis is that left is north and right is south value. Y-axis is that top is west and bottom is east value. If you will go the city hall station, a visit stands on west side. If you will go the Gangnam station, a visit stands on east side.

The first hypothetical scenario is that visit would like to go from HYU to city hall station. The visit stands on in front of exit 2 at HYU station. The scenario is that visit walks through every beacon in the HYU station: EB2 -> NB3 -> GB4 -> NB8.

The second hypothetical scenario is visit's movement at EB3 -> NB1 -> GB1 -> NB5. The visit would like to go from HYU to Gangnam station. The visit stands

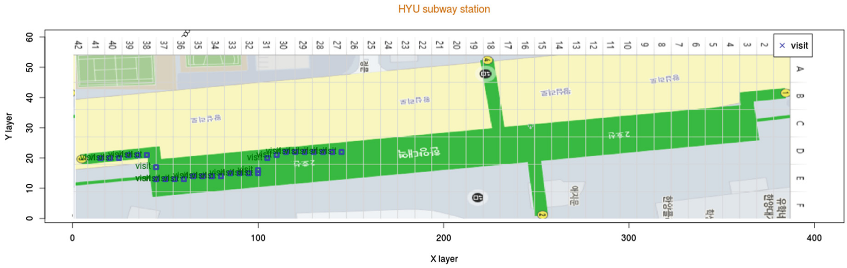


Fig. 4. Result of data mining for the second hypothetical scenario.

on in front of exit 3 at HYU station. Figure 4 plots the visit's tracking when visit used the apps that it shows the direction which route is best path.

5 Conclusion

In this paper, we designed a navigation tracking systems based on BDD algorithm. This paper shows an example that the location is determined directly on smartphone of the use via beacons like indoor GPS within station. The suggested BDD algorithm is a new concept that deal with server-based and client-based architecture. It provides more flexible and easy to use. However, current limitation of the system are not fast in aspect of real-time to detect people and not enough to be accuracy ratio. Moreover, it is not convenience to usage for people on apps. In future version of BDD, we plan to improve the station map and beacons in order to be accuracy ratio and deploying beacons effectively. Also, it need the 3D station map using unity engine on apps for user experiment.

References

1. Sunggyun J, Guangqiu J, Inwhee J (2016) A dual-mode beacon profile for normal and disaster environments. In: Mobile and wireless technologies. Springer, Singapore, pp 59–68
2. Rantakokko J, Rydell J et al (2011) Accurate and reliable soldier and first responder indoor positioning: Multisensor systems and cooperative localization. IEEE Wirel Commun 18(2): 10–18
3. Ijaz F, Yang HK, Ahmad AW, Lee C (2013) Indoor positioning: A review of indoor ultrasonic positioning systems. In: 2013 15th international conference on advanced communication technology (ICACT) pp 1146–1150
4. El Arby AA, Thiare O (2016) Voice over LTE performance evaluation. In: Mobile and wireless technologies 2016. Springer, Singapore, pp 25–33

A Distributed Gossip Optimization Algorithm for Wireless Multi-hop Networks

Jain-Shing Liu[1] and Wan-Ling Chang[2(✉)]

[1] Department of Computer Science and Information Engineering, Providence University,
200, Section 7, Taiwan Boulevard, Shalu District, Taichung City 43301, Taiwan, R.O.C.
chhliu@pu.edu.tw

[2] Department of Insurance, Chaoyang University of Technology,
168, Jifeng East Road, Wufeng District, Taichung 41349, Taiwan, R.O.C.
wlchang@cyut.edu.tw

Abstract. In this paper, we study a constrained network lifetime maximization problem in wireless sensor networks, and introduce a cross-layer formulation with general NUM (network utility maximization) that accommodates routing, scheduling and stream control from different layers of network. Specifically, for this problem, we derive a gossip-based formulation for the consensus agreement on the variables involved, and develop an asynchronous decentralized algorithm specific to the optimization problem. Our numerical experiments exhibit its results, showing that the gossip-based consensus algorithm can actually achieve the optimization objective by means of the simple and robust asynchronous operations developed.

1 Introduction

Recently, the services in wireless multi-hop networks have created large scale demands for transmission of traffic requiring, which continuously intensifies the interest of researchers in the development of utility optimal wireless transmission schemes. For this problem, the cross-layer optimization reveals a good direction for its solution to incorporate the different layers regarding the system utility so that the overall network performance can be improved. Given this potential, a cross-layer optimization scheme is usually implemented in a centralized manner, which would be unsuitable for the wireless networks. In contrast, a distributed algorithm specific to the optimization is more practical for the multi-hop wireless networks, and in fact it had been employed to address a wild range of problems arising in, e.g., wireless sensor networks. Nevertheless, many of distributed algorithms still require a synchronous implementation and a globally known order of stations to be given in advance, which can hardly result in a fully distributed solution [1–7]. To get rid of the limitation, we combine the ideas of average consensus technique and distributed programming model to conduct a decentralized asynchronous gossip algorithm for the challenging cross-layer optimization in wireless multi-hop networks, wherein the scheduling sub-problem involved is already a NP-hard problem [8]. When compared with the programming-based methods, e.g., [9–12], which

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_5

usually resort to certain decomposition techniques for their optimizations and most rely on centralized implementations, our work introduces an asynchronous distributed algorithm without a synchronous clock and a globally known order usually required in these methods. Specifically, our algorithm allows a subset of edges in a network to be randomly activated and makes the nodes independently compute for their own local objectives to communicate with neighbors, leading to a practical fully-distributed approach for such multi-hop networking.

The rest of this paper is organized as follows. In Sect. 2, the cross-layer utility optimization problem is introduced. Then, our primal-dual algorithm for solving the optimization problem with gossip primal-dual algorithm is developed in Sect. 3. Following that, the experiment results on the algorithm with varying parameters are exhibited in Sect. 4. Finally, our conclusions are drawn in Sect. 5.

2 Cross-Layer Utility Optimization Problem

In this work, we consider a wireless network $G = (N, L)$ with a set of multicast sessions, and assume that each session denoted by its source node $s{\in}S \subset N$ multicasts packets to its destination node set t_S at a rate of x_S. In addition, we consider intra-session network coding, wherein the actual physical flows on each link need only to be the maximum of the individual destination's flows. For this, let f_{ij}^{S} denote the information flow rate from source S to its destination node t_S over link (i,j). In addition, let $N(i)$ be the set of 1-hop neighbors of node i. Given that, (i,j), $j \in N(i)$, represents an outgoing link, and (j, i) an incoming link, of node i. Further, we define transmission mode $\xi_k{\subset}L$ as a set of hyperlinks that can be concurrently activated, and scheduling matrix as an indexed collection of these modes, $\Xi = \{\xi_k\}$, where index $k \in K = \{1, \ldots |\Xi|\}$ Given that, p_k is defined to specify the possibility with which the transmission mode ξ_k can happen as the scheduling variable for the cross-layer optimization, while x_S is thought of as the stream control variable and f_{ij}^{S} as the routing variable. Besides, the capacity of link $l = (i,j)$ within transmission mode ξ_k is denoted by r_{ij}^{k}.

With the above, we aim here to adopt a gossip-based formulation for the consensus agreement on the variables involved, and develop a corresponding asynchronous decentralized algorithm. For this aim, we first introduce the concept of utility on resource allocation, and use the utility U_i, which represents a general form helping us to formulate the performance objective we expect to obtain in the network. Then, we solve the sum of aggregated utility maximization problem max $\sum_{i \in N} U_i$ with respect to the metric to be considered. Now, for a data session where its source node wants to transmit on a rate of x_S to its destination node, we have the flow conservation law as

$$\sum\nolimits_{j \in N(i)} \left(f_{ij}^{S} - f_{ji}^{S} \right) \geq x_{i,s}, \forall i \in N, \forall s{\in}S \tag{1}$$

where x_{is} is x_S if i is the source of session S, $-x_S$ if i is the sink, and 0 otherwise.

Obviously, the session rate and then the flow rate in the upper layers should be realized by the link capacity to be scheduled in the MAC layer and the data rate in the physical layer. Thus, we should proceed to establish the relationship between the upper layers and the lower layers. This is done by using a constraint that the physical flow accounting for all sessions S ∈ S should be upper bounded by the physical capacity r^k_{ij}, scheduled by the hyperlink with the transmission mode with p_k. That is,

$$\sum\nolimits_{j\in N(i)} f^S_{ij} \le \sum\nolimits_{k\in K} p_k r^k_{ij}, \forall i \in N, \forall S \in S \tag{2}$$

Finally, we can derive a scheduling constraint to exhibit the fact that the sum of scheduling probabilities p_k should be equal to 1 for the MAC involved, from the viewpoint of the whole network, represented by.

$$\sum\nolimits_{k\in K} p_k = 1 \tag{3}$$

Now, let $\boldsymbol{x}, \boldsymbol{f}, \boldsymbol{p}$ be the vectors of the above variables involved, we can formulate the cross-layer utility optimization problem (CLUOP) as follow:

$$\begin{array}{lll} \textbf{maximize} & \sum_{i\in N} U_{\mathrm{i}} & \text{(a)} \\ \textbf{subject to} & \sum_{j\in N(i)} \left(f^S_{ij} - f^{\mathrm{s}}_{\mathrm{ji}}\right) \ge x_{i,s}, \forall i \in N, \forall S \in S & \text{(b)} \\ & \sum_{j\in N(i)} f^S_{ij} \le \sum_{k\in K} p_k r^k_{iJ}, \forall i \in N, \forall S \in S & \text{(c)} \\ & \sum_{k\in K} p_k = 1 & \text{(d)} \\ & \boldsymbol{x} \succcurlyeq 0, \boldsymbol{f} \succcurlyeq 0, \boldsymbol{p} \succcurlyeq 0 & \text{(e)} \end{array} \tag{4}$$

3 Solving CLUOP with Gossip Primal-Dual Algorithm

In this section, we introduce our gossip-based primal-dual algorithm to solve the programming model just formulated. As shown in above, the difficulties of our problem come from not only the variables in the objective function that should be obtained under the same consensuses by each node, but also the other variables in the constraints that cooperatively contribute to the performance metrics required to simultaneously reach their consensuses as well. To resolve these difficulties in a distributed manner, we treat in the sequel CLUOP as a local problem of node *i*, using the same formulation of (4) which, for notational simplicity, ignores a more specific subscript *i* in each variable with respect to node *i* supposed to be given to identify itself being considered in the local problem, and replaces to concern only node *i*'s objective, U_i. For the local problem, we relax the first constraint of (4) to form the partial Lagrange as follows:

$$L(\boldsymbol{x}, \boldsymbol{f}, \boldsymbol{p}, \boldsymbol{u}) = U_i + \sum\nolimits_{i\in N-\{t_S\}, s\in S,} u^S_i \left(\sum\nolimits_{j\in N(i)} f^S_{ij} - \sum\nolimits_{i\in N(j)} f^S_{ji} - x_{i,s} \right) \tag{5}$$

where $\{u^S_i\}, \forall i \in N - \{t_S\}, \forall S \in S,$ are the Lagrange multipliers corresponding to (4(b)). Given that, we can now proceed to resolve it by finding the saddle points of $L(\boldsymbol{x}, \boldsymbol{f}, \boldsymbol{p}, \boldsymbol{u})$ with the following min-max problem.

$$\begin{array}{l}\mathbf{min}\ \Big(\mathbf{max}\big[U_i - \sum_{i\in N,s\in S} u_i^S x_{i,s}\big]\ + \mathbf{max}\ \Big[\sum_{(i,j)\in L,s\in S,} f_{ij}^S\Big(u_i^S - u_j^S\Big)\Big]\Big]\Big) \\ \text{subject to } (4(\mathrm{c})) - (4(\mathrm{e}))\end{array} \tag{6}$$

This problem can be solved successively in $\boldsymbol{x}$,$\boldsymbol{f}$,$\boldsymbol{p}$. When proceeding, the key challenge is to solve its scheduling sub-problem, i.e.,

$$\begin{array}{l}\max\boldsymbol{f},\boldsymbol{p}\succcurlyeq 0 (i,j) \in L, s \in S fijs(uis - ujs)) \\ \text{subject to } (4(\mathrm{c})) \ - \ (4(e))\end{array} \tag{7}$$

Now, after solving the above on $\boldsymbol{f}$, this problem can be simplified to the maximization problem on $\boldsymbol{p}$, as follows:

$$\begin{array}{c}\max_{\boldsymbol{p}\succcurlyeq 0} \sum_{k\in K} p_k \sum_{(i,j)\in L} r_{ij}^k \omega_{ij} \\ \text{subject to } \sum_{k\in K} p_k = 1\end{array} \tag{8}$$

where ω_{ij} is the maximum differential backlog to be introduced next. Clearly, this is a maximum weight independent set problem and $\sum_{(i,j)\in L} r_{ij}^k \omega_{ij}$ is the weight for each independent set or transmission mode, $\xi_k \in \Xi$. While such a problem is NP-hard and difficult to be approximated even in a centralized way [6], we can still solve it with a distributed approach based on the log-sum-exp approximation. Specifically, the maximization can be approximated by the log-sum-exp function with a coefficient β as

$$\max_{k\in K} \sum_{(i,j)\in L} r_{ij}^k \omega_{ij} \approx \frac{1}{\beta} \log \Big(\sum_{k\in K} \exp\Big(\beta \sum_{(i,j)\in L} r_{ij}^k \omega_{ij}\Big)\Big) \tag{9}$$

Then, we are led to solve an approximated version of the above, off by an entropy term $-\frac{1}{\beta} \sum_{k\in K} p_k \log p_k$, as shown as follows:

$$\begin{array}{c}\max_{\boldsymbol{p}\succcurlyeq 0} \sum_{k\in K} p_k \sum_{(i,j)\in L} r_{ij}^k \omega_{ij} - \frac{1}{\beta} \sum_{k\in K} p_k \log p_k \\ \text{subject to } \sum_{k\in K} p_k = 1\end{array} \tag{10}$$

The optimal solution to the above can be obtained by

$$p_k^* = \frac{exp\Big(\beta \sum_{(i,j)\in L} r_{ij}^k \omega_{ij}\Big)}{\sum_{\hat{k}\in K} exp\Big(\beta \sum_{(i,J)\in L} r_{ij}^{\hat{k}} \omega_{ij}\Big)} \tag{11}$$

3.1 Backpressure Scheduling

In this subsection, we introduce the backpressure scheduling algorithm to be adopted in the cross-layer optimization. Specifically, a node i in the network should decide, for each link (i,j), its target session by

$$s^*_{ij} = \arg\max\nolimits_{s\in \mathcal{S}} \left[u_i^S - u_j^S\right]_+ \tag{12}$$

and the maximum differential backlog over (i,j) by

$$\omega_{ij} = \left[u_i^{s^*_{ij}} - u_j^{s^*_{ij}}\right]_+ \tag{13}$$

The above formulates a distributed scheduling algorithm that uses backpressure to equalize differential backlog. Accordingly, each node maintains a separate queue for each destination, and the algorithm can ensure that each of the per-destination queues is stable [7]. In particular, this algorithm has the merit that it is not required to have a predefined set of routes. Instead, the flow data rates or routing variables for the cross-layer design can be dynamically decided by

$$f_{ij}^S = \begin{cases} \sum_{k\in K} p_k r_{ij}^k, & \text{if } s = s^*_{ij}, \text{and } u_i^S - u_j^S > 0 \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

3.2 Asynchronous Gossip Primal-Dual Algorithm

Based on the above, we introduce an asynchronous gossip-based primal-dual algorithm that can update the primal and dual variables at the same time and moves together toward the optimal solutions asymptotically. For this, we assume that each node has a clock based on Poisson distribution and a single virtual clock can tick whenever any of the local Poisson clock ticks to facilitate the theoretical representation [9]. Further, let V_m denote the m-th tick of the virtual clock, I_m be the index of the node whose local clock actually ticked at that instant. Then, we let J_m be the random index of the node communicating with node I_m. Now, by equipping the variables with the time index $(m-1)$ that node i iterates at time immediately before the virtual clock V_m, we conduct our primal variables and dual variables to be updated at the same time with the aid of gossip operations as follows:

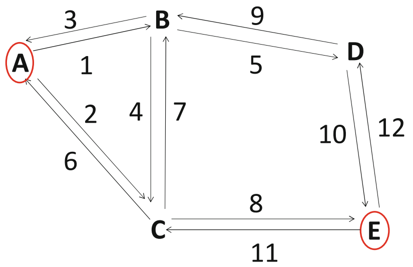
$$\begin{cases} x_S(m) = \left[\bar{x}_S + \epsilon_1(m)\dfrac{\partial L(\bar{\boldsymbol{x}},\boldsymbol{f},\boldsymbol{p},\boldsymbol{u})}{\partial x_S}\right]^+, & \forall S \in \mathcal{S}, \text{if } i \in \left\{I_m, J_m\right\} \\ x_S(m) = x_S(m-1), & \text{if } i \notin \left\{I_m, J_m\right\} \\ u_i^S(m) = \left[\bar{u}_i^S - \epsilon_2(m)\dfrac{\partial L(\boldsymbol{x},\boldsymbol{f},\boldsymbol{p},\bar{\boldsymbol{u}})}{\partial u_i^S}\right]^+, & \forall S \in \mathcal{S}, \text{if } i \in N - \left\{t_S\right\} \text{and } i \in \left\{I_m, J_m\right\} \\ u_i^S(m) = u_i^S(m-1), & \text{if } i \notin \left\{I_m, J_m\right\} \\ u_{t_S}^S(m) = u_{t_S}^S(m-1) = 0, & \forall S \in \mathcal{S} \end{cases} \tag{15}$$

where $[\cdot]_+$ is the projection operator defined as $\max(\cdot, 0)$. Moreover, $\epsilon_i(m), i \in \{1,2\}$, is the step size at time m, and

$$\begin{cases} \bar{x}_S = \dfrac{1}{2}\left(x_{s_{I_m}}(m-1) + x_{s_{J_m}}(m-1)\right) \\ \bar{u}_i^S = \dfrac{1}{2}\left(u_{I_m}^S(m-1) + u_{J_m}^S(m-1)\right) \end{cases} \tag{16}$$

4 Numerical Results

In this section, we provide our numerical results on the asynchronous distributed optimization resolved by the primal-dual gossip algorithm with simulation experiments. To this end, we let $U_i = \sum_{s \in S} \log\left(x_S\right)$ as the utility of node $i \in N$ to be an example to examine. Our simulation environment is then arranged to consist of 5 nodes, $\{A, B, C, D, E\}$, and 12 (directional) links, $\left\{l_1 = (A, B), \ldots, l_{12} = (E, D)\right\}$, as shown in Fig. 1(a), along with 10 transmission modes,$\Xi = \left\{\xi_1 = \left\{l_1, l_8\right\}, \ldots, \xi_{10} = \left\{l_{11}, l_3\right\}\right\}$, as tabulated in Fig. 1(b). With these transmission modes, we particularly conduct the experiment so that the transmission probabilities of certain modes $\left(p_1, p_2, p_9, p_{10}\right)$ are nonempty, for the two sessions $s_1(\text{from} A \text{to} E)$ and $s_2(\text{from} E \text{to} A) \in S$, to reflect the possibility that a link may be involved in two or more transmission modes (such as link 1 in modes 1 and 2, and link 3 in modes 9 and 10, as shown in Fig. 1(b). In addition, to focus on the optimization framework expected to be general enough, we do not consider a specific physical layer, and instead assign each $r_l^k, \forall k \in K, \forall l \in L$, with a real numbers $\in (0, 1)$ to represent the possible value when normalized with respect to a specific model. Next, we let step size be $\epsilon_i(m) = \Gamma_i^m / \gamma + c, \forall i \in N, \forall m > 0$, where Γ_i^m denotes the total number of node i updates up to the virtual clock tick m, and c is a positive constant to avoid a significantly unstable fluctuation at the very beginning that may cause the gossip operations on the sub-gradients with respect to x_{s_i}, u_i^S, or both, to fluctuate with unreasonable values (e.g., dividing by zero). Similarly, γ is used to avoid a fast decade on ϵ_i before reaching a neighborhood region of the optimal.



Number	Tx. mode	Number	Tx. mode
1	{1,8}	6	{9}
2	{1,10}	7	{7}
3	{2}	8	{6}
4	{4}	9	{12,3}
5	{5}	10	{11,3}

Fig. 1. Simulation environments: (a) the topology consisting of 5 nodes and 12 links, and (b) the transmission modes.

4.1 Results on Varying β

In the first set of experiments, we examine how the log-sum-exp function parameter β can impact the convergence behavior of our primal-dual gossip algorithm. To this end, we vary β to change its value among the four scales {50, 100, 200, 300}, while fixing $\gamma = 100$ and $c = 20$. Specifically, to see the convergence behaviour, we let continuously 10 times when the normalized errors on $x_S, s \in \{1, 2\}$, are less than 1% be the convergence achieved in a simulation experiment. Given that, in Fig. 2, we exhibit the results from a typical scenario of $\beta = 300$ along with $\gamma = 100, c = 20$. As shown there, the five nodes can independently converge to the global optimal session rates $x_S, s \in \{1, 2\}$, by individually adopting our primal-dual gossip algorithm. Apart from the example, similar results can be also observed from the others with different parameters. Specifically, in Fig. 3, we adopt an error bar representation to show both mean and standard deviation for the five nodes involved on their numbers of virtual clock ticks to converge with different β.

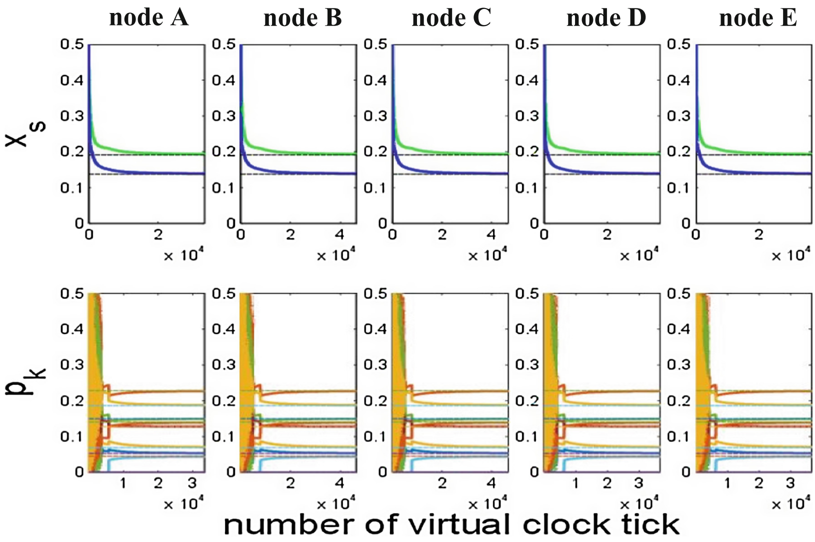


Fig. 2. Convergence behavior of a typical scenario.

From this figure, we can see that increasing β could increase the convergence speed owing to a better approximation to $p^*_{k_i}$. However, its effect is not obvious, and a higher β might only marginally decrease the number of virtual clock tick to converge while potentially increasing the numerical difficult for implementation. So, we restrict ourselves to consider only $\beta = 300$ in the next set of experiments.

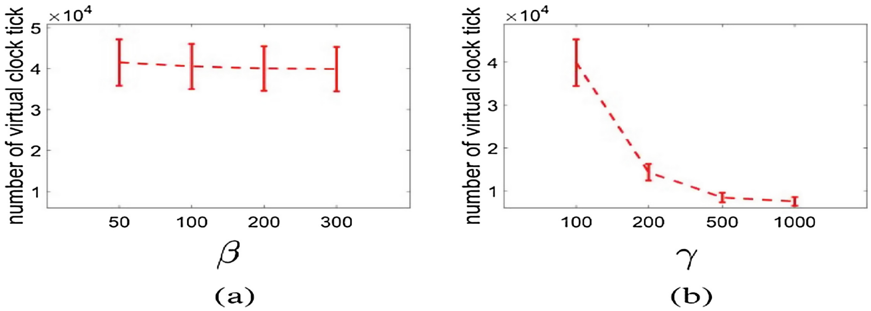


Fig. 3. Experiment results: (a) varying β, and (b) varying γ.

4.2 Results on Varying γ

In the second set of experiments, we vary γ with the four scales {100, 200, 500, 1000}, while fixing $\beta = 300$ and $c = 20$ to see its impact on the convergence behavior of $x_S, s \in \{1, 2\}$. The results are now summarized in Fig. 3, showing that a slower decreasing step size, or say, a higher γ could significantly increase the convergence speed in this case when comparing with β that might only marginally contribute to the speedup in the previous set of experiments. However, when the step size decreases too slowly, it will eventually approach an uncoordinated constant step size that can only guarantee a limiting error bound on the optimal value.

5 Conclusion

In this work, we have developed a distributed optimization algorithm specific to the cross-layer utility maximization problem by using a gossip-based formulation for the consensus agreement on the variables involved. Our numerical results have exhibited the correctness of this programming model and readily showed that the gossip-based algorithm can achieve the optimization objective by means of the simple and robust asynchronous operations just developed. Specially, by varying the two major parameters, β and γ, we have examined and found the prefer values that can result in a better convergence performance in the typical scenario of wireless networks.

Acknowledgment. This work was supported by the National Science Council, Republic of China, under grant MOST 105-2221-E-126-001-MY3.

References

1. Tsitsiklis JN (1984) Problems in decentralized decision making and computation. Ph.D. Dissertation, Mass. Institute of Technology, Cambridge, MA, pp 1–272
2. Jadbabaie A, Lin J, Morse S (2003) Coordination of groups of mobile autonomous agents using nearest neighbor rules. IEEE Trans Autom Control 48(6):988–1001

3. Xiao L, Boyd S, Kim SJ (2005) Distributed average consensus with least-mean-square deviation. J Parallel Distrib Comput 67:33–46
4. Lobel I, Ozdaglar A Distributed subgradient methods for convex optimization over random networks. LIDS, MIT, Technical report 2800
5. Leinonen M, Codreanu M, Juntti M (2012) Distributed consensus based joint resource and routing optimization in wireless sensor networks. In: Proceedings of ASILOMAR 2012, pp 811–815, November 2012
6. Mota JFC, Xavier JMF, Aguiar PMQ, Puschel M (2012) ADMM for consensus on colored networks. In: IEEE Annual conference on decision and control (CDC), pp 5116–5121, December 2012
7. Wei E, Ozdaglar A (2013) Distributed alternating direction method of multipliers. In: Proceedings of the 46th annual conference on information sciences and system, pp 5445–5450
8. Chen M, Liew S, Shao Z, Kai C (2010) Markov approximation for combinatorial network optimization. In: Proceedings of IEEE INFOCOM 2010, pp 1–9
9. Cruz RL, Santhanam AV (2003) Optimal routing, link scheduling and power control in multihop wireless networks. In: Proceedings of IEEE INFOCOM 2003, vol 1, pp 702–711
10. Madan R, Cui S, Lal S, Goldsmith A (2006) Cross-layer design for lifetime maximization in interference-limited wireless sensor networks. IEEE Trans Wireless Commun 5(11): 3142–3152
11. Nama H, Chiang M, Mandayam N (2006) Utility-lifetime trade-off in self-regulating wireless sensor networks: a cross-layer design approach. In: Proceedings of IEEE international conference on communications, vol 8, pp 3511–3561
12. Palomar DP (2006) Chiang M A tutorial on decomposition methods for network utility maximization. IEEE J Sel Areas Commun 24(8):1439–1451

Call Admission Control for Real-Time and Non-real-time Traffic for Vehicular LTE Downlink Networks

Mamman Maharazu(✉), Zurina Mohd Hanapi, Azizol Abdullah, and Abdullah Muhammed

Department of Communication Technology and Networking,
Faculty of Computer Science and Information Technology, University Putra Malaysia,
UPM Serdang Selangor, 43400 Seri Kembangan, Malaysia
maharazu2003@yahoo.com, {zurinamh,azizol,abdullah}@upm.edu.my

Abstract. Provision Quality of Service (QoS) is a crucial challenge in any vehicular multimedia wireless application. In wireless applications, maintaining QoS requirements of various calls is a more challenging issue because of the concurrent need to prioritize the handoff calls and new calls trying to access the network. In this paper, a novel call admission control scheme is proposed which prioritizes calls and differentiates the calls types into real-time and non-real-time traffic. By doing so, the scheme reduces both the handoff blocking probability, new call dropping probability, and improves data throughput utilization. Experimental results reveal the outstanding performance of the proposed scheme as it is able to achieve a better call blocking and call dropping probabilities compared to Non-prioritize scheme.

Keywords: Call admission control · Call blocking probability · Call dropping probability · Handoff · Quality of service

1 Introduction

The next generation wireless network support different types of services with diverse QoS requirements. These services are confronted with severe issues because of the increasing exponential growth of mobile users and different types of applications [1]. In response to this, the third Generation Partnership Project (3GPP) adopted the Long Term Evolution (LTE). LTE employs Orthogonal Frequency Multiple Access (OFDMA) and Multi-user Multiple-Input Multiple-Output (MU-MIMO) technologies to increase users high data rate, provide wide area coverage, and improved spectral efficiency [2]. The fundamental objective of LTE is to guarantee QoS requirements and minimize network congestion for different types of users; therefore Radio Resource Management (RRM) is needed in such situation.

Radio Resource Management (RRM) consists of important functionalities such as scheduling algorithm and Call Admission Control (CAC). CAC is the process of accepting new call or handoff call in the network while regulating QoS of the existing calls without degrading any call drops. Handoff call refers to the method of transmitting a User Equipment (UE) from one Evolved NodeB (enodeB) to another enodeB without

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_6

compromising users QoS requirements [3]. Researchers and vendors in the LTE community are left to design and implement their own CAC scheme as the LTE standard did not specify any standard for CAC scheme and scheduling algorithm [4].

CAC is an active research area, many CAC schemes haven been proposed for LTE networks to handle many challenging issues from different outlooks. For instance, the work in [5] proposed a connection access control scheme and resource allocation algorithm for LTE systems with heterogeneous services. The algorithm introduces a transmission guard interval technique which gives high priority to real-time (RT) service packets approaching the delay deadline. The algorithm uses an adaptive threshold to adjust the network condition based on CAC criteria on the available resources per time slot. However, handoff call (HC) and new call (NC) are treated in the same way. An adaptive QoS oriented CAC scheme for the 4G wireless network was proposed in [6]. The algorithm uses a hybrid approach in considering vertical handoff function, service type differentiation, and call origination point to maximize the utilization of available resources. It achieves and guarantees QoS requirements for both RT users and non-real-time (NRT) users.

A novel CAC scheme which guarantees QoS for all users was proposed in [7]. The algorithm uses adaptive modulation to provide good utilization of system resources. To satisfy user satisfaction and a number of slots needed by each user, the algorithm scheduled users based on user priority function. However, the algorithm ignored the effect of buffer status which increases the number of users in the network and hence resulted in network congestion. Previous work in the literature tried to optimize the handover performance and CAC scheme. The authors in [8] suggested self-optimisation of CAC scheme and handover in the LTE environment. The algorithm uses guard channel strategy to reserve resources for HC. It uses a dynamic threshold parameter to achieve an optimum value for call blocking probability (CBP) and call dropping probability (CDP). However, the strategy used to set the dynamic threshold parameter ignored minimum QoS requirement.

In [9] an efficient channel state based CAC for NRT traffic in LTE networks is proposed. The algorithm categorized call requests into HC, NC, Voice over Internet Protocol (VoIP) call and video call which are prioritized. The HC and NC are distinguished based on call classification, channel status estimation and call admission. The algorithm uses Received Signal Strength (RSS) for channel status estimation. At initial point, an optimal RSS value is considered as a minimum threshold value. RSS of the channel is calculated repeatedly and is compared with the minimum threshold value. When the calculated RSS value is greater than the minimum threshold value then the channel status is considered a good channel state otherwise bad channel state. However, the algorithm favours NCs over HCs this resulting in its inability to guarantee QoS requirement. The work in [10] proposed a CAC scheme that depend on adaptive multi-level bandwidth allocation of NRT calls. The algorithm utilizes the available radio resources to provide QoS. However, channel control allocation is ignored which resulted in high CBP and low resource utilization.

An adaptive CAC algorithm for 3GPP LTE networks is proposed in [11]. The algorithm reduces the CBP by means of an adaptive strategy which prioritizes the HC over NC. The scheme achieves better performance in term of low CBP. However, the QoS requirement is only guaranteed for HC while that of NC was neglected. The work in [12] suggested CAC schemes with new call reattempts. The algorithm allowed NC reattempts if a caller fails to successfully establish a connection with the main network after a given period of time. However, the algorithm generates extra load on the system performance, which results in an increase of CBP, CDP, and Call Reattempt Probability (CRP). In [13] the authors proposed an efficient CAC scheme based on adaptive resource reservation. The scheme dynamically adjusted users' priority based on the present network status. Users are categorized as Golden users and Silver users, and the type of service per user is classified as RT and NRT services. The scheme achieved a better balance between users' privileges, QoS provisioning and system utilization compared to other scheme.

The work in [14] proposed CAC algorithm for QoS provisioning for multimedia traffics in cellular networks. For the purpose of this paper henceforth we call Ref. [14] as Non-prioritize scheme. The scheme maintains the CDP under a certain threshold value to attain maximum channel utilization. But, both HC and NC are given equal treatment this causes an increase of CBP and CDP which degraded the performance of the network. In this paper, the proposed novel CAC scheme named "Call Admission Control for Real-Time and Non-Real-Time Traffic in Vehicular LTE Downlink Networks" aims to remedy the inefficiency of [14].

The rest of this paper is organized as follows: Sect. 2 describes the LTE system architecture and frame structure. Section 3 presents the proposed algorithm. Experimental results are illustrated in Sect. 4. Section 5 concludes the paper and outlines of the future work.

2 LTE System Architecture and Frame Structure

The LTE architecture is described in Fig. 1 which generally consists of two parts: (1) the Radio Access Network (RAN) and (2) core networks also called Evolved Packet Core (EPC). The EPC connects the RAN with other IP network such as the Internet and is responsible for routing, scheduling, mobility, handover processes, and CAC [15]. The LTE RAN has two types of nodes: enodeB that serve as base station and UE. Each enodeB offers access for a certain number of UEs over a wireless channel, which creates a cell. The CAC schemes are used to manage the calls for diverse UEs connected to the same enodeB. The UEs are uniformly distributed in the cell, and each UE can be classified into different kinds based on their QoS requirements and call type. In this work, the network consists of call types which include Handoff a call and new call. Every call type can be classified to serve different traffic classes which are categorized into RT and NRT. We assume that the requests from RT users have the highest priority, while NRT user requests are considered as the lowest priority.

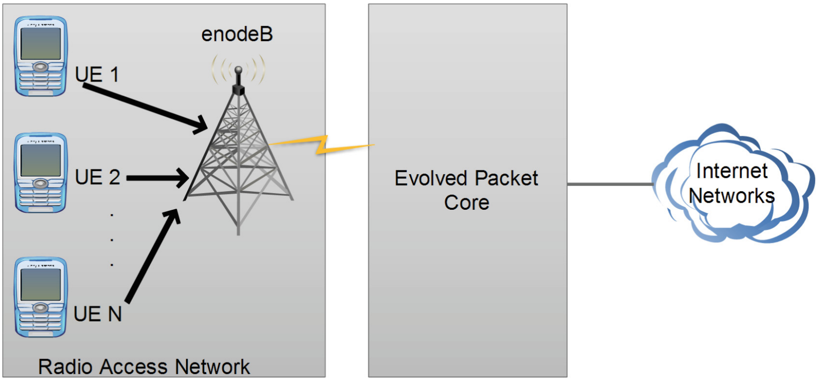


Fig. 1. LTE system architecture

The frame structure for LTE downlink consists of 10 ms radio frames, each frame is divided into ten subframes of 1 ms, while each of the subframes is divided into 0.5 ms slots. Each slot contains 7 Orthogonal Frequency Division Multiplexing (OFDM) symbols for normal cyclic prefix and 6 for extended cyclic prefix [15]. Moreover, the frequency domain consist of resources which are grouped into 12 subcarriers (180 kHz) and subcarrier of 15 kHz spacing; each one unit of the 12 subcarriers for the duration of one slot (0.5 ms) is called a resource block. A Resource block is the minimum element of resource allocation assigned by the enodeB. The resource element corresponds to one complex value modulation symbol in OFDM.

3 Proposed Algorithm

In this paper, we proposed a novel CAC scheme “Call Admission Control for Real-Time and Non-Real-Time Traffic in Vehicular LTE Downlink Networks” which is an improvement of Non-prioritize scheme. Firstly, the limitations of Non-prioritize scheme are outlined. The scheme maintains the CDP under a certain threshold value to attain maximum channel utilization. But, both HC and NC are given equal treatment this causes an increase of CBP and CDP which degraded the performance of the network. Furthermore, to address the shortcoming of the Non-prioritize scheme, our scheme prioritized the calls into HC and NC. The HC has the highest priority and in this case we considered the RT traffic class. On the other hand, the NC has the lowest priority hence NRT are the best example. When calls arrive into the network, the algorithm categorizes them into HC and NC. At this stage, prioritization takes place in which the HC gets the highest priority. For both HC and NC, the intensity of the traffic is calculated which can be in twofold: low and high. A threshold value is computed based on Eq. 1. When the threshold value is less or equal to the intensity of traffic, then HC or NC calls are accepted, otherwise, the calls are rejected. Figure 2 presents the block diagram on how the proposed algorithm operates.

$$th_{value} = \beta * P_{HC} + P_{NC}. \tag{1}$$

where β is equal to 10, P_{HC} and P_{NC} denotes the probabilities of HC and NC respectively.

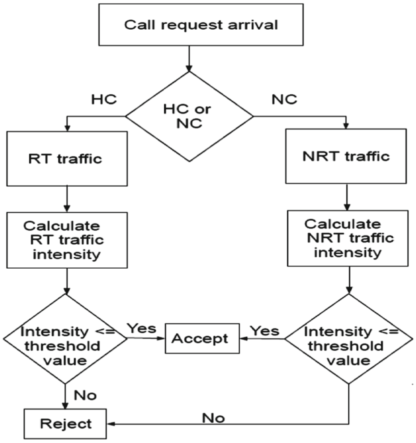

Call request arrival
HC
HC or NC
NC
RT traffic
NRT traffic
Calculate RT traffic intensity
Calculate NRT traffic intensity
Intensity <= threshold value
Yes
Accept
Yes
Intensity <= threshold value
No
Reject
No

Fig. 2. Block diagram of the proposed algorithm

4 Experimental Results

This section analyses the findings and experimental results to evaluate the performance of the proposed scheme. The experimental results are obtained with the help of system level simulator. The performance of the proposed scheme is measured in terms of call dropping probability, call blocking probability and data throughput. The simulation scenario consists of one hexagonal cell with 500 m radius. The total bandwidth used is 5 MHz with 25 resource block per slot of 12 subcarriers spacing. The classifications of traffic are allocated between RT and NRT users. The arrival rate for both RT and NRT is a Poisson distribution with the service mean exponentially distributed. Table 1 summarizes the simulation parameters.

Table 1. Simulation parameters

Description	Value
Bandwidth frequency	5 MHz
Number of RBs	25
Distance between enodeB	500 m
User distrubution	Uniform
Simulation time	50 TTIs (50 ms)
Average user speed	4.16 m/s

Figures 3 and 4 depict the results of call dropping probability and call blocking probability. It can be observed that the proposed scheme has the lowest probabilities due to the prioritizing of user call request by applying the formula (1) which gives a higher priority to HC while NC has the lowest priority. At arrival rate of 0–0.1 (call/sec), both scheme indicate similar performance. But when the arrival rate was increased from 0.1–0.8 (call/sec) the proposed algorithm outperformed the Non-prioritize scheme with a decrease of CDP of 22.88% and CBP of 16.01%. Hence, we can conclude that the proposed scheme guarantee the QoS requirement for vehicular multimedia wireless systems.

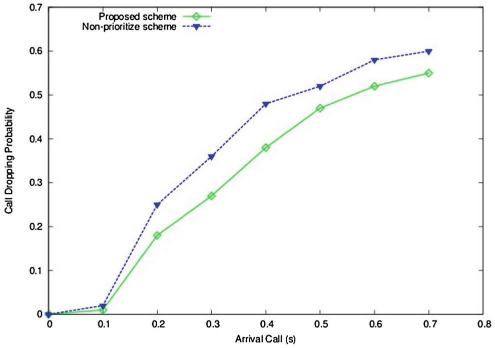


Fig. 3. Call dropping probability

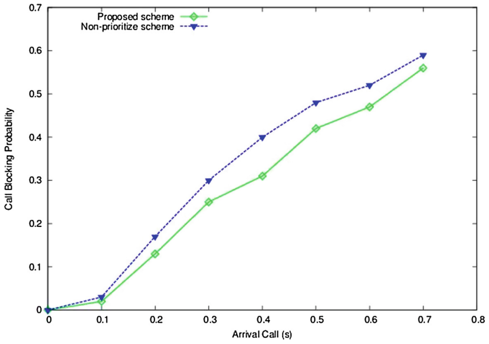


Fig. 4. Call blocking probability

Data throughput is compulsory for a fruitful communication. Figure 5 shows the data throughput for the two schemes. From the figure we can observe that the Non-prioritize scheme cannot fully utilize the available resources efficiently due the Non-prioritization of user calls. This condition resulted in high CDP, CBP and low data throughput. The proposed scheme adopted prioritization of user call; hence this causes low CDP (22.88%), CBP (16.01%) and an increase in data throughput with 22.67% improvement compared with Non-prioritize scheme.

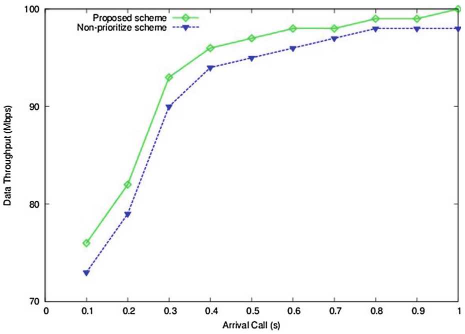


Fig. 5. Data throughput

5 Conclusion and Future Directions

In this work, we propose a novel call admission control scheme named "Call Admission Control for Real-Time and Non-Real-Time Traffic in Vehicular LTE Downlink Networks" that aims at accepting or rejecting user based on prioritization of calls. In this proposed scheme, the types of calls are classified into HC and NC calls, and the traffic requests are categorized into RT and NRT. Extensive experimental results using the system level simulator shows that the proposed scheme outperformed the Non-prioritize scheme with decrease of CBP (16.01%), CDP (22.88%), and improved data throughput (22.67%). As part of future work, we plan to extend the work using different load scenarios and test its performance in an analytical study.

Acknowledgements. This work was supported by research grant UPM-FRGS-08-02-13-1364FR.

References

1. Ramkumar MV, Nielsen RH, Stefan LA, Prasad NR, Prasad R (2013) A joint allocation, assignment and admission control (AAA) framework for next generation networks. Wirel Pers Commun 73(3):1245–1267
2. Semiconductor, F (2008) Long term evolution protocol overview, in white paper, documents NO LTEPTCLOVWWP, Rev 0, pp 1–18
3. Ghaderi M, Boutaba R (2006) Call admission control in mobile cellular networks: a comprehensive survey. Wirel Commun Mob Comput 6(1):69–93
4. 3GPP (2012) 3GPP, Tech. Speci_c. Group Services and System Aspects - Policy and charging control architecture (Release 11) 3GPP 23.203
5. Lei H, Yu M, Zhao A, Chang Y, Yang D (2008) Adaptive connection admission control algorithm for LTE systems. In: Vehicular technology conference (VTC Spring), pp 2336–2340. IEEE, Canada
6. Bejaoui T, Nasser N (2009) Efficient call admission control scheme for 4G wireless networks. Wirel Commun Mob Comput 9(4):489–499

7. Ramkumar MV, Anggorojati B, Stefan AL, Prasad NR, Prasad R (2010) QoS guaranteed admission control for OFDMA-based systems. In: IEEE international workshop on management of emerging networks and services, pp 606–610. IEEE, San Francisco
8. Sas B, Spaey K, Balan I, Zetterberg K (2011) Self-optimisation of admission control and handover parameters in LTE. In: 73rd IEEE Vehicular technology conference (VTC Spring), pp 1–6, Budapest, Hungary
9. Franklin JV, Paramasivam K (2012) Efficient channel state based call admission control for non real time traffic in LTE (3GPP) networks. Int J Comput Sci 9(2):231–237
10. Chowdhury MZ, Jang MZ, Haas ZJ (2013) Call admission control based on adaptive bandwidth allocation for wireless networks. J Commun Networks 15(1):15–24
11. Zarai F, Ali KB, Obaidat MS, Kamoun L (2014) Adaptive call admission control in 3GPP LTE networks. Int J Commun Syst 27(10):1522–1534
12. Abdulova V, Isik A (2015) Performance evaluation of call admission control schemes with new call reattempts in wireless cellular. Wirel Pers Commun 84(4):2859–2879
13. Alqahtani SA (2016) Users classification-based call admission control with adaptive resource reservation for LTE-A networks. J King Saud Univ – Comput Inf Sci 29(1):103–115
14. Konte R, Kumar R, Singh J, Tiwari M (2014) Performance analysis of handoff calls for quality of service using call admission control algorithm for data, video and voice traffic. Int J Comput Appl 95(2):19–22
15. Stefania S, Issam S, Matthew B (2011) LTE: The UMTS Long Term Evolution from Theory to Practice. Wiley, Hoboken

Using Weighted Based Feature Selection Technique for Android Malware Detection

Nurul Hidayah Mazlan and Isredza Rahmi A. Hamid(✉)

Information Security Interest Group (ISIG), Faculty Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Johor, Malaysia
hi150018@siswa.uthm.edu.my, rahmi@uthm.edu.my

Abstract. Recently, the popularity of mobile devices has risen drastically due to the increased functionality of the devices. This matter forces a large number of security challenges that need high consideration. Android malware detection method can be divided into two types, which are static and dynamic analysis. Static techniques are often prone to high false negative rates due to evolution in code basis and code repacking, although fast and efficient. While dynamic and behavior based analysis aims to provide methods for effectively and efficiently extracting unique patterns of each malware family based on its behavior. To address some of those shortcomings, the study uses permission-based Android malware feature as a basis for malware detection using weighted based technique.

Keywords: Android malware · Detection · Weighted-based · Term Frequency-Inverse document frequency

1 Introduction

At present, most of users prefers to use Android based mobile devices due to easy access to many applications such as web browsing, social networking and online banking transaction [1]. There are 3.79 billion mobile devices users out of 7.395 billion total global populations reported [2]. Meanwhile in Malaysia, there are 43.43 million mobile from 30.54 million total populations. This shows that a person may have more than one mobile device where 59% is mobile Internet user. This phenomenon forces a large number of security challenges such as leak of sensitive data, network or file system to be addressed. As the rapid growth of smartphones, malware that target popular mobile devices platform also widespread [3].

Smart mobile devices have an operating system just like a computer with additional capabilities and capacities. Moreover, the mobile device has combination of network connectivity with high-speed data networking capabilities and geo-location services [4]. Users can easily being forced to subscribe message or calls, remote control of money transfer and extortion to ransomware. G Data reported 440,000 new Android malware strains in the first quarter of 2015 which shows that each 18 s, a new mobile malware strain for Android is discovered [5]. As a result, mobile devices become main target for the malware attackers to seeking confidential information. This paper uses permission

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_7

based analysis of Android malware as basis for malware detection. Objectives of this study are:

1. to propose the Android Malware detection model using weighted-based approach,
2. to develop weighted based feature selection approach to detect Android malware, and
3. to analyze the proposed feature selection approach and existing approaches by using performance metric to classify Android malware.

The remainder of this paper is organized as follows. Section 2 describes related research regarding Android malware detection approaches. Section 3 examines the Android malware feature selection approach pertaining the data and feature set used in the experiment and Term Frequency-Inverse Document Frequency (TF-IDF) algorithm as well. Section 4 gives the performance analysis result and the effectiveness of the proposed weighted-based feature selection. Section 5 concludes the work and direction for future work is discussed.

2 Related Work

This section will examine two feature selection approaches for Android malware, namely, static analysis and dynamic analysis.

2.1 Static Analysis

Static analysis will inspect mobile application and disassemble the code [6]. Two main techniques in static analysis are decompiling and data flow tracking. Thus, this analysis is fast and fairly easy, static analysis requires regular updates of threat databases and it may be evaded by complicated techniques. Varsha *et al.* [7] proposed a broad static analysis system to classify the Android malware application. To generate vector space model, hardware components, permissions, application components, filtered intents, op-codes and number of small files per application are used as features which are selected using Entropy based Category Coverage Difference. Support Vector Machine (SVM), Rotation Forest and Random Forest are used for system performance evaluation. They achieved 98.14% accuracy using Random Forest classifier tested on 198 of feature length. However, opcodes are prone to obfuscation which could be handled by implementing a normalizer.

Other static analysis approach is DeDroid which investigated botnet-specific properties used to detect mobile applications with botnet intensions [8]. Command and control features associated with four well-known malware families including DroidKungFu, Plankton, GoldDream, and Geinimi has been examined. Static analysis is performed using reverse engineering applications by taking five samples from each malware family. The first evaluation was run on 5064 malware binaries belong to 20 malware families. The result shows that 1795 malware samples have been detected having command and control features. Top six malware families with the highest detection ratio have been taken to validate the results where FakeRun achieved the highest accuracy with 100% value.

2.2 Dynamic Analysis

Dynamic analysis provides new methods for extracting the malware patterns effectively. This method focused on time when the Android application are being executed either

Table 1. Comparisons of related works

Work	Approach	Technique	Features	Sample	Accuracy
Varsha [7]	Static Analysis	Entropy based Category Coverage Difference	Hardware components, Requested permissions, Application components, Filtered intents, Op-codes	Drebin, Google Play Store	98.14%
Karim [8]	Static Analysis	Reverse Engineer	Permissions, API calls	Drebin	**First evaluation:** FakeRun: 100%, FakeDoc: 98%, DroidKungFu:78%, Plankton: 84%, Geinimi: 90%, GoldDream 80% **Second evaluation:** Geinimi: 93%, GoldDream: 89%, Plankton: 82%, Kungfu: 78%, DroidKungfu: 69%
Zhou *et al.* [9]	Dynamic Analysis	Permission based behavioral footprinting, Heuristics based filtering	Permissions	Android Market	Not available
Enck *et al.* [10]	Dynamic Analysis	System-wide dynamic taint tracking	System calls, network access	Android Market	Not available
Li *et al.* [6]	SVM based (Machine Learning)	Weight Calculation	Permissions, API calls	Drebin, Google Play Store	API calls: 81% API calls and risky permission combinations: 86%
Wu *et al.* [11]	Dynamic analysis	Reverse Engineer	API Calls	Google Play Store	86.1%

accessing private data or using Application Program Interface (API) calls [6]. Dynamic analysis undergoes offline analysis because of the large amount of computational overhead. To detect new samples of known Android Malware families, DroidRanger [9] propose a permission based behavioral foot-printing scheme. The DroidRanger identified certain inherent behaviors of unknown malicious families on 204,040 applications collected from five different Android Markets.

TaintDroid [10] used System Call information and other network access to do real-time analysis by leveraging Android's virtualized execution environment. TaintDroid is capable to track multiple sources of sensitive data and monitor the behavior of 30 popular third-party Android applications simultaneously. TaintDroid manage to identify misbehaving applications by monitoring flow of privacy data.

Li *et al.* [6] integrated risky permission combinations and vulnerable API calls as features in the Support Vector Machine (SVM) algorithm. The small files are analyzed and the weight of every dangerous API in the feature vector is calculated using Term Frequency-Inverse Document Frequency (TF-IDF) algorithm. The SVM-based malware detection that contribute dangerous API calls achieved 81% accuracy, while both dangerous API calls and risky permission combinations achieve 86% accuracy.

DroidDolphin [11] extracted useful static and dynamic features to detect malicious applications. This analysis involved GUI-based testing, big data analysis, and machine learning. DroidDolphin Architecture code collects the run-time logs of an Android application and decides whether it is a malware or not using machine learning techniques. The preliminary experiment used 32,000 benign and 32,000 malicious applications as training data, and 1,000 benign and 1,000 malicious applications as testing data. The results showed that the prediction accuracy reaches 86.1%.

Although there are clear advantages to detect Android malware as shown in Table 1, there are at present not many methods specifically designed based on dynamic approach. Our study differs from the previous work on feature selection in several ways. We propose a dynamic feature selection by using weighted based technique on permission-based features. We considered analyzing permission-based Android malware features in order to evaluate the malware behaviors. We then choose to use Random Forest algorithm as our classifier.

3 Android Malware Feature Selection Approach

In this section, we discuss the proposed weighted based technique for Android malware detection approach. We will first introduce the model and feature selection and extraction process.

3.1 Android Malware Detection Model

Figure 1 shows the Android malware detection model and general processing steps. The processing phases includes: preprocessing of the Android malware dataset, feature extraction and selection, dynamic feature selection, classification using machine learning algorithm, and finally the evaluation of the detection result. The model is generated by following general data mining approach which aims to build a classifier for Android malware. The classifier should be able to class correctly all the sample either malware or benign.

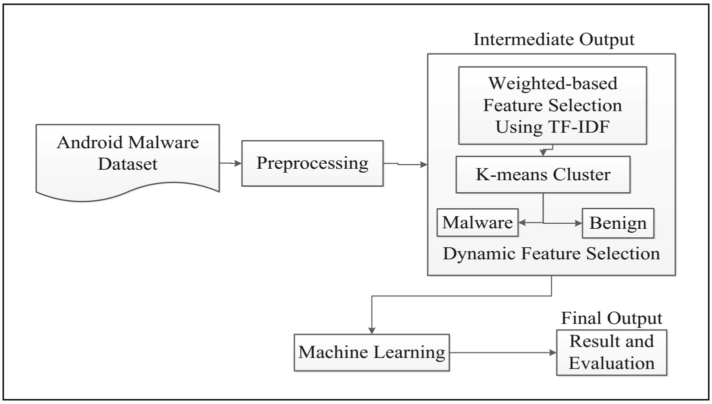


Fig. 1. Android malware detection model

The proposed model is evaluated using Drebin [12] dataset. We consider both samples from benign and malicious application. There are various Android dangerous features such as *read_calendar, read_contacts, access_fine_location, receive_sms* and *read_external_storage*. All samples are extracted into human readable format (.xml). Then, the samples will undergo the preprocessing process where all data are cleaned up and normalized. Figure 2 shows the data will go through feature selection process by implementing dynamic feature selection approach using weighted-based feature selection technique. Data will be classified into malware or benign using K-means cluster.

In machine learning phase, we used Random Forest [19] as a classification algorithm. The models will be generated and trained using Waikato Environment for Knowledge Analysis, WEKA [13]. Random Forest algorithms are selected because the algorithm works based on combination of tree predictors. Class with the highest vote is considered the best output by considering the voted classes of all individual trees. Moreover, Random Forest algorithm is good for complex classification tasks. It has methods for balancing error in class population that unbalanced. Finally, two datasets are constructed namely as *Experiment 1* and *Experiment 2*. We use the same sets of data as our main focused in this paper is to propose the weighted based feature selection approach.

3.2 Feature Selection and Extraction

Sample of Android data that have been extracted into.xml file will undergo feature selection process. The next step in the process is to generate components of a feature vector by analyzing the database. In feature selection, irrelevant and redundant features will be removed [7]. The dangerous permission is selected as feature because each application installation will request the permission to use certain systems data and features [14]. Functionality of Android application will be exposed to other application through request permissions. Moreover, application with dangerous permission can access user private information and affect the stored data or operation of others application. Table 2 shows the description of 18 Permission-based Android malware features that will be used in the experiment. Then, the permission-based feature is selected derived from weight value calculated using TF-IDF and the proposed feature selection algorithm.

Table 2. Types of dangerous permission-based feature [14]

Type	Permission	Description
Calendar	*read_calendar*	Allows an application to read the user's calendar data.
	write_calendar	Allows an application to write the user's calendar data.
Camera	*camera*	Required to be able to access the camera device.
Contacts	*read_contacts*	Allows an application to read the user's contacts data.
	write_contacts	Allows an application to write the user's contacts data.
	get_accounts	Allows access to the list of accounts in the Accounts Service.
Location	*access_fine_location*	Allows an application to access precise location.
	access_coarse_location	Allows an application to access approximate location.
Microphone	*record_audio*	Allows an application to record audio.
Phone	*read_phone_state*	Allows read only access to phone state, including the phone number of the device, current cellular network information, the status of any ongoing calls, and a list of any PhoneAccounts registered on the device.
	call_phone	Allows an application to initiate a phone call without going through the Dialer user interface for the user to confirm the call.
	process_outgoing_calls	Allows an application to see the number being dialed during an outgoing call with the option to redirect the call to a different number or abort the call altogether.
SMS	*send_sms*	Allows an application to send SMS messages.
	receive_sms	Allows an application to receive SMS messages.
	read_sms	Allows an application to read SMS messages.
	receive_mms	Allows an application to monitor incoming MMS messages.
Storage	*read_external_storage*	Allows an application to read from external storage.
	write_external_storage	Allows an application to write to external storage.

3.3 Term Frequency Inverse Document Frequency (TF-IDF) Algorithm

Based on vector space model, a textual file is represented as bag of word in text categorization [15]. A vocabulary for the whole collection of word is constructed after analyzing the text and extracting the words. A vector of terms extracted from the whole set of documents, thus the vocabulary is maintained. Term Frequency (TF) creates document for each vector of terms that match the vocabulary. Each cell in the vector will represents the TF in the corresponding document as shown in Eq. 1. Document Frequency (DF) is characterized term in the vocabulary using its frequency in the whole collection. TF and DF are two important terms in TF-IDF algorithm. The evolution of TF representation from counting words was replaced by other methods for representing executable files such as byte-sequence n-grams, in malware classification analysis.

$$TF = \frac{term\ frequency}{\max\ (term\ frequency\ in\ document)} \tag{1}$$

$$TF - IDF = TF \times \log\left(\frac{N}{DF}\right) \tag{2}$$

Next, TF extended into Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF combine term's frequency in the document (TF) and its frequency in the document's collection, called as Document Frequency (DF) as shown in Eq. 2. Then, the normalized TF value is multiplied by $IDF = \log\left(\frac{N}{DF}\right)$ where N is number of documents in the entire file collection, while DF is number of files in which the term appears. By using TF-IDF, the weight of specification behavior is calculated.

SVM-based approach [6] altered the TF-IDF formula to analyze and calculate the weight of dangerous API calls. TF_{ij} defined in the Eq. 3, where N_{ij} is the number of times that Android application D_j calls specific dangerous API i. M_j is the total number of times D_j calls all different dangerous API.

$$TF_{ij} = \frac{N_{ij}}{M_j} \tag{3}$$

IDF_i in Eq. 4 represent D as the total number of malware in the training dataset. D_i is the number of times that a certain dangerous API is called.

$$IDF_i = \log\left(\frac{D}{D_i + 1}\right) \tag{4}$$

3.4 Weighted Based Feature Selection Algorithm

The proposed weighted based feature selection technique is modified based on the TF-IDF algorithm where our approach focused on both sample and feature. The following equations explain the modified TF-IDF where Eq. 5 shows that tf_{ij} is the value of Android malware feature's i in sample j. While, $\max(tf_i)$ is a maximum value of Android malware feature's i in all sample.

$$TF_{ij} = \frac{tf_{ij}}{\max(tf_i)} \tag{5}$$

IDF in Eq. 6 represent N as the total number of Android malware feature in dataset, where n_j is number of occurrence of Android malware feature appear in sample j.

$$IDF_j = \log\left(\frac{N}{n_j}\right) \tag{6}$$

Therefore, the weight equation is defined in Eq. 7 where W_i is weighted calculation by multiply the amount of TF and IDF.

$$W_i = TF_{ij} \times IDF_j \tag{7}$$

4 Performance Analysis

This section explained the experimental setup of Android malware detection using feature selection methods.

4.1 Experimental Setup

In our study, the detection was performed using WEKA. We used 500 permission-based samples of dataset consist of malware and benign from Drebin [12] for the experiments. The permission-based sample consists of 18 features tested on two types of experiment denoted as *Experiments 1* and *2*. The *Experiment 1* is permission-based features tested using TF-IDF algorithm while *Experiment 2* use weighted-based Feature Selection algorithm as feature selection approach.

4.2 Performance Metric

In order to measure the effectiveness of the detection approach, we refer to four possible outcomes as: Accuracy, Precision, Sensitivity and F-score. Accuracy in Eq. 8 shows the probability of the class label value to assess the effectiveness of the algorithm. To assess the predictive power of the algorithm, precision estimate the predictive value of a label depending on the class. Main evaluation parameter, F-score, is a composite measure which benefits algorithms with higher sensitivity. It is a weighted average of precision (P) in Eq. 10 and Recall (R) in Eq. 11 calculated using Eq. 9. During the result analysis, True Positive Rate (TPR) and False Positive Rate (FPR) are measured. True positive (TP) is malware classified as malware, while false positive (FP) is benign being misclassified as malware. True Negative

(TN) is benign classified as benign while false negative (FN) is malware misclassified as benign.

$$Accuracy = (TP + TN) / (TP + FN + TN + FP) \tag{8}$$

$$F - score = (2PR) / (P + R) \tag{9}$$

$$Precision\ (P) = (TP) / (TP + FP) \tag{10}$$

$$Recall\ (R) = (TP) / (TP + FN) \tag{11}$$

4.3 Result and Discussion

The experiment is tested on 500 dataset, which is then constructed two sets of experiment. Data for *Experiment 1* consists of 10 features that have the highest TF-IDF value. *Experiment 2* contains 10 features with highest weighted-based Feature Selection algorithm value. Random Forest classifier with 10 folds cross-validation has been used for both experiments. Table 3 shows list of permission-based features selected by TF-IDF algorithm and weighted-based feature selection algorithm.

Table 3. Permission-based features selection approaches

Rank	TF-IDF	Value	Rank	Weighted-based feature selection	Value
1	*process_outgoing_calls*	2.096910013	1	*access_coarse_location*	1.255272505
2	*receive_mms*	2.000000000	2	*access_fine_location*	1.255272505
3	*read_external_storage*	1.657577319	3	*call_phone*	1.255272505
4	*write_calendar*	1.657577319	4	*camera*	1.255272505
5	*read_calendar*	1.619788758	5	*get_accounts*	1.255272505
6	*record_audio*	1.619788758	6	*read_phone_state*	1.255272505
7	*get_accounts*	1.585026652	7	*record_audio*	1.255272505
8	*write_contacts*	1.468521083	8	*send_sms*	1.255272505
9	*read_sms*	1.397940009	9	*write_external_storage*	1.255272505
10	*receive_sms*	1.376750710	10	*process_outgoing_calls*	0.954242509

Table 4. Evaluation performance using random forest classifier

Dataset	Class	TP rate	FP rate	Precision	Recall	F-Measure	Accuracy
Experiment 1	Benign	1	0.12	0.994	1	0.997	99.4%
	Malware	0.88	0	1	0.88	0.936	
Experiment 2	Benign	0.998	0	1	0.998	0.999	99.8%
	Malware	1	0.002	0.989	1	0.994	

Table 4 shows the evaluation performance for *Experiments 1* and *2* with accuracy value of 99.4% and 99.8% respectively. There is a slight increased on detecting Android malware using our proposed feature selection algorithm. Furthermore, *Experiment 2* achieved 100% TP rate to classify malware as compare to 88% for *Experiment 1*. Thus, this shows that *Experiment 2* has better features selection which correctly classified the malware. We consider to use F-measure value because it is a robust measure. The larger F-measure values correspond to better predictability of the classes. *Experiment 2* achieved higher F-measure value than *Experiment 1* for both benign and malware classes. As for Benign class, *Experiments 1* and *2* achieved 99.7% and 99.9% respectively. Also, for malware class, *Experiments 1* and *2* gets 93.6% and 99.4% respectively. As a result, the feature selection for *Experiment 2* is more generalized and managed to detect benign and malware sample accurately.

5 Conclusion

This paper proposed dynamic feature selection by using weighted-based feature selection algorithm to detect Android malware. The importance of feature selection methods such as TF-IDF will be experimentally justified using machine learning algorithm. The proposed features selection approach and existing approaches has been analyzed using performance metric. We used permission-based feature for the experiment. For future research, we need to analyze other Android malware feature to expand the Android malware detection research area.

Acknowledgement. The authors express appreciation to the Universiti Tun Hussein Onn Malaysia (UTHM). This research is supported by Postgraduate Research Grant vot number U610, Short Term Grant vot number U653 and Gates IT Solution Sdn. Bhd. under its publication scheme.

References

1. Abdullah Z, Saudi MM, Anuar NB (2014) Mobile botnet detection: Proof of concept. In: 2014 IEEE 5th control and system graduate research colloquium, pp 257–262
2. Kemp S (2016) We Are Social, “Digital in 2016”. www.wearesocial.com, http://wearesocial.com/sg/special-reports/digital-2016. Accessed 05 Nov 2016
3. Abawajy J, Kelarev A (2017) Iterative classifier fusion system for the detection of android malware. IEEE Trans Big Data PP(99):1
4. Vidas T, Votipka D, Christin N (2011) All your droid are belong to us: a survey of current android attacks. In: WOOT 2011, pp 81–90
5. Data G (2016) G Data Releases Mobile Malware Report for the Fourth Quarter of 2015. https://www.gdata-software.com/g-data/newsroom/news/article/g-data-releases-mobile-malware-report-for-the-fourth-quarter-of-2015. Accessed 05 Nov 2016. (Search Date 19/4/2016)
6. Li W, Ge J, Dai G (2016) Detecting malware for android platform: An SVM-based approach. In: Proceedings - 2nd IEEE international conference on cyber security and cloud computing, CSCloud 2015 - IEEE international symposium of smart cloud, IEEE SSC 2015, pp 464–469
7. Varsha MV, Vinod P, Dhanya KA (2015) Heterogeneous feature space for Android malware detection. In: Eighth international conference on contemporary computing, (IC3) 2015, Noida, India, 20–22 August 2015, pp 383–388

8. Karim A (2016) On the analysis and detection of mobile botnet. J Univ Comput Sci 22(4): 567–588
9. Zhou Y, Wang Z, Zhou W, Jiang X (2012) Hey, you, get off of my market: detecting malicious apps in official and alternative android markets. In: NDSS 2012, vol. 25(4), pp 50–52
10. Enck W, Gilbert P, Han S, Tendulkar V, Chun B-G, Cox LP, Jung J, McDaniel P, Sheth AN (2014) TaintDroid: An information-flow tracking system for realtime privacy monitoring on smartphones. ACM Trans Comput Syst 32(2):51–529
11. Wu W-C, Hung S-H (2014) DroidDolphin: A dynamic android malware detection framework using big data and machine learning. In: Proceedings of the 2014 conference on research in adaptive and convergent systems, pp 247–252
12. Arp D, Spreitzenbarth M, Hubner M, Gascon H, Rieck K (2014) DREBIN: Effective and explainable detection of android malware in your pocket. In: NDSS
13. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. ACM SIGKDD Explor Newsl 11(1):10–18
14. Android Developers. Permissions. https://developer.android.com/index.html, https://developer.android.com/guide/topics/permissions/index.html. Accessed 22 Dec 2016
15. Shabtai A, Moskovitch R, Elovici Y, Glezer C (2009) Detection of malicious code by applying machine learning classifiers on static features: A state-of-the-art survey. Inf Secur Tech Rep 14(1):16–29

Providing Differentiated Services for Full-Duplex Wireless LANs

Zhijie Ma, Qinglin Zhao(✉), and Huan Zhang

Faculty of Information Technology, Macau University of Science and Technology,
Avenida Wei Long, Taipa, Macau, China
mazhijie0000@hotmail.com, qlzhao@must.edu.mo, huan@buu.edu.cn

Abstract. Wireless full-duplex can significantly improve the system efficiency of a wireless LAN. Most existing full-duplex MAC protocols select a pair of full-duplex transmissions (uplink transmission and downlink transmission) only based on the interference between the uplink sender and the downlink receiver. In this paper, we propose a novel requirement-based full-duplex (RB-FD) MAC protocol, which aims at providing the differentiated services for all nodes in WLAN. In our RB-FD protocol, according to the demands for uplink transmission and downlink transmission of each node, we differentiate uplink-dominant (UD) node and downlink-dominant (DD) node. For an UD node, we target to provide more uplink resource, while for a DD node, we target to provide more downlink resource. Extensive simulations are performed to validate that the proposed RB-FD can achieve the desired objectives.

Keywords: Full-duplex · Differentiated service · Requirement · QoS · WLANs

1 Introduction

Recently, wireless full-duplex has been an attractive topic and is discussed in IEEE 802.11ax task group [1] for the next generation wireless LAN (WLAN) standard, because it may potentially double the system throughput [2]. In order to maximize the system efficiency, some existing works [3–7] proposed several feasible full-duplex MAC protocols to select a pair of full-duplex transmissions (i.e., uplink and downlink transmissions). However, most of them select a pair of full-duplex transmissions only based on the interference between the uplink sender and the downlink receiver.

The authors in [5] selected a pair of full-duplex transmissions based on the relationship between the interference and the threshold value. For example, when a node executes an uplink transmission to AP, AP randomly chooses a node as its downlink receiver. If the interference between this node and the selected node is less than the interference threshold, the selected node would notify AP to continue its downlink transmission, and then the full-duplex transmission happens. However, if the interference is larger than the threshold, the selected node would notify AP to stop its transmission, and then the transmission reduces to half-duplex transmission.

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_8

The authors in [3, 4] focus on minimizing the interference, and introduce the priority setting in their MAC protocols, so that the node with the smallest interference has the highest priority to initiate a transmission.

The authors in [6, 7] focus on avoiding the interference, and proposed the full-duplex MAC protocol to achieve the full-duplex transmission among the uplink sender, AP, and the downlink receiver. In their studies, the uplink sender and the downlink receiver are hidden terminal with each other, so there is no interference between them.

In short, all of these related works proposed the mechanisms only depending on one perspective. The similar study method also exists in another research direction. For example, the authors in [8] proposed a social incentive mechanism. However, in [8], what a participant gains only depends on his/her friends. Different from all of the above related works, in this paper, we proposed a novel requirement-based full-duplex (RB-FD) MAC protocol to achieve the different resource allocation for the node with different requirements. Our RB-FD protocol chooses a pair of full-duplex transmission depending on two perspectives: the interference value and the requirements of each node. By doing so, we can well guarantee the quality of service (QoS) requirements of all nodes.

2 RB-FD Protocol Overview

In this section, we overview the design idea of the RB-FD protocol. In our design, all nodes and AP have a single antenna with the full duplex capability. According to the demands for the uplink transmission and downlink transmission of each node, we divide all nodes into two types: uplink-dominant (UD) node, and downlink-dominant (DD) node, where we regard a node as UD node if its uplink traffic ≥ its downlink traffic, and regard a node as DD node if its uplink traffic < its downlink traffic.

The main goal of RB-FD protocol is to provide differentiated services for both UD node and DD node. Specifically, for a UD node, the system will provide more resources to its uplink than to its downlink; for a DD node, the system will provide more resources to its downlink than to its uplink. To this end, we introduce a dynamic priority assignment in the contention process, so that the nodes with different types can acquire different services. Figure 1 shows the transmission procedures of our RB-FD protocol. In RB-FD, TDC contention follows the legacy RTS/CTS protocol. FDC contention contains three-round process: round 1 (R1), round 2 (R2), and round 3(R3), and the dynamic priority assignment is in R1. DTX transmission contains a pair of full-duplex transmission: uplink and downlink transmissions.

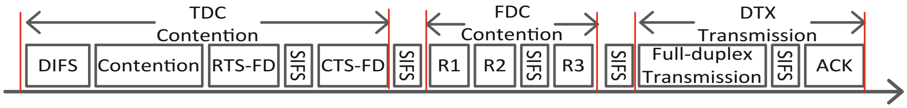


Fig. 1. The transmission procedure of the RB-FD protocol.

One salient feature of our design is to introduce two-type of contention (i.e., TDC and FDC), the purpose is to either choose the right uplink sender via TDC contention and the right downlink receiver via FDC contention, or choose the right downlink receiver via TDC contention and the right uplink sender via FDC contention. The rationale behind such design is explained in Sect. 3.2.3. Once both the uplink sender and the downlink receiver are determined, in DTX transmission, our design allows simultaneous transmissions of uplink and downlink to be executed, and then improve the system utilization.

3 RB-FD Design Details

3.1 TDC Contention

TDC contention is similar to the legacy contention in 802.11, as shown in Fig. 1. In the legacy 802.11 protocol, after sensing channel idle for DIFS, all nodes and AP generate a random backoff counter value among [0,CW-1], and contend for the channel using the binary exponential backoff (BEB) algorithm. The backoff counter is decremented as long as the channel is sensed idle, and frozen when the channel is sensed busy. Once its backoff counter decreases to 0, a node reserves the channel by sending an RTS frame. However, different from the legacy protocol, in this paper, in order to differentiate the UD node and DD node, we introduce a new type of RTS and CTS frames, termed as RTS-FD frame and CTS-FD frame, respectively.

3.1.1 RTS-FD Frame

We introduce an RTS-FD frame to differentiate the RTS-sender type (i.e., AP, or UD node, or DD node). This RTS-FD frame differs from the legacy RTS frame by 2 bits. In particular, the Frame Control field of the legacy RTS control message contains two distinct sub-fields: a 2-bit ‘Type’ field and a 4-bit ‘Sub-Type’ field. At present, in 2-bit Type field, three values of ‘00’, ‘01’, and ‘10’ have been used, while the value of ‘11’ is still reserved. Meanwhile, all corresponding 16 Sub-Type values (0000–1111) for 2-bit ‘11’ are also reserved. Hence, when we set the value of 2-bit ‘Type’ to ‘11’, we can use the last 2-bit of 4-bit Sub-Type field to represent the RTS-sender type. For example, the last 2-bit value is 00 if the RTS-sender is AP, 01 if the RTS-sender is a UD node, and 10 if the RTS-sender is a DD node. By doing so, the receiver can easily determine the RTS-sender type by checking this 2-bit value.

3.1.2 CTS-FD Frame

The CTS-FD frame not only needs to inform the RTS-sender like the legacy CTS frame that the channel has been reserved successfully, but must need to inform all nodes that (1) the current TDC transmission type; (2) which node can attend the FDC contention; and (3) the interference value between the RTS-sender and RTS-receiver (this value will be discussed in Sect. 3.2.1).

Figure 2 illustrates the frame structure of a CTS-FD frame. Three red dashed line boxes respectively denote three new added functions. In the following, we explain these three functions one by one.

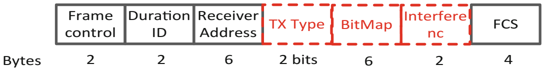


Fig. 2. The structure of the proposed CTS-FD frame.

TX Type: In our design, we define four TDC transmission types: Case UD → AP, i.e., a UD node sends an RTS-FD frame to AP, and becomes an uplink sender. Case DD → AP, i.e., a DD node sends an RTS-FD frame to AP, and becomes an uplink sender. Case AP → UD, i.e., AP sends an RTS-FD frame to a UD node, and this UD node becomes a downlink receiver. Case AP → DD, i.e., AP sends an RTS-FD frame to a DD node, and this DD node becomes a downlink receiver.

These four TDC transmission types are mapped to two binary bits. The value of 00 and 01 respectively denote that the TDC transmission types are Case UD → AP and Case DD → AP, and the TX Type value of 10 and 11 respectively denote that the TDC transmission types are Case AP → UD and Case AP → DD. The CTS-sender directly sets the TX Type value to the corresponding binary bits in a CTS-FD frame once the TDC transmission type is determined, and then all nodes can determine the current TDC transmission type by checking the TX Type value.

Bitmap: We use Bitmap to notify the node which needs to attend the FDC contention. In Bitmap [11], the i-th bit value of 1 represents that the i-th node needs to attend FDC contention, and the bit value of 0 represents that the node does not attend the contention. In this paper, the length of Bitmap is 6 Bytes, which means AP can support 48 nodes at most.

When the current TDC transmission type is Case UD → AP or Case DD → AP, AP can directly set the i-th bit value to 1 if AP has packet to node i, and set it to 0 otherwise. This is because once a node wins the FDC contention, it would become the downlink receiver. When the current TDC transmission type is Case AP → UD or Case AP → DD, this downlink receiver directly set all bits in Bitmap field to 1, means all nodes can attend the contention. This is because once a node wins FDC contention, it would become the uplink sender. However, in these two cases, the downlink receiver does not know which node has packet to AP, so it set all bits in Bitmap field to 1.

3.2 FDC Contention

After TDC contention, the nodes perform FDC contention to choose the undetermined uplink sender or downlink receiver. In FDC contention, three-round process are included: R1, R2, and R3, as shown in Fig. 2. In R1, a node first judges that whether it satisfies the FDC contention conditions. If yes, it performs the FDC contention according to the dynamic priority rules. For example, it randomly selects a subcarrier from a specified subcarrier range.

In R2, a node, whose selected subcarrier number in R1 is the smallest, transmits its signature to AP through the whole channel. Note that if multiple nodes select this smallest subcarrier, they would send their signatures in R2 simultaneously.

In R3, AP detects the received signatures through cross-correlation method [10]. If only one signature is detected, it indicates that there is no collision occurs in R1. AP directly broadcasts this signature in R3. The node which sends signature in R2 becomes the downlink receiver or the uplink sender. If AP detects multiple signatures, it indicates that there must be a collision occurs in R1, and then AP sends nothing in R3. As a result, all nodes fail in FDC contention, and the undetermined uplink sender or downlink receiver in TDC contention is still unknown.

Below, we specify the contention condition and the dynamic priority assignment in R1. After that, we explain the rationale behind such contention design.

3.2.1 FDC Contention Contentions

If the current TDC transmission type is Case UD → AP or Case DD → AP, it implies that in the incoming DTX transmission, the uplink sender has been determined, and FDC contention is used to determine the potential downlink receiver. Assume that node i is the determined uplink sender. Node j is allowed to perform FDC contention, if both of the following two conditions are satisfied: (1) Node j receives a CTS-FD and finds that the j-th bit in the Bitmap is set to 1; (2) Node j's signal-to-interference ratio, SIR_j, is larger than a threshold γ, namely, $SIR_j > \gamma$. Note that $SIR_j > \gamma$ means that node i's uplink transmission will not interfere with node j's reception. We define SIR_j as follows.

- When $j = i$, node j's transmission to AP will interfere with node j's reception and therefore such interference is called a self-interference. Assume that after cancellation, the remaining self-interference (RSI_j) at node j can be expressed as a constant fraction of the transmitted power, namely, $RSI_j = g_j P_{j\to AP}$, where g_j is a constant determined by the hardware [12], and $P_{j\to AP}$ is node j's transmission power to AP. We define SIR_j as follows.

$$SIR_j = \frac{P_{AP\to j}|h_{AP\to j}|^2}{g_j P_{j\to AP}}, \tag{1}$$

  where $P_{AP\to j}$ is AP's transmission power to node j; $h_{AP\to j}$ is the channel coefficient between AP and node j, and this value can be measured based on the power of the signal (CTS-FD frame) received from AP [5] [7].
- When $j \neq i$, node i's transmission to AP will interfere with node j's reception. We define SIR_j as follows.

$$SIR_j = \frac{P_{AP\to j}|h_{AP\to j}|^2}{P_{i\to AP}|h_{i,j}|^2}, \tag{2}$$

  where $P_{i\to AP}$ is node i's transmission power to AP;$\mathrm{h}_{\mathrm{i,j}}$ is the channel coefficient between node i and node j, and this value can be measured based on the power of the signal (RTS-FD frame) received from node i.

If the current TDC transmission type is Case AP → UD or Case AP → DD, it implies that in the incoming DTX transmission, the downlink receiver is determined, and FDC contention is used to determine the potential uplink sender. Assume that node j is the downlink receiver. Node i is allowed to perform FDC contention, if both of the following two conditions are satisfied: (1) node i has packets to AP. This is because once node i wins the FDC contention, it will execute an uplink transmission to AP in the DTX transmission. Therefore, node i must check itself whether it has data to AP at first. (2) Node i computes the SIR at node j, SIR_j, and makes sure that $SIR_j > \gamma.SIR_j > \gamma$ means that the node i's uplink transmission will not interfere with node j's reception. The calculation of SIR_j can refer to (1) and (2).

3.2.2 Dynamic Priority Rules

In our design, like the T2F scheme in [9], each node picks a random subcarrier from a specified subcarrier range, and transmits a signal on the corresponding subcarrier. After compare one's own subcarrier number with others, a node with the smallest subcarrier number could proceed with the contention in R2, and possibly becomes the FDC contention winner. Unlike the T2F scheme, to achieve the differentiated services, assuming three integers a, b, c where a < b < c, we define 3 priorities as follows.

- Prio_0: the highest priority whose subcarrier range is [1,a].
- Prio_1: the middle priority whose subcarrier range is [1,b].
- Prio_2: the lowest priority whose subcarrier range is [1,c].

Table 1 lists the dynamic priority rules in our design. Next, we focus on Case UD → AP to specify the priority rules, and other cases can refer to Case UD → AP.

Table 1. Dynamic priority rules

	Current UD node	Current DD node	Other UD nodes	Other DD nodes
Case UD → AP	Prio_1	Null	Prio_2	Prio_0
Case DD → AP	Null	Prio_0	Prio_2	Prio_1
Case AP → UD	Prio_0	Null	Prio_1	Prio_2
Case AP → DD	Null	Piro_1	Prio_0	Piro_2

When the transmission type is Case UD → AP, all nodes (including the current uplink sender, termed as current UD node) which satisfy the FDC contention conditions, will contend for becoming the downlink receiver. The priority setting of each node is listed as follows.

- For the DD nodes, because they have more requirements for downlink resource, so their priorities are the highest.
- For current UD node i, even though it does not require more downlink resource, it is the TDC contention winner, so its priority is middle.
- For the other UD nodes (except UD node i), because they neither need more downlink resource, nor win TDC contention, so their priorities are the lowest.

3.2.3 The Rationale Behind Such Contention Design

In our design, we adopt both TDC and FDC contentions. The reason can be summarized as follows. Compared with the TDC contention, the FDC contention can significantly reduce the contention overhead, and then improve the system efficiency. So we introduce the FDC contention in RB-FD. However, although FDC contention can improve the system efficiency, it requires the stringent signal synchronization, and this is much difficult to achieve when the node number is large. In this regard, the TDC contention can well address this problem since TDC allows large amount of nodes to contend for the channel with lax synchronization. On this account, we first adopt the TDC contention to ensure one transmission direction is determined, and then only allow few nodes to attend FDC contention.

3.3 DTX Transmission

In RB-FD, if the contentions in both TDC and FDC succeed, i.e., both the uplink sender and the downlink receiver are determined. Then, at the time of the uplink sender executes an uplink transmission to AP, AP also executes a downlink transmission to the downlink receiver. Hence, full-duplex transmission happens.

If only the TDC contention succeeds, while the FDC contention fails, namely, either the uplink sender or the downlink receiver is determined in TDC contention. In this case, there only exists an uplink transmission or a downlink transmission. Therefore, the data transmission reduces to half-duplex transmission.

Finally, the uplink receiver and downlink receiver send back the ACKs to the corresponding senders simultaneously, and the procedure of RB-FD is finished.

4 Performance Evaluation

In this section, we verify the performance of RB-FD protocol using our C++-based simulator, which has been adopted in our previous works [13]. Concentrating on verifying the main design objectives (e.g., for an UD node, the system needs to provide more resources to its uplink than to its downlink; and for a DD node, the system needs to provide more resources to its downlink than to its uplink), we consider a saturated wireless LAN (e.g., all nodes and AP always have packets to send), and assume that all nodes satisfy the FDC contention conditions, and therefore have qualifications to attend the consequent FDC contention. In the network, there are one AP, N = 2 UD nodes, and M = 2 DD nodes. Each node has the same contention window size. The packet size is set to 1500 bytes. The other default protocol parameter values are listed in Table 2. Each simulation run lasts for 200 s.

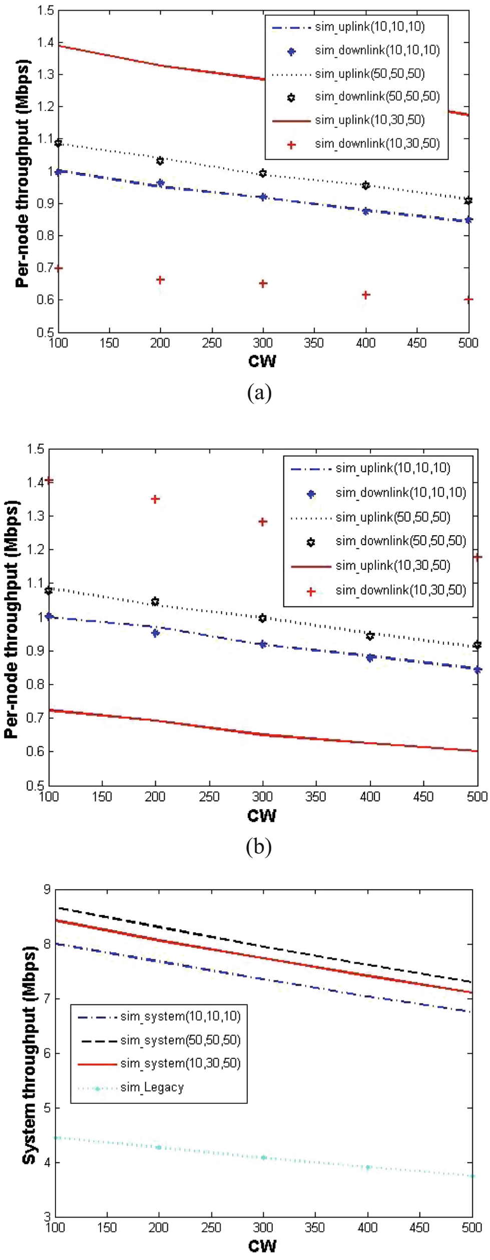
Figure 3(a) plots the average uplink throughput and downlink throughput of a UD node when CW varies from 100 to 500, where we respectively set the priority value of (a, b, c) to (10, 10, 10), (50, 50, 50), and (10, 30, 50). From this figure, we can see that, when we set the priority value to the same number (i.e., (10, 10, 10) or (50, 50, 50)), the uplink throughput is almost equal to the downlink throughput. In contrast, when we set the priority to the different values (i.e., (10, 30, 50)), the uplink throughput of UD node is always higher than its downlink throughput. This manifests that the proposed RB-FD can well provide the differentiated service for a UD node, namely, allocate more resource to its uplink transmission than to its downlink transmission.

Table 2. Parameters for performance evaluation.

Parameter	Value
DIFS	50 us
SIFS	10 us
Slot time	20 us
RTS/RTS-FD	38/38 bytes
CTS/CTS-FD	44/52 bytes
ACK	38 bytes
T_{R1}	10 Us
T_{R2}/T_{R3}	8 us
P_{hdr}	26 bytes
M_{hdr}	28 bytes
Payload	1500 bytes
Rbasic	1 Mbps
R_{data}	11 Mbps
T_{DATA}	$(P_{hdr} + M_{hdr})$@R_{basic} + Payload@R_{data}

Figure 3(b) plots the average uplink throughput and average downlink throughput of a DD node. The simulation results almost repeat the results of Fig. 3(a), except that the case when the priority value is (10, 30, 50). In such case, the downlink throughput of DD node is always higher than its uplink throughput. This manifests that the proposed RB-FD can well provide the differentiated service for a DD node, namely, allocate more resource to its downlink transmission than to its uplink transmission.

Figure 3(c) plots the system throughput of the proposed RB-FD protocol and the legacy RTS/CTS protocol. From this figure, we can see that, compared with the legacy RTS/CTS protocol, the proposed RB-FD protocol can achieve average 79.86%, 89.35%, and 94.64% improved throughput for the priority value of (10, 10, 10), (10, 30, 50), and (50, 50, 50), respectively. This manifests that our proposed RB-FD protocol can significantly improve the system efficiency.



(a)

(b)

(c)

Fig. 3. The average uplink throughput and downlink throughput of (a) an UD node, (b) a DD node; (c) the system throughput of the RB-FD and legacy protocols, when we set (a, b, c) to (10, 10, 10), (50, 50, 50), and (10, 30, 50), respectively.

5 Conclusion

In this paper, we propose a novel full-duplex MAC design, called RB-FD, to provide the differentiated services according to the requirement of each user. We validate our design through extensive simulations, and the simulation results verify that our design is feasible.

Acknowledgment. This work is supported by the Macao FDCT-MOST grant 001/2015/AMJ, and Macao FDCT grants 056/2017/A2 and 005/2016/A1.

References

1. IEEE 802.11-Task Group AX (2016). http://www.ieee802.org/11/Reports/tgax_update.htm
2. Sabharwal A, Schniter P, Guo D, Bliss W et al (2014) In-band full-duplex wireless: challenges and opportunities. IEEE JSAC 32(9):1637–1652
3. Choi W, Lim H, Sabharwal A (2015) Power-controlled medium access control protocol for full duplex WiFi networks. IEEE TWC 14(7):3601–3613
4. Qu Q, Li B, Yang M, et al (2015) FuPlex: a full duplex MAC for the next generation WLAN. In: IEEE QSHINE 2015, pp 239–245
5. Goyal S, Liu P, Gurbuz O, Erkip E, Panwar S (2013) A distributed MAC protocol for full duplex radio. In: Asilomar conference on signals, systems and computers. IEEE, pp 788–792
6. Kim J, Kim W, Kim J (2015) A new full duplex MAC protocol to solve the asymmetric transmission time. In: IEEE Globecom. IEEE, pp 1–5
7. Kim C, Kim C (2016) A full duplex MAC protocol for efficient asymmetric transmission in WLAN. In: International conference on computing, networking and communications. IEEE, pp 1–5
8. Yang G, He S, Shi Z et al (2017) Promoting cooperation by the social incentive mechanism in mobile crowdsensing. IEEE Commun Mag 55(3):86–92
9. Sen S, Choudhury RR, Nelakuditi S (2010) Listen (on the frequency domain) before you talk. In: Proceedings of the 9th ACM SIGCOMM workshop on hot topics in networks. ACM, p 16
10. Misra S, Khatua M (2015) Semi-distributed backoff: collision-aware migration from random to deterministic backoff. IEEE TMC 14(5):1071–1084
11. IEEE 802.11 - wireless LAN medium access control (MAC) and physical layer (PHY) specifications, IEEE Std. 02.11-2007, June 2007
12. Marasevic J, Zhou J, Krishnaswamy H et al (2015) Resource allocation and rate gains in practical full-duplex systems. ACM SIGMETRICS Perform Eval Rev 43(1):109–122
13. Ma Z, Zhao Q, Zeng Y, Zhang H, Dai H (2016) AT-MAC: a novel full duplex MAC design for achieving asymmetric transmission. In: Proceedings of the 3rd international conference on mobile and wireless technology, Jeju Island, Korea, pp 41–49

Speeding up the Montgomery Exponentiation with CMM-SDR Over GPU with Maxwell and Pascal Architecture

Xian-Fu Wong[1], Bok-Min Goi[1](✉), Wai-Kong Lee[2], and Raphael C.-W. Phan[3]

[1] Lee Kong Chian Faculty of Engineering and Science,
Universiti Tunku Abdul Rahman, Sungai Long, Malaysia
wongxf92@1utar.my, goibm@utar.edu.my

[2] Faculty of Information and Communication Technology,
Universiti Tunku Abdul Rahman, Kampar, Malaysia
wklee@utar.edu.my

[3] Faculty of Engineering, Multimedia University,
Cyberjaya, Malaysia
raphael@mmu.edu.my

Abstract. RSA is an algorithm widely used in protecting the key exchange between two parties for secure mobile and wireless communication. Modular exponentiation is the main operation involved in RSA, which is very time consuming when the bit-size is large, usually in the range of 1024-bit to 4096-bit. The speed performance of RSA comes to concerns when thousands or millions of authentication requests are needed to handle by the server at a time, through a massive number of connected mobile and wireless devices. The performance of RSA can be improved by utilizing parallel computing architecture or enhancing existing modular exponentiation algorithm. In this paper, we exploit the massively parallel architecture in GPU to perform RSA computations. Various optimization techniques were proposed in this paper to achieve higher throughput in RSA computation in two GPU platforms. Moreover, we also incorporated signed-digit recoding to further improve the performance. To allow a fair comparison with existing implementation techniques, we proposed to evaluate the speed performance in the best case (least '0' in exponent bits), average case (random exponent bits) and worse case (all '1' in exponent bits). The overall throughput achieved by our implementation is about 12% higher in random exponent bits and 50% higher in all 1's exponent bits compared to the implementation without signed-digit recoding technique. Our implementation is able to achieve 17713 and 89043 1024-bit modular exponentiation per second on random exponent bits in GTX 960 M and GTX 1080, which represent the two state of the art GPU architecture.

Keywords: RSA · GPU · Signed-digit recoding · Montgomery exponentiation

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_9

1 Introduction

Mobile and wireless communication technologies are growing by leaps and bounds in the past decade, which foster the emergence of Cloud Computing. One of the key aspects of these technologies is the security features offered to protect the user's privacy during communication. RSA is a public key cryptosystem widely used for encrypting messages or generating digital signatures to provide authentication feature in secure communication. The core operation in RSA is modular exponentiation, which can be represented by the equation $C = M^e \bmod N$, where M is the plaintext, C is the ciphertext, e is the exponent and N is a large prime number. Computing modular exponentiation is non-trivial when the integer size is large. For example, RSA is consider insecure if the bit-size is smaller than 1024-bit; this implies that the implementation of modular exponentiation in RSA needs to handle integers of at least 1024-bit.

A straightforward implementation of 1024-bit modular exponentiation with the exponent as large as 1024-bit will require a lot of modular multiplication to be performed. Moreover, each modular multiplication involves a 2048-bit (2049-bit if the carry present) product followed by an expensive division. To simplify the computation, Montgomery Multiplication [7] was introduced to avoid the expensive division by replacing it with cheaper shift operations. Binary method [8] is also widely used to reduce the number of modular multiplication needed to compute a modular exponentiation.

Graphics Processing Unit (GPU) is massively parallel processors capable of computing thousands of threads in parallel. GPU has been used in various applications to accelerate cryptographic algorithms [1–4]. This motivates us to explore the possibility to compute RSA in parallel using GPU, which is very useful for server applications that need to handle millions of authentications from the clients simultaneously, such as assessing personal and business information through mobile devices.

GPU was used in computing cryptographic algorithms including RSA. Neves et al. [1] analyzed different methods to interleave Montgomery multiplication on GPU with Tesla architecture and found FIOS and FIPS method is better than the most commonly used CIOS method. They were able to achieve throughput of 41426 512-bit modular exponentiations per second. On the other hand, Leboeuf et al. [2] mentioned that CIOS method is more suitable for GPU with Fermi architecture; their implementation was able to achieve throughput of 1.24 to 1.72 times greater than the fastest implementation on the same GPU. Recently, Emmart and Weems [3] introduced four methods to perform multiply-accumulate instruction across multiple generations of GPU (Tesla, Fermi, Kepler, and Maxwell) and compared their performance. They found that every GPU architecture achieved its best performance with different methods, as the clock per operation for the multiply-accumulate instruction are different for some GPU. In another work, Emmart et al. [4] discovered the optimal use of the instructions, memory, registers and threads on the GPU with Maxwell architecture. In addition, they claimed that their implementation is much faster than the state of the art by using the row-oriented multiply and reduce. However, the work from Emmart et al. [3, 4] do not consider the signed digit recoding technique, which can further improve the performance of Montgomery exponentiation. Wu et al. [5] proposed a technique CMM-SDR which improve the conventional signed-digit recoding method to accelerate modular exponentiation. However, no experimental result was presented in their paper.

Based on our observation of previous works [1–4], we found that the experiments were only performed based on random bit exponent. Since the number of bit "1" occurs in the exponent can greatly reduce the performance of modular exponentiation, it is not fair to evaluate the performance solely based on random bit exponent that is not reproducible. Hence, we proposed to evaluate the performance by including the best case (least '0' in exponent bits) and worst case (all '1' in exponent bits). This gives a better understanding of the actual performance of an implementation as well as providing a fair and reproducible comparison.

In this paper, we implemented the CMM-SDR method proposed by Wu et al. [5] in GPU. We evaluated the performance in two GPUs, GTX 960 M (Maxwell) and GTX 1080 (Pascal). GTX 960 M is similar to the GPU used by Emmart et al. [3, 4], while GTX 1080 represent the state of the art GPU architecture available in the market as of 2017. Our implementation can achieve 19528 (best case), 17713 (random exponent) and 17822 (worst case) 1024-bit modular exponentiations per second in GTX 960 M. On GTX 1080, the performance is 102379 (best case), 89043 (random exponent) and 90562 (worst case) 1024-bit modular exponentiations per second.

The layout of this paper is organized as follows. In Sect. 2, we introduce the background of Montgomery multiplication, binary exponentiation, signed digit recoding algorithm [5] and the CMM-SDR Montgomery algorithm [5] and followed by our proposed GPU implementation. In Sect. 3, the experimental setup and result, then Sect. 4, analysis and discussion. Lastly, the conclusion of our work in Sect. 5.

2 Background

To compute arithmetic involving large integer size, a common method is to represent the large integer in radix form. The coefficients of the radix form are stored in an array; the arithmetic computations are then performed on these coefficients. If one of these coefficients overflows, we need to perform carry propagation to avoid error. For example, the number 12345 can be represented in radix form of base 10:

$$12345 = 1 * 10^4 + 2 * 10^3 + 3 * 10^2 + 4 * 10^1 + 5 * 10^0$$

We store 1, 2, 3, 4, 5 in an array in this case. In this paper, the large integer is represented in radix 2^{32} as the GPU is a 32-bit processor. Hence, a 1024-bit integer can be represented in radix form with 32 coefficients (which is also referred to as *limbs* in the literature). Similarly, 2048-bit can be represented using 64 *limbs*, each *limb* is 32-bit.

2.1 Montgomery Multiplication

The conventional way to perform modular multiplication requires expensive division operation. Instead of using expensive division, Montgomery multiplication is able to perform the reduction by using addition and bit shifting with a base to the power of two, which is optimized for majority hardware architectures. Notice that, this requires conversion from radix form to Montgomery form at the beginning of computation; it also requires another conversion back to radix form at the end of the computation. These two conversions are expensive, but it is still beneficial to use Montgomery multiplication for modular exponentiation because most of the computations can be done in Montgomery form. Algorithm 1 shows the operations involved.

Note that if the b is selected as a power of two, the modular reduction in line 6 can be replaced by bitwise shifting, which is very fast in most of the computer hardware.

Algorithm 1. Montgomery Multiplication

```
Input: $x = (x_{n-1} \dots x_1 x_0)_b, y = (y_{n-1} \dots y_1 y_0)_b, m = (m_{n-1} \dots m_1 m_0)_b$ with
$0 \le x, y < m$, $R = b^n$ with gcd$(m, b) = 1$, and $m' = -m^{-1}$ mod $b$
// $m^{-1}$ is the inverse modular of b
Output: A = $xyR^{-1}$mod $m$ or Mont$(x, y)$
// $R^{-1}$ is the inverse modular of $m$
1.  begin
2.      $A = 0$;                                   // $A = (a_n a_{n-1} \dots a_1 a_0)_b$
3.      for $i = 0$ to $n - 1$ do
5.          begin
6.              $u_i = (a_0 + x_i y_0) m'$mod $b$;
7.              $A = (A + x_i y + u_i m)/b$;
8.          end;
9.      if $A \ge m$ then $A = A - m$;
10.     Return($A$);
11. end.
```

2.2 Binary Montgomery Exponentiation

Binary method (Algorithm 2) can be used in conjunction with Montgomery multiplication to perform modular exponentiation. The algorithm begins by scanning the exponent bits from right to left; if the bit is ‘0’, only squaring is performed; if the bit is ‘1’, an additional Montgomery multiplication is performed.

Algorithm 2. Binary Montgomery Exponentiation

```
Pre-compute: $\tilde{x} = \mathrm{Mont}(x, R^2 \bmod m)$, $A = R \bmod m$
// $x, m$, and $R$ are the inputs of Algorithm 1
Input: $\tilde{x}, m$, and $e = (e_t \ldots e_1 e_0)_2$ with $e_t = 1$
// $(e_t \ldots e_1 e_0)_2$ is the binary representation of $e$
Output: $x^e \bmod m$
begin
    for i = t to 0 do                               // Scan exponent bits from
        begin                                       // right to left
            $A = \mathrm{Mont}(A, A)$;                       // Squaring
            if $e_i = 1$ then $A = \mathrm{Mont}(A, \tilde{x})$;       // Multiplication
        end;
    $A = \mathrm{Mont}(A, 1)$;
    Return($A$);
end.
```

2.3 Signed Digit Recoding

In binary method for exponentiation, the speed performance is determined by the number of ‘1’ bit in the exponent, as additional multiplication is required for every ‘1’ bit. Signed-digit recoding (Algorithm 3) reduce the number of ‘1’ bit in the exponent. The output of this recoding method always has extra one digit than the binary representation. For example, decimal 31 is represented as [1, 1, 1, 1, 1] (5 digits) in binary, but represented as [1, 0, 0, 0, 0, -1] after signed digit recoding (6 digits). The number of zero has been increased compared to the binary representation, but the number of one is reduced.

Algorithm 3. Signed Digit Recoding

```
Input: $E = (r_{m-1} r_{m-2} \ldots r_0)_2$ where $r_i$ is the binary representation of $E$
Output: $E_{SD} = (e_n e_{n-1} \ldots e_0)_{SD}$
begin
    $c_0 = 0$; $r_{m+1} = 0$; $r_m = 0$;
    for i = 0 to m do
        begin
            $c_{i+1} = \lfloor (c_i + r_i + r_{i+1}) / 2 \rfloor$;   // $\lfloor x \rfloor$ : the largest integer $\leq x$
            $e_i = c_i + r_i - 2(c_{i+1})$;
        end;
end.
```

2.4 CMM-SDR Montgomery Algorithm

Wu et al. [5] proposed an improvement to the conventional signed-digit recoding technique, named CMM-SDR (Algorithm 3). For each iteration, if the scanned bit is "1" or "−1", then it is a multiply and square, if "0", then it is only a squaring.

Algorithm 4. CMM-SDR Montgomery Algorithm

```
Input: $\boldsymbol{M, N, E_{SD}, R}$ where $\boldsymbol{M}$ is the message, $\boldsymbol{N}$ is the modular and $\boldsymbol{E_{SD}}$ is the exponent in signed digit form
// $\boldsymbol{R}$ is the input of Algorithm 1 and $\boldsymbol{E_{SD}}$ is the output of Algorithm 3
Output: $\boldsymbol{C} = \boldsymbol{M}^{\boldsymbol{E_{SD}}} \bmod \boldsymbol{N}$
begin
    $\boldsymbol{S} \equiv \boldsymbol{MR} \bmod \boldsymbol{N}$, $\boldsymbol{C} \equiv \boldsymbol{R} \bmod \boldsymbol{N}$, $\boldsymbol{D} \equiv \boldsymbol{R} \bmod \boldsymbol{N}$
    for $i = 0$ to $m$ do
      begin
        if $(\boldsymbol{e_i} = \boldsymbol{1})$ then $\boldsymbol{C} = \mathrm{Mont}(\boldsymbol{S}, \boldsymbol{C})$;      // Multiplication
        if $(\boldsymbol{e_i} = \boldsymbol{-1})$ then $\boldsymbol{D} = \mathrm{Mont}(\boldsymbol{S}, \boldsymbol{D})$;   // Multiplication
        $\boldsymbol{S} = \mathbf{Mont}(\boldsymbol{S}, \boldsymbol{S})$;                          // Squaring
      end;
    $\boldsymbol{C} = \mathrm{Mont}(\boldsymbol{C}, \mathbf{1})$; $\boldsymbol{D} = \mathrm{Mont}(\boldsymbol{D}, \mathbf{1})$;
    $\boldsymbol{C} \equiv \boldsymbol{C} \times \boldsymbol{D}^{-1} \bmod \boldsymbol{N}$                       // $C \equiv M^{E_{SD}} \bmod N$
end.
```

Based on the theoretical complexity analysis by Wu et al. [5], as the probability of computing REDC (S, C), REDC (S, D) and REDC (S, S) are same with the occurrence of the signed digit "1", "−1" and "0", then together with the respective n number of single precision multiplications. Thus, in averagely:

REDC (S, C) requires $\frac{1}{6}\left(2n^2 + n\right)$ single-precision multiplications

REDC (S, D) requires $\frac{1}{6}\left(2n^2 + n\right)$ single-precision multiplications

REDC (S, S) requires $\frac{2}{3}\left(n^2 + 2n + 2\right)$ single-precision multiplications

With CMM-SDR, the occurrence of "0" digit is higher and able to save the number of multiplication compared to the original method. For example, to compute the exponent of decimal 31 (scan from right to left):

In binary, [1, 1, 1, 1, 1], requires $\mathbf{5}\left(2n^2 + n\right) + \mathbf{5}\left(n^2 + 2n + 2\right) = \mathbf{15}\boldsymbol{n}^2 + \mathbf{15}\boldsymbol{n} + \mathbf{10}$ single-precision multiplications.

In signed digit, [1, 0, 0, 0, 0, −1], requires $\mathbf{2}\left(2n^2 + n\right) + \mathbf{6}\left(n^2 + 2n + 2\right) = \mathbf{10}\boldsymbol{n}^2 + \mathbf{14}\boldsymbol{n} + \mathbf{12}$ single-precision multiplications single-precision multiplications.

Notice that, there is an operation involved modular inverse at the end of this algorithm. In order to reduce the extra cost of computing expansive in the modular inverse, we can perform the inverse modular multiplication with the technique introduced by Koc et al. [6] which still utilize the usage of cheap division in reduction.

3 Proposed GPU Implementation

GPU has deep memory architecture with various memory types; each of them has their own strength and limitation. We implemented CMM-SDR Montgomery multiplication based on coarse-grained parallelism, whereby each thread is assigned to compute one modular exponentiation. Since each thread is independent of each other, there is no intense communication between threads, so shared memory does not provide significant benefits to our implementation. At the same time, the computations within one thread are somehow more intensive compared to fine-grained implementation. Thus, we do not limit the number of registers used per thread and let the compiler allocates as much as it could (Fig. 1).

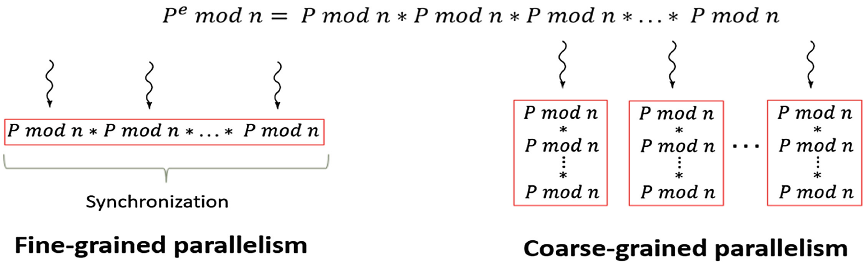

$P^e \bmod n = P \bmod n * P \bmod n * P \bmod n * ... * P \bmod n$
$P \bmod n * P \bmod n * ... * P \bmod n$
Synchronization
Fine-grained parallelism
$P \bmod n * P \bmod n \vdots * P \bmod n$
Coarse-grained parallelism

Fig. 1. Fine-grained parallelism vs. Coarse-grained parallelism

First, we pre-compute the values of *R, C, D* and *S*, then copy these pre-computed values, together with *M*' (required to compute Montgomery multiplication), *M* and E_{SD} to global memory in GPU. Notice that all the values are represented in multi-limbs (32-bit each) and store in the form of arrays, except *M*' which is store in register.

Next, 32000 threads are launched to perform 32000 modular exponentiations; the threads are organized as 125 blocks per grid, and 256 threads per block. Each thread has to load the values of R, M, E_{SD}, C, D and *S* into local memory and *M*' into register. During the computations, C, D and S will be used to store the intermediate values. The results of Montgomery exponentiation are stored in global memory and copied to the host memory after the computations are completed.

4 Experimental Setup and Result

Most of the available works evaluate the performance based on random bit patterns on the exponents. However, these random bit patterns are difficult to reproduce by others as no information is provided regarding the random seed and algorithm used for generating random numbers. In order to perform a fair comparison with other available works, we proposed to evaluate the performance based on three different bit patterns. The first bit pattern is the smallest exponent (prime number with least number of ‘0’ in the exponent), the second bit pattern is random exponent and third bit pattern is the largest exponent (prime number with the most number of ‘1’ in the exponent). This corresponds to the best case, average case and worst case respectively.

We evaluated the performance of 1024-bit and 2048-bit modular exponentiation on GTX 960 M (Maxwell) and GTX 1080 (Pascal). We design and setup three different scenarios to compute the modular exponentiations, with the largest, smallest and random exponent bits. Each scenario is performed for 20 times and the average result is reported. Besides, we only record the time taken for memory transaction within GPU and the computation of modular exponentiation. The time for pre-computation, copy data in between CPU and GPU and result verification are not recorded. The throughput is calculated as the number of modular exponentiation computed per second.

Figures 2 and 3 show the results of our experiment on GTX 960 M and GTX 1080. The results are compared with conventional Montgomery Multiplication without CMM-SDR technique.

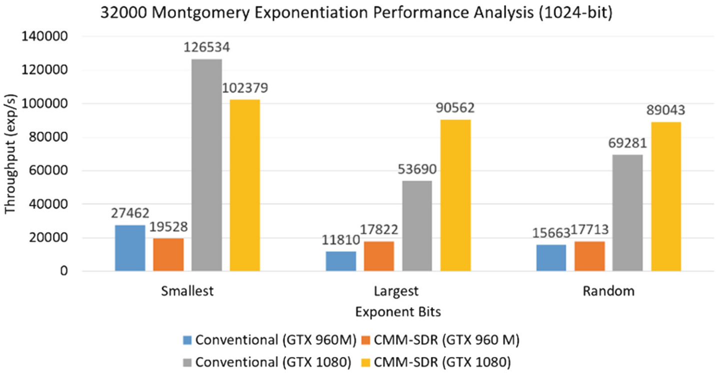


Fig. 2. Average throughput for 1024-bit montgomery exponentiation

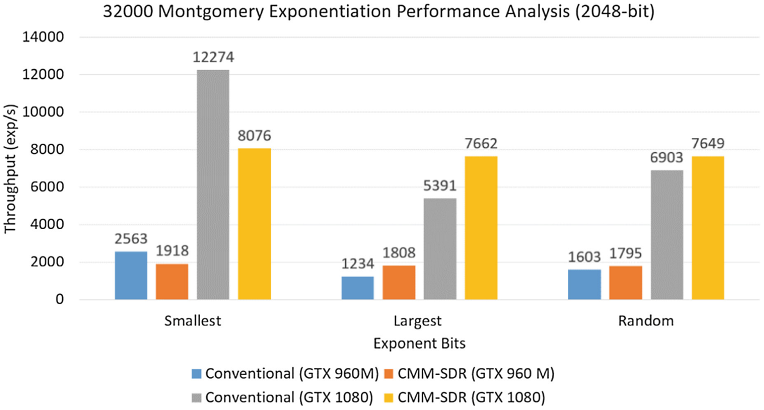


Fig. 3. Average throughput for 2048-bit montgomery exponentiation

5 Analysis and Discussion

5.1 The Smallest Exponent Bits (Best Case)

From Figs. 2 and 3, we can see that the throughputs for the conventional method are always higher than the CMM-SDR method in this particular case. Referring to Table 1, the number of non-zero remains the same as in CMM-SDR; instead, CMM-SDR method has an extra computation of zero in both 1024 bits and 2048 bits. In fact, CMM-SDR requires extra one additional Montgomery Multiplication (compare line 7, Algorithm 2 and line 9, Algorithm 4) and the extra computation of modular inverse and modular multiplication (line 10, Algorithm 4). CMM-SDR suffers the computation overhead in this case. As a result, the conventional method is more efficient than CMM-SDR for the case of smallest exponent bits.

Table 1. Numbers of non-zero and zero in smallest exponent bits for conventional and CMM-SDR (1024-bit and 2048-bit modular exponentiation)

	Conventional	CMM-SDR
Non-zero	1	1
Zero	1023/2047	1024/2048

5.2 The Largest Exponent Bits (Worst Case)

The CMM-SDR method started to shine in this scenario as its throughput is around 50% higher than the conventional method. From the Table 2, we can see that the number of non-zero is greatly reduced in CMM-SDR. In this scenario, the conventional method needed to compute 1024 (1024 bits) and 2048 (2048 bits) times of squaring and multiplication, whereas CMM-SDR method only needed to compute 2 times (1024 bits and 2048 bits). Thus, the CMM-SDR method is more efficient in this case.

Table 2. Numbers of non-zero and zero in largest exponent bits for conventional and CMM-SDR (1024-bit and 2048-bit modular exponentiation)

	Conventional	CMM-SDR
Non-zero	1024/2048	2
Zero	0	1024/2048

5.3 Random Exponent Bits (Average Case)

In random exponent bits, the CMM-SDR method still able to outperform the conventional method, the overall throughput is about 12% higher than the conventional method. From the Table 3, we can see that the number of non-zero is still greatly reduced in CMM-SDR, from 497 to 7 in 1024-bit modular exponentiation and 1015 to 16 in 2048-bit modular exponentiation. However, the computation overhead as mentioned in Sect. 5.1 which limit the maximum achievable throughput.

Table 3. Numbers of non-zero and zero in random exponent bits for conventional and CMM-SDR (1024-bit and 2048-bit modular exponentiation)

	Conventional	CMM-SDR
Non-zero	497/1015	7/16
Zero	527/1033	1017/2032

5.4 Performance Comparison of Our Work with Recent Work

The work from Emmart and Weems [3] is the fastest among all coarse grain modular exponentiation in GPU. They are using the GTX 750Ti (640 cores) from Maxwell architecture, which is the same with GTX 960 M (640 cores) we used. We are able to achieve 17.17 k modular exponentiation per second (random exponent bits) which is slower than the achievement by Emmart and Weems [3] (22.72 k). However, our implementation is not fully optimized compared to them. Firstly, they used fixed window exponentiation which scans multiple bits per iteration, but our method only scans one bit per iteration. Secondly, we have not fully optimized the operation to store and load the message, modulus, and exponent in GPU, which involves optimized usage of local memory and registers. Thirdly, they also used CUDA PTX assembly code to fully optimized the implementation. In fact, our work can be integrated with the techniques proposed by Emmart and Weems [3] to further improve the throughput of modular exponentiation in GPU.

On the other hand, we also evaluated the same implementation in GTX 1080 with the latest GPU architecture, Pascal. GTX 1080 consists of 2560 cores, which is four times more than GTX 960 M (640 cores). From our experiments, the throughput achieved by GTX 1080 is 3.5–4.2 times more than GTX 960 M, which is coherent with the hardware capability of both hardware platforms. GTX 960 M is more widely used for low-end mobile computing system like laptops; we selected this platform to perform a direct comparison with Emmart and Weems [3]. Conversely, GTX 1080 can be used in a server environment to handle massive digital signatures (RSA) in parallel.

6 Conclusion

In this paper, we have shown that our GPU implementation is able to achieve high throughput by incorporating the CMM-SDR method by Wu et al. [5]. Although our proposed implementation does not show good result in smallest exponent bits, it eventually shows good result in random exponent bits which is more closely related to the real world scenarios. By integrating the technique proposed in this paper to the work from Emmart and Weems [3], the modular exponentiation can achieve higher throughput in GPU platforms.

Acknowledgements. This research is partially supported by the Malaysia Ministry of Science, Technology & Innovation (MOSTI) eScience fund 01-02-11-SF0201 and 01-02-11-SF0202.

References

1. Neves S, Araujo, F (2011) On the performance of GPU public-key cryptography. In: 2nd IEEE international conference on application-specific systems, architectures and processors, ASAP 2011
2. Leboeuf K, Muscedere R, Ahmadi M (2013) A GPU implementation of the montgomery multiplication algorithm for elliptic curve cryptography. In: 2013 IEEE international symposium on circuits and systems (ISCAS 2013) (2013)
3. Emmart N, Weems C (2015) Pushing the performance envelope of modular exponentiation across multiple generations of GPUs. In: 2015 IEEE international parallel and distributed processing symposium
4. Emmart N, Luitjens J, Weems C, Woolley C (2016) Optimizing modular multiplication for NVIDIA's maxwell GPUs. In: 2016 IEEE 23nd symposium on computer arithmetic (ARITH) (2016)
5. Wu C-L, Lou D-C, Chang T-J (2008) An efficient montgomery exponentiation algorithm for public-key cryptosystems. In: 2008 IEEE international conference on intelligence and security informatics
6. Savas E, Koc C (2000) The montgomery modular inverse-revisited. IEEE Trans Comput 49:763–766
7. Montgomery P (1985) Modular multiplication without trial division. Math. Comput. 44:519
8. Koc CK, Acar T, Kaliski B (1996) Analyzing and comparing montgomery multiplication algorithms. IEEE Micro 16:26–33

Robust Object Tracking via Improved Mean-Shift Model

Liqun Wang[1(✉)], Xuenan Shi[1], Sunyi Han[1], and Jinchi[2]

[1] School of Information Engineering, Northeast Electric Power University Jilin,
Jilin 132012, China
guoshuqiang@gmail.com

[2] School of Computer Science and Technology, University of South China, Hengyang, China

Abstract. In this paper we propose a robust object tracking algorithm using a improved Mean-Shift model. As the traditional Mean-Shift algorithm for object tracking uses a single histogram. Because the traditional Mean-Shift lacks spatial distribution information, so it is difficult to track non-rigid object especially. With a focus on this problem, an improved Mean-Shift algorithm based on the shape feature and color of the target is presented. The results show that the algorithm can track the moving vehicles in real time, and it has a preferable adaptability and robustness to the irregular motion and deformation of the target.

Keywords: Object tracking · Mean shift · Machine vision · Color feature

1 Introduction

Target tracking is a key issue in the research of computer vision, and it has a broad development prospects in the field of robot vision navigation, medical diagnosis, traffic monitoring and other fields. Because of the deformation of the target, the nonlinearity of the motion, the occlusion of the object, the non-rigid target tracking is a difficult problem. Mean-Shift algorithm is widely applied to non-rigid the target tracking field due to non-parameters and fast mode matching, and real-time performance.

Mean-Shift algorithm is a data analysis method based on kernel density estimation proposed by Fukunaga and Hosteter in 1975 [1]. After 1995, it was widely used in image segmentation, image smoothing, medical image analysis, object tracking and other image processing fields [2–5]. Mean-shift algorithm uses color histogram, sets the weight coefficient according to the contribution to the mean shift vector of the sample point, and then obtains the maximum offset of the solution using the mountain climbing algorithm to track the non-rigid target. Therefore, the mean-shift algorithm can effectively track the target without a large range of fluctuations in illumination conditions, and the target can be tracked effectively even if the target is not beyond the range of the image. With the development of image processing technology, Mean-Shift algorithm has been further improved in many fields. There have been many new applications. In the Nummiaro's algorithm, the weights of the mean-shift algorithm are extended to negative space, and the non-rigid object can be tracked in the continuous variation condition [6]. Kwon and Lee proposed a method for the non-rigid object tracking combining illumination space and the image space [7].

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_10

The tracking algorithm based on mean-shift has the characteristics of less computation and real time. However, when the illumination conditions change, the color histogram will changes, and the tracking error will occur. In order to improve the recognition rate of this kind of situation, Oshima and Collins proposed a modified algorithm for monochrome target tracking, and the peak value of the histogram can be moved to improve the tracking effect caused by the change of illumination conditions. However, this method would change the gray value of the target in the image, so it is very difficult to distinguish between the tracking target and the gray level change from the RGB value.

In this paper, a Mean-Shift tracking algorithm based on block color histogram is proposed, which uses the block color histogram as the target mode, and introduces the target rotation and scaling matrix. The block color histogram contains the spatial information of object, which makes the target with the same color distribution but different spatial structure. So it can improve the recognition ability and robustness of the target.

2 Object Representation

2.1 Block Color Histogram

Color histogram does not contain the spatial information of the target, and it has the property of rotation invariance. Different targets may have the same color histogram. As shown in Fig. 1(b) and (c), the two targets have the same color histogram, so these two targets cannot be distinguished according to the color histogram, and the rotation of the target will not cause the color histogram changes. So we use the block color histogram to represent the object. As shown in Fig. 1 (a), the image is split in to several blocks. The blocks are given a serial numbers. A total color histogram can be obtained using the color histogram of each sub block in a sequence order. The sub block color histogram contains the spatial information of the target, and it has no longer the rotation invariance. So this method can improve the recognition ability of the target, and can determine the rotation of the target according to the change of the sub block color histogram. In this paper, we divide the target area into 4 blocks, of course, different block number and different cutting method can also be used in accordance with different objectives. The more number of sub blocks make, the representation of the target more precise, but it also increases the amount computation.

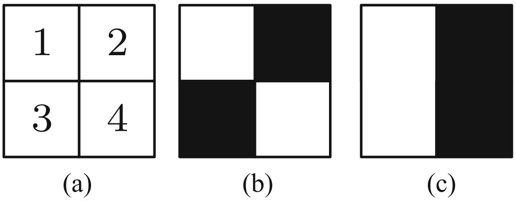


Fig. 1. Block color histogram

2.2 Object Mode

The object is an rectangle region in the image. In the case of non-universality, the target can be considered as the center of the space coordinates. The target mode $\widehat{q}$ is a normalized weighted color histogram.

$$\sum_{u=1}^{M} \hat{q}_u = 1 \qquad \hat{q} = \left\{\hat{q}_u\right\}_{u=1,2,\ldots m} \tag{1}$$

$$\hat{q}_u = C\sum_{i=1}^{n} k(\|Rx_i^*\|^2)\delta\left[b(x_i^*) - u\right] \tag{2}$$

Where *{xi***}i=1… N* is the pixel position in target area. The weighted value of the target area is generated by the isotropic kernel *K(x)*. The *K(x)* is constructed by monotone increasing function. This method can make the weighted value of target center larger. It increases the robustness of the density estimation, because the pixels in target edge are usually unstable and are often subject to occlusion and background interference.

R is the target rotation and scaling matrix:

$$R = L(\beta)R(\theta)S(\varphi) = \begin{pmatrix} \beta & 0 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & \tan\varphi & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{3}$$

where *L(β)* is the amplification of the transformation matrix, β is the zoom factor. *R(θ)* is rotation matrix, *θ* is the rotation angle. *S(φ)* is the migration transformation matrix, *φ* is horizontal offset angle.

δ is the delta Kronecker function. Function *b* correlates the position of *x* and histogram quantization feature space index. Normalized constant *C* can be derived according to the conditions $\sum_{i=1}^{n} \hat{q}_u = 1$:

$$C = \frac{1}{\sum_{i=1}^{n} k(\|Rx_i^*\|^2)} \tag{4}$$

2.3 Candidate Target Mode

If the center position of the candidate target in the current frame is *y*, then the candidate target mode $\widehat{p}$ is the normalized weighted color histogram:

$$\sum_{u=1}^{M} \hat{p}_u = 1 \qquad \hat{p} = \left\{\hat{p}_u\right\}_{u=1,2,\ldots m} \tag{5}$$

$$\hat{p}_u = C_h \sum_{i=1}^{n} k(\|R(y - x_i)\|^2)\delta[b(x_i) - u] \tag{6}$$

Where *x* is the pixel position of the candidate target in the current frame, *K(x)* is the kernel, *R* is rotation and scaling matrix; C_h is the normalized constant:

$$C_h = \frac{1}{\sum_{i=1}^{n} k(\|R(y - x_i)\|^2)} \tag{7}$$

2.4 Similarity Measurement

The similarity between the histogram of the target template and the candidate target region is measured with a coefficient of the discrete form, it is defined as follows:

$$\rho(y) = \rho[p_u(y), q_u] = \sum_{i=1}^{n} \sqrt{p_u(y)q_u} \tag{8}$$

2.5 The Optimality of Bhattacharyya Coefficient

The minimization of the distance is equivalent to the maximum Bhattacharyya coefficients. Starting from the target position of the last frame, the target location of the *Y* is searched for the new target location in the current frame. Therefore, it is necessary to calculate the candidate target mode *y* in the current frame, using Taylor expansion, the Bhattacharyya coefficients are linear approximation to Eq. (9):

$$\rho(y) = \rho[\hat{p}(y), \hat{q}] = \frac{1}{2}\sum_{u=1}^{n} \sqrt{\hat{p}_u(\hat{y}_0)\hat{q}_u} + \frac{1}{2}\hat{p}_u(y)\sqrt{\frac{\hat{q}_u}{\hat{p}_u(\hat{y}_0)}} \tag{9}$$

We put Eq. (6) into the Eq. (10), the approximation calculation can be obtained:

$$\rho(y) = \rho[\hat{p}(y), \hat{q}] \approx \frac{1}{2}\sum_{u=1}^{n} \sqrt{\hat{p}_u(\hat{y}_0)\hat{q}_u} + \frac{C_h}{2}\sum_{u=1}^{n_k} w_i k\left(\|R(y - x_i)\|^2\right) \tag{10}$$

where w_i is $w_i = \sum_{i=1}^{n} \sqrt{\frac{\hat{q}_u}{\hat{p}_u(\hat{y}_0)}}\delta[b(x_i) - u]$.

So in order to minimize distance value in Eq. (8), the second item of Eq. (11) must be max, and the first term of Eq. (8) has nothing to do with the *y*. The second term is based on the kernel *K(x)* density estimation in the current frame, and the gradient direction of the density estimation can be obtained using the Shift Mean method. The core from the current position *Y* to the new location *y*:

$$\hat{y}_1 = \frac{\sum_{i=1}^{n_k} x_i w_i g\left(\left\|R(y - x_i)\right\|^2\right)}{\sum_{i=1}^{n_k} w_i g\left(\left\|R(y - x_i)\right\|^2\right)} \tag{11}$$

3 Experiment and Result Analysis

3.1 Experimental Conditions

Using MV-EM200CM color industrial camera which is produced by the Beijing Weishi digital image Co., Ltd., and the M160-MPW2 lens whose focal length is 16 mm, the viewing angle of the field is 37.7° × 30.7° × 17.1°, the tracking system constructed shown in Fig. 2. This system is always aiming at the vehicle on the road.

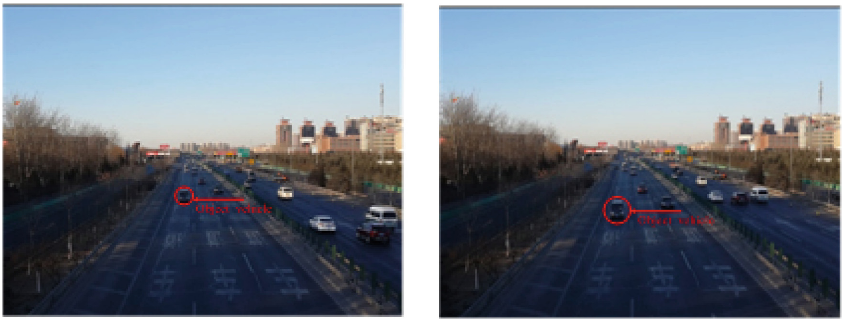
Fig. 2. The tracking system

This experiment cannot be used in the field for long time, so the vehicle motion image is saved in video format, and then the video is analyzed by using the personal computer. The experiments were implemented on a computer with Intel i3 processor 2.7 GHz CPU, and 4G memory capacity.

3.2 Processing Method

This study is a sub project of Jilin Province Education Bureau project whose project name is intelligent transport systems of engineering and technology based on image processing technology. The vehicle motion along the road can be predicted by processing technology. In this method, the camera aims at the vehicle, so the vehicle motion image can be obtained. Using the adjacent two frames of the vehicle image and the aforementioned Mean-Shift non-rigid object tracking method, the velocity field can be calculated. Finally, the moving direction and speed of the vehicle are computed based on the motion field.

Feature points matching based on mean-shift. In this method, the feature points are matched in two frames of the video image, and the matching method is above-mentioned. The original image is obtained by using aforementioned experimental conditions as shown in Fig. 3. In the original image, the similarity principle of Eq. (8) is used, and the mean-shift algorithm is used to improve the matching accuracy, so as to the motion field of vehicle can be obtained as shown in Fig. 4.



(a) The first frame (b) The frame after 30s

Fig. 3. The image captured by the sun tracking system

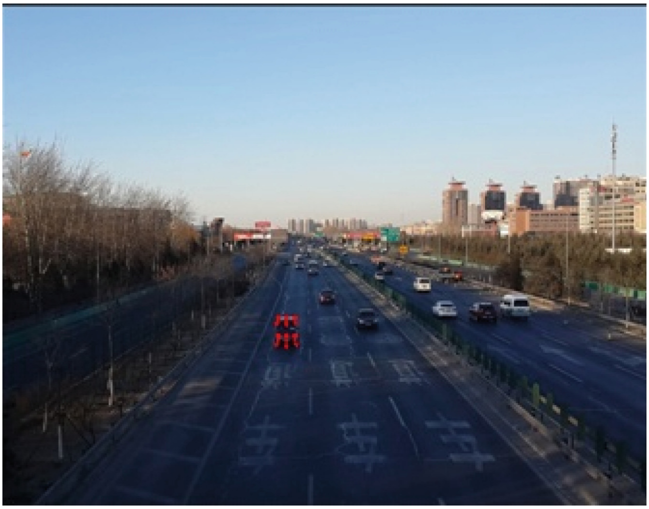


Fig. 4. The motion field of vehicle

Calculation method for moving velocity of vehicle. According to the image of the moving field obtained from Fig. 4, the moving speed of the matching points can be calculated, and the whole moving speed of the vehicle is obtained by using the average moving speed. At this point, the movement speed of each point is different; the distribution of the weight can be used to reduce the calculation error of the vehicle velocity.

3.3 Experimental Results

Using the aforementioned experimental conditions, the object tracking result are shown in Fig. 5. Figure 5 (a) (b) (c) (d) are the original image in video scene. Figure 5 (e) (f) (g) (h) are object tracking result that are expressed by red rectangle. As shown in Fig. 6, the maximum error of our method is 7.6%, and the average error is 6.5% of all frames.

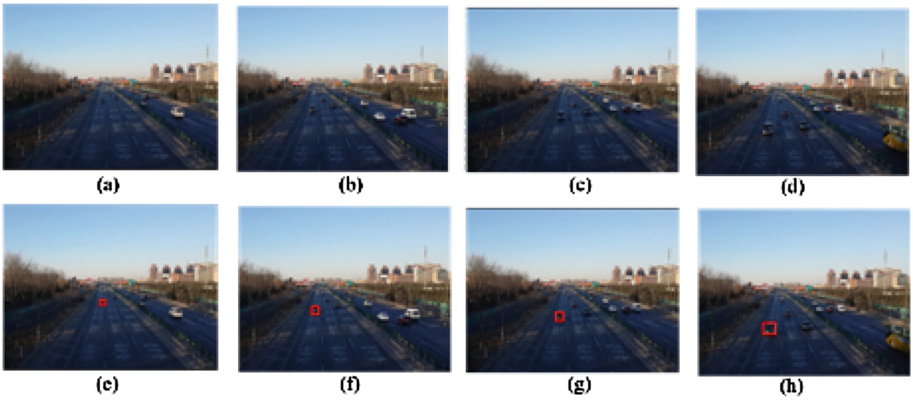


Fig. 5. Object tracking result

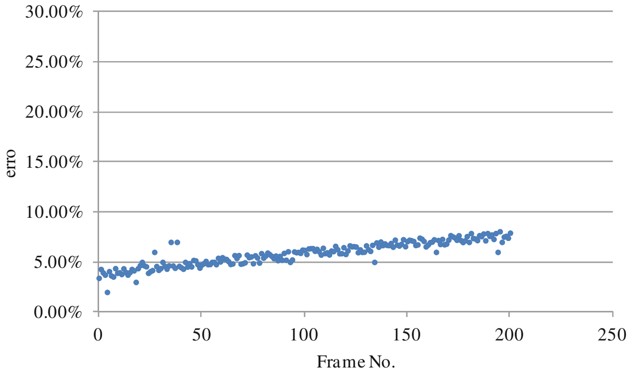


Fig. 6. Object tracking result

4 Conclusion

In this paper, a novel object tracking method based on Modified Mean-Shift is presented. According to the tracking results, the error of our method is less than 8%.

At present, the object tracking method can guarantee a smaller error in the sunny weather conditions. But in the fog and other weather conditions, the tracking error significantly increases. In future research, we would use tensor voting feature extraction method to improve the tracking precision in non-ideal weather such as fog.

Acknowledgment. This work was supported by the Research Foundation of Education Bureau of Jilin Province under the Grant Number: 2015254.

References

1. Fukunaga K, Hostetler L (1975) The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Trans Inf Theory 21(1):32–40
2. Tao W, Jin H, Zhang Y (2007) Color image segmentation based on mean shift and normalized cuts. IEEE Trans Syst Man Cybern Part B Cybern 37(5):1382–1389
3. Shan C, Wei Y, Tan T, Ojardias FEDE (2004) Real time hand tracking by combining particle filtering and mean shift, pp 669–674
4. Paris S, Durand F (2007) A topological approach to hierarchical segmentation using mean shift, pp 1–8
5. Carreira-Perpinan MA (2006) Acceleration strategies for gaussian mean-shift image segmentation, pp 1160–1167
6. Nummiaro K, Koller-Meier E, Van Gool L (2003) Color features for tracking non-rigid objects. ACTA Automatica Sinica 29(3):345–355
7. Kwon J, Lee KM (2009) Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive basin hopping monte carlo sampling, pp 1208–1215
8. N. Oshima, T. Saitoh and R. Konishi. Real Time Mean Shift Tracking using Optical Flow Distribution.ed. 2006:4316-4320
9. Collins RT (2003) Mean-shift blob tracking through scale space, pp 234–240

Phase Rotation Codebook Precoding for Space Shift Keying MIMO Systems

Mohammed Al-Ansi[1(✉)], Syed Alwee Aljunid[1], and Essam Sourour[2]

[1] School of Computer and Communication Engineering, Universiti Malaysia Perlis (UNIMAP), Arau, Perlis, Malaysia
m.ahmed@studentmail.unimap.edu.my

[2] Department of Electrical Engineering, College of Engineering at Wadi Al Dawaser, Prince Sattam bin Abdulaziz University, Wadi AL-Dawaser 11991, Saudi Arabia

Abstract. Transmit Precoding (TPC) techniques based on maximizing the Minimum Euclidean Distance (MED) of the received constellation have been proposed to improve the performance of Space Shift Keying (SSK) MIMO systems. These methods have limitations, including the necessity of full Channel State Information (CSI) at the transmitter, high complexity since they generate arbitrary precoders, and the requirement of infinite rate feedback channel. On the other hand, Codebook based precoding approach have the competency to tackle these difficulties. Hence, this paper proposes Full-Combination (FC) codebook with a systematic structure constructed based on Phase Rotation Precoding (PRP) where the CSI is known to the receiver only. Also, two effective codebooks called Factorized Full-Combinations (FFC) and Statistically Filtered Full-Combinations (SFF) are designed to solve the exhaustive search problem in FC codebook needed to find the best codeword that maximizes MED. Results illustrate the considerable performance gain achieved through applying the FC codebook. In addition, we show the capability of FFC and SFF codebooks to reduce the complexity and provide almost the same BER as FC codebook.

Keywords: Space Shift Keying (SSK) · Spatial Modulation (SM) · Phase Rotation Precoding · Minimum Euclidean Distance · Codebook precoding

1 Introduction

Space-Shift Keying (SSK) is presented in [1] as a special form of the novel technique called Spatial Modulation (SM) [2]. SSK trades off receiver complexity with data rate in Multiple Input Multiple Output (MIMO) systems. SSK carries data bits through the index of Transmit Antenna (TA) only, whereas SM conveys part of the data over the index of TA and another part by the standard way of selecting Amplitude and Phase Modulation (APM) symbols. Therefore, many benefits can be achieved, such as mitigating the inter-channel interference and reducing the complexity of transceiver due to applying single Radio Frequency (RF) chain [3].

Transmit Precoding (TPC) techniques with possible feedback from the receiver have been introduced to enhance the SSK/SM system performance where the downlink

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_11

Channel State Information (CSI) is known at the transmitter or the receiver. It is based on precoding the TAs by a specific codeword to maximize the Minimum Euclidean Distance (MED) between pairs of channel vectors and thus minimize the Bit Error Rate (BER) [1]. The TPC methodologies for constructing the precoders found in the literature can be grouped into three types.

The first TPC methodology is based on a full knowledge of CSI to design the precoder and apply an identical algorithm on both sides of the transceiver [4–7]. It results in a noteworthy minimization of the BER, since the selected codeword corresponds to the instantaneous channel. However, the solutions of these methods are typically iterative and the complexity may increase in the case of high number of TAs. Besides, these methods are not practical in the limited feedback systems, particularly in Frequency Division Duplex (FDD) systems due to the enormous overhead required to obtain full CSI. *The second TPC methodology* is based on creating a precoding codebook and the CSI is known to the receiver only [8, 9]. This constructed codebook contains many precoding codewords and is known by both transmitter and receiver. The receiver selects the best codeword that satisfies the MED criteria and feeds-back its index to the transmitter. This approach is suitable for the FDD systems. However, a loss of the system performance may occur since the chosen codeword does not match the exact channel. *The third TPC methodology* also establishes a codebook and the CSI is known to the receiver only, but this codebook is designed to match large training set of the channel *on the average* as in the first method of [9]. This approach cannot be applied in the finite rate feedback channel since the whole codebook is fed-back from the receiver to the transmitter.

This research fits in the above-mentioned second methodology where the main contributions are: Firstly, Full-Combinations (FC) codebook-based Phase Rotation Precoding (PRP) for SSK-MIMO system is investigated and compared to the literature. It improves the performance and the codebook has a permanent and systematic structure. Secondly, two codebooks termed as Factorized Full-Combinations (FFC) and Statistically Filtered Full-Combinations (SFF) are designed to solve the exhaustive search problem in the FC codebook to find the best codeword. FFC and SFF codebooks offer significant reduction of processing time and maintain similar BER to FC codebook.

The remainder of this paper is organized as follows: in Sect. 2, SSK-MIMO system model is introduced. Precoding method with FC codebook, and the alternative FFC and SFF codebook designs are explained in Sect. 3. Simulation results are presented in Sect. 4. Finally, Sect. 5 concludes the paper.

2 SSK-MIMO System Model

In this research, we consider SSK-MIMO system with N_t TAs and N_r receive antennas as shown in Fig. 1. The number TAs N_t is a power of 2. Single TA is active in each transmission time. Information bits are divided into parts of length $log_2\left(N_t\right)$. Each part selects the unique active antenna and the complex received signal can be expressed as:

$$\mathbf{y} = \mathbf{HPx}_k + \mathbf{w} \tag{1}$$

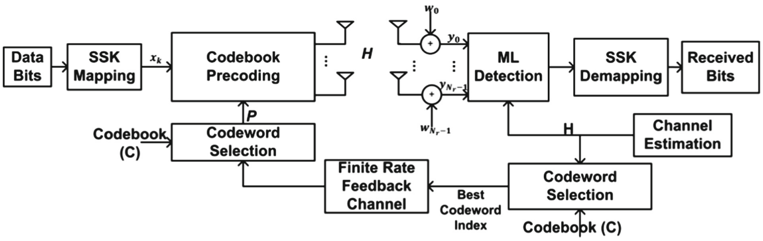


Fig. 1. SSK-MIMO system model based codebook precoding.

Where **H** is the $(N_r \times \mathrm{N}_t)$ channel matrix known to the receiver only, with entries h_{ij} that are assumed to be identical and independently distributed (i.i.d.) complex Gaussian random variables with zero mean and unit variance. $\mathbf{w} \approx \mathbb{CN}(0, I/\gamma)$ is the $(N_r \times 1)$ i.i.d. complex Gaussian noise and γ is the Signal to Noise Ratio (SNR) per receive antenna. $\mathbf{x}_k = [0, \ldots, \underset{kth}{1}, \ldots, 0]^T$ is the transmitted vector. The $(N_t \times \mathrm{N}_t)$ PRP matrix **P** has a diagonal structure to satisfy the SSK requirement of single active TA at a time. Hence, $\mathbf{P} = diag(\mathbf{p}_q)$, where $\mathbf{p}_q$ is the best codeword of length N_t, selected among all possible combinations in the codebook. The codebook $\mathbf{C} = [\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_q, \ldots, \mathbf{p}_{N_c}]^T$ consists of N_c codewords (i.e. N_c is the codebook length). Without precoding, all elements in $\mathbf{p}_q$ are ones. Figure 1 shows the SSK system with codebook- based precoding.

3 Codebook Design with Phase Rotation Precoding

When the transmitter has information about the downlink channel, the transmitted signal can be pre-processed to reduce the BER. Conditioned on the channel gains h_{ij}, the BER is upper bounded by [1]:

$$P_{bit}(e) \leq \frac{2}{N_t \log_2 N_t} \sum_{i=0}^{N_t-1} \sum_{j=i+1}^{N_t-1} D(\mathbf{h}_i,\ \mathbf{h}_j) Q\left(\sqrt{\gamma \left\| \mathbf{h}_i -\ \mathbf{h}_j \right\|^2 / 2}\right) \tag{2}$$

where $D(\boldsymbol{h}_i, \boldsymbol{h}_j)$ is the number of bit errors when the channel vector $\boldsymbol{h}_i$ is wrongly detected as $\boldsymbol{h}_j$, and $Q(.)$ is the Q-function. The upper bound (2) involves a weighted sum of Q-functions whose arguments are functions of the squared Euclidean distances between pairs of $\boldsymbol{h}_i$. Since $Q(x)$ rapidly decays with increasing x, the upper bound (2) is dominated by the MED. When precoding the i^{th} transmit antenna by the i^{th} element of a codeword $\mathbf{p} = \left[p_0, p_1, \ldots, p_{N_t-1}\right]^T$, the *precoded* $\left\| p_i \mathbf{h}_i - p_j \mathbf{h}_j \right\|^2$, $i{\neq}j$ will be increased and BER in (2) decreased. For that reason, the best codeword $\mathbf{p}_k$, $k = 0, 1, \ldots N_c$–1, that maximizes the MED is selected as follows:

$$\mathbf{p}_k = \arg\max_{\mathbf{p} \in \mathbf{C}} \left(\min_{i \neq j} \left(\left\| p_i \mathbf{h}_i - p_j \mathbf{h}_j \right\|^2 \right) \right),\ i\,\&\,j = 0,\ 1,\ \ldots,\ N_t - 1. \tag{3}$$

The index k of this selected codeword is fed back to the transmitter using N_{fb} bits. In the following, all codebooks are constructed based on phase-only precoding where the codeword elements are chosen from the set $\{\exp(j2\pi m/M), m = 0, 1, \ldots, M-1\}$, where $M \geq 1$ is the number of possible phases. The first antenna is always taken as a reference with $p_{k,0} = 1$.

3.1 Full-Combinations (FC) Codebook

Full-Combination (FC) codebook contains N_c codewords, the first element of each codeword is taken as a reference with $p_{k,0} = 1$. The remaining codeword elements $p_{k,i}, i = 1, 2, \ldots, \left(N_t - 1\right)$, can take any of the possible M phases. Hence, the codebook size is $FC_{N_c} = M^{(N_t-1)}$ codewords and every codeword is unique, i.e., no codeword is a phase rotation of another codeword. The best codeword is selected using (3) with $k = 0, 1, \ldots, (N_c - 1)$. The number of feedback bits is $\left\lceil N_{fb} = (N_t - 1)\log_2 M \right\rceil$. However, this FC codebook is practical only for small N_t and M due to the exhaustive search needed to find the best codeword among all possible codewords and therefore, the processing time complexity is exponentially increased with N_t. Furthermore, it is important to mention that although codewords are unique, several codewords yield the same MED value. In this case, the earlier in the codebook is selected. Thus, this redundancy motivates us to find more efficient codebooks while FC codebook is used as a reference.

3.2 Factorized Full-Combinations (FFC) Codebook

Factorization Full-Combination (FFC) method assumes the number of phases M is a power of 2 and the FC (mother) codebook is factorized into $N_p = log_2 M$ (children) codebooks, each containing $2^{(N_t-1)}$ codewords. The Factorization is such that the mother codebook is equal to all possible element-by-element multiplication of the codewords of all the N_p children codebooks. We first define the phasor $x_n = exp(j\pi/2^n), n = 0, 1, \ldots, N_p - 1$. Then, the codewords in a child codebook n have 1 at the first element, and the remaining elements consist of all $2^{(N_t-1)}$ combinations of 1 and x_n. Let's take an example with $N_t = 4$ and $M = 4$ phases. The FFC mother codebook consists of $N_c = M^{(N_t-1)} = 64$ codewords that needs to be searched. This mother codebook is factorized into $N_p = 2$ children codebooks, each containing $2^{(N_t-1)} = 8$ codewords in Table 1. The receiver selects the best codeword through using the channel $\mathbf{H}$ to search the first codebook ($n = 0$) using (3) and finds the best codeword. This codeword is element-by-element multiplied by all rows of $\mathbf{H}$ to generate an updated (i.e., precoded) channel. This updated $\mathbf{H}$ is used in (3) to search the next codebook ($n = 1$) to find the best codeword where it is a fine-tuning to the first codeword. This process is repeated for all children N_p codebooks. The final precoder that will be used by the transmitter is the element-by-element multiplication of the selected codewords from all codebooks. As in Table 1, the final codeword could be the element-by-element multiplication of the shaded codewords. Thus, the number of code words that should be tested is

$FFC_{N_c} = 2^{(N_t-1)} \log_2 M$, which is a considerable saving in comparison to $FC_{N_c} = M^{(N_t-1)}$. However, the number of feedback bits is the same as in the FC codebook case.

Table 1. Example of FFC codebook, $N_t = 4$ and $M = 4$ phases.

Child codebook $n = 0, x_0 = -1$					Child codebook $n = 1, x_1 = j$				
$k/p_{k,i}$	$p_{k,0}$	$p_{k,1}$	$p_{k,2}$	$p_{k,3}$	$k/p_{k,i}$	$p_{k,0}$	$p_{k,1}$	$p_{k,2}$	$p_{k,3}$
0	1	1	1	1	0	1	1	1	1
1	1	1	1	–1	1	1	1	1	j
2	1	1	–1	1	2	1	1	j	1
3	1	1	–1	–1	3	1	1	j	j
4	1	–1	1	1	4	1	j	1	1
5	1	–1	1	–1	5	1	j	1	j
6	1	–1	–1	1	6	1	j	j	1
7	1	–1	–1	–1	7	1	j	j	j

3.3 Statistically Filtered Full-Combinations (SFF) Codebook

The objective of Statistically Filtered Full-Combinations (SFF) codebook is to filter-out the redundant codewords in the FC codebook to minimize the complexity while maintaining acceptable performance. This redundancy comes due to the fact that many different codewords yield the same MED value. This redundancy increases the processing time due to testing many codewords that yield the same MED. The idea of SFF codebook is based on gathering statistical information about each codeword of FC codebook. A weight is set for each codewords based on how many times it is selected as the best to satisfy MED criteria of (3) after testing with a large number of trials with independent channels. Then, SFF codebook is constructed through sorting the codewords of the FC according to their weights, from most used to least used. Therefore, the weightiest codewords are indexed earlier. As a result, the top part of the SFF codebook contains the more significant codewords of the FC codebook. Thus, we can select the SFF codebook as the top N_c codewords. The value of N_c can be selected at will, preferably a power of 2. The same performance of the FC codebook can be achieved by $N_c \ll M^{(N_t-1)}$. Decreasing N_c, results in a slight performance loss, but with the benefit of lower complexity. Hence, the SFF codebook can be used to compromise between the performance and complexity.

4 Numerical Results and Discussions

In this section, we present an evaluation of the proposed methods through numerical results applied in un-correlate Rayleigh fading channel. SSK performance without precoding is termed as SSK in all the subsequent figures. Perfect channel estimation to get CSI is assumed at the receiver. Figure 2 shows the SSK performance improvement gained by applying the FC codebook with 2, 4 and 8 phases, corresponding to $N_c = 8$, 64, and 512 codewords, respectively. The results are also compared to the Maximum Minimum Distance (MMD) and non-convex Guaranteed Euclidean distance (GED) methods [4, 5]. MMD and GED are iterative methods that belong to the first methodology explained in Sect. 1 and they are expected to outperform the FC codebook. However, we can see that the performance gap is only about 0.5 dB at BER of 10^{-3}, which suggests that the codebook-based approach presented in this paper, with PRP, is very promising. Also, Fig. 2 shows a gain of almost 5 dB between the FC codebook with $M = 2$ phases and the SSK at the BER of 10^{-3}. Then, 1 dB additional gain is attained when $M = 4$. However, minor gain is achieved with increasing the phases from $M = 4$ to 8. Therefore, this clearly demonstrates that 4 phases are adequate for performance improvement.

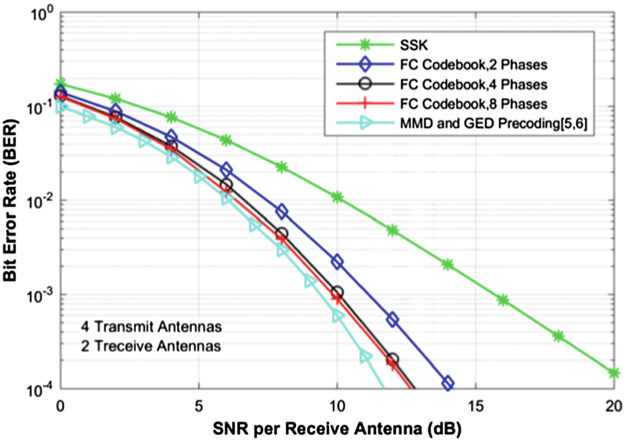


Fig. 2. Performance evaluation of FC codebook, $M = 2, 4$ and $8, N_t = 4$ and $N_r = 2$.

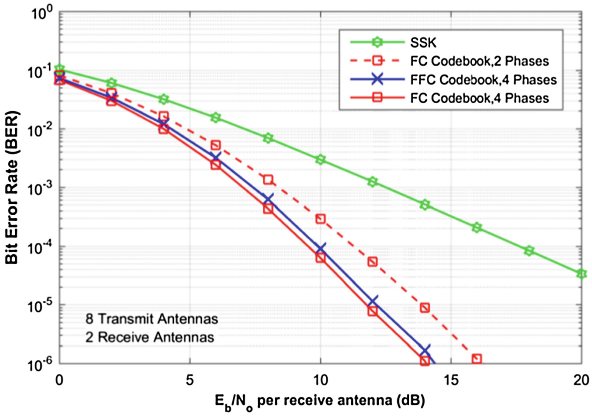


Fig. 3. BER performance of the FC and FFC codebooks, $M = 2$ and $4, N_t = 8$ and $N_r = 2$.

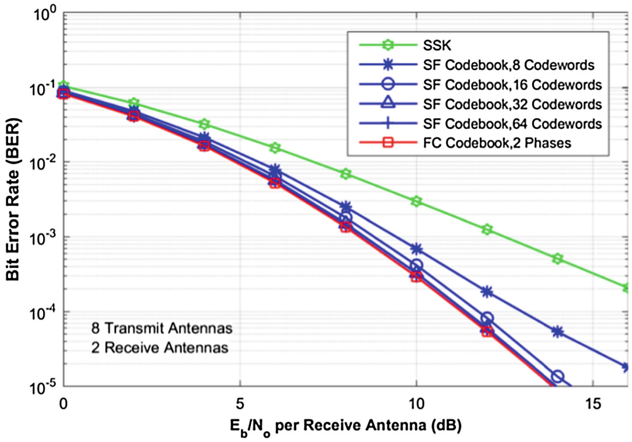


Fig. 4. BER comparison between the FC and SFF codebooks, $M = 2, N_t = 8$ and $N_r = 2$.

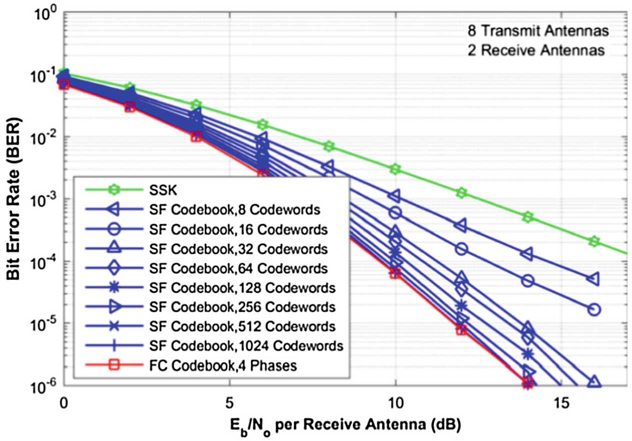


Fig. 5. BER comparison between the FC and SFF codebooks, $M = 4, N_t = 8$ and $N_r = 2$.

Figure 3 compares the performance of the FC and FFC codebooks. At $M = 4$, The much smaller FFC codebook ($N_c = 256$) provides close BER to FC codebook ($N_c = 16384$). This saving means lower searching complexity for the best codeword and less required number of feedback bits. Figures 4 and 5 portray the capability of SFF codebook to attain close BER performance to FC codebook. In Fig. 4, $M = 2$ phases and the FC codebook size is $N_c = 128$. The SFF codebook is shown with $N_c = 8$, 16, 32 and 64. Size $N_c = 32$ is enough to get the same BER of the FC codebook. Similarly, in Fig. 5, $M = 4$ and the FC codebook size is $N_c = 16384$. The SFF codebook with $N_c = 512$ provides the same performance as FC codebook. Hence, SFF codebook reduces complexity and maintains performance. Moreover, with the SFF codebook are able to trade off performance with complexity through selecting smaller codebook size. For example, in Fig. 5, the performance with SFF codebook increases gradually from $N_c = 8$ to 1024.

5 Conclusion

In this research, we have proposed systematic, structured Full-Combination (FC) codebook based Phase Rotation Precoding (PRP) to enhance the conventional SSK system performance. Two other codebook designs (i.e. Factorization Full-Combination (FFC) and Statistically Filtered Full-Combinations (SFF) codebooks) are introduced to minimize the codebook lengths and, consequently, the processing time required to find the best codeword. This smaller size also decreases the number of feedback bits. Simulation results illustrate that FFC and SFF approaches have the capability to provide almost the same performance of FC codebook.

Acknowledgements. This work was supported by the Deanship of Scientific Research at Prince Sattam Bin Abdulaziz University, Kingdom of Saudi Arabia, research project No. 2016/01/6637.

References

1. Jeganathan J, Ghrayeb A, Szczecinski L, Ceron A (2009) Space shift keying modulation for MIMO channels. IEEE Trans Wireless Comm 8(7):3692–3703. doi:10.1109/twc.2009.080910
2. Mesleh RY, Harald H, Sinanovic S, Ahn CW, Yun S (2008) Spatial Modulation. IEEE Trans Veh Tech 57(4):2228–2241. doi:10.1109/TVT.2007.912136
3. Garcia-Rodriguez A, Masouros C, Hanzo L (2015) Pre-scaling optimization for space shift keying based on semidefinite relaxation. IEEE Trans Comm 63(11):4231–4243. doi:10.1109/TCOMM.2015.2470656
4. Lee MC, Chung WH, Lee TS (2015) Generalized precoder design formulation and iterative algorithm for spatial modulation in MIMO systems with CSIT. IEEE Trans Comm 63(4):1230–1244. doi:10.1109/TCOMM.2015.2396521
5. Masouros C (2014) Improving the diversity of spatial modulation in MISO channels by phase alignment. IEEE Comm Letters 18(5):729–732. doi:10.1109/LCOMM.2014.031414.140233
6. Yang P, Guan YL, Xiao Y, Renzo MD, Li S, Hanzo L (2016) Transmit precoded spatial modulation: maximizing the minimum euclidean distance versus minimizing the bit error ratio. IEEE Trans Wireless Comm 15(3):2054–2068. doi:10.1109/TWC.2015.2497692
7. Lee MC, Chung WH, Lee TS (2014) Precoder design for space shift keying in MIMO systems with limited feedback. In: 25th IEEE Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Washington DC, USA, 2–5 September
8. Lee MC, Chung WH, Lee TS (2015) Limited feedback precoder design for spatial modulation in MIMO systems. IEEE Commun Lett 19:1909–1912. doi:10.1109/LCOMM.2015.2475265

Novel UE RF Condition Estimation Algorithm by Integrating Machine Learning

Yupu Dong[1](✉), Zhenni Pan[1], Mohamad Erick Ernawan[1], Jiang Liu[1],
Shigeru Shimamoto[1], Ragil Putro Wicaksono[2], Seiji Kunishige[2],
and Kwangrok Chang[2]

[1] Graduate School of Fundamental Science Engineering,
WASEDA University, Tokyo 169-8050, Japan
{Yupu.dong,z.pan,erick.ernawan}@fuji.waseda.jp,
liujiang@aoni.waseda.jp, shima@waseda.jp
[2] MOTiV Research Co., Ltd., Tokyo 158-0097, Japan
{ragil.wicaksono,seiji.kunishige,
kwangrok.chang}@motiv-research.com

Abstract. By 2020, 5G era will be commercially available. The smart city construction will also make great progress. Compared to current situation, more than thousand times of devices will connect to the cellular networks. For the operators, in order to analyze overall network performance, it is a key factor to estimate the user equipment (UE) radio frequency (RF) condition. However, practical RF estimation scheme is based on UE data log which can only observe UE that is at the top-serving cell with good RF condition. However, according to the comparison of actual UE data log and the scanner data log, potential RF problems may still exist since the UE will not always be served by the top-1 cell. In this paper, we propose a novel estimation scheme by integrating machine learning (ML) algorithm to analyze the scanner data logs from the target estimation zones where the mobility problems may occur. A hypothesis is obtained from learning step by various kinds of RF condition as input features. The numerical results show that the proposed estimation algorithm integrated ML can estimate probability of the potential mobility problems accurately.

Keywords: Machine learning · Estimation · RF condition · Mobility problem

1 Introduction

In the future smart city era, it is very important for the operators to analyze overall network performance. However, it is very difficult to analyze the ubiquitous networks with continuous data stream by the manual method.

Machine learning (ML) can be utilized as a tool to process a very large data samples (ex: UE RF condition) and analyze the network performance. RF condition can be estimated by using different data sets. One is UE data log while the other is scanner data log. Sometimes, the UE data log cannot reflect the real RF condition at that target area, since the actual UE serving cell may not be in a good RF condition, which will create

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_12

some UE RF potential problems. Normally, the scanner data log will contain more information than the UE data log, however, if assuming UE always at TOP 1 cell with good RF condition, we will miss some UE RF potential problems [1]. Therefore, in this paper, we will utilize the TOP 1 to TOP 3 cells data sets from the scanner data log.

Due to the format similarity between UE data and scanner data, we apply ML to estimate UE RF condition based on the scanner data log (ex: possibility whether UE served by TOP 1 cell or not) and develop the "learn" function or hypothesis [2] to "learn" the RF condition based on the N inputs (fundamental parameter & derived parameters) from scanner data log and the probability of potential mobility problem as well.

The aim of this article is to introduce the basic idea of the RF estimation algorism with scanner log data by integrating ML method. In the rest, we organize as following: in Sect. 2, the related works will be summarized, in Sect. 3, the proposed algorithm will be explained, in Sect. 4, the simulation result will be analyzed, and in Sect. 5, this paper will be concluded.

2 Related Works

These days, many artificial intelligence (AI) theory can also be applied to cognitive radio field, like artificial neural networks (ANNs), metaheuristic algorithms, hidden Markov models (HMMs), rule-based systems, ontology-based systems (OBSs), and case–based systems (CBSs), which can help the wireless network operate better in the way of observation, analysis and prediction [3]. [4] introduced the major families of machine learning algorithms and discussed their applications in the context of next-generation networks, including massive MIMOs, the smart grid, cognitive radios, heterogeneous networks, small cells, D2D networks, and so on. By investigating the analysis of the spatial-temporal information of cellular traffic flow and the prediction of cell-station traffic volumes, based on the integration of K-means clustering, Elman Neural Network (Elman-NN), and wavelet decomposition methods, [5] characterize the performance comparison of traffic volume prediction. [6] applies three well-known ML algorithms combined with a non-intrusive cost-sensitive classification (CSC) scheme and predicts of the proposed time series model successfully which reaches false negative rates (FNRs) below 2%. In [7], a distributed Q-learning mechanism that exploits prior experience has been proposed to address the channel selection functionality that decides the most appropriate channel in the unlicensed band to set-up a LTE-U carrier for supplemental downlink as a means to facilitate the coexistence. In order to improve the Quality of Experience (QoE) of the user in the presence of obstacles, a ML based handover management scheme for LTE is presented by [8]. [9] introduced how utilizing ML algorithms along with software defined networking (SDN) and network function virtualization (NFV) on a diverse set of use cases and scenarios, for instance, through ML, a predictive model can be built of certain aspects of the environment in order to optimize the resources utilization for a better performance of the network.

3 The Proposed Algorithm

3.1 System Architecture

In our supervised learning system, firstly we obtain a learning dataset including the inputs and define output indicating the potential mobility problem.

By using the scanner machine, we get the various kinds of scanner data logs, and select nine parameters of them as shown Table 1 as the inputs.

Table 1. .

PCI of top1 cell	RSRP of top1 cell	RSRQ of top1 cell
PCI of top2 cell	RSRP of top2 cell	RSRQ of top2 cell
PCI of top3 cell	RSRP of top3 cell	RSRQ of top3 cell

To consider about the probability of UE mobility problem, we define as following: if Serving-PCI is equal to TOP1-PCI, the output will be “0”, which means UE has no mobility problem; else output will be “1”, which means UE has mobility problem.

Therefore output is a metric set {(0, 1)}. After the ML process, we can learn a function H(x) as a “good predictor” of the output.

In the future, when input the scanner data only and change the RF condition, the potential problem (low through put, VoLTE mute, radio link failure) could be estimated.

The flowchart of ML based learning system is shown as bellow in Fig. 1:

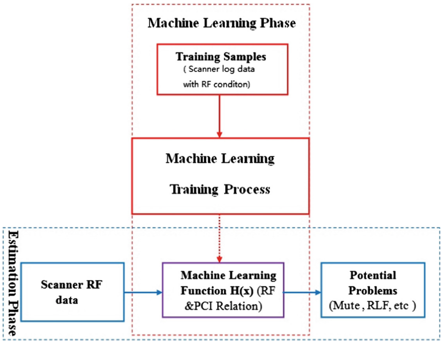


Fig. 1. .

3.2 Some Definitions and Basic Knowledges

(a) Target hypothesis $H(x)$ of linear regression

$$H(x) = \theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_n x_n = \sum\nolimits_{i=1}^{n} \theta_i x_i = \theta^T \vec{X} \tag{1}$$

where x_n is n^{th} feature and θ_n is the weight of the x_n. At final formula, we set parameters and arguments in the matrix of $\theta^T \vec{X}$, since it is easier for the matrix calculation.

(b) Cost function $J(\theta)$ of linear regression

$$J(\theta) = \frac{1}{2m}\sum\nolimits_{i=1}^{m}\left(h_{\theta}\left(x^{i}\right) - y^{i}\right)^{2} \tag{2}$$

where $J(\theta)$ describes the distance between the output Y and the estimated $H(x)$. $h_{\theta}\left(x^{i}\right)$ is n^{th} estimation, and y^{i} stands for n^{th} output of the learning samples. Certainly, the smaller J(θ) we can get, the better $H(x)$ would be.

(c) Gradient descent algorithm of linear regression

$$\widehat{\theta}_{J} = \theta_{j} - \alpha\sum\nolimits_{i=0}^{m}\left(h_{\theta}\left(x^{i}\right) - y^{i}\right)x_{i}^{i} \tag{3}$$

$$h_{\theta}\left(x^{i}\right) = \theta^{T}\vec{X} \tag{4}$$

Gradient descent algorithm can be used to calculate θ, and α is the learning rate in (3). Learning rate is a very important parameter at the iteration step. If the value of α is too small, iteration step will cost more time, however, if α is set too large, it could skip the expected minimum value at the learning step. Therefore, α needs to be adjusted according to actual situation.

In ML process, there are 2 main kinds of gradient descent schemes called stochastic gradient descent (SGD) and batch gradient descent (BGD). In our simulation, we choose SGD by considering the length of learning sample and the learning speed of the simulator. However, the result of BGD is close to the local optimum, not the global optimum.

(d) Target hypothesis $H(x)$ of Logistic regression

$$H(x) = g\left(\theta^{T}\vec{X}\right) = \frac{1}{1 + e^{-\theta^{T}\vec{X}}} \tag{5}$$

$$\theta^{T}\vec{X} = \theta_{0}x_{0} + \theta_{1}x_{1} + \cdots + \theta_{n}x_{n} = \sum\nolimits_{i=1}^{n}\theta_{i}x_{i} \tag{6}$$

(e) Cost function $J(\theta)$ of Logistic regression

$$J(\theta) = -\frac{1}{m}\left[\sum\nolimits_{i=1}^{m} y^{i} log\, h_{\theta}\left(x^{i}\right) + \left(1 - y^{i}\right) log\left(1 - h_{\theta}\left(x^{i}\right)\right)\right] \tag{7}$$

where $J(\theta)$ can be built by using Maximum Likelihood scheme.

(f) Gradient descent algorithm of Logistic regression

$$\widehat{\theta}_{J} = \theta_{j} - \alpha\sum\nolimits_{i=0}^{m}\left(h_{\theta}\left(x^{i}\right) - y^{i}\right)x_{i}^{i} \tag{8}$$

$$h_\theta\left(x^i\right) = \frac{1}{1 + e^{-\theta^T \vec{X}}} \tag{9}$$

3.3 The Proposed UE RF Condition Estimation Algorithm by Integrating ML

(a) Compare the results of linear regression and logistic regression

By applying serving RSRP and serving RSRQ respectively from the UE logs, after the learning step based on the liner regression scheme and the logistic regression scheme, two figures are shown as Figs. 2 and 3. We find that the probability is larger than 100% in some situations based on the liner regression scheme. As a result, we select the logistic regression scheme in the further simulation.

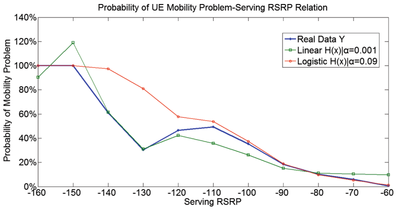


Fig. 2. .

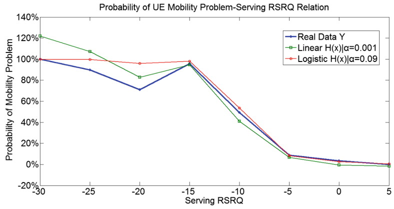


Fig. 3. .

(b) Select the learning rate "α"

In Fig. 4, we set α to different values. Based on the Fig. 4, we can know that the red curve is better than the other ones. As a result, in our simulation, learning rate α will be set as 0.01.

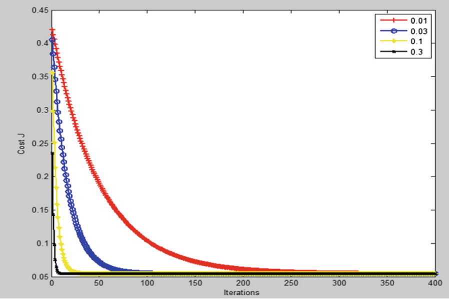


Fig. 4. .

(c) Input Data

In our simulation, by using the scanner, we get the data from two different areas and will apply them into three groups listed in Table 2 in order to prove the correctness of the proposed algorithm.

Table 2. .

Learning data group	Area A
Testing data group	Area B
Estimating data group	Area A

(d) Define features

By utilizing the nine parameters from Table 1, we create 24 derived features for the ML algorithm.

4 Numerical Simulation Result

4.1 Learning Step

At this step, we utilize scanner data log of area A, which contains 322,866 samples.

In Fig. 5, the relationship between the probability of the UE mobility problems and RSRP of top1 cell is shown. The trend of the two curves for the real data and learning data is almost the same. And at the zone [−88 dB, −76 dB] where most data samples are located at, the curve of the learning data is close to that of the real data.

In Fig. 6, the relationship between the probability of the UE mobility problems and RSRQ of top1 cell is shown. The trend of the two curves for the real data and learning data is almost the same. And at the zone [−11 dB, −5 dB] where most data samples are located at, the curve of the learning data is close to that of the real data.

In Fig. 7, the relationship between the probability of the UE mobility problems and the difference between RSRP of top1 and top2 cells is shown. The trend of the two curves for the real data and learning data is almost the same. And at the zone [5 dB, 15 dB]

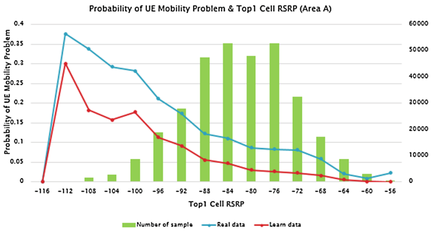


Fig. 5. .

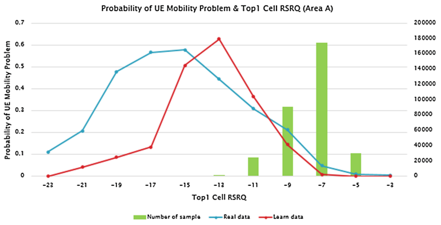


Fig. 6. .

where most data samples are located at, the curve of the learning data is close to that of the real data.

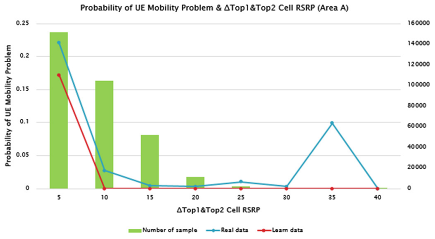


Fig. 7. .

Based on Figs. 5, 6 and 7, at the learn step, our algorithm works well. The learning curve can reflect the real data's situation mostly.

4.2 Testing Step

At this step, we change to use another area's scanner data log from area B, which contains 113,234 samples.

In Fig. 8, the relationship between the probability of the UE mobility problems and RSRP of top1 cell is shown. The trend of the two curves for the real data and testing data is almost the same. And at the zone [−96 dB, −68 dB] where most data samples are located at, the curve of the testing data is close to that of the real data.

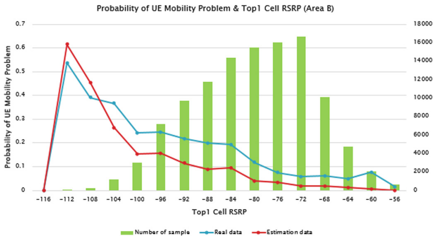


Fig. 8. .

In Fig. 9, the relationship between the probability of the UE mobility problems and RSRQ of top1 cell is shown. The trend of the two curves for the real data and testing data is almost the same. And at the zone [−11 dB, −5 dB] where most data samples are located at, the curve of the testing data is close to that of the real data.

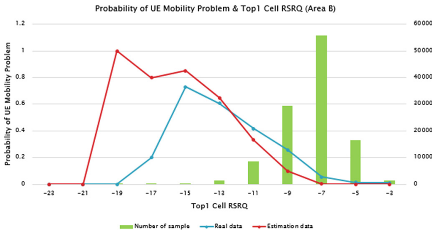


Fig. 9. .

In Fig. 10, the relationship between the probability of the UE mobility problems and the difference between RSRP of top1 and top2 cells is shown. The trend of the two curves for the real data and testing data is almost the same. And at the zone [5 dB, 15 dB] where most data samples are located at, the curve of the testing data is close to that of the real data.

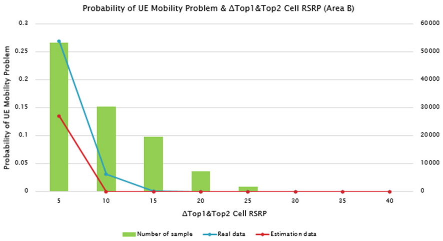
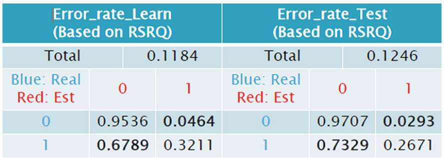


Fig. 10. .

At the testing step, based on Figs. 8, 9 and 10, the result of our algorithm is acceptable. The testing curve can reflect the real data's situation mostly.

As shown below in Table 3, we also calculate the error rate of the data based on the RSRQ of top1 cell.

Table 3. .



Error_rate_Learn (Based on RSRQ)			Error_rate_Test (Based on RSRQ)		
Total	0.1184		Total	0.1246	
Blue: Real Red: Est	0	1	Blue: Real Red: Est	0	1
0	0.9536	0.0464	0	0.9707	0.0293
1	0.6789	0.3211	1	0.7329	0.2671

At the learning step, the total error rate is 11.84%, and the error rate of "0" estimation is only 4.64%. At the testing step, the total error rate is 12.46%, and the error rate of "0" estimation is only 2.93%. This result is acceptable, as a result, this algorithm will be applied to the estimating step.

4.3 Estimating Step

At this step, we utilize scanner data log of area A and real UE data log.

In Fig. 11, the relationship between the probability of the UE mobility problems and RSRP of top1 cell is shown. The trend of the most parts of these two curves for the real UE data and estimating data is almost the same. Although at the zone [−116 dB, −96 dB], these two curves are not close to each other, however, at the zone [−84 dB, −68 dB] where most data samples are located at, the curve of the estimating data is close to that of the real data.

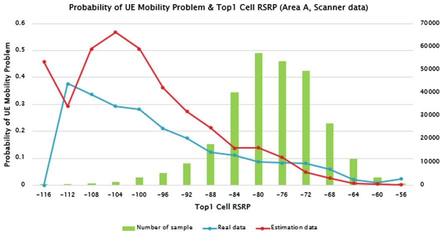


Fig. 11. .

In Fig. 12, the relationship between the probability of the UE mobility problems and RSRQ of top1 cell is shown. The trend of the two curves for the real data and estimating data is almost the same. And at the zone [−11 dB, −7 dB] where most data samples are located at, the curve of the estimating data is close to that of the real data.

In Fig. 13, the relationship between the probability of the UE mobility problems and the difference between RSRP of top1 and top2 cells is shown. The trend of the two curves for the real data and estimating data is almost the same. And at the zone [5 dB, 25 dB] where most data samples are located at, the curve of the estimating data is close to that of the real data.

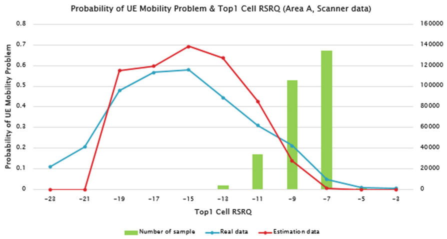


Fig. 12. .

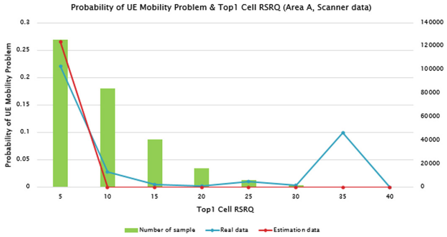


Fig. 13. .

At the estimating step, based on Figs. 11, 12 and 13, the result of our algorithm is very well. The estimating curve can reflect the real data's situation mostly.

4.4 Mapping Result

At the previous estimating step, the numerical simulation result is acceptable. In particular, the estimating curve based on the RSRQ of top1 is very good, as a result, we show the top1 cell's RSRQ map in Fig. 14. And as shown in the Fig. 14, different colors stands for different RF conditions, the darker green mean the better RF condition of top1 cell.

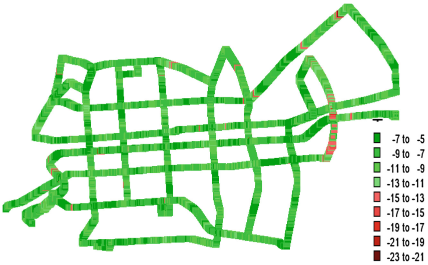


Fig. 14. .

In Fig. 15, we show the places with probability of UE mobility problems higher than 50% based on the real UE data on the map. Totally, there are 170 places. In Fig. 16, based on the proposed estimation algorithm, 772 places with probability of UE mobility problems higher than 50% are found.

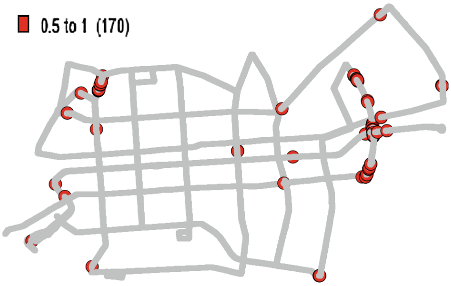


Fig. 15. .

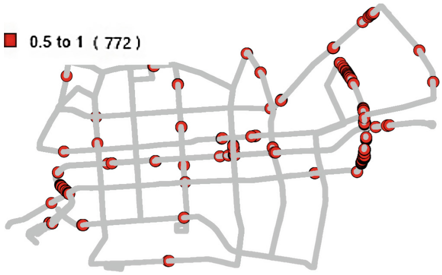


Fig. 16. .

By comparing Figs. 15 and 16, although the prediction places are more than the real places, considering the great number of the data samples, the estimation result is acceptable.

5 Conclusion

The proposed algorithm can estimate the potential UE mobility problems well. If the "machine" can keep learning more data samples, we believe that the result will became more accurate.

For the future research, we are considering to apply machine learning algorithms to help the LTE-U system learn the radio environment, such as the WiFi networks in the unlicensed band and set a suitable data transmission power upper band dynamically by apply the dynamic resource allocation scheme.

Acknowledgment. This paper is sponsored by China Scholarship Council.

References

1. Bagubali A, Prithiviraj V, Mallick PS, Krishnan KV (2012) Handover problem for integrating LTE with femtocell network. J Theor Appl Inf Technol 46(2):1007
2. Shwartz SS, David SB (2014) Handbook of "understanding machine learning: from theory to algorithms". Cambridge University Press, New York
3. He A, Bae KK, Newman TR, Gaeddert J (2010) A survey of artificial intelligence for cognitive radios. IEEE Trans Veh Technol 59:1578–1592
4. Jiang C, Zhang H, Ren Y, Han Z, Chen KC, Hanzo L (2016) Machine learning paradigms for next-generation wireless networks. IEEE Wirel Commun 20:2–9
5. Zang Y, Ni Y, Feng Z, Cui S, Ding Z (2015) Wavelet transform processing for cellular traffic prediction in machine learning networks. In: 2015 IEEE China summit and international conference on signal and information processing, 12–15 July 2015
6. Sue JA, Hasholzner R, Brendel J, Kleinsteuber M, Teich J (2016) A binary time series model of LTE scheduling for machine learning prediction. In: IEEE international workshops on foundations and applications of self* systems, 12–16 September 2016
7. Sallent O, Pérez-Romero J, Ferrús R, Agustí R (2015) Learning-based coexistence for LTE operation in unlicensed bands. In: IEEE international conference on communication workshop (ICCW 2015), 8–12 June 2015
8. Ali Z, Baldo N, Mangues-Bafalluy J, Giupponi L (2016) Machine learning based handover management for improved QoE in LTE. In: IEEE/IFIP network operations and management symposium (NOMS 2016), 25–29 April 2016
9. Buda TS, Assem H, Xu L (2016) Can machine learning aid in delivering new use cases and scenarios in 5G? In: IEEE/IFIP network operations and management symposium (NOMS 2016), 25–29 April 2016

Fairness-Aware Hybrid Resource Allocation with Cross-Carrier Scheduling for LTE-U System

Yupu Dong[1](✉), Zhenni Pan[1], Kang Kang[1], Jiang Liu[1],
Shigeru Shimamoto[1], Ragil Putro Wicaksono[2], Seiji Kunishige[2],
and Kwangrok Chang[2]

[1] Graduate School of Fundamental Science Engineering, WASEDA University,
Tokyo 169-8050, Japan
{Yupu.dong, z.pan, kangkang0611}@fuji.waseda.jp,
liujiang@aoni.waseda.jp, shima@waseda.jp
[2] MOTiV Research Co., Ltd., Tokyo 158-0097, Japan
{ragil.wicaksono, seiji.kunishige,
kwangrok.chang}@motiv-research.com

Abstract. Recently, the 3rd Generation Partnership Project (3GPP) proposes to extend the Long Term Evolution Advanced (LTE-A) to the unlicensed spectrum, named Long Term Evolution Unlicensed (LTE-U), which enables LTE to operate in both the licensed band and the unlicensed band. In this paper, we consider how to make LTE-U get even higher throughput in the high traffic unlicensed band. In order to achieve this target, there is a big challenge to make LTE-U a good neighbor to the existing wireless communication technologies in the unlicensed band, such as WIFI system in the 5 GHz band. In our research, we assume two carriers aggregated, one is from licensed band, and the other is from unlicensed. We also define two kinds of frequency resources in the unlicensed band. One is the normal frequency resource that has not been utilized by the WIFI systems, and the other is the special frequency resource that has been utilized by the WIFI systems already. We propose a novel hybrid resource allocation algorithm by combining two different frequency sharing schemes (Underlay and Interweave) and apply different resource allocation algorithm to achieve higher throughput for all kinds of user equipment (UE) from LTE-U when the WIFI system's traffic is heavy. We also consider the fairness for all UEs from LTE-U system and guarantee that the interference to the WIFI UEs (WUEs) is acceptable.

Keywords: Hybrid · Resource allocation · LTE-U · Unlicensed band · Carrier aggregation · Fairness · Cognitive radio

1 Introduction

The radio spectrum is a naturally limited resource. According to the Cisco's forecast study [1], the global mobile data traffic will be more than 8 exabytes per month by 2018 via the fourth-generation (4G) wireless communication networks.

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_13

In order to meet the explosive demand of various wireless application and services such as pictures, voices, and videos, especially the 4 K&8 K videos in the near future, 3GPP proposed carrier aggregation (CA) in Release 10 and beyond, which enables LTE-A to aggregate maximum 5 carriers [2] from licensed band with intra or inter CA schemes [3].

However, by only utilizing carriers from the remaining licensed band, LTE-A networks will become capacity constrained, thus impacting the user experience and preventing the mobile operators from offering the mobile data plans and services that are available today [4].

Recently, 3GPP proposes to extend LTE-A to the unlicensed spectrum, named LTE-U, which enables LTE to operate in both the licensed band and the unlicensed band. Although the licensed band remains 3GPP operators' top priority to serve QoS-guaranteed UE, the unlicensed band can be an alternative for operators to offload their traffic. Therefore, the utilization of the unlicensed band is an important complement to meet the ultra-high need especially in the Fifth-generation (5G) wireless communication networks [5].

One of the major ongoing discussion of 3GPP is how to make LTE-U a good neighbor with other existing wireless communication technologies in the unlicensed band. According to latest news, in some countries like Republic of Korea, the free WIFI hot-spot systems are constructed very quickly because of the government support. Therefore, due to the heavy traffic of the WIFI system, the Smart UEs (SUEs) cannot be allocated enough resource from the unlicensed band, although LTE-U has the capability to utilize the unlicensed band.

The unlicensed bands of current interest by 3GPP are mainly the 2.4 GHz band and the 5 GHz band. However, in the 2.4 GHz band, there already exist many wireless communication systems such as WIFI, Bluetooth, ZigBee, etc., whereas the environment of the 5 GHz band is relatively simpler with mainly WIFI system and LTE-U system deployed [6]. Therefore, we will utilize the 5 GHz band to apply the proposed hybrid resource allocation algorithm in the simulation.

The rest of the paper is organized as follows: in Sect. 2, the related works will be summarized; in Sect. 3, the proposed algorithm will be explained; in Sect. 4, the simulation result will be analyzed; in Sect. 5, this paper will be concluded.

2 Related Works

It is different with the CA in LTE-A, which just aggregates several carriers in the licensed band. In LTE-U, the carriers will be aggregated from both licensed band and unlicensed band, so the scheme about coexisting with the existing UE (like WUEs etc.) in the unlicensed band has to be considered. Currently, there are some works related to this topic. In [7], the effectiveness of listen-before-talk and interference-aware channel selection in LTE-U networks for coexisting with legacy indoor WIFI and other LTE-U networks deployed by multiple operators are analyzed in two major realistic coexistence scenarios (indoor LTE-U femtocell and outdoor LTE-U picocell deployments); in [8], an algorithm that is adaptive to the interference and channel conditions in both licensed and unlicensed bands is proposed for femto BSs to assign traffic to the licensed and unlicensed bands in a way that improves the overall utilities of all macro, femto and non-cellular WIFI users; in [9], the algorithm searches for the optimal power allocation in the licensed band and the optimal channel time usage in the unlicensed band; [10] shows that proportional fairness is achieved by assigning equal channel times to every competing entity including idle periods, successful transmissions and collisions for the WIFI network; [11] investigates the usage of the channel sensing method for LTE in unlicensed bands and proposed two channel sensing schemes (periodic sensing and persistent sensing); in [12], a distributed Q-learning mechanism that exploits prior experience is proposed; [13] quantitatively analyzes the inter-system interference between LTE and WLAN systems in the same unlicensed frequency based on the developed interference analysis technique. In [14–16], Listen before Talk (LBT) and Talk before Listen (TBL) in the LTE-U system are studied.

However, these works focused on the topic as to how to efficiently allocate the resource to SUEs and WUEs, respectively. In our search, we are considering to make SUE and WUE able to utilize the same frequency at the same transmission time interval (TTI) in the unlicensed band, while guaranteeing that the interference to each other is limited. Therefore, SUE form LTE-U system can get even high throughput when the unlicensed band traffic is heavy.

3 The Proposed Hybrid Resource Allocation Algorithm

3.1 Some Definitions and Basic Knowledges

(a) SUEs are served by the eNodeB (eNB) over the licensed band and the Smart Base Station (SBS) over the unlicensed band.
(b) WUEs, which are served by the WIFI Router (WiR) over the unlicensed band.
(c) Cellular UEs (CUEs), which are served only by the eNB over the licensed band.
(d) CA Scheduling schemes

In order to apply the CA in LTE-U, we use the simplest case, assuming only two carrier aggregated, one is from the licensed band and the other is from the unlicensed band. According to the 3GPP standard, there are two scheduling schemes for CA, which are the same carrier scheduling (Independent Scheduling) and cross-carrier scheduling. Although via both of these two scheduling schemes, the CA UE can increase their throughput, considering the fairness between CA UE and non-CA UE [17], the cross-carrier scheduling is better than the independent scheduling. Therefore, we will apply cross-carrier scheduling in the simulation.

(e) Current Spectrum sharing schemes

Normally, there are two spectrum sharing schemes, one is frequency domain sharing scheme, the other one is time domain sharing scheme [12].

Usually, in the same TTI, different UEs cannot utilize the same subcarrier. In other words, in different TTI, different UEs can utilize the same subcarrier.

(f) Current coexisting schemes in the frequency domain

According to the cognitive radio theory, there are three coexisting schemes in the frequency domain: Underlay, Overlay and Interweave [18]. In our research, we will utilize Underlay scheme and Interweave scheme. The characteristics of the two schemes are summarized as follows:

- Underlay, shown in Fig. 1:

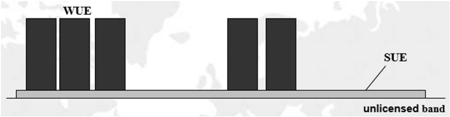


Fig. 1. Underlay coexisting scheme

(1) Channel Side Information: Cognitive (secondary) transmitter knows the channel strengths to non-cognitive (primary) receiver(s).
(2) Cognitive user can transmit simultaneously with non-cognitive user as long as the interference caused by cognitive user is below an acceptable limit.
(3) Cognitive user's transmit power is limited by the interference constraint.

- Interweave, shown in Fig. 2:

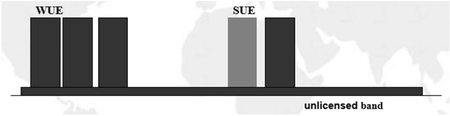


Fig. 2. Interweave coexisting scheme

(1) Activity Side Information: Cognitive user knows the spectral holes in space, time or frequency when the non-cognitive user is not using these holes.
(2) Cognitive user transmits simultaneously with a non-cognitive user only in the event of a false spectral hole detection
(3) Cognitive user's transmit power is limited by the range of its spectral hole's sensing.

(g) Proportional Fair (PF) Scheduling scheme

PF scheduling is trade-off between the throughput and fairness. It considers the UE's past average throughput and the current achievable throughput, so that each UE can be served with an appropriate priority. The calculations of PF scheduling can be written as [19]:

$$m_{i,j} = \frac{R_{i,j}(t)}{\bar{R}_i(t-1)} \tag{1}$$

$$\bar{R}_i(t) = \left(1 - \frac{1}{t_c}\right)\bar{R}_i(t-1) + \frac{1}{t_c} r_i(t-1) \tag{2}$$

where $R_{i,j}(t)$ is the current achievable throughput for the i^{th} UE on the j^{th} PRB. $\overline{R_i}(t)$ is the averaged throughput of the i^{th} UE for the past t_c TTI, $r_i(t)$ is referred as the instantaneous achievable throughput of the i^{th} UE in the TTI t.

Although from the aspect of the system throughput, some other scheduling schemes such as Round Robin scheduling and Best CQI scheduling may have better performance than the PF scheduling. However, from the aspect of the fairness of all the UEs' throughput, the PF scheduling has better performance than the other two. Therefore, we choose to apply PF scheduling in the simulation.

3.2 The Proposed Hybrid Resource Allocation Algorithm

Current studies are mostly based on the interweave scheme, which means two UEs from different wireless system cannot utilize the same frequency. For example, in the unlicensed band, if the UE from WIFI system hold 80% of frequency resources at an instantaneous TTI, then maximally only 20% of frequency resources can be allocated to SUE.

However, if we treat WUE as the primary user in the unlicensed band and SUE as the secondary user, then, according to Underlay scheme, even though the frequency resources has been utilized by WUE, SUE can still utilize the same frequency resources as long as the interference to WUE is acceptable. Thus, in the previous example, SUE can utilize 100% frequency resources from the unlicensed band. However, only 20% can be utilized freely by SUE, while for the other 80%, which has been utilized by WUE, SUE has to use a very limited power to do the data transmission in order to guarantee that the interference to WUE is under a special threshold ω_j. The system will learn the environment via the spectrum sensing technology from cognitive radio theory and set the value of ω_j accordingly.

In order to introduce the proposed algorithm a little bit easier, we are going to introduce it by three different level scenarios, which are link level scenario, cell level scenario and system level scenario.

(a) Link level scenario

In the link level scenario, we assume only one CUE and one SUE. In this step, we explain the algorithms of each kind of UE's throughput in the normal situation, respectively. The architecture is shown in Fig. 3:

Cellular UE:

$$Throughput_{c,l} = b\ \log\left(1 + SNR_{c,l}\right) \tag{3}$$

$$SNR_{c,l} = \frac{P_{c,l,r}}{N_{c,l,0}} = \frac{P_{c,t} \cdot |H_c|^2}{N_{c,l,0}} \tag{4}$$

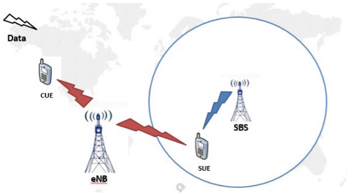


Fig. 3. Link level scenario

where $Throughput_{c,l}$ refers to the throughput of the c^{th} CUE in the licensed band, $P_{c,l,r}$ is the receive power, $P_{c,t}$ is the transmission power, H_c is the channel gain, and $N_{c,l,0}$ is the White Gaussian Noise.

Smart UE:

- Licensed Band

$$Throughput_{s,l} = b\ \log\left(1 + SNR_{s,l}\right) \tag{5}$$

$$SNR_{s,l} = \frac{P_{s,l,r}}{N_{s,l,0}} = \frac{P_{s,l,t} \cdot \left|H_{s,l}\right|^2}{N_{s,l,0}} \tag{6}$$

- Unlicensed Band

$$Throughput_{s,ul} = b\ \log\left(1 + SINR_{s,ul}\right) \tag{7}$$

$$SINR_{s,ul} = \frac{P_{s,ul,r}}{I_{s,ul,w} + N_{s,ul,0}} = \frac{P_{s,ul,t} \cdot \left|H_{s,ul}\right|^2}{P_{w,ul,t} \cdot \left|H_{ws,ul}\right|^2 + N_{s,ul,0}} \tag{8}$$

- Both licensed and unlicensed (normal)

$$Throughput_{s,l,ul} = Throughput_{s,l} + Throughput_{s,ul} \tag{9}$$

where $Throughput_{s,l}$ refers to the throughput of the s^{th} SUE in the licensed band, $Throughput_{s,ul}$ refers to the throughput of the s^{th} SUE in the unlicensed band, and $Throughput_{s,l,ul}$ refers to the throughput of the s^{th} SUE in both the licensed and unlicensed band totally. Because of the interference from the WiR, SINR is used here and $I_{s,ul,w}$ is the interference from WiR to SUE in the unlicensed band.

As mentioned in III-A-g, we apply the PF scheduling in the research, in case there are some SUEs with very bad channel quality between the SBS, which could make the SBS keep allocate the subcarriers to this bad SUE. In order to make SUE utilize the unlicensed band more efficiently, we set a threshold based on CQI at this step. Only the SUE, whose CQI value is larger than the CQI threshold, can utilize the unlicensed band. And this threshold is constrained to the unlicensed band traffic in each TTI.

(b) Cell level scenario

In the cell level scenario, we assume that there are multiple CUEs and multiple SUEs in the system. In this step, we explain the algorithms of PF-factors in different CA scheduling schemes as mentioned in III-A-d, respectively. The architecture is shown in Fig. 4:

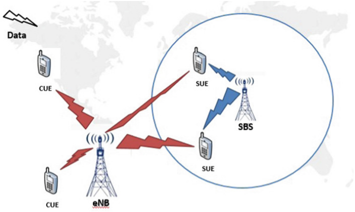


Fig. 4. Cell level scenario

- Proportional Fairness Factor (PF-factor) for Independent Scheduling scheme

$$P_{k,i,m} = \frac{R_{k,i,m}(t)}{R_{k,i}(t-1)} \tag{10}$$

- Proportional Fairness Factor for Cross-carrier Scheduling scheme

$$P_{k,i,m} = \frac{R_{k,i,m}(t)}{R_{k,total}(t-1)} \tag{11}$$

where $P_{k,i,m}$ refers to the PF-factor of the k^{th} UE at the m^{th} subcarrier in the i^{th} carrier.

As mentioned in III-A-d), we will only deduce (11) for CUE and SUE, respectively. At this step, we set a time block t_c to calculate the UE's average throughput in past t_c TTI.

<u>Cellular UE:</u>

$$P_{c,l,m}(t) = \frac{Throughput_{c,l,m}(t)}{\sum_{T=t-t_c}^{T=t-1} Throughput_{c,l}(T)/t_c} \tag{12}$$

where $P_{c,l,m}(t)$ is the PF-factor of the c^{th} CUE at the m^{th} subcarrier at the t^{th} TTI in the licensed band, $Throughput_{c,l}(t)$ is the achievable throughput of the c^{th} CUE at the m^{th} subcarrier at the t^{th} TTI in the licensed band, and $\sum_{T=t-t_c}^{T=t-1} Throughput_{c,l}(T)/t_c$ is the average throughput of the c^{th} CUE during the time block t_c, which are from $(t-t_c)$ to $(t-1)$ in the licensed band.

Smart UE:

- Licensed Band

$$P_{s,l,m}(t) = \frac{Throughput_{s,l,m}(t)}{\sum_{T=t-t_c}^{T=t-1} Throughput_{s,l,ul}(T)/t_c} \tag{13}$$

- Unlicensed Band

$$P_{s,ul,n}(t) = \frac{Throughput_{s,l,n}(t)}{\sum_{T=t-t_c}^{T=t-1} Throughput_{s,l,ul}(T)/t_c} \tag{14}$$

where $P_{s,l,m}(t)$ is the PF-factor of the s^{th} SUE at the m^{th} subcarrier at the t^{th} TTI in the licensed band, $P_{s,ul,n}(t)$ is the PF-factor of the s^{th} SUE at the n^{th} subcarrier at the t^{th} TTI in the unlicensed band, and $\sum\limits_{T=t-t_c}^{T=t-1} Throughput_{s,l,ul}(T)/t_c$ is the average throughput, which is the sum throughput from both the licensed band and the unlicensed band, of the s^{th} SUE during the time block t_c from $(t-t_c)$ to $(t-1)$.

And in the simulation, for the t^{th} TTI, the specific subcarrier will be allocated to the UE with the largest PF-factor, as shown in Eqs. (15) and (16),

- Licensed Band

$$\mathrm{MAX}\left(P_{c,l,m}(t), P_{s,l,m}(t)\right), \mathrm{c} = 1, 2, \cdots, \mathrm{C}; \mathrm{s} = 1, 2, \cdots, \mathrm{S}. \tag{15}$$

- Unlicensed Band

$$\mathrm{MAX}\left(P_{s,ul,n}(t)\right), \mathrm{s} = 1, 2, \cdots, \mathrm{S}. \tag{16}$$

till all the subcarrier are allocated to the highest priority UE, respectively.

(c) System level scenario

In the system scenario, we assume multiple CUEs, multiple SUEs and one WUE. In this step, the intra-band interference will be considered, and we only consider the downlink of both LTE-U system and WIFI system so that SUEs will get the interference from the WiR. Meanwhile, WUE will get interference from the SBS as in Fig. 5.

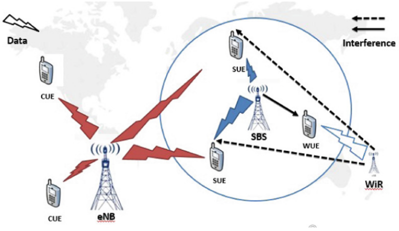


Fig. 5. System level scenario

For the free frequency, SUE's throughput can be calculated by Eq. (7), but for the special frequency, the normal throughput algorithm is not suitable because the interference to WUE has to be limited.

In order to guarantee WUE's throughput, the SINR of WUE should be larger than $Threshold_w$ as follow:

$$SINR_w = \frac{P_{w,r}}{\omega_j + N_0} = \frac{P_{w,t} \cdot |H_w|^2}{\omega_j + N_0} > Threshold_w \tag{17}$$

Therefore, ω_j should be below $Threshold_I$, as follow:

$$\omega_j < Threshold_I \tag{18}$$

where $Threshold_I$ is the threshold for the interference from LTE-U to WUEs. In order to guarantee that the ω_j is limited, we set an upper band δ_m for the transmission power of the LTE-U system allocated at the subcarriers from special frequency.

We define $P^r_{\sigma,\rho}$ as the minimum required receive power at any subcarrier to make sure that bit error rate (BER) is below the threshold ρ, and σ is the number of bits for each symbol [20]. For different modulation type, different transmission power is required at each subcarrier. For example, for the BPSK modulation (σ = 1), the minimum required power is as Eq. (19):

$$P^r_{1,\rho} = \frac{N_\varnothing}{2}\left[Q^{-1}(\rho)\right]^2 \tag{19}$$

where $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-t^2/2}dt$ and $N_\varnothing$ is the single sided noise power spectral density (PSD). For QPSK (σ = 2), M – ary QAM ($\sigma = 4, 6, \cdots$), $P^r_{\sigma,\rho}$ is given by [21]. For transmitting even number of bits per symbol can be written as:

$$P^r_{\sigma,\rho} = \frac{(2^\sigma - 1)N_\varnothing}{3}\left\{Q^{-1}\left[\frac{\rho\sqrt{2^\sigma}}{4\left(\sqrt{2^\sigma} - 1\right)}\right]\right\}^2 \tag{20}$$

For 16-QAM (σ = 4), the minimum required power for the BER threshold ρ can be written as:

$$P^r_{4,\rho} = 5N_{\varnothing}\left[Q^{-1}\left(\frac{\rho}{3}\right)\right]^2 \tag{21}$$

In order to satisfy the BER requirement, at the SBS, the minimum transmitter power $P^t_{\sigma_{s,m},\rho}$ assigned to the m^{th} subcarrier for the s^{th} SUE is written as [21]:

$$P^t_{\sigma_{s,m},\rho} = \frac{P^r_{\sigma_{s,m},\rho}}{H_{s,m}} \tag{22}$$

where $H_{s,m}$ is the magnitude of the channel gain of the s^{th} SUE over the m^{th} subcarrier.

Assuming $P^t_{0_{s,m},\rho} = 0$, the power allocated to the m^{th} subcarrier can be written as:

$$\delta_m = \sum\nolimits_{s=1}^{S} P^t_{\sigma_{s,m},\rho} \tag{23}$$

In order to get the maximum total throughput via the special frequency, as shown in (24)

$$\text{maximum total throughput}_{Spe} = \text{Max}(rate) \tag{24}$$

where *rate* is the total throughput via the special frequency, which can be written as:

$$\text{rate} = \sum\nolimits_{m=1}^{M} \sum\nolimits_{s=1}^{S} \sigma_{s,m} \tag{25}$$

subject to:

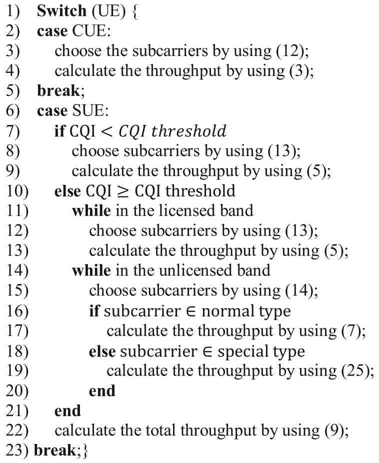
$$\sum\nolimits_{m=1}^{M} \sum\nolimits_{s=1}^{S} P^t_{\sigma_{s,m},\rho} \le P_{max} \tag{26}$$

$$\sum\nolimits_{m=1}^{M} \delta_m G^W_{j,m} \le \omega_j, j = 1, 2, \cdots, J \tag{27}$$

where P_{max} is the total power available at the SBS, and regarding to the number of bits per symbol, we just consider four modulation types: BPSK, 4-QAM, 8-QAM, 16-QAM, which means $\sigma_{k,m} = 0, 2, 3, 4$, and $\sigma_{s,m} = 0$ means s^{th} SUE transmit 0 bit over the m^{th} subcarrier.

The pseudo code of the proposed hybrid resource allocation algorithm is shown is Table 1:

Table 1. Pseudo code of the proposed hybrid resource allocation algorithm



```
Switch (UE) {
case CUE:
    choose the subcarriers by using (12);
    calculate the throughput by using (3);
break;
case SUE:
    if CQI < $CQI\ threshold$
        choose subcarriers by using (13);
        calculate the throughput by using (5);
    else CQI ≥ CQI threshold
        while in the licensed band
            choose subcarriers by using (13);
            calculate the throughput by using (5);
        while in the unlicensed band
            choose subcarriers by using (14);
            if subcarrier ∈ normal type
                calculate the throughput by using (7);
            else subcarrier ∈ special type
                calculate the throughput by using (25);
            end
    end
    calculate the total throughput by using (9);
break;}
```

4 Numerical Result

In order to analyze the proposed algorithm’s performance, the following results will be presented:

(a) Average throughput for all kinds of UEs in LTE-U system

(b) Growth percentage of average throughput

$$\text{Growth percentage of average throughput} = \frac{\text{throughput}_{pa} - \text{throughput}_{ca}}{\text{throughput}_{ca}} \times 100\% \quad (28)$$

where throughput_{pa} is the average throughput of the proposed algorithm, throughput_{ca} is the average throughput of the current algorithm.

(c) Fairness index for the average throughput of all UEs

We utilize Eq. (29) to calculate the fairness index:

$$Fairness\,Index = \frac{\left(\sum_{i=1}^{I} Average\,throughput_i\right)^2}{I \cdot \sum_{i=1}^{I} Average\,throughput_i^2} \quad (29)$$

where I is the number of all UEs.

In the simulation, the main simulation parameters are set as follows in Tab. 2:

Simulation results are shown in Figs. 6, 7 and 8:

In Fig. 6, when the WIFI system has already utilized 64% to 96% frequency resources of the unlicensed band, the SUEs from LTE-U system can get less and less frequency resource from the unlicensed band. As a result, the average throughputs of the CUEs and the SUEs from LTE-U system decrease respectively. However, compared to the current algorithm, both two kinds of UEs from LTE-U system can get benefit by applying the proposed algorithm.

Table 2. Main simulation parameters

Time	4000TTI
Time block	200TTI
Carrier from licensed bandwidth	5 MHz
Carrier from unlicensed bandwidth	5 MHz
CUE number	40
SUE number	40
WUE number	20

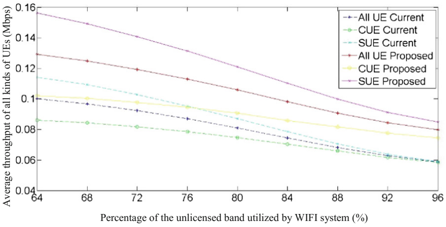


Fig. 6. .

In Fig. 7, after applying the proposed algorithm, the growth percentage of the average throughput of both CUE and SUE increase, when the percentage of unlicensed band utilized by the WIFI system is increased from 64% to 96%.

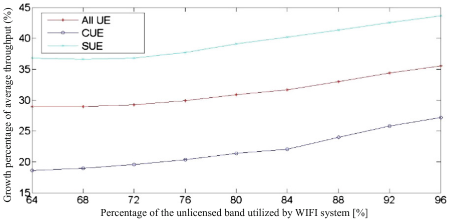


Fig. 7. .

In Fig. 8, the Fairness Index of the average throughput of all UEs in the LTU-U under the proposed algorithm and the current algorithm is calculated and compared. It is shown that the fairness index of the average throughput under the proposed algorithm is lower compared to the Fairness Index under the current algorithm. However, since the proposed algorithm has increased tremendously the average throughput of all kinds of UEs according to Fig. 6, the decrease of the fairness is acceptable.

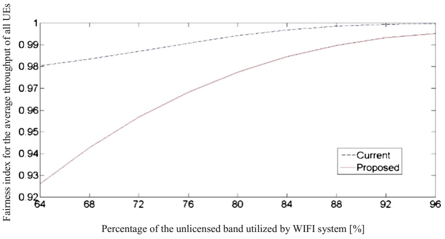


Fig. 8. .

5 Conclusion

In this paper, the proposed algorithm aggregated two carriers, which are from the licensed band and the unlicensed band respectively, and defined two kinds of frequency (normal frequency and special frequency). For the normal frequency, we applied the Interweave frequency sharing scheme, while for the special frequency we applied the Underlay frequency sharing scheme. Hybrid resource allocation algorithm is also applied to these two different kinds of frequency in the unlicensed band.

The proposed algorithm increased the average throughput of both CUEs and SUEs from LTE-U system. Meanwhile, based on the proposed algorithm, the higher the percentage of the unlicensed spectrum is utilized by WIFI system, the higher growth percentage of average throughput will be achieved for all kinds of UEs from LTE-U system.

However, there appears to be a trade-off between the average throughput and the fairness for all UEs from LTE-U. Under the proposed algorithm, the Fairness Index of the average throughput is lower. Nevertheless, since the proposed algorithm has increased tremendously the average throughput of all kinds of UEs, the decrease of the fairness is acceptable.

These days, artificial intelligence (AI) theory such as artificial neural networks (ANNs), metaheuristic algorithms, hidden Markov models (HMMs), rule-based systems, ontology-based systems (OBSs), and case–based systems (CBSs) h applied to cognitive radio field. The application of AI theory can help to assist in way of observation, reconfiguration and cognition of cognitive radio networks [22]. In future research, we will consider to apply AI algorithms to help the LTE-U system learn the radio environment and set a suitable data transmission power upper band dynamically.

Acknowledgment. This paper is sponsored by China Scholarship Council.

References

1. Cisco White Paper (2014) Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013–2018
2. 3GPP TR 36.814 v. 1.2.1 (2009) Further Advancements for EUTRA: Physical Layer Aspects, Technical Specification Group Radio Access Network, Rel. 9, June 2009
3. Yuan G, Zhang X, Wang W, Yang Y (2010) Carrier aggregation for LTE-advanced mobile communication systems. IEEE Commun Mag 48:88–93
4. Signals Research Group, The Prospect of LTE and WIFI sharing unlicensed Spectrum, February 2015
5. Al-Dulaimi A, Al-Rubaye S, Ni Q, Sousa E (2015) 5G communications race: pursuit of more capacity triggers LTE in unlicensed band. IEEE Veh Technol Mag 10:43–51
6. Zhang R, Wang M, Cai LX, Zheng Z (2015) LTE-unlicensed: the future of spectrum aggregation for cellular networks. IEEE Wirel Commun 22:150–159
7. Voicu AM, Simic L, Petrova M (2015) Coexistence of pico- and femto-cellular LTE-unlicensed with legacy indoor Wi-Fi deployments. In: 2015 IEEE international conference on communication workshop (ICCW), June 2015, pp 2294–2300
8. Liu F, Erkip E, Beluri MC, Yang R (2012) Dual-band femtocell traffic balancing over licensed and unlicensed bands. In: 2012 IEEE international conference on communications (ICC), June 2012, pp 6809–6814
9. Liu F, Bala E, Erkip E, Beluri MC (2015) Small-cell traffic balancing over licensed and unlicensed bands. IEEE Trans Veh Technol 64:5850–5865
10. Cano C, Leith DJ (2015) Coexistence of WiFi and LTE in unlicensed bands: a proportional fair allocation scheme. In: 2015 IEEE international conference on communication workshop (ICCW), June 2015, pp 2288–2293
11. Jia B, Tao M (2015) A channel sensing based design for LTE in unlicensed bands. In: 2015 IEEE international conference on communication workshop (ICCW), June 2015, pp 2332–2337
12. Sallent O, Perez-Romero J, Ferrus R, Agusti R (2015) Learning-based coexistence for LTE operation in unlicensed bands. In: 2015 IEEE international conference on communication workshop (ICCW), June 2015, pp 2307–2313
13. Jeon J, Li QC, Niu H, Papathanassiou A (2014) LTE in the unlicensed spectrum: a novel coexistence analysis with WLAN systems. In: 2014 IEEE global communications conference (GLOBECOM), December 2014, pp 3459–3464
14. Jeon J, Niu H, Li Q, Papathanassiou A (2015) LTE with listen-before-talk in unlicensed spectrum. In: 2015 IEEE international conference on communication workshop (ICCW), June 2015, pp 2320–2324
15. Xia P, Teng Z, Wu J (2015) How loud to talk and how hard to listen-before-talk in unlicensed LTE. In: 2015 IEEE international conference on communication workshop (ICCW), June 2015, pp 2314–2319
16. Chen C, Ratasuk R, Ghosh A (2015) Downlink performance analysis of LTE and WiFi coexistence in unlicensed bands with a simple listen-before-talk scheme. In: Vehicular technology conference (VTC Spring), May 2015, pp 1–5
17. Frias Z, Ventura M, Perez J (2013) Analysis of a transportation problem based scheduling algorithm for inter-band carrier aggregation. In: 2013 IEEE 24th international symposium on personal indoor and mobile radio communications (PIMRC), pp 2434–2438
18. Goldsmith A, Jafar SA, Maric I, Srinivasa S (2009) Breaking spectrum gridlock with cognitive radios: an information theoretic perspective. Proc IEEE 24:894–914

19. Lin Y-C, Lai WK, Yang K-T, Lin M-T (2012) An novel scheduling algorithm for video stream in LTE. In: Proceedings of IEEE 2012 conference on genetic and evolutionary computing (ICGEC), pp 107–110
20. Mao Z, Wang X (2008) Efficient optimal and suboptimal radio resource allocation in OFDMA system. IEEE Trans Wirel Commun 7:440–445
21. Proakis JG (1995) Digital Communications. McGraw-Hill, New York
22. He A, Bae KK, Newman TR, Gaeddert J (2010) A survey of artificial intelligence for cognitive radios. IEEE Trans Veh Technol 59:1578–1592

An IoT Implementation for Vacancy State of Public Coin Operated Washing Machine Using Vibration Level Sensors in an Apartment Setting in Thailand

Pisal Setthawong[1](✉), Anan Osothsilp[1], and Tuul Triyason[2]

[1] Department of Business Information Systems, Assumption University, Bangkok, Thailand
{pisalstt,ananstth}@msme.au.edu
[2] School of Information Technology, King Mongkut's University of Technology, Thonburi, Bangkok, Thailand
tuul.tri@sit.kmutt.ac.th

Abstract. Public coin operated washing machines are common in many apartment complexes in Thailand. As many tenants are from lower socioeconomic standing, many tenants do not own a washing machine. Due to that, public coin operated washing machines serve as an affordable solution for tenants when doing their laundry. However, in many apartments, the number of public coin operated washing machines are usually inadequate compared with the number of tenants, which can lead to long queues. This research aims to help alleviate the issue, by utilizing an IoT solution using vibration level sensors to provide tenants the ability to be able to know if the public washing machines are vacant. The IoT solution proposed can be easily installed in many apartment complexes using the existing WiFi network infrastructure that is available in many apartment complexes.

Keywords: IoT · Washing machines · IoT applications · State machines

1 Introduction

Public coin operated washing machines are considered as an important service offered by many apartments complexes, student dormitories, and lower-end condominium in Thailand. Most of the tenants usually do not own a washing machine due to many reasons. For most tenants, many of them are from lower socioeconomic standing therefore many of the tenants cannot afford washing machines. For other tenants, there is the possibility of the limited space of the rented room makes the installation of washing machine difficult and may be prohibited by the apartment regulations. In other cases, some tenants choose not to procure washing machines due to the difficulty of moving the washing machines when the tenant decides to move to a new place. As most tenants in apartments do not own washing machines and laundry services are considerably more expensive, many tenants resort to using public coin operated washing machines.

For apartment complexes and student dormitories in Thailand, it is common to see many coin operated washing machines installed with a wide range of configurations. Many apartments will allocate an area on the ground floor in which is accessible for the

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_14

tenants to install the washing machines. However typical apartments allocate most of the space towards rooms for rent, the public area for the washing machines are usually cramped and do not have basic amenities.

Many of the tenants usually wait till their laundry bin is full before doing their laundry. Due to that behavior, many tenants carry huge baskets of laundry to the washing machines from their rooms which could be at different floors. This can be an arduous endeavor as in lower-end low-rise apartments, elevators are considered as an option. Once the tenants reached the laundry room, it is possible that all the washing machines are occupied and a queue could be formed. As the public room for washing machines are usually cramped, it is not comfortable to wait for an extended period of time. The tenant may choose to wait, or return to their rooms and try their luck later.

Based on tenants' feedback, many tenants mention that they would like to know if the washing machines are vacant or not. By knowing the state of the washing machines, it would be possible for tenants to plan if they should carry their laundry from their room. Based on that feedback, the research aims to help alleviate the issue by providing an IoT solution that can help tenants to know the vacancy of the coin operated washing machines that are offered in the apartment. The IoT solution should also be relatively cheap, be able to detect the vacancy of the washing machines, be accessible with web browsers both desktop and mobile, and should fit in the existing network infrastructure of typical apartments.

2 Background

In this section, a background is provided on several topics that are of interest regarding the domain. The first subsection discusses about IoT applications in the domain that are like the research that have been proposed. After the first subsection, background on the environmental factors such as the typical apartment infrastructure in Thailand, and the nature of typical public coin operated washing machines are provided.

2.1 IoT Applications

Internet of Things (IoT) are a series of interconnected devices such as sensors, software, electronics which are connected via a network infrastructure to allow the objects to collect and exchange data. With better network infrastructure and more robust and affordable microcontrollers, IoT applications are becoming more common and are used in many different domains. IoT applications have been deployed in many domains, and can be varied [4] ranging from home appliances, work domain, exercise, and even in the production of food [10].

In the example of appliances, IoT provide interesting avenue for research. As traditional electronic appliances are standalone, their usage is limited by the traditional design of the appliances. As home appliances are well defined consumer products, there is less avenue for product improvement. For example, improvements for a traditional washing machine will focus on improving and optimizing the controller for the washing cycles [3].

By utilizing IoT in home appliances and providing the interconnectivity to the home appliances, there are many applications and extensions that could be built on existing electronic appliances. By building on home appliances, it is possible to use IoT to extend and supersede traditional appliances [2]. By providing connected devices, instead of standalone appliances, IoT appliances can word in tandem and provide interconnected features which can greatly add to the functionality of the device. In cases of IoT appliances, some approaches even propose that appliances can move from embedded microcontroller to cloud based controllers [5]. Based on the potential of IoT appliances, it is observed that traditional washing machines can be improved by IoT approaches.

2.2 Typical Apartment Infrastructure in Thailand

Apartments in Thailand typically serve people of lower socioeconomic standing. As many of the tenants have lower disposable income, the apartments usually have smaller sized room, and little amenities in the apartment. Typical configurations usually contain a public coin operated zone where there are washing machines and water kiosks, and occasionally a small minimarket in the premise. Figure 1 displays an example of a public coin-operated washing machine zone in an apartment.

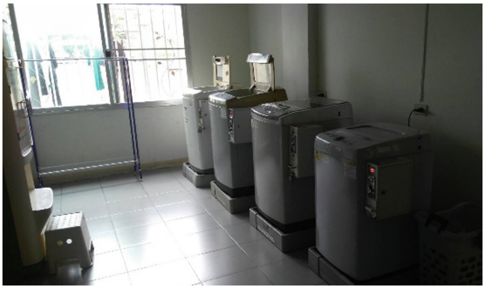


Fig. 1. Coin operated washing machine in apartments

Due to building and zoning laws, apartments are typically five stories tall since they are not in high rise zones. Due to the limited number of floors in apartment buildings, it is not mandated by law to have elevator systems setup in the apartment. For apartments without elevator systems and a limited supply of publicly coin operated washing machines, many tenants complain about lifting heavy laundry via the stairs to long queue.

Though apartments usually are not well equipped in facilities, one sector that is well equipped are usually the WiFi network infrastructure of the building. As many Thais are addicted to their mobile phones and use the Internet extensively [6], the availability of Internet services is considered as one of the important factors in selecting an apartment. The Internet must be fast and stable, and there should be strong WiFi access within the building premises.

In any publicly shared Internet, based on Thai law [9], all service providers are required to audit all users' Internet usage and browsing patterns. Due to that, many apartments have installed access gateways to log all the Internet usage of the tenants that are using the shared infrastructure. Due to that, it is observed that many apartments have a rather advanced network infrastructure.

2.3 Coin Operated Washing Machine

Coin operated washing machines that are found in Thailand are typically converted from consumer washing machines. In the conversion, cheaper and lower-end washing machines are converted as cheaper washing machines typically contain less complex internal circuits that in turn makes the conversation process simpler [7]. Another curious feature is that the settings of the coin operated machine are fixed to a default wash, rinse, and spin cycle and tenants are unable to change the settings of the machine. As there exists certain wash modes that require significantly more time than the default settings, owners of coin operated washing machines do not want tenants to use more time-consuming modes. Tenants will fill in coins to the amount between ~$0.56 to ~$1.42 in order to start the cycle, in which the cost is usually defined by the capacity of the washing machine, which was predetermined. There also exist variations of the coin operated washing machines in which settings can be changed after the tenant has input the amount of credits to start the cycle, but this is less common and usually more expensive than fixed cycle washing machines. An example of a washing machine that is converted into a coin operated washing machine is displayed in Fig. 2.

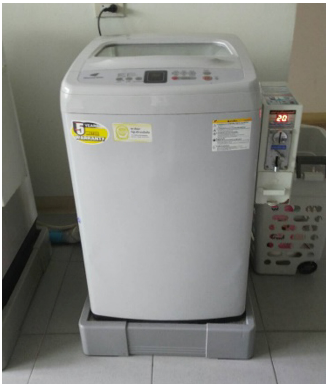


Fig. 2. Example of coin operated washing machine

3 Proposed Platform

The discussion from the last section provides the background on the requirements of proposed system. The proposed IoT platform should be able to detect if the washing machine is active or not via an IoT device. The device should upload the data to a server that is present in the apartment WiFi network infrastructure. With the data uploaded to the server, the tenants would be able to check to the status of the washing machine by accessing the web interface of the application to check the vacancy state of all the washing machine.

3.1 Microcontroller Node

The first part of the proposed platform is the microcontroller node that would be used to check if the washing machine is active or not. As established in the background, washing machines have vibration patterns during the wash cycle. The microcontroller node should be able to detect the vibration and transfer the data to the server for later processing.

The microcontroller node is based on Node NCU (ESP8266-12e) microcontroller. The microcontroller is a small and cheap microcontroller that is capable of WiFi transmission and connection to external sensors. With WiFi access, it is possible to connect via the local WiFi network infrastructure of the apartment to connect to the server node. A vibration sensor is attached to the node to detect the vibration emission from the washing machine. The microcontroller node is enclosed in the case to complete the node. The cost of the microcontroller node is displayed in Table 1 and the completed microcontroller node is shown in Fig. 3.

Table 1. Approximate cost of microcontroller node

Microcontroller Node	
Node NCU (ESP8266-12e)	$10
Vibration Sensor	$2
Cabling	$1
Case	$1
Total	$14

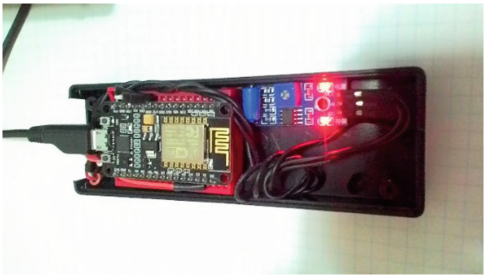


Fig. 3. Completed microcontroller node

With the completed microcontroller node completed, the node is configured with the washing machine ID, to connect to the local cloud, and server details before being attached to its assigned washing machine. The details of the microcontroller deployment are displayed in Fig. 4.

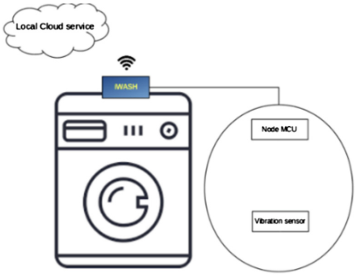


Fig. 4. Controller node setup

3.2 Software Backend/Frontend and Topology

The next component of the proposed system is the backend. Once the microcontroller node is configured, setup, and deployed on the machine, the backend is created and setup to fit with the setup in the apartment. Based on the discussion earlier regarding the legal requirements in Thailand to log all shared Internet user activities [9], many apartments have an access gateway that has a web server stack as part of their shared WiFi network infrastructure. To save cost, the backend application is deployed on the same server as the access gateway.

The backend is created on the web stack. The data is saved in a MySQL database, and the web application is built based on PHP. The microcontroller node updates data into the web database, in which populates the details of the machines.

To allow the tenants to connect, a front end is developed to display the useful information of the washing machines to the tenants. The tenants can choose to connect via the web application or mobile application to examine the system. The first choice is to connect to the web application via their desktop browsers or mobile browsers to access the service. Alternatively, the tenants can also connect to the mobile application designed specifically for the apartment.

Once connected to the web application, the tenants will be able to view the vacancy state of the machines, and examine for those that are in use, when the machine was used in the current wash cycle to provide a good estimate for the next available time.

Regarding the setup, the system could be configured to be publicly accessible via Internet, or the system would be only accessible via the local network. Though most apartment would deploy the system only in the local network to support tenants that are in the same network, there exists cases in which the apartment may also provide the coin operated washing service to the general public in close vicinity. In this scenario, making the system publicly available via cloud [8] will also fit the use scenario better. The details of the topology are displayed in Fig. 5.

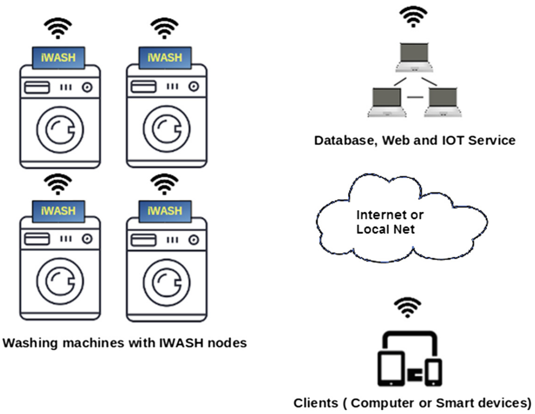

IWASH
IWASH
IWASH
IWASH
Washing machines with IWASH nodes
Database, Web and IOT Service
Internet or
Local Net
Clients (Computer or Smart devices)

Fig. 5. Topology of proposed system

4 Experiment and Results

The proposed platform after been implemented and been tested with a selected washing machine for a pilot project. The selected washing machine selected was Electrolux Washing Machine EWF1074 [1], a 7 kg front loading washing machine. The micro-controller node was placed on the machine, and recorded with the default wash cycle.

The default wash cycle consists of the wash, rinse, and spin modes. Before the wash cycle, in the default settings, the washing machine does a watering and sensing process before the washing cycle. The pre-wash cycle consisting of the watering and sensing process lasts ~240 s. The washing machine then does the wash and rinse cycle, before

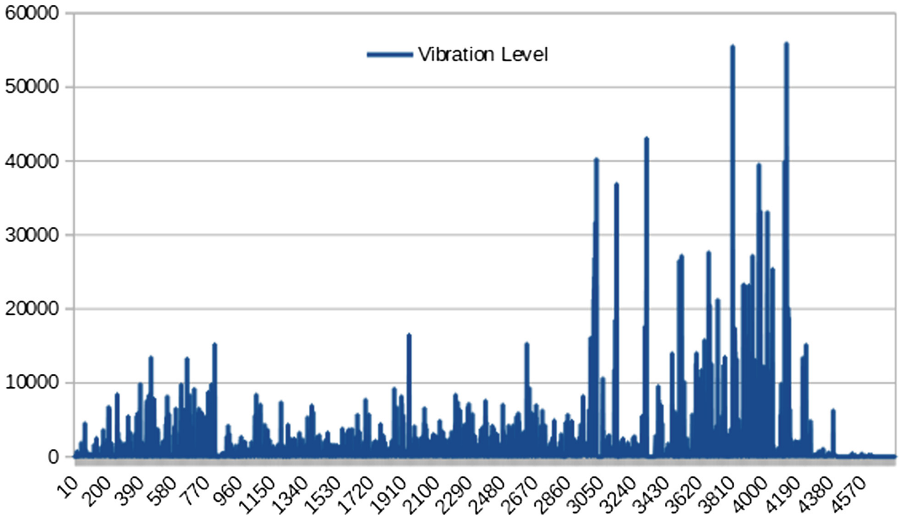


Fig. 6. Complete cycle - vibration level over time (s)

spinning the clothes. The wash, rinse, and spinning cycle lasts about ~4,340 s. After the spinning is done, the washing machine goes to the post spin period, and shuts off, which lasts about ~50 s. The whole default wash cycle of the test machine lasts approximately 4,630 s, which is close to ~77 min long.

The microcontroller node was setup and collects the vibration level data from the washing machine. For the experiment, 20 wash cycles were used, and the results of the vibration at every time interval in the wash cycle were submitted to the server node. The results of the data collection are shown in Fig. 6 where the timing have been normalized and the vibration cycle is averaged over the number of cycles that have been experimented on.

Based on the vibration level patterns, there are specific patterns in which could be used to detect the start and end of the washing cycle by using a threshold function over a sliding window. Based on the data gained from the wash cycle, specific patterns were observed. The details of the working states and its associated data reading and notification mechanisms were created by a rule-based system. The details of the wash cycle working states and data reading and notification mechanism are displayed in Table 2.

Table 2. Working States and Data Reading and Notification Mechanism

Working states	Data reading and notification mechanism
Initial	- start data reading at Tsampling = 1 s - set state = stop
Stop	if vibration value > threshold value - send “possible start cycle” message to server - set data Tsampling = 100 ms - set state = “possible start”
Possible Start	- read vibration value for about 30 s if the average value > the threshold value - send “confirm start cycle” message to the sever - set data Tsampling = 2 s - set state = “start” else - send “fail possible start cycle” message to the sever - set state = “stop”
Start	- read vibration value if the vibration value < the threshold value - read vibration value continuously at 100 ms for 10 s if the average value remain < threshold value then - send “possible stop” message to server - set state = “possible stop”
Possible Stop	- read vibration value for 1 min if the average value still < the threshold value - set Tsampling = 1 s - send “confirm stop” message to server - set state = “stop”

After defining the rules, further experiments with the wash cycles were tested, and the proposed system manages to detect the current state of the washing machine accurately over the remaining wash cycles. Overall the experiment shows that it is possible to detect the start and end of the wash cycle with a high degree of accuracy, and is suitable for the problem domain. Though the proposed system manages to detect the current wash state of the machine, it is observed that the proposed notification mechanisms cannot be applied to other washing machines due to variations of the vibration settings, and individualized profiles will have to be created for each washing machine.

5 Conclusions and Future Work

The proposed platform provides an IoT solution that will allow tenants to quickly figure the vacancy of the public coin operated washing machines in the apartment. Coupled with the low cost of the microcontroller nodes with utilizing the existing network infrastructure of the apartments, the system can be deployed cheaply and quickly providing a good, cheap, and reliable solution to the problem domain.

Though the proposed platform has solved many issues related with the problem domain, there are several potential areas of future work. One of the area is to deploy the proposed platform to a commercial apartment setting for feedback. Though the proposed system has worked in an experimental setting, it would be interesting to see the platform at work in a real setting, in which the system is currently being installed in a test run at a local apartment. Another area to expand the system is to provide a queuing system for the machines. This can be useful in busy apartments, but there are many issues that queuing may not work especially if users do not use the system. Another issue that could be of use is to keep track of the washing machine utilization over a longer period. With longer tracking, it is possible to see the trend which could be useful for tenants, or for the apartment owner to know if they require more machines if the utilization rate is too high. Another area that could be explored is exploring if other forms of sensors could be used to detect the wash cycle in the machine. Usage of light detecting sensors, or optical sensors can be potential approaches, in which can provide more contextual data, though at the cost of price and complexity of the system. Another area of exploration would be examining the difference between different washing machine vibration patterns. Though the washing machine vibration patterns in each cycle are similar, due to the build quality, material, model, and size, the patterns can slightly be different, which may require custom parameters for different builds.

References

1. Electrolux Washing Machine EWF1074. https://www.appliancesonline.com.au/manuals/ewf1074/ewf1074_user_manual.pdf
2. Gubbi J, Buyya R, Marusic S, Palaniswami M (2013) Internet of Things (IoT): a vision, architectural elements, and future directions. Future Gener Comput Syst 29(7):1645–1660
3. Gu J., Qiang D (2015) The design of intelligent washing machine controller based on FPGA. In: Fifth international conference on instrumentation and measurement, computer, communication and control (IMCCC), Qinhuangdao, pp 1529–1532

4. Hodges S, Taylor S, Villar N, Scott J, Bial D, Fischer PT (2013) Prototyping connected devices for the internet of things. Computer 46(2):26–34
5. Kovatsch M, Mayer S, Ostermaier B (2012) Moving application logic from the firmware to the cloud: towards the thin server architecture for the internet of things. In: Sixth international conference on innovative mobile and internet services in ubiquitous computing, Palermo, pp 751–756
6. NBTC Internet Statistics Report. http://webstats.nbtc.go.th/netnbtc/INTERNETUSERS.php
7. SmileVending – Coin Operated Machines. http://www.smilevending.com/
8. Soliman M, Abiodun T, Hamouda T, Zhou J, Lung CH (2013) Smart home: integrating internet of things with web services and cloud computing. In: IEEE 5th international conference on cloud computing technology and science, Bristol, pp 317–320
9. Thailand Computer Crime Act Criminal Law. http://www.ratchakitcha.soc.go.th/DATA/PDF/2560/A/010/24.PDF
10. Triyason T, Setthawong P (2016) PLAKDA - an IoT platform for the production of Mekong basin styled fermented fish (plara). In: 2nd international conference on science in information technology (ICSITech), Balikpapan, Indonesia, pp 203–208

Developing Specific Usability Heuristics for Evaluating the Android Applications

Roobaea Alroobaea(✉)

Department of Information Technology, College of Computers
and Information Technology, Taif University, Taif, Saudi Arabia
r.robai@tu.edu.sa

Abstract. The lack of specific heuristics evaluation method for improvement of the quality and usability assessment process for Android applications (Apps) on mobile devices represents a missing piece in usability testing. This paper aims to present specific usability heuristic to measure and improve the efficiency and effectiveness of the Android applications, to assess user satisfaction, and ultimately to improve their quality. The results show that the developed method provide evaluators with insights into how interfaces can be assessed to be effective, efficient and satisfying. It also support more uniform problem descriptions and can guide evaluators in finding real usability problems. Moreover, it can facilitate the problem-matching process, thereby facilitating the evaluation process by assessing each area and page in the Android applications.

1 Introduction

The growth of the Internet and related technologies, such as mobile devices, has enabled the development of many of mobile applications that is growing rapidly in use and that has had a huge impact on our life and different area businesses. Thus, mobile application developers need an appropriate usability evaluation method that can help them to improve the quality aspects of their applications, and ultimately to success of their products. Evaluation (HE) and User Testing (UT) are the most important traditional usability evaluation methods for ensuring system quality and usability [1]. However, most of such studies have described these methods not directly applicable to the product being tested, not directly related to the context of the tested product, and not able to identify specific areas and types of usability problems [2]. This paper aims to develop specific heuristics for Android Apps on mobile devices for enhancing their quality and facilitate experts in performing evaluation. These heuristics are characterized as being pertinent to the context and specific target for Android Apps. Also, they combine the inputs from users and experts. This paper is developed as follows. After the introduction, the related work to the usability of mobile Apps will be reviewed to identify the research gap. Next, the methodologies are used will be discussed. Then, the findings will be presented. Finally, the future work and conclusion will be discussed.

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_15

2 Literature Review

Since the 1980s, the user testing (UT) method has become the major method for evaluating a new and improved interface. However, [3] pointed out four different shortcomings in UT; the first limitation is that the testing session is always a fabricated circumstance (i.e. not real); the second limitation is that the results of UT do not mean that the product works; the third limitation is that the sample of users may not fully represent the target population; and the fourth limitation is that choosing UT is not always the best approach. For those reasons, developers in the 1990s started to search for other methods that are low in terms of cost, consume less time, and can be used in the earlier stages of the design process. As a result, expert-based inspection methods grew in popularity to fulfil those requirements. Some of these methods are still popular, such as Heuristic Evaluation (HE) which appears to be the most popular form of inspection method [4]. However, it is claimed to be a general, subjective assessment, does not cover the specific aspects of a targeted product, does not guide evaluators during the evaluation, and can miss some of the real problems [5, 6].

The literature review shows that many researchers have attempted to enhance the traditional heuristics through assessing them to identify those heuristics that do not work, and remove them. Then they develop new heuristics to cover areas not covered by the traditional heuristics. Finally, these new heuristics are added to the ones remaining from the traditional heuristics [7]. Other researchers went further than that through developing customized heuristics for areas such as games [8]. In terms of mobile applications, [9] developed heuristics for map application. [10] developed heuristics for IOS applications. Furthermore, Google lists seventeen Android design principles, however, [11] claims that they are general and short rules. Also, novice evaluators might not be able to uncover as many problems as the experienced ones. Consequently, there is a needed for context heuristics for Android Apps that plays a vital role in influencing the evaluation results better than abstract ones.

3 Methodology

To achieve the aim of this paper, hybrid data collection methods and multiphase designs ware adopted. Thus, there are three development steps, as outlined below, for gathering together suitable components to obtain inputs from users and experts for developing new and novel heuristic sets for Android Apps. There is an adopted method in each step, and the limitations of each method are complemented by the strengths of the others. These steps are; *Step 1:* (a) Identifying the problem and its scope; (b) Reviewing the published material and analysing the gathered data; (c) Content analysis method is adopted when there is very limited literature or not enough information through using an emerging coding approach with two researchers [1]; *Step 2:* (a) Field studies method is used to elicit feedback on Android Apps from the real users in the natural environments. [12] stated that "it is especially so for the evaluation of new technologies, such as mobile devices". It is used to obtain user input for identifying user requirements by using questionnaire to capture their experience with the tested App, understanding the effect of Apps design on user experience in real circumstances, learning from the errors made by the users during the observation, ultimately to providing reliable results for identification of usability problem areas and formulation specific heuristics from the usability problems that were found. The results of step 1 and 2 are used as the starting point for next step; *Step 3:* (a) Focus group method is used to obtain expert input through a discussion amongst experts, who are mobile applications developers and usability experts, on all issues arising from the field studies results and the literature review. This method leads to a better design and presenting new ideas for the new heuristic sets. Then, it entails to identify the areas of usability problems related to the Android Apps from the overall results and a reliability evaluation of these usability areas is computed until a satisfactory level is achieved. Finally, this step helps to formulate specific heuristics for each usability problem area.

4 Recruiting Participants

For field studies, 30 users were chosen carefully to reflect the real users of the Android Apps. The criteria that were considered to recruit these users were: (a) real users for the Android applications; (b) willingness to participate; (c) having good experience in the Android Apps by using similar Apps in their daily life. For the focus group, five experts people were chosen carefully in which three experts in usability (i.e. having a certificate in the HCI field) and two in the mobile Apps developers (i.e. having certificate in the HCI field and Android Apps). The expert evaluators were invited to participate based on their availability and experience. All of them have knowledge of usability evaluation through teaching or studying HCI courses, and some of them have evaluated many mobile Apps for usability. These combined facts confirmed that the experienced evaluators were chosen in the hope of maximising the benefits of using expert evaluators in an efficient manner.

5 Results and Discussion

5.1 Construct Specific Heuristics for Android Apps

Based on the first step, an extensive literature survey was conducted on the materials relating to usability of mobile and Android Apps, such as and not limited, [9–11]. However, limited studies were found. Consequently, content analysis was conducted by using an emerging coding approach with two independent experts. Next, second step was started by observing users, taking notes and asking questions. The gathered data of this step was analysed and reported. The results from previous steps were discussed by experts in the third step, including their points of view. Before end of the session, they identified the list of usability problem areas in Android Apps. Also, a set of heuristics with their explanations were identified. Cohen's kappa coefficient was used to enable a calculation of the reliability quotient on the result of the Likert questionnaire. The intra-observer test-retest using Cohen's kappa yielded a reliability value of 0.8, representing satisfactory agreement between the evaluators in the focus group session. Finally, the usability problem areas were merged and grouped into nine usability problem areas. Also, the identified heuristics were classified according to the

Table 1. Specific usability heuristics for evaluating the android applications

Usability problem area	Specific heuristics for Android applications
Layout and formatting (LF)	Consistency in UI Design • The look and feel of the App should be properly defined and maintained throughout the App. The user should be able to see similarity in details like buttons, links, pages, etc • The layout, formatting and navigation controls should be proper and describing themselves so that user know its importance and relate to the detail • Visuals (such as labels, colors, and icons) and touch also gives a feedback to clearly define the details
	Information Architecture • The information which are important actions would be placed at the top or bottom of the screen so that user get it in a glance and is reachable all the time. Place the related items of a similar hierarchy of main information next to each other • Use short menu, paths and buttons for main functionalities and secondary information. This should be placed in recognizable positions. Overall experience should consist of clear contrast between visual elements, balanced layout and informative colors so that it won't became harmonious • Avoid clutter visual, visual elements should be user friendly and make their gaze to important elements. Navigation controls and the main information should not be mixed. A proper clarity and format and hierarchy should be maintained
	Consistency in UI Layout: Consistency increases App usability, since users don't have to learn new ways as they move around the App *Page layout:* • The page layout should be divided in the section so that the user wouldn't scroll much for the relevant content. Section give him the ability to choose the content he\she wants to select. The section consists of the clear browsing information for the user. The menu could in the form of pull downs, pop-up menu, accordion, column grid layout. The App should provide complete freedom to the user to make customizations based on user choice (like creating one's own template or page layout) *Text Layout:* • The text should be in the form of paragraph if the text is long so that the interest could be maintained and should be in the readable form. By using the techniques of consistency and explicit step by step formation, it would be easy to grasp the information. Also, the group information visually can be enhanced by using the suitable colors, texts and topics smartly
	Simple but Complete Design: • The App should have content categorization and hierarchal information layout (such as primary and secondary) with hidden on-demand content. Minimalistic design with fewer clicks, scrolling and pop-ups as well as highlighting important features helps users to minimize their memory load • Easy and corrective actions (like undo, redo options) help users to rectify errors
	Navigation Bar: • The overall navigation and top-list information (with search and help options and easy bookmarks) facilitate the user in retrieving the required screen/information quickly in the App • Keep the user interface navigation structure narrow, simple and straightforward. This is important in an App design that the navigation back button and other interaction controls behave predictable as it can make user experience good or worst. A navigation bar appears at the top of an App screen, below the status bar, and enables navigation through a series of hierarchical App screens. When a new screen is displayed, a back button, often labeled with the title of the previous screen, appears on the left side of the bar. Sometimes, the right side of a navigation bar contains a control, such as an ''Edit" or a "Done" button for managing content within the active view. In a split view, a navigation bar may appear in a single pane of the split view. Navigation bars are translucent, may have a background tint, and can be configured to hide when appropriate action is done, such as when the keyboard is on screen, a gesture occurs, or a view resizes
App Functionality (AP)	Search Support for User Queries: • The App should facilitate users with clear functions that allow them to conduct any related search without having to leave the current working environment. Users should have an accurate search engine for basic and advance search support (e.g. groups, people, interests, content, suggestions and companies) with clear and relevant search result pages that allow them to view, edit and resubmit their search queries

(*continued*)

Table 1. (*continued*)

Usability problem area	Specific heuristics for Android applications
Content Quality (CQ)	Controls and Content Alignment: • App should be properly aligned as it provides ease scanning, highlight hierarchy and define organization. Alignment makes an App look neat and organized, helps people focus while scrolling, and makes it easier to find information. Indentation and alignment can also indicate how groups of contents are related
	Handling Change in Orientation: • Avoid gratuitous layout changes. Just because someone rotates a device doesn't mean the entire layout needs to change. For example, if your App shows a grid of images in portrait mode, it doesn't have to present the same images as a list in landscape mode. Instead, it might simply adjust the dimensions of the grid. Try to maintain a comparable experience in all contexts. If possible, support both portrait and landscape orientations. People prefer to use Apps in different orientations, so it's best when you can fulfill that expectation
	Key information discernible at a glance: • The App should provide easy readable and understandable content that is placed in separated blocks. Content used with familiar vocabulary, terminology and graphical symbols facilitate and ease the tasks of the users • Use the standard back button. Consider temporarily hiding the navigation bar when displaying full-screen content
	Appropriate & approachable content: • Approachable and appropriate amount of information provisioned with FAQ will help users to achieve their primary goal. Content blocks, icons and different colors will help users to take further actions in the App
	Ease of Access to your App: • Include appropriate content labeling to accommodate users who experience a text-only version of your App. Avoid flashing large central regions of the screen. While using the App the system should support different multimedia channels to represent the information
Standard behaviour and Interaction (SBI)	Behaviour: • Avoid crowding a navigation bar with too many controls. As the navigation bar, major action is to redirect the user to proper information. In general, use a tab bar to organize information at the App level. A tab bar is a good way to flatten your information hierarchy and provide access to several peer information categories or modes at once • Provide relevant toolbar buttons. A toolbar should contain frequently used commands that make sense in the current context. To improve readability, users might increase font size. User can zoom in out the screen size • Animations and transitions should be displayed smoothly
	Accuracy of information: • The App should only provide available, concise, relevant, reliable, non-repetitive and frequently updated information that is suitable to the page length
	Conciseness of the content: • Maintain focus on the current content during context changes. Content is your highest priority. Changing focus when the environment changes can be disorienting, frustrating, and make people feel like they've lost control of the App
	Content objectivity and Visibility: • Ensure primary content is clear at its default size. People shouldn't have to scroll horizontally to read important text, or zoom to see primary images, unless they choose to change the size
	Coverage of up to date Content: • Correct, relevant, up to date and reliable information regarding the content details must be present as user need those most frequently
	Content Segmentation: • This also applies to media, which must be fully exhibited, unless the user opts to hide them. The elements on the screen must be adequately aligned and contrasted
	Avoid Redundant Information: • Avoid the use of scrolling. Make information easy to read, skim (or) and scan. This also applies to media, which must be fully exhibited, unless the user opts to hide them. The elements on the screen must be adequately aligned and contrasted • The App should only provide available, concise, relevant, reliable, non-repetitive and frequently updated information

(*continued*)

Table 1. (*continued*)

Usability problem area	Specific heuristics for Android applications
User Usability (UU)	Grouping of related items: • Keeping related items in proximity to one another is helpful for those who have low vision or may have trouble focusing on the screen. App shouldn't use too much CPU, memory, screen space, or other system resources. It should respond well to sudden interruptions and audio from other Apps, transition to and from the background quickly and smoothly, and behave responsibly when operating in the background • Make sure your interface works with a double-high status bar. Certain features, such as in-progress phone calls, audio recording, and tethering display an additional status bar at the top of the screen. In unprepared Apps, this added height can cause layout problems by covering or pushing down other interface elements
	Controlling Notification: • Appears at the top of the screen for a few seconds while the device is in use, then disappears. Don't send multiple notifications for the same thing, even if the user hasn't responded. People attend to notifications at their convenience. If you send multiple notifications for the same thing, you fill up Notification Center, and users may turn off notifications from your App • A notification detail view provides more information about a notification, as well as the ability to take immediate action without leaving the current context to open your App. Provide a sound to supplement your notifications. The App should use e-mail notifications to encourage members. But user have the option to change it on and subscription for email notification
	Controlling the Animations: • Use animation and motion effects judiciously. Don't use animation for the sake of using animation. Excessive or gratuitous animation can make people feel disconnected or distracted, especially in Apps that don't provide an immersive experience. Use consistent animation. A familiar, flowing experience keeps users engaged • Make animations optional. When the option to reduce, motion is enabled in accessibility preferences, your App should minimize or eliminate App animations
	Controlling the Branding Impact: • Don't let branding get in the way of great App design. Ensure that it's intuitive, easy to navigate, easy to use, and focuses on content. Even if your App is available on other platforms, avoid diluting your design by focusing too much on consistent branding. Resist the temptation to display your logo throughout your App. Avoid displaying a logo throughout your App unless it's necessary for providing context. This is especially important in navigation bars, where a title is more helpful
	Choosing Proper Colour Scheme: • Use complementary colors throughout your App. The colors in your App should work well together, not conflict or distract. If pastels are essential to your App's style, for example, use a coordinating set of pastels • Consider choosing a key color to indicate interactivity throughout your App. Avoid using the same color for interactive and no interactive elements. If both elements have same color, it's hard for people to realize where to tap • Be aware of colorblindness and how different cultures perceive color. People see colors differently. Many colorblind people, for example, find it difficult to distinguish red from green (and either color from gray), or blue from orange. Avoid using these color combinations as the only way to distinguish between two states or values. For example, instead of using red and green circles to indicate offline and online, use a red square and a green circle. Some image-editing software includes tools that can help you proof for colorblindness. Also consider how your use of color might be perceived in other countries and cultures. Make sure the colors in your App send the appropriate message
	Using Proper Font scheme: • App should be prepared so that user can change text size as per needs. User expects most Apps to respond appropriately when they choose a different text size in settings. Font size: the font size is not always consistent and often results in being too small causing reading difficulties. To accommodate some text-size changes, you might need to adjust the layout • Emphasize important information by font weight, size, and color to highlight the most important information in your App
	Manageable personal profile & user-driven content: • The App should facilitate the user with easy registration, managing the personal profile (create, modify) and password recovery options. Users must have overall control and ease to perform any activity • User's complaints and reports should also be taken seriously. The user-driven content management system (such as edit/delete, or liked/marked content) should facilitate the user
	Freedom in User Access functionality of App: • Privileges for users to perform various activities (such as public or private messaging, adding/blocking friends or their connections etc.) • User should have complete freedom to create groups, fan clubs, bands, etc. & to choose the friends, groups, etc. they want. Supporter of users' skills & freedom, such as the customization of users' content/messaging and notifications. The App should facilitate users in initiating actions (messaging, contents, notifications, etc.) on their profile page
	Provide Components for important tasks: • Reinforce important information through multiple visual and textual cues. Use color, shape, text, and motion to communicate what is happening. Display error messages in a language familiar to the user, indicating the issue in a precise way and suggesting a constructive solution. The designed App should consist the emergency exit which make user able to leave the App anytime. When such interruptions occur, the App should save its current state and still be able to give the needed navigation instructions • Make links and buttons clearly visible and distinguishable from other user interface elements. Make sure the user interface is scalable for different screen sizes of mobile devices • Visual Identity: some pictures are displayed, while others are not visible. Use clear, intuitive, commonly known symbols. Icon consistency: several icons and symbols are not immediately recogniz-able and visible. Use simple and meaningful icons. Avoid the use of interaction timeouts and provide ample time to read information

(*continued*)

Table 1. (*continued*)

Usability problem area	Specific heuristics for Android applications
Interface Elements and design (IED)	User Centric Design: • Apps should provide easy ways to input data. A well-designed App is accessible to users of all abilities, including those with low vision, blindness, hearing impairments, cognitive impairments, or motor impairments. Improving your App's accessibility enhances the usability for all users. It's also the right thing to do. The user to use both hands. Screen content should be easy to read and navigate through notwithstanding different light conditions. Ideally, the user should be able to quickly get the crucial information from the App by glancing at it • The metaphor of each component or feature must be unique throughout the App, to avoid misunderstanding. The interface should be designed so that the items are neither too distant, nor too stuck. Margin spaces may not be large in small screens to improve information visibility. The more related the components are, the closer they must appear on the screen. Interfaces must not be overwhelmed with many items • The main features of the App must be easily found in a single interaction. Most-frequently-used functionalities may be performed by using shortcuts or alternative interactions. All input components should be easily assimilated
	Search Bar Design Features: • Use a search bar rather than a text field to implement search. Enable the Clear and Cancel buttons. Most search bars include a 'Clear' button that erases the contents of the field, and a 'Cancel' button that immediately terminates the search. In App if necessary, provide hints and context in a search bar. Consider providing helpful shortcuts and other content below a search bar • The user can touch the microphone icon to initiate a voice search. Use persistent search when search is the primary focus of your App
	Scope Bar for Controlling Scope of the Search: • A scope bar can be added to a search bar to let people refine the scope of a search. A scope bar adopts the appearance of its search bar. Favor improving search results over including a scope bar. A scope bar can be useful when there are clearly defined categories in which to search. However, it's best to improve search results so scoping isn't necessary
	Offers of informative feedback - action & reaction: • The App provides clear goals or supports user-created goals. The App should support the performance tools provided to mimic users' real-world counterparts. The user must not have to remember information from one screen to another to complete a task. Also consist, multilingual lessons should help users who have physical impairments • The App should provide timely, meaningful, easy to understand and informative overviews (such as current level of achievement or profile status) with action confirmation. The App must provide the user's current task-related feedback (e.g. error messages) in an appropriate manner (not too long not or too short). Users should be provided with the opportunity to access extended feedback from instructors through email and internet communication as well as FAQ (page, help and other additional guidance). The information of the interface must be sufficient for the user to complete the current task • Keep the background simple and avoid transparency. Make sure your icon is opaque, and don't clutter the background. Give it a simple background so it doesn't overpower other App icons nearby. You don't need to fill the entire icon with content • You must supply high-resolution images for all artwork in your App
	Transparency and Security Policies for User Data: • The App should protect his/her personal data by using privacy and security settings. All data should be protected, fully inaccessible or accessible as per authentication. Users should be aware of the information they have stored within the App • Transparency of transactions helps in building and maintaining users' trust (e.g. personal information and uploaded data will not be used or displayed without the user's permission). 'Privacy policies' and 'terms and conditions' should be displayed clearly (and be clear) to the user • Users should be informed for any promotional or marketing communication. Also, the facility to report (to developer or the manager) any suspicious activity or inappropriate data posted by others
Accessibility, Navigation and compatibility (ANC)	Accessibility: • Support assistive technologies specific to your platform, just as you support the input methods of touch and keyboard. For example, ensure your Android App works with Google's screen reader, TalkBack Apps • The Apps should increase, maintain, or improve the functional capabilities of individuals with disabilities
	Accessibility and compatibility of hardware devices: • The App should be working and compatibility on different hardware devices. Also, it must have satisfactory performance and be able to load content quickly • The App should have the option of disable inputs when required. App must be properly load-tested (allow multiple users at a time) and have a proper Disaster Recovery system. The user should be assisted with clear contact details (using multiple contact formats, like email, forms, etc.) and it should resume incomplete work left off
	Easy access through universal design: • The App has a universal design and structure (not too tight, not too loose) to facilitate diversified user groups. User able to customize the App. Make sure that the main functions of the map application (e.g. exploring, route guidance, zooming, panning, POI selection) are easily accessible. App can also use calendar if needed. Give easy access to additional information (metadata, links, user-generated content) • Accessibility of topics or links should be clear where to get the right information. So, design in that manner
	Adequacy of the component to its functionality: • The user should know exactly which information to input in a component, without any ambiguities or doubts. Metaphors of features must be understood without difficulty. Indicate clearly the reasons for why the searched locations are not found. Save the user's previous searches for fast repetition • Adequacy of the message to the functionality and to the user i.e. The App must speak the user's language in a natural and non-invasive manner, so that the user does not feel under pressure. Instructions for performing the functionalities must be clear and objective. Provide both: fast guidance focused on the user's task and more detailed documentation with search functions

(*continued*)

Table 1. (*continued*)

Usability problem area	Specific heuristics for Android applications
Error Handling and Help (EHH)	Error Detection and Recovery: • If there is no solution to the error or if the error would have negligible effect, enable the user to gracefully cope with the error • When an error occurs, the App should quickly warn the user and return to the last stable state of the App. In cases in which a return to the last stable state is difficult, the system must transfer the control to the user, so that he decides what to do or where to go
Business Support (BS)	Advertising or sales pitches mechanism: • The factors should be keep in mind such that: Is it enjoyable, disturbing or undesirable (like pop-up ads.)? Can the user take part in it (e.g. 'like' or comment option)? Users must be aware of paid membership features, benefits and available hot offers if any • The App should help users in easily classifying advertisements
	Trust & credibility of information sources and company advertising: • An advertisement must lead the user to a trusted site or App in a separate window. The company must have legal rights to publish their product advertisement
	Frequent posting & updating: • Users must have authentication (modify, update and remover own post and group). The App assists the user in participating in various facilities (e.g. posting text, or single or multiple chat) as frequently (and as much) as they want

agreed usability areas. Thus, the specific heuristics was created (as shown in Table 1), closely focused on the Android Apps. Moreover, the results of the three steps were reported for further analysis in the future.

5.2 Pilot Study

The developed heuristics need to be tested before using them to improve their quality. The pilot study was conducted twice to assess the revised heuristics as recommended by [13]. Thus, two independent evaluators were recruited and two Android Apps were chosen. Then, the evaluators were asked to evaluate the chosen Apps based on the developed heuristics. The results led to improve the developed heuristics by editing spelling and grammatical errors and removing ambiguous words. Furthermore, a post-test questionnaire was developed to gather feedback after evaluation. It was contained two open ended questions. The first one was "Do you think that the developed heuristics helped you to discover more usability problems?", and the second one was "Did you find the developed heuristic useful, and was it easy to use?". The evaluators concluded that the developed heuristics was helpful and it was easy to use compared with the traditional heuristics based on their previous experience. Also, they discovered more usability problems.

6 Conclusion and Future Work

In relation to the previous of the three steps, this research has generated specific heuristics which were specific for Android Apps. These heuristics can be used as a tool that affords designers, developers, instructors, and evaluators the facility to design an interactive interface or assess the quality of existing applications. Furthermore, it has identified the usability problem areas in the Android Apps (nine areas as shown at Table 1). These areas provide designers and developers with insights into how interfaces can be designed to be more effective, efficient and satisfying. Also, they support a more uniform problem description and can guide expert evaluators in finding more usability problems, thereby, facilitating the evaluation process by assessing each area and page in the target application. In addition, it allows adoption any area of usability or different principles to determine the usability problems related to the nine specific areas in the Android applications. The future work is to validate this heuristics against other methods. Thus, the developed heuristics needs to be tested intensively through rigorous validation methods (e.g. analytically against traditional heuristics and empirically against user testing or filed studies) and many of usability metrics (e.g. satisfaction, efficiency, effectiveness, thoroughness, validity, and reliability) to verify the extent to which it achieves the identified goals. This validation study will be published in the further research.

References

1. Lazar J, Feng JH, Hochheiser H (2010) Research methods in human-computer interaction. Wiley, London
2. Henninger S (2000) A methodology and tools for applying context-specific usability guidelines to interface design. Interact. Comput. 12:225–243
3. Rubin J, Chisnell D (2008) *Handbook of usability testing: howto plan, design, and conduct effective tests*. Wiley, New York
4. Hollingsed T, Novick DG (2007) Usability inspection methods after 15 years of research and practice. In: Proceedings of the 25th annual ACM international conference on design of communication, October 22, 2007, Texas, USA. ACM, pp 249–255
5. Molich R, Dumas JS (2008) Comparative usability evaluation (CUE-4). Behav. Inf. Technol. 27:263–281
6. Chattratichart J, Lindgaard GA (2008) Comparative evaluation of heuristic-based usability inspection methods. In: CHI'08 extended abstracts on Human factors in computing systems, Italy. ACM, pp 2213–2220
7. Ling C, Salvendy G (2005) Extension of heuristic evaluation method: a review and reappraisal. Ergonomia Int J Ergon Hum Factors (IJE&HF) 27:179–197
8. Pinelle D, Wong N, Stach T (2008) Heuristic evaluation for games: usability principles for video game design. In: CHI'08 proceedings of the SIGCHI conference on human factors in computing systems, Italy. ACM
9. Kuparinen L, Silvennoinen J, Isomäki H (2013) Introducing usability heuristics for mobile map applications. In: Proceedings of the 26th international cartographic conference, Vancouver
10. Nayebi F, Desharnais J-M, Abran A (2013) An expert-based framework for evaluating IOS application usability. In: 2013 Joint Conference of the 23rd International Workshop on Software Measurement and the 2013 Eighth International Conference on Software Process and Product Measurement (IWSM-MENSURA), Ankara. IEEE, pp 147–155
11. Thitichaimongkhol K, Senivongse T (2016) Enhancing usability heuristics for android applications on mobile devices. In: Proceedings of the world congress on engineering and computer science, vol 1
12. Sharp H, Rogers Y, Preece J (2007) Interaction design: beyond human-computer interaction
13. Ruxton G, Colegrave N (2011) Experimental design for the life sciences. Oxford University Press, Oxford

Monitoring Environmental Variables Through Intelligent Lamps

Martin Pies(✉) and Radovan Hajovsky

Department of Cybernetics and Biomedical Engineering, Faculty of Electrical Engineering
and Computer Science, VSB-Technical University of Ostrava, 17. Listopadu 2172/15,
70833 Ostrava, Czech Republic
{martin.pies,radovan.hajovsky}@vsb.cz

Abstract. The article describes the design and realization of a smart streetlight prototype. The lamp is designed to implement a comprehensive system of Smart Cities and Smart Lighting in the research and development of IoT issues. In addition to the possibility of wireless switching, the lamp contains switching and dimming, as well as sensors for measuring temperature, humidity, and pressure, and it is equipped with a PIR sensor for switching the lamp when a person passes through its active field. All communication with the parental system, including the collection and processing of data is performed by IQRF technology. It also allows the creation of an extensive IQMESH network of these lamps.

Keywords: IQRF · IQMESH · Street lamp · IoT · Measurement · Wireless communication

1 Introduction

In recent years, there has been a big boom in the introduction of new technologies into the daily life of man. This includes the implementation of various microelectronic systems in commonly used items such as watches, wearable clothes, health care - see [12], automobiles – see [2], home appliances – see [11] and others. It is closely related to the development of the Internet of Things. Various studies indicate that by 2020, up to 50 billion devices will be connected to the Internet. The so-called SMART complex solutions, such as Smart Cities, Smart parking, Smart Lighting and others, are also undergoing a big development. The basis for communication between the individual components of these large systems should be low power technology that enables the transmission of necessary data from the deployment location to the parental system. Currently, technologies for wireless communications, such as LoRa, SIGFOX or NarrowBand-IoT, have been emerging globally. These solutions have their advantages, but also disadvantages that hamper the deployment within SMART projects. One disadvantage is the energy intensity when transmitting 1B data. Another disadvantage is the amount of transmitted data for one device. The advantage may be high reach, which, however, loses its sense in urban development. An alternative to the above-mentioned solution may be IQRF technology, see [13]. This technology allows communication at a lower direct distance than, for example, LoRa SIGFOX; however, energy intensity

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_16

and the possibility of communication between sensors within the MESH topology can successfully eliminate this disadvantage. IQRF technology has recently been becoming recognized by companies, dealing with street lighting and with the area of the Internet of Things in general. The topic of intelligent street lighting control has resulted in several publications, directly from the IQRF technology manufacturer – see [14], or based on other communication standards such as ISO/IEC 14908 – see [7, 15].

One important area in which technology companies in the Czech Republic have recently been involved is Smart Cities. Different areas of urban infrastructure are associated with this concept. This includes, in particular, the smart car park, monitoring the air quality, smart lighting, reading of meters of water consumption, but also the rate of filling dustbins. Intelligent lighting is based on the use of smart lamps that communicate with each other; it is possible to read operating data from them, such as electricity consumption or failure, or they can be switched on and off depending on the traffic around the intelligent lamp. In many cities, smart lamps are already used. For example, the publication [3] associates intelligent lamp control with photovoltaic panels. Another publication [5] describes communication using the G3-PLC standard. Generally, intelligent lamps of the future must be based on LED technology, which is still the most convenient alternative to currently used sodium lamps, whose production is in decline based on the European Union's commitment under the Kyoto Protocol, see [6, 8].

The basis of our proposal is to create a prototype of an intelligent street lighting lamp, which implements the possibility to measure basic environmental variables such as temperature and relative humidity, atmospheric pressure, and also the possibility to control the brightness depending on the detection of persons under the lamp.

2 Intelligent Lamp

General structure of the lamp

The basis of the intelligent lamp is a diode module, whose luminosity is 140 lm/W. The diode module is powered by a voltage transformer with an output of 40 W. The drive has an external 12 V DC power supply and control lines for controlling the voltage transformer power. The control wire for power control marked as DIMM + closes along with the power GND of the external supply 12 V DC electrical circuit. The voltage transformer output power can be controlled in two ways. Either by applying a voltage of 0–10 V DC between inputs DIMM+ and GND or by connecting the potentiometer with a range of 5–100 kΩ between the same inputs. The block diagram of the lamp is shown in Fig. 1.

The IQRF module is powered by the power supply with a nominal value of 3.3 V. The power supply can be supplemented by a backup battery, whose use is described below. The voltage transformer output power is controlled by an electronic potentiometer, which is controlled via the I2C bus. It is a less demanding method of voltage transformer output control compared to generating voltage in the range from 0 to 10 V DC. The voltage transformer output power cannot be reduced to zero, therefore, a relay is used to completely disconnect the light source from the voltage transformer. Figure 2 is a photograph of the

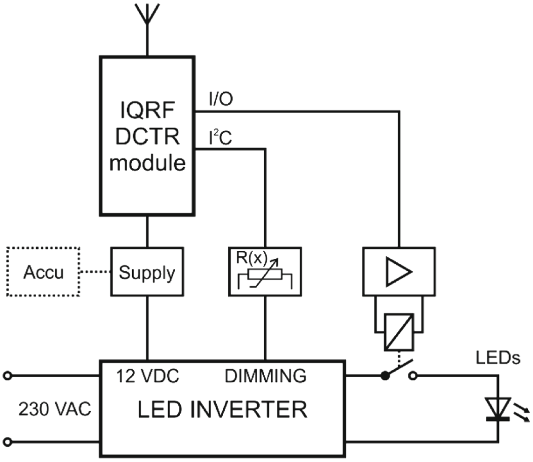


Fig. 1. The internal structure of the intelligent lamp without sensors with optional backup

electronics of an intelligent lamp. All electronic circuits including the IQRF transceiver module are located on the underside of the PCB.

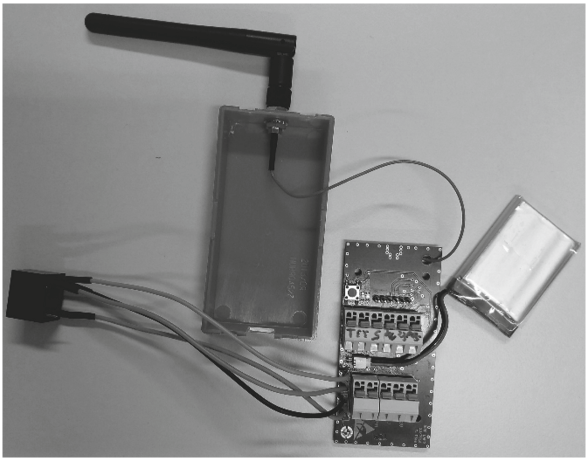


Fig. 2. Control electronics of an intelligent lamp

Sensory part of the lamp

The sensory part of the intelligent lamp is equipped with a combined sensor for measuring the temperature and relative humidity of the surrounding air. Simultaneously, this part is supplemented by the measurement of atmospheric pressure. Both of these sensors are connected to the IQRF module via the I2C bus. The pyroelectric sensor (hereinafter referred to as PIR) provides information about the presence of a person close to the lamp. The sensor is connected to the IQRF module through the general I/O pin. The structure of the intelligent lamp with the sensory part is shown in Fig. 3.

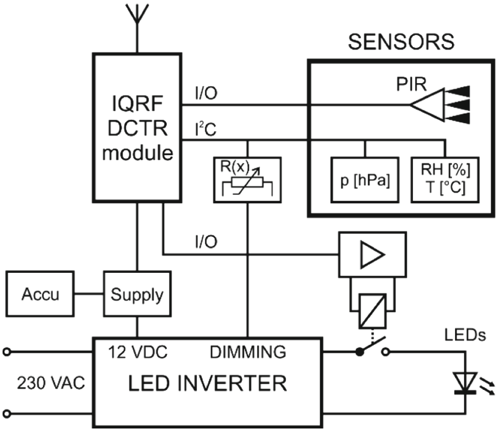


Fig. 3. Internal structure of the intelligent lamp with the sensory part

To enable the intelligent lamp to provide information about the surrounding air, the supply circuit of the IQRF module is supplemented by a Li-Pol battery. Figure 4 shows a photograph of the intelligent lamp interior including the outside part with sensors.

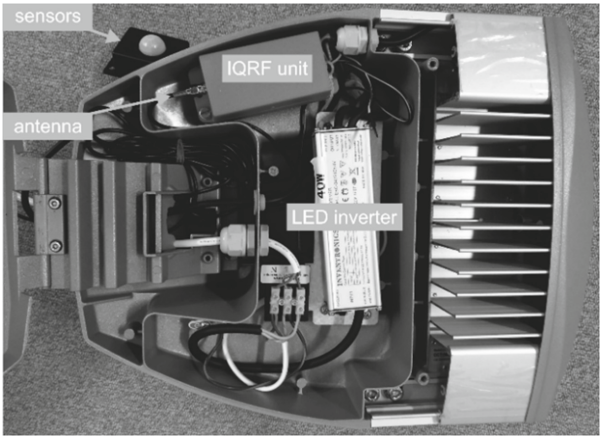


Fig. 4. The inner space of the intelligent lamp

Sensors are placed in a separate box. In Fig. 4, this can be seen in the upper left corner. The part with the sensors is connected by a shielded cable with control electronics. Since the cover of the intelligent lamp is made of aluminium alloy, the antenna is taken out of the inner space of the intelligent lamp.

IQRF technology

The intelligent lamp is complemented by the IQRF module from the company Microrisc. IQRF is a platform for low speed, low power, reliable and easy-to-use wireless connectivity e.g. for telemetry, industrial control and building automation. It can be used with any electronic equipment. It can be used whenever wireless information transfer is needed, e.g. remote control, monitor, alarm, displaying of remotely acquired data or connection of more devices to a wireless network. IQRF is a complete ecosystem from one brand including hardware, software, development support and services. The IQRF network can easily be connected to the Internet via a Cloud server. IQRF is ideal to implement the Internet of Things.

IQRF parameters:

- RF bands: world-wide ISM 433 MHz, 868 MHz and 916 MHz
- Based on transceiver modules with built-in operating system (OS)
- Fully open functionality depends solely on user-specific application written in C language
- Packet-oriented communication, max. 64 B per packet
- Range up to several hundred metres per hop, up to 240 hops per packet
- Extra low power consumption: 380 nA standby, 25 μA receiving
- Low bit rate: 19.2 kb/s. Preliminary: 1.2 kb/s, 57.6 kb/s and 86.2 kb/s
- Supporting MESH, IQMESH protocol implemented in OS
- No licence fees
- Very easy to implement

Data controlled transceivers (DCTR) enable applications even without programming [4].

The Mesh topology consists in connecting each node to all the other ones (Full mesh), thus ensuring high reliability. So, if any of the routes are currently not available, there is an alternative path to deliver the message (usually multiple paths).

In the case of a wireless IQRF network, in the mesh topology, each node has its defined time transmission slot, which enables that transmission does not interfere with other nodes. The node continuously receives messages and forwards them in its transmission slot to all the nearby devices. Thanks to alternative paths, there is a high probability that in the case of damage to multiple nodes, the message arrives at the destination.

IQMESH is a protocol for wireless IQRF networks developed by the company Microrisc. It is based on the mesh topology. IQRF transceivers are deployed in the area, which ensures better coverage and a wider range thanks to mesh topology. The neighbours are those transceivers that are mutually in the transmission range.

In the case of the deployment of, e.g. 10 transceivers in the area in mesh topology, each of them will be transmitting in its time interval; in this case, there are 10 intervals. Transceivers transmit in a synchronized way; therefore they do not interfere mutually when transmitting.

The main control transceiver – the so-called coordinator, sends data when necessary, ambient transceivers receive this data and gradually send it to the neighbourhood in their time interval, where other transceivers receive it; these forward it again gradually in their time interval, and the data spreads within the network in this manner [10].

3 An Example of Using the Intelligent Lamp in Real Traffic

A possible scenario for use of the intelligent lamp is shown in Fig. 5. The individual lines between the network nodes represent the created communication paths within IQMESH.

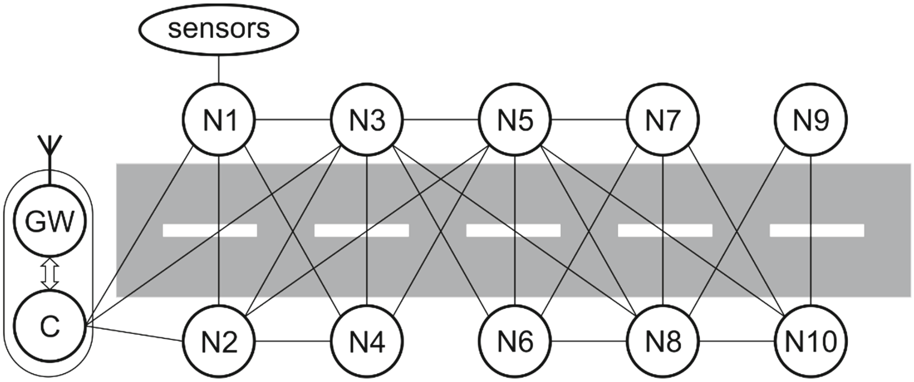


Fig. 5. Possible connection topology of intelligent lamps

The network coordinator, indicated by the letter C in Fig. 5, has a line of sight to nodes 1, 2 and 3. These wireless network nodes are thus in zone 0. If it is needed to communicate with a lamp marked as N9, the shortest path would be communication over nodes marked as N3 and N8. This means that the lamp is marked as N9 in zone 2 and the coordinator may communicate with the lamp at a distance of 3 hops – the defined time slots. In Fig. 5, some of the links do not occur. This situation may be represented by a static obstacle between the lamps, such as a tree with leaves. However, network topology with alternative paths can deliver a message even to these weakly interconnected nodes. Creating links between individual network nodes is provided by the network coordinator by means of the command “Discovery”, which is one of the functions of the coordinator operating system.

In Fig. 5, the node marked as N1 is equipped with a sensor portion, whose block diagram is shown in Fig. 3. Other lamps have a block structure as indicated in Fig. 1. In this case, the sensor part of the lamp can serve for providing information on the immediate surroundings of the intelligent lamps. Users living in the particular street can find out the temperature, humidity, pressure, or any other monitored variables directly in the area where there are or where they live. The PIR sensor can serve to enhance the brightness of a light line in the particular street if a person is going along the street. Obviously, it is possible to place multiple sensors in a row, and the light thus illuminates progressively, depending on the current position of the person.

The actual control of the brightness of the lamp is implemented directly in the node of the intelligent lamp with the sensory part. The requirement to reduce/increase the brightness of lamps in the street depending on the occurrence of a person would be addressed by the wireless network communication. These issues would not be controlled by the parental system. At maximum, this property could be enabled or disabled from the parental system.

For the monitoring of environmental variables, it is necessary to have these sensors supplemented by a backup power source – a battery providing energy at a time when the lamps are not powered from the grid. If wireless lamps with sensory parts are often placed outside the range of the network coordinator (i.e. outside the zone 0), it is necessary to place a lamp which also has a backup power supply between the network coordinator and the lamp with the sensory part – see Fig. 1. This lamp would serve a packet router during the day. In essence, this extends the range of the wireless network.

One network coordinator can serve up to 239 wireless nodes. These can be placed in up to 239 zones, which is the least preferred network architecture, since it represents a chain architecture. Failure of any node of the network would mean a loss of connectivity to the other network nodes. Overall, the wireless network IQMESH can serve up to 65,000 nodes, where there are nodes of the type coordinator-node (abbreviated CN), which create subnetworks. On the side of the subnetwork, the node of the CN type is thus in the role of the coordinator of the subnetwork, but on the side of the superior network, it is in the role of a network node.

A superior system is the gateway, whose second communication party is either Ethernet, Wi-Fi or a GSM modem. The gateway usually contains the IQRF module, which serves as the network coordinator. It is possible to use commercial gateways delivered by the IQRF technology manufacturer or our own solution based on the platform Raspberry-Pi, along with the IQRF module. Our team is also engaged in our own gateways to mediate communication between the wireless sensor network and the Internet. This gate is described in the article [9]. It also deals with the area of reducing energy intensity in the operation of these wireless networks, as described in the article [1].

The gateway can send data on the traffic to the parental monitoring system. This superior system would then provide status information for each intelligent lamp, the value of measured variables, enabling it to manage their brightness or enable brightness control based on the presence of a person in the vicinity of the lamp.

4 Conclusion and Future Work

The paper has dealt with the design of the intelligent lamp, which will be, in addition to such basic functions as switching on and off, also able to measure basic environmental variables, including temperature, humidity, atmospheric pressure, the concentration of selected gases and concentrations of airborne dust. The latter two variables are planned to be implemented in the next stage of development. Furthermore, the lamp is mounted with a PIR sensor used to control the brightness of the intelligent lamp according to the detection of the passage of persons under the lamp. The intelligent lamp is constructed with a view to integration into the Smart Cities/Smart Lighting system, the idea being to control and monitor lamps in one street or in a single location individually. Communication with other lamps is solved by the IQMESH network based on IQRF. Transferring data to a superordinate monitoring system is solved through the GSM gateway. A software solution is implemented here, based on cloud storage, from which it is possible to visualize the measured data.

Acknowledgments. This work was supported by the project of the Ministry of Industry and Trade - MPO FV10706 "Research and development of intelligent sensor network to monitor the safety barriers for geotechnical use" and by project SP2017/158, "Development of algorithms and systems for control, measurement and safety applications III" of the Student Grant System, VSB-TU Ostrava.

References

1. Prauzek M, Krömer P, Rodway J, Musilek P (2016) Differential evolution of fuzzy controller for environmentally-powered wireless sensors. Appl Soft Comput J 48:193–206. doi:10.1016/j.asoc.2016.06.040 Elsevier Science BV, Amsterdam
2. Slanina Z, Docekal T (2016) Energy monitoring and managing for electromobility purposes. In: Proceedings of SPIE - the international society for optical engineering, art. no. 100311P. doi:10.1117/12.2247799
3. Kovács A, Bátai R, Csáji BC, Dudás P, Háy B, Pedone G, Révész T, Váncza J (2016) Intelligent control for energy-positive street lighting. Energy 114:40–51. doi:10.1016/j.energy.2016.07.156 Elsevier Ltd
4. Hajovsky R, Pies M (2015) Use of IQRF technology for large monitoring systems. IFAC-PapersOnLine 28(4):486–491. doi:10.1016/j.ifacol.2015.07.082
5. Mlynek P, Misurec J, Kolka Z, Slacik J, Fujdiak R (2015) Narrowband power line communication for smart metering and street lighting control. IFAC-PapersOnLine 28(4): 215–219. doi:10.1016/j.ifacol.2015.07.035
6. Ciriminna R, Albanese L, Meneguzzo F, Pagliaro M (2015) LED street lighting: a looking ahead perspective. Green 5(1–6):83–94. doi:10.1515/green-2015-0020
7. Ozadowicz A, Grela J (2015) The street lighting control system application and case study. In: Proceedings of 1st international conference on event-based control, communication and signal processing, art. no. 7300701. doi:10.1109/EBCCSP.2015.7300701
8. De Souza IH, Denardin GW, De Pelegrini Lopes J, Vargas DR (2013) Street lighting system based on led modular drivers. In: IEEE 13th brazilian power electronics conference and 1st southern power electronics conference, art. no. 7420214. doi:10.1109/COBEP.2015.7420214
9. Pies M, Hajovsky R, Latocha M, Ozana S (2014) Radio telemetry unit for online monitoring system at mining dumps. Appl Mech Mater 548-549:736–743. doi:10.4028/www.scientific.net/AMM.548-549.736
10. Pies M, Hajovsky R, Ozana S, Haska J (2014) Wireless sensory network based on IQRF technology. In: 4th international workshop on computer science and engineering-winter
11. Behan M, Krejcar O (2013) Vision of smart home point solution as sustainable intelligent house concept. IFAC-PapersOnLine 12(PART 1):383–387. doi:10.3182/20130925-3-CZ-3023.00057
12. Stankus M, Prauzek M, Jirka J (2013) Universal wireless system for advanced biomedical applications. Appl Mech Mater 330:1049–1053. doi:10.4028/www.scientific.net/AMM.330.1049
13. Seflova P, Sulc V, Pos J, Spinar R (2012) IQRF wireless technology utilizing IQMESH protocol. In: 35th international conference on telecommunications and signal processing, pp 101–104. doi:10.1109/TSP.2012.6256261

14. Spinar R, Spiller M, Seflova P, Sulc V, Kuchta R, Vrba R (2011) IQRF street lighting - A case study. In: 4th international conference on advances in mesh networks, pp 33–38. ISBN 978-161208147-2
15. Krejcar O (2009) Full scale software support on mobile lightweight devices by utilization of all types of wireless technologies. In: Social-informatics and telecommunications engineering. Lecture notes of the institute for computer sciences, vol 13, Springer, pp 173–184. doi: 10.1007/978-3-642-03819-8_17

Options of Monitoring the State of Protection Networks

Radovan Hajovsky(✉), Martin Pies, and Martin Stankus

Department of Cybernetics and Biomedical Engineering, Faculty of Electrical Engineering and Computer Science, VSB-Technical University of Ostrava, 17. Listopadu 2172/15, 70833 Ostrava, Czech Republic
{radovan.hajovsky,martin.pies,martin.stankus}@vsb.cz

Abstract. This article describes the conceptual proposal of measuring erosion of rock formations using protection networks. It outlines the concept of two types of wireless sensors of a steel retaining network tension and a data concentrator unit, which cyclically reads the measured data from these wireless sensors. The measured data is then sent to the parental monitoring system, which is used for archiving, visualization and data processing. Communication between the wireless nodes and the data concentrator is based on the IQRF technology.

Keywords: IQRF · MESH network · Monitoring system · Retaining network · Vibrating-wire strandmeter

1 Introduction

In the Czech Republic, as well as abroad, there are a large number of rock formations, whose damage poses a direct threat to objects or traffic on the roads situated under these formations. For these reasons, the stability of many of such unstable rock formations is ensured using protection networks or dynamic barriers. Various examples of rehabilitation using barriers and protection networks are described in the article [13], which deals with the rehabilitation of a rock massif in Germany. Another example is the use of a retaining wall and retaining networks in Albania, see [9] and in the USA, see [8, 10].

Generally, for the rehabilitation of rock formations, high-strength steel mesh is used for two basic systems of protection.

The first type is used for simply ensuring safe trajectory of the fall of loose stones from their original location to the foot of the wall. The network is anchored in the upper part to the bedrock; the individual strips are interconnected in a single unit and they either hang freely under their own weight, or using a suitable load. The aim of such an installation is not to secure falling rocks in place, but to ensure the safe transport space for naturally falling loose rocks beneath this network.

The second system that uses high-strength steel mesh, which dominates in the Czech Republic, is sheathing the rehabilitated massif. The network follows the morphology of the rock wall to the maximum possible extent. The individual strips are fastened together as a whole and anchored to the ground in the entire area as shown in Fig. 1. The density and anchoring system mainly depends on the specific state of the bedrock. The strength of such measures can be further enhanced by adding the help of ropes (system surface

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_17

snubbing), or by covering with stronger networks constructed from steel cables, so-called HEA panels.

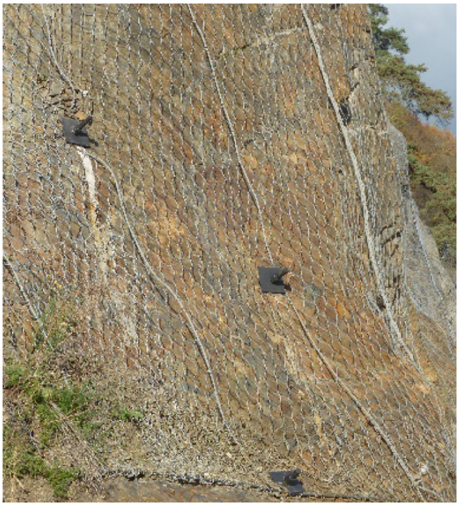


Fig. 1. Anchoring high-strength steel mesh to the bedrock in the Štěchovice area, Czech Republic

The primary objective of such (either basic or reinforced) installation is to stabilize and maintain loose stones and blocks in their original position. Ideally, stabilizing blocks in the original stable position without the risk of destabilization and collapse.

A secondary goal in the case of the destabilization and collapse of the blocks is to maintain loose, already falling blocks in the network, and to prevent their falling into the protected area. Here the inevitable factor of gradual natural filling and loading the nets and anchor systems arises. Natural evolution has to lead to reaching the limit capacity of filling with the possibility of overloading and collapse of the system. The measures originally installed for security purposes can suddenly become a risk in themselves.

Such limit states can be achieved within decades, depending on the state of weathering processes, type and dimensioning of protective measures. In principle, however, it can be safely stated that there is a high level of probability of achieving such a condition [11].

Due to the fact that currently there is no system of inspections and maintenance almost anywhere, and with respect to placement in difficult terrain, installations are basically inaccessible or they cannot be accessed easily, risk prediction using newly developed autonomous measuring systems seems to be necessary and the only possible option. The first installations of this type in the Czech Republic have been more widely implemented especially during reconstruction of railway corridors after the year 2000. The need for the above-mentioned predictions already starts to be relevant, see [6]. The aim of this article is to publish the idea of monitoring obsolete, but also newly constructed retaining networks and barriers through advanced monitoring capabilities, which fall within the area of the Internet of Things.

In cooperation with a major company in the Czech Republic, which deals with the installation of the aforementioned protective networks, a research project with the aim of developing a comprehensive monitoring system to monitor the condition of protective nets, as well as dynamic barriers has been implemented in the years 2016–2019. The intended result is continuous monitoring of the condition of retaining networks and indicating the occurrence of events, such as dislocation in the retaining network by motion of a rock block, fallen rocks, etc.

2 Proposal for the Monitoring System

Based on the analysis of the current state of the art in the field of monitoring systems for protection nets and dynamic barriers, a block diagram of a comprehensive measuring system that is shown in Fig. 2 has been proposed.

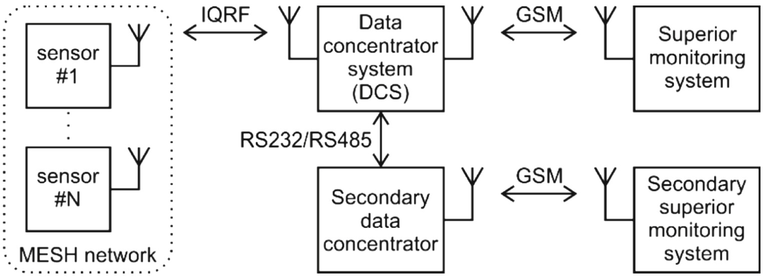

sensor #1
sensor #N
MESH network
IQRF
Data concentrator system (DCS)
GSM
Superior monitoring system
RS232/RS485
Secondary data concentrator
GSM
Secondary superior monitoring system

Fig. 2. A block diagram of the monitoring system

A part of the system placed in the field consists of wireless sensors connected using the IQRF wireless network in the MESH topology and a data concentrator (hereinafter referred to as the Data Concentrator Station - DCS).

The data concentrator further communicates with the parental monitoring system, which consists of data storage, a software application providing data processing and visualization. The data concentrator acts as an interface between the network of wireless sensors – nodes and the parental monitoring system.

Within the analysis of the requirements for the complex measuring system it revealed that, besides the implementations of the communication interface, the following requirements are placed on the DCS:

- optimization at the lowest possible power consumption, possibly with battery recharging using a solar cell or other suitable means for energy harvesting. Servicing the battery may be difficult to achieve, for example because of poor access to the DCS in the field. For the same reason, the device must be operational without recharging the battery.
- maintenance-free condition. The equipment must be able to work in aggravated conditions and resist physical influences (moisture, dust).
- wide range of operating temperatures.
- the possibility of extending functionality. Prospectively, it is supposable that there will be a possibility to implement communication with the secondary data concentrator or other data source connected by asynchronous serial interface (RS232,

RS485). This may include the system GEMON from the company STRIX Chomutov, a.s., or another one.

Based on the analysis, it was found that no product completely meeting the above-mentioned requirements is currently available on the market. It was therefore decided to initiate the development of a complete system, including individual sensor modules and data concentrator in accordance with the specified requirements.

To implement the wireless communication of the data concentrator with sensors, it was decided to use the IQRF technology that allows realization of the network with MESH topology; this decision was based on positive experience in dealing with similar types of applications. The maximum number of sensors in the network is not limited by the structure of the measuring system (but may be restricted by the wireless network parameters – maximal count of wireless nodes per one coordinator is limited to 239). The data concentrator is used in the role of the coordinating node of the wireless network.

The data concentrator communicates with the parental monitoring system. GSM technology is used for communication. The data transfer assumes the use of TCP/IP and FTP protocol. The data concentrator provides the parental monitoring system with periodic data about the values measured by the sensors, or information about alarm conditions, if they are detected by the individual sensors. The data concentrator is equipped with an asynchronous communication interface of RS232 or RS485 type, which is primarily designed to connect to another (secondary) data concentrator (for example, type GEMON). This is an element designed for the future expansion of the measuring system. The purpose is to enable the sharing of a single GSM interface by more data concentrators, or general data exchange between concentrators.

Implementation of a reserved external memory (memory card, etc.) to store data from the sensors in the data concentrator is not expected. The data stream generated by the sensors is small (it is a device optimized for low power consumption); therefore, using the internal memory of the data concentrator in the role of a buffer is assumed. The purpose of the data concentrator is not to create a backup copy of the measured data. The research team has several years of experience in the development of wireless monitoring systems. The publication [3] describes a telemetry station and the parental monitoring system. The publication [4] describes the use of wireless IQRF technology for monitoring mine heaps.

2.1 Data Concentrator Architecture

The central element of the data concentrator is the microcontroller (MCU). Using PCB, the microcontroller is connected with the other functional elements of the device using standard communication interfaces. The individual function blocks of the concentrator are shown in Fig. 3.

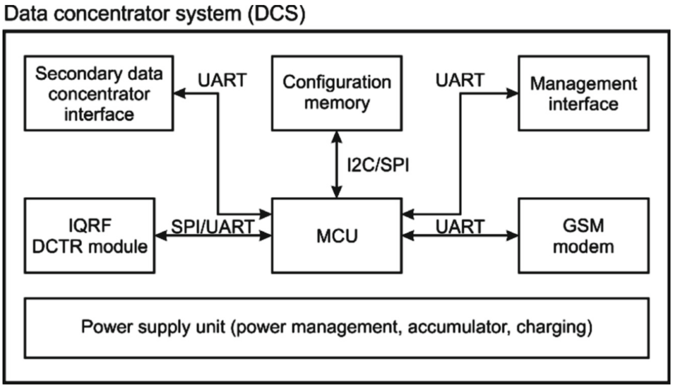

Data concentrator system (DCS)
Secondary data concentrator interface
UART
Configuration memory
UART
Management interface
I2C/SPI
IQRF DCTR module
SPI/UART
MCU
UART
GSM modem
Power supply unit (power management, accumulator, charging)

Fig. 3. The considered data concentrator structure

The used microcontroller is based on the ARM Cortex-M0+ technology in low power version, in which the manufacturer guarantees current consumption of 12 mA at an operating frequency of 75 MHz and 4 μA in stop mode. For communication with the parental system, the GSM modem will be used, which contains its own implementation of TCP/IP and FTP data transfer. Therefore, there will be no need to implement these services within the microcontroller. To communicate with the individual nodes of the sensor network, the IQRF module will be used, whose basal operating system includes a complete set of commands for creating, managing, and communicating in a wireless network of the MESH topology. The DCS unit will also be equipped with advanced technology of the power supply management, whose analysis is described in [2].

2.2 Sensory System

Currently, there are not enough studies for measuring the state of stretching of the retaining steel mesh. An approach closest to our requirement on the solution is presented in the publication [12], where the researchers measured deflection on the network in a laboratory model. Other publications [1, 5] are focused on applications of mathematic models of retaining networks and restraining walls on unstable slopes.

The aim of the sensory system will be to measure the state of tension of the retaining network on a long-term basis. As already mentioned in the introduction, the retaining networks may sag over time due to the release of rocks that the given retaining network holds in place. Releasing rocks can thus be monitored by monitoring the state of stretching the retaining network. When the retaining network is stretched, its shape is changed and its surface is increased. There should be two possible ways to monitor this change:

- Wire linear potentiometer
- Vibrating-wire strandmeter

Wired linear potentiometer. A possible way of installing this type of sensor on the retaining network is shown in Fig. 4. The sensor is placed in the desired location in the protection networks so that the base of the sensor is mounted to the static portion of a retaining network, such as the anchor. A steel wire stretched over the entire active area of the retaining network is fixed to the movable part of the sensor.

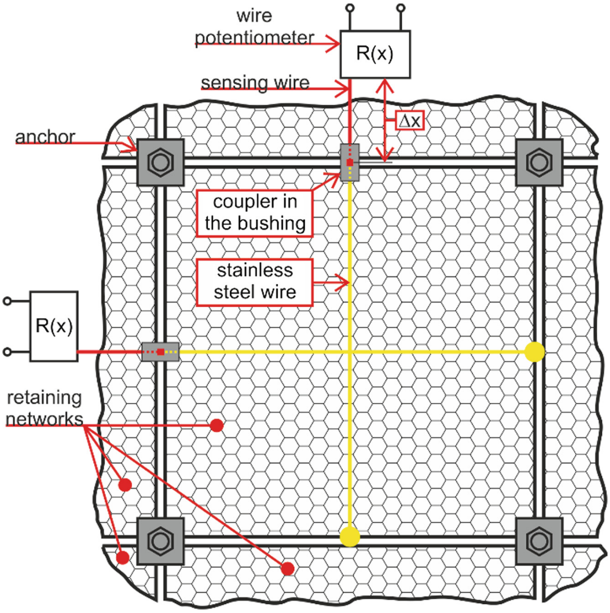


Fig. 4. Possible attachment of wired linear potentiometers to the retaining net.

The active area of the retaining network is an area where loose rocks and thus e tension of the network can occur. Therefore, the degree of tension of the protective network is directly reflected by the extent of pushing the wire out of the wire linear potentiometer. The wired linear potentiometer is connected to a wireless measurement node that sends information about current situation to the data concentrator (DCS) and later to the parental system. Power supply is ensured using 3.3 V lithium battery with sufficient capacity to ensure operation for several years, depending on the frequency of requests about the state of the protection network tension. For monitoring the states of the protective networks, a 4-channel measurement module – a wireless node will be designed and realized, meaning that it will be possible to connect up to 4 linear potentiometers to one node.

As an alternative to analogue processing of the information from these potentiometers, there are wired linear potentiometers, where the angle sensor is realized by an encoder, IRC sensor. This variant is less susceptible to interference and temperature-compensation is not needed, as in the case of analogue linear potentiometers. Another advantage is the lower price. The disadvantage is more complicated processing of the information from this sensor.

Vibrating-wire strandmeter. Another option for monitoring the tension of the protective network is the use of the vibrating-wire strandmeter. In this case, however, the use is more complex due to the necessity to develop and implement the supporting electronics for processing signals from the vibrating-wire strandmeter. In this case, it is necessary to design and implement a unit for processing the signals from this sensor, as is briefly described in [7]. The developed system would have the support of wireless technology IQRF, as in the case of monitoring with a wired potentiometer. Fixing this type of sensor would be carried out through wire ropes, to which the given sensor would be attached. The degree of tension of the steel cable would respond to the tension of the protection network. A sketch of the attachment of the sensor is shown in Fig. 5.

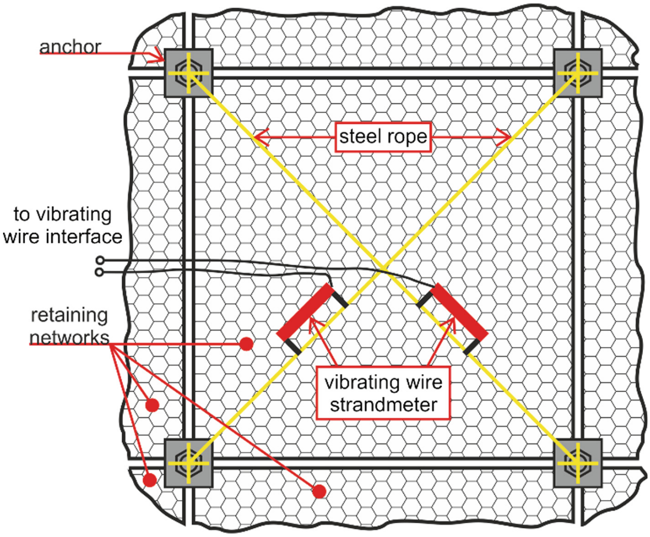


Fig. 5. Possible attachment of vibrating-wire strandmeters to the retaining net.

As in the case of the first measurement method and the second method of measuring the tension of the retaining network, the measurement would not take place continuously, but for the necessary time in milliseconds. The measured data would be stored in the transmitter memory. This arrangement of time limited measurement is due to higher energy consumption than in the previous measuring method. In the event of a query from the network coordinator each wireless transmitter would then send the pre-measured data. After the end of the transmission, a new measurement would be carried out.

2.3 Parental Monitoring System

The parental monitoring system will provide the processing function, archiving and data visualization. It will also be its responsibility to inform about achieving limit states of the chosen measurement. The use of modern technologies and services for commercial cloud storage is intended along with a parallel archiving of data in a data-base system for web hosting. To visualize, our own solution based on dynamic web pages using PHP/MySQL technology will be applied, and also connections of visualization tools with Cloud services will be used, for example, solutions from Microsoft - Azure + Power BI.

3 Conclusion

The contribution focused on describing the design of a comprehensive monitoring system for the measurement of tension of protection networks and barriers serving as protection against falling rocks from eroded rock formations. The basis of the system is the use of wireless MESH networks of measuring points that communicate with the data concentrator as was the mentioned on the beginning of Sect. 2 in this paper. The DCS subsequently transfers this measured data to the parental monitoring system. All communication is performed wirelessly using IQRF technology. System development is funded in the framework of the R&D project and pilot deployment at the selected location is planned in the next two years. Currently, laboratory tests are being carried out on various types of sensor and communication between them. The portfolio of sensors and actuators in the field of IoT will therefore be complemented with other interesting sensors monitoring the erosion of the rock formations.

Acknowledgement. This work was supported by the project of the Ministry of Industry and Trade - MPO FV10706 "Research and development of intelligent sensor network to monitor safety barriers for geotechnical use" and by project SP2017/158, "Development of algorithms and systems for control, measurement and safety applications III" of Student Grant System, VSB-TU Ostrava.

References

1. Wendeler C, Bühler Y, Bartelt P, Glover J, Luis R (2016) Application of three-dimensional rockfall modeling to rock-face engineering. Rock Mech Rock Eng From Past Future 2:1303–1306
2. Prauzek M, Krömer P, Rodway J, Musilek P (2016) Differential evolution of fuzzy controller for environmentally-powered wireless sensors. Appl Soft Comput J 48:193–206. doi:10.1016/j.asoc.2016.06.040 Elsevier Science BV, Amsterdam
3. Pies M, Hajovsky R, Latocha M, Ozana S (2014) Radio telemetry unit for online monitoring system at mining dumps. Appl Mech Mater 548-549:736–743. doi:10.4028/www.scientific.net/AMM.548-549.736
4. Pies M, Hajovsky R, Ozana S, Haska J (2014) Wireless sensory network based on IQRF technology. In: 4th international workshop on computer science and engineering-winter
5. Lin H, Ling HI, Li L, Collin JG, Leshchinsky D, Rimoldi P (2013) Centrifuge modeling of gabion walls reinforced with geosynthetics. Design Practice of Geosynthetic-Reinforced Soil Structures, pp 93–102
6. Talend D (2013) Going to extremes: Expertise and slope stabilization systems handle some of the toughest sites imaginable. Erosion Control 20(4):28–35
7. Simonetti A (2012) A measurement technique for the vibrating wire sensors. NORCHIP 2012, art. no. 6403147. doi:10.1109/NORCHP.2012.6403147

8. Mandavkar S, Stralla A, Kanati NV (2012) Construction of MSE walls at St. Anthony Falls (I-35 W) bridge. Geotechnical Special Publication (225 GSP), pp 3634–3643. doi: 10.1061/9780784412121.372
9. Jayakrishnan PV., Suntharathevan, Vicari M, Uguccioni G, Budhbhatti R (2012) Composite soil reinforcement system in construction of high retaining walls for Albania Highway Project. In: GA 2012 - 5th Asian regional conference on geosynthetics: geosynthetics for sustainable adaptation to climate change, pp 939–944
10. Mandavkar S, Stralla A (2011) Maccaferri terramesh system, gabion face MSE walls at ST. Anthony Falls (I-35 W) bridge. Geotechnical Special Publication (211 GSP), pp 3449–3458. doi:10.1061/41165(397)353
11. Benessiuti MF, Bernardes GP, Ananias EJ (2010) Segmental retaining wall: Comparison between predicted and observed slip surface. In: 9th international conference on geosynthetics - geosynthetics: advanced solutions for a challenging world, pp 1773–1776
12. Castro-Fresno D, López Q. L, Bianco-Fernandez E, Zamora-Barraza D (2009) Design and evaluation of two laboratory tests for the nets of a flexible anchored slope stabilization system. Geotech Test J 32(4):315–324
13. Züger M, Haller B (2005) High-energy rockfall barriers for 3,000 KJ impact energies to protect B4 highway at Ilfeld, Germany. Pol Geol Inst Spec Pap 20:125–130

Performance Enhancement of ℓ0-LS Approximation in Sparse Underwater Channel Estimation

S. Jimaa[1](✉), L. Weruaga[2], T.A. Shimamura[3], A. AlAli[1], D. AlAkil[1],
S. Albraiky[1], and J. AlAli[1]

[1] ECE Department, Khalifa University, Abu-Dhabi, United Arab Emirates
saj@kustar.ac.ae
[2] Proactivaudio, Vienna, Austria
luis.weruaga@proactivaudio.pro
[3] ICS Department, Saitama University, Tokyo, Japan
shima@sie.ics.saitama-u.ac.jp

Abstract. Finding efficient methods to communicate over underwater acoustic channels that suffer from multipath, attenuation, noise, frequency selectivity and Doppler effect has been a challenge. A multipath underwater acoustic channel can be seen as a sparse system, whose impulse response contains many zeros and very few large coefficients. As stated in the compressive sensing theory, the actual measure of sparsity is the ℓ0-norm. Though, minimizing the ℓ0-norm is an untraceable problem. Consequently, the ℓ0-LS approximation with major reduction in computational complexity is proposed. This paper study the approximation of the ℓ0-LS on sparse underwater channels' estimation. Also it presents the performance enhancement of using such approximation on the underwater sparse channel estimation.

1 Introduction

The underwater digital communication researches have been developing over the last three decades leading to achievements and improvements since it started. This field of communication would support so many sectors such as defense, marine research, marine commercial operations, oceanography, and the offshore oil industry. The underwater nature has significant amount of complications compared to the one in the air. One to mention is that a regular EM waves will suffer from large attenuation. In addition, optical waves will lead to a large amount of scattering. Consequently, the most suitable type of signals to use underwater will be the acoustic signals. Even though, using a high speed acoustic channel is challenging due to the limited bandwidth, severe fading, extended fading, large Doppler shift and refractive properties [1]. The underwater acoustic channel is considered sparse such that it can consist of one direct path and number of multipath. That indicates the impulse response of the channel will have few large coefficient and many small near to zero coefficients. Taking into consideration the sparse nature of the acoustic channel, the project target will be to use the sparse adaptive algorithms in order to enhance underwater acoustic communication.

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_18

So far Compressive Sensing (CS) theory has been applied to perform sparse channel estimation [6]. In [7] authors showed that the Matching Pursuit (MP) and Basis Pursuit (BP) are well suited for underwater acoustic (UWA) channel, since this channel is often very sparse [8]. Thus, those algorithms and their variants have been used in [10–14] for a multicarrier system to estimate the channel. It is notable that the $\ell 1$-norm optimization problem is now largely popular under the name BP as introduced in [9].

The authors in [10] developed a time domain channel estimation technique based on MP which gives a sub-optimal estimate by detecting the best aligned signal subspace and canceling the effect of the detected subspace iteratively. They used Orthogonal MP (OMP) algorithm to solve the slow convergence problem of re-selected taps. OMP basically utilizes the subspace chosen by MP to re-compute the taps according to the LS estimate. Authors proved that the OMP algorithm has a superior performance over the original MP and that the time-domain MP is superior to conventional frequency-domain CE using interpolation methods. Also, authors showed that the performance of conventional time-domain Least Square (LS) channel estimation is poor comparing with OMP. This is mainly due to its inability to exploit the channel's sparsity feature.

Various channel estimators have been investigated in [11] that exploit channel sparsity in the time and/or Doppler domain for a multicarrier underwater acoustic system. Root-MUSIC and ESPRIT subspace algorithms were used to estimate the underwater channel that have limited Doppler spread. On the other hand, for channels with Doppler spread, compressed sensing approach was adopted, mainly OMP and BP algorithms. The proposed algorithms were compared to the conventional LS channel estimator. Results showed that compressed sensing algorithms uniformly outperformed the LS and subspace methods.

In [12] authors deployed OMP and BP along with the conventional LS to estimate the UWA channel. Numerical simulations and field results demonstrated the substantial benefits of compressive sensing for underwater acoustic communications over long dispersive channels with large Doppler spread.

An efficient and low complexity channel estimation algorithm [13] has been proposed by exploiting the sparseness of the channel impulse response and the prior information for the non-Gaussian channel gain which is modeled by an exact continuous Gaussian mixture (CGM). By combining the MP and the maximum a posteriori probability (MAP) based space-alternating generalized expectation-maximization technique, the estimation of channel taps and their locations is improved in an iterative manner. Results showed that the proposed algorithm exhibits excellent symbol error rate and estimated the UWA channel very effectively.

Finally the authors in [14] proposed a virtual time reversal processing (VTRP) for OFDM systems to reduce the ISI caused by delay spread resulted from the underwater channel. They used two sparse channel estimation methods which are matching pursuit (MP) and basis pursuit de-noising (BPDN) to estimate the channel impulse response (CIR). Results showed that VTRP with BPDN channel estimation is slightly better than using MP channel estimation.

2 System Model

An underwater acoustic channel is a time varying channel, therefore a time varying filter can be used in which the coefficients of the filter are constantly updated each time interval. Adaptive filters are self-designing filters because they calculate the coefficients from pervious samples [15]. The main aim of using an adaptive filter is to minimize the cost function by updating the coefficients every time interval, as can be seen from the following equation:

$$\mathrm{W}(\mathrm{n}+1)=\mathrm{W}(\mathrm{n})+\Delta \mathrm{W}(\mathrm{n}) \tag{1}$$

Where W(n+1) denotes the new coefficients of the current time step, W(n) is the previous time step coefficients, and ΔW(n) is the adaptive algorithm connection. The impulse response is sparse for the underwater acoustic channel. That is the system contains many negligible small coefficients with only few large ones. Taking the sparse nature of the system helps in utilizing those large coefficients in order to enhance the estimation performance. In Fig. 1, sparse system identification is shown. The input signal u(n) undergoes to unknown sparse system which the output signal of it is the desired signal d(n). Additionally, When the output error e(n) is minimized, the adaptive filter becomes a model for the unknown sparse system. Consequently, the main problem relies in finding an adaptive algorithm that is able to identify and exploit the sparse nature of the unknown sparse system [16].

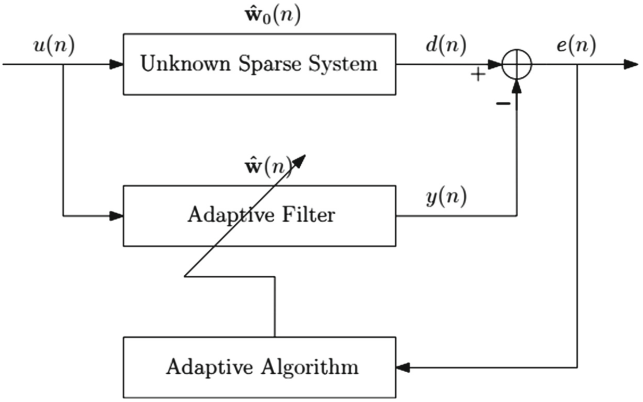


Fig. 1. System identification model using linear transversal adaptive filter

3 Adaptive Filtering and Sparse Adaptive Algorithms

This section illustrates the effective approximation of ℓ0-LS algorithm in order to improve the underwater acoustic channel estimation performance using an OFDM system. It has been proven that ℓ0-LS approximation has better performance in estimating an OFDM channel [21, 22, 25]. Generally, an underwater channel suffers from a large amount of multipath and echo. Therefore, a sparse system exists in an underwater acoustic channel. The system is considered to be sparse when most of its coefficients are insignificant, while few coefficients are significant. Figure 2 illustrates a sparse system. It can be noticed that the number of significant coefficients is equal to 3 where the length of the channel is 20.

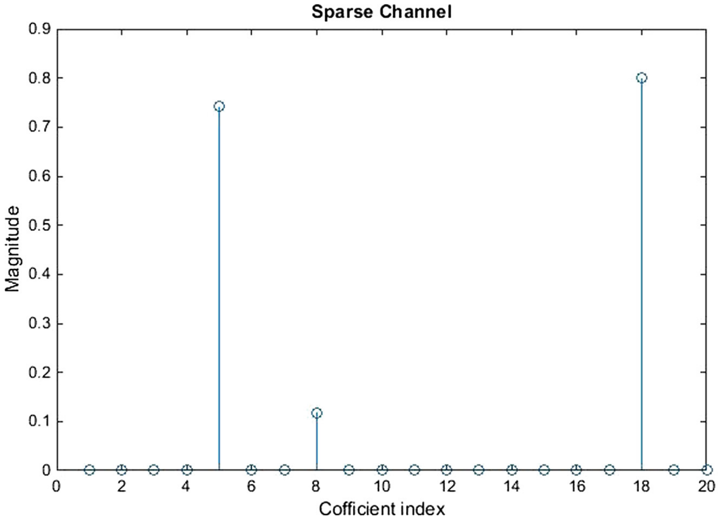


Fig. 2. A sparse channel

3.1 Sparsity Channel Estimation

The regular adaptive filtering algorithms [15–24] such as LS, NLMS, and RLS do not take into consideration the nature of the acoustic channels. In other words, those algorithms do not consider the sparsity nature of the channel. To overcome this problem Compressive Sensing (CS) theory has been developed [25]. The ℓp-norm ($0 \leq p \leq 1$) has the ability to take advantage of the sparse nature [20]. It will attract the insignificant taps or the small tapes to zero. The ℓ0-norm measures the exact sparsity of the channel, hence it gives the sparsest solution. However that raise a penalty for non-Polynomial hard problems [26]. In [23] an alternative approach has been developed to overcome the penalty of the ℓ0-norm. They have come with an algorithm that is close to the ℓ0-norm that is called approximation of ℓ0-LS.

3.2 Approximation of ℓ0-LS Channel Estimation Method

Approximation of ℓ0-LS method is created as several algorithms that complete each other gaps. Starting with the conventional LS estimation, to applying ℓ1-norm, then employs a suitable threshold, and ending with re-computing techniques to reach the final ℓ0-LS approximation. Figure 3 below shows a diagram that illustrates the approximation ℓ0-LS algorithm.

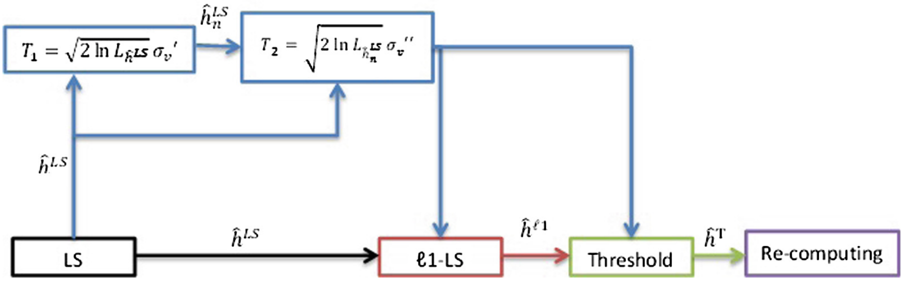


Fig. 3. Block diagram of the ℓ0-LS approximation algorithm

3.3 Conventional LS Estimation

An OFDM system that has N number of subcarriers and M number of equally spaced pilots is considered. At the receiving end, the received pilots can be expressed as

$$\mathrm{Y_p} = \mathrm{X_P}\,\mathrm{H_P} + \mathrm{W_P} \tag{2}$$

where Y_p is the received pilot vector, $Y_p = [y_{M0}, y_{M1}, \ldots \ldots ., y_{M-1}]T$, and X_P is the transmitted pilots diagonal matrix, $X_p = diag(x_0, x_1, \ldots \ldots ., x_{M-1})$. To approximate the channel, the square error cost function is minimized based on the frequency domain LS as

$$J = \left\| Y_p - X_P \hat{H}_P \right\|_2^2 \tag{3}$$

As a result the solution of the frequency domain LS is

$$\hat{H}_p^{LS} = \left(X_p^H X_P\right)^{-1} \left(X_p^H Y_P\right) = X_p^{-1} Y_P \tag{4}$$

where H is the transpose and complex conjugate. An important point to mention is that the estimation done here is applicable only for the pilot's locations. That will allow the usage of several interpolation techniques, such as Discrete Fourier Transformed (DFT). DFT in this paper is used because the pilots are equally spaced, to obtain the channel frequency response at data sub-carrier.

3.4 LS Analysis

The frequency domain LS is known by its simplicity, nevertheless it can be further enhanced by considering the sparsity nature of the channel. While the channel is sparse, the noise will extremely influence the LS estimation. The channel estimation in this case can be expressed as

$$\hat{h}^{LS} = h + v \tag{5}$$

where h is the actual channel and v is the noise vector. Then the estimated elements of the channel can be sorted as significant and noisy elements as illustrated in the following formula:

$$\hat{h}_i^{LS} \sim \left\{ \begin{array}{cc} CN\left(0, \sigma_h^2 + \sigma_v^2\right) & i \epsilon \Omega \\ CN\left(0, \sigma_v^2\right) & i \notin \Omega \end{array} \right\}. \tag{6}$$

where Ω represents the significant tapes, σ_h^2 represents the significant tap variance and σ_v^2 represents the noisy tap variance. This analysis will help in extracting information and remove the noise form the LS estimated channel.

4 Simulation Results

4.1 Underwater Channel Settings

The following simulations were made by taking into consideration the Mediterranean specifications [27]. The physical characteristics of Mediterranean channel such as salinity and temperature influence the acoustic signal attenuation. The ambient noise and Doppler spread were calculated accordingly to the signal's frequency. The specifications are listed in Table 1.

Table 1. Underwater channel characteristics values

Parameter	Value
Frequency (KHz)	10
Distance (Km)	1
Temperature (C)	30
Attenuation (dB)	0.3695
Ambient noise (dB)	24.1
Doppler spread (dB)	1.25

4.2 Sparse Estimation (ℓ1-LS)

In this section, the sparse adaptive algorithm ℓ_1-LS performance is compared with the conventional LS in terms of estimating the actual channel impulse response of the underwater acoustic channel. The system parameters that are used in the simulation are shown in Table 2.

The system uses the minimum condition of a full rank problem, where the number of pilot carriers is 16. Figure 4 shows a comparison between ℓ_1-LS and conventional LS in estimating the impulse response of the channel, where the actual impulse response is estimated using the conventional LS and ℓ_1-LS. The figure shows that the ℓ1-LS estimation algorithm is better in estimating the impulse response of the channel than the LS estimation algorithm as it appears closer to the actual impulse response.

Table 2. Simulation Parameters for conventional vs ℓ1-LS sparse estimation

Parameter	Value
Number of symbols	10000
Number of subcarriers	256
Channel length	16
Cyclic Prefix length	16
Modulation scheme	16-QAM

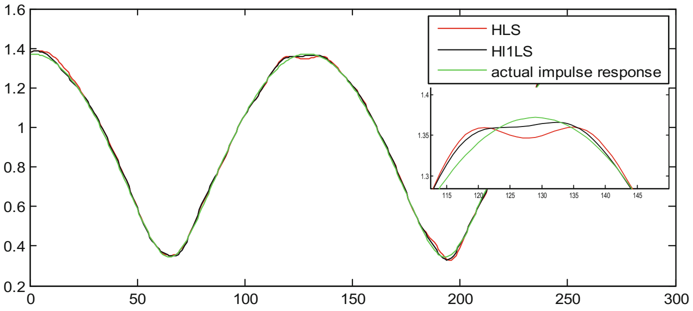


Fig. 4. Channel impulse response estimation using LS and ℓ_1-LS

4.3 Comparison Between Conventional LS and Sparse Estimations

The aim of this section is to compare between different algorithms used in estimating the channel, and to confirm the performance of the approximated ℓ_0-LS algorithm in estimating the underwater channel. These algorithms are time domain LS, ℓ_1-LS, L_1 + threshold, and ℓ_0-LS. The details of the OFDM scenario selected in the simulation are illustrated in Table 2 above. Details on these algorithms are in Refs. [23, 25, 26]. The channel model used in these tests has two significant taps so the sparsity degree is 12.5%.

The comparison between these estimation algorithms was done according to the MSE and SER versus the SNR of the OFDM system that has 16 pilot carriers, which correspond to the minimum condition of a full rank problem. Moreover, the SNR vector has a range of 15 to 55 and simulated with a step of 5. The simulation results are shown in Figs. 5 and 6.

Figure 5 shows the MSE comparison between various algorithms. The best estimation is achieved by ℓ_0-LS approximation as it has the lowest MSE rates. Nevertheless, the worst estimation is the Time Domain LS. The ℓ_1-LS estimation is slightly better than the Time Domain LS. Otherwise, advancement estimation is achieved by ℓ_1 + threshold and ℓ_0-LS approximation. However, for higher SNR values the performance of ℓ_0-LS approximation outperform the ℓ_1+threshold.

It can be noticed from Fig. 6 that the closest algorithm to the known channel is the ℓ_0-LS approximation algorithm. Thus, it achieves the best estimation in comparison with the rest. Actually, the SER of the ℓ_0-LS approximation is the nearest to the SER of the known channel. On the other hand, the farthest algorithm from the known channel is the Time Domain LS, as it achieves a very high SER. Even though ℓ_1-LS provides a slightly better estimation, still it is not good enough. The Time Domain LS + threshold approximately has similar performance compared with the ℓ_0-LS approximation. However, for larger SNR values ℓ_0-LS approximation outperform the Time Domain LS + threshold.

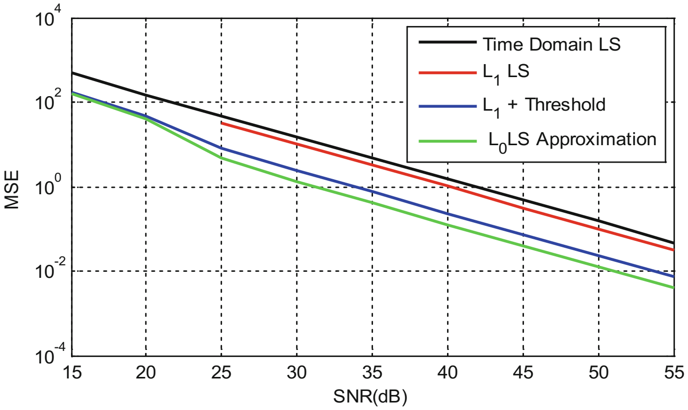


Fig. 5. Comparison between different adaptive algorithms in terms of MSE

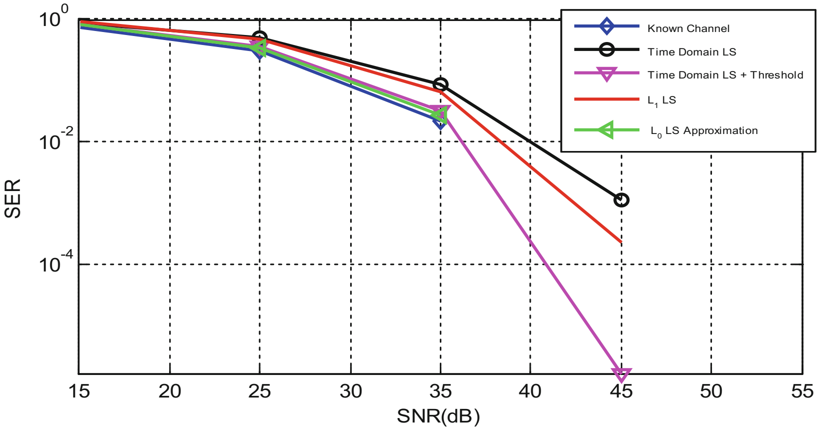


Fig. 6. Comparison between different adaptive algorithms in terms of SER

5 Conclusion

Channel estimation, using various sparse algorithms, was tested for an underwater acoustic channel that is considered to be a sparse channel. An underwater acoustic sparse channel characteristic has been defined. A literature survey of recent publications on the underwater sparse channel estimation has been presented. A very close approximation of the ℓ0-LS with reduction in computational complexity is used and simulation results are presented and discussed that are obtained under relatively low-noise conditions. Results conclude that the ℓ0-LS approximation algorithm, which supports sparsity, has better MSE and SER performances in the estimation of underwater acoustic channels.

References

1. Chitre M, Shahabudeen S, Stojanovic M (2008) Underwater acoustic communications and networking: recent advances and future challenges. Mar Technol Soc J 42(1):103–116
2. Berger CR, Zhou S, Preisig J, Willett P (2009) Sparse channel estimation for multicarrier underwater acoustic communication: from subspace methods to compressed sensing. In: MTS/IEEEOCEANS conference, Bremen, Germany
3. Li B, Zhou S, Stojanovic M (2006) Pilot-tone based ZP-OFDM demodulation for an underwater acoustic channel. In: OCEANS 2006
4. Heidemann J, Mitra U, Preisig JC, Stojanovic M, Zorzi M (2008) Guest editorial—underwater wireless communication networks. IEEE J Sel Areas Commun 26(9):1617–1619
5. Karabulut GZ, Yongacoglu A (2004) Sparse channel estimation using orthogonal matching pursuit algorithm. In: Vehicular technology conference, Los Angeles, CA, September 2004
6. Zhou S, Wang Z (2014) OFDM for underwater acoustic communications. In: OFDM for underwater acoustic communications, 3 June 2014. Wiley
7. Stojanovic M (2015) Underwater acoustic communication. Northeastern University, Boston
8. Kocic M, Brady D, Stojanovic M (1995) Sparse equalization for realtime digital underwater acoustic communications. In: MTS/IEEE OCEANS, vol 3, p 1417–1422
9. Chen SS, Donoho DL, Saunders MA (2001) Atomic decomposition by basis pursuit. SIAM J Sci Comput 43(1):129–159
10. Kang T, Iltis R (2008) Iterative carrier frequency offset and channel estimation for underwater acoustic OFDM systems. IEEE J Sel Areas Commun 26(9):1650–1660
11. Berger C, Zhou S, Preisig J, Willett P (2010) Sparse channel estimation for multicarrier underwater acoustic communication: from subspace methods to compressed sensing. IEEE Trans Signal Process 58(3):1708–1720
12. Berger C et al (2010) Application of compressive sensing to sparse channel estimation. IEEE Commun Mag 48(11):164–174
13. Panayirci E et al (2016) Sparse channel estimation and equalization for OFDM-based underwater cooperative systems with amplify-and-forward relaying. IEEE Trans Signal Process 64(1):214–228
14. Yin Y, Liu S, Qiao G (2015) OFDM demodulation using virtual time reversal processing in underwater acoustic communications. J Comput Acoust 23
15. Takekawa H, Shimamura T, Jimaa SA (2008) An efficient and effective variable step-size NLMS algorithm. In: Proceedings of the 42nd ASILOMAR conference on signals, systems and computers, USA, 26–29 October 2008
16. AlShabili A, Weruaga L, Jimaa S (2016) Adaptive sparsity tradeoff for ℓ1-constraint NLMS algorithm. In: Proceeding of the IEEE ICASSP 2016, Shanghai, China, May 2016, 7472570, pp 4707–4711
17. Jimaa SA, Jadah ME, Sharif BS (2004) Least-Mean-Mixed-Norm adaptive filtering for impulsive DS-CDMA channels. In: Proceedings of the 4th IEEE ISSPIT international conference, Rome, Italy, 18–21 December 2004

18. Jimaa SA, Al-Araji SR, Al-Kaabi A, Shimamura T (2012) Impulsive noise reduction using adaptive receiver structure technique. In: Proceedings of the 11th IEEE international conference on signal processing ICSP, Beijing, China, October 2012, pp 119–122
19. Jimaa SA, Al Simiri A, Shubair RM, Shimamura T (2009) Convergence evaluation of variable step-size NLMS in adaptive channel equalization. In: IEEE ISSPIT 2009, Ajman, UAE, December 2009
20. Weruaga L, Jimaa S (2015) Exact NLMS algorithm with ℓp-norm constraint. IEEE Signal Process Lett 22(3):366–370
21. AlShabili A, Weruaga L, Jimaa S (2016) Optimal sparsity tradeoff in ℓ0-NLMS algorithm. IEEE Signal Process Lett 23(8):1121–1125
22. AlShabili A, Weruaga L, Jimaa S (2016) Modal analysis of the ℓ0-LMS and ℓ0-NLMS algorithms. In: The IEEE MWSCAS 2016, Abu-Dhabi, UAE, October 2016
23. Al-Ogaili F, Elayan H, Alhalabi L, AlShabili A, Taha B, Weruaga L, Jimaa S (2015) Leveraging the ℓ1-LS criterion for OFDM sparse wireless channel estimation. In: Proceeding of the 11th IEEE WiMob 2015, Abu Dhabi, UAE, 19–21 October 2015
24. AlShabili A, Taha B, Al-Ogaili F, Elayan H, Alhalabi L, Weruaga L, Jimaa S (2015) Sparse NLMS adaptive algorithms for multipath wireless channel estimation. In: Proceeding of the 11th IEEE WiMob 2015, Abu Dhabi, UAE, 19–21 October 2015
25. Pathak S, Sharma H (2013) Channel estimation in OFDM systems. Int J Adv Res Comput Sci Softw Eng 3(3):312–327
26. Zhou H, Wang Z (2014) OFDM for underwater acoustic communications. Wiley, New York
27. Zaïbi G, et al (2007) Underwater channel coding and physical characteristics variation in Mediterranean Sea. In: Fourth international multi-conference on systems, signals & devices, Hammamet, Tunisia

Improving the Intra-prediction of H.264 and H.265 Video Coding Standards Using Adaptive Weighted Least Squares Based Predictor

Yenewondim Biadgie[1], Jung-Ju Choi[2], and Kyung-Ah Sohan[1](✉)

[1] Department of Software and Computer Engineering, Ajou University, Suwon, South Korea
{wondim,kasohn}@ajou.ac.kr
[2] Department of Digital Media, Ajou University, Suwon, South Korea
jungju@ajou.ac.kr

Abstract. Advanced Video Coding (H.264/AVC) and its recent extension High Efficiency Video Coding (H.265/HEVC) are the current industrial video coding standards that have directional and planar intra-predictors. The intra-prediction scheme of both standards is dominated by directional predictors. Ordinary least squares (OLS) based adaptive predictor is superior around edge pixels without considering the direction of an edge explicitly. In OLS based predictor, each pixel in the local area contributes equal weight to estimate prediction parameters. We observed that this does not hold true due to noise and the existence of different correlations between the causal context of each neighboring pixel. To address this problem, we proposed an adaptive weighted least squares (AWLS) based predictor that assigns different weights to neighboring pixels to reflect its relative contribution. Experimental results show that the proposed method outperforms the OLS based intraprediction for directional images and marginal improvements are obtained for other tested images and videos.

Keywords: H.264/AVC · H.265/HEVC · Intra-prediction · Contextual pattern matching · Least squares based prediction

1 Introduction

In both H.264/AVC and H.265/HEVC video coding standards, each image of a video sequence is partitioned into macroblocks (MBs). Each MB is encoded based on the spatial extrapolation of samples from previously decoded image blocks directly above and to the left of the current block. In H.264/AVC [1], each MB is 16 × 16 samples. The intra-prediction of a MB can be done by three different block sizes, namely Intra 16 × 16, Intra 8 × 8 and Intra 4 × 4. For Intra 4 × 4, the MB is divided into sixteen 4 × 4 sub-blocks, and each block is predicted separately by using 8 directional modes to predict edge pixels and one planar mode to predict smooth blocks by using the average of reference pixels. For Intra 8 × 8, the MB is divided into four 8 × 8 sub-blocks, each sub-block is predicted by extending the concept of Intra 4 × 4 prediction modes. The Intra 16 × 16 has two directional modes (vertical and horizontal) and two non-directional

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_19

modes (DC and plane). A detailed review of H.264/AVC intra-prediction process can be found in [2].

However, in H.265/HEVC standard [3], each MB is 64 × 64 samples instead of 16 × 16 samples. The H.265/HEVC standard has 35 intraprediction modes compared with the 9 modes of H.264/AVC. Thirty three of them are directional modes and the remaining two modes are non-directional modes. Similar to H.264/AVC, the intra-prediction of H.265/HEVC can be done at different block sizes, namely Intra 64 × 64, Intra 32 × 32, Intra 16 × 16, Intra 8 × 8 and Intra 4 × 4. Both standards support 4 × 4 and 8 × 8 sub-block sizes with dominate number of directional predictors. A detailed review of H.265/HEVC intra-prediction process can be found in [4].

In this paper, we improved the intraprediction of H.264 standard for 4 × 4 and 8 × 8 blocks by using adaptive weighted least squares (AWLS) based predictor, and similar results can be obtained for the intraprediction of H.265/HEVC for 8 × 8 and 4 × 4 sub-blocks. In the reference software of H.264/AVC, a full search is used to examine each of the 3 block sizes and each of the 4 or 9 intra-prediction modes to find the smallest rate-distortion (RD). The same idea holds for H.265/HEVC. Since computation of RD is expensive, a number of fast algorithms have been proposed to select few blocks and few intra-prediction modes [5–8]. However, the reduction in computational complexity increases the bit rate and reduce the image quality (PSNR).

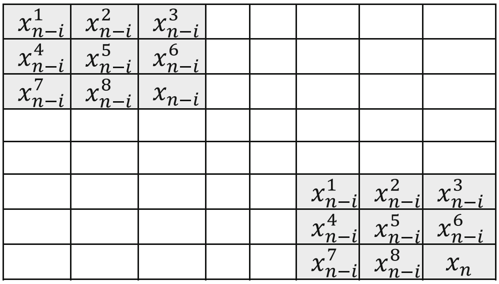
There is few related work which increases the PSNR value as well as reduces the bitrate of the intraprediction mode of H.264/AVC by using OLS based predictor. In [9], it is assumed that each pixel in the local area contributes equal weight to estimate prediction parameters. However, we identified three factors that can violate this assumption. We defer the detailed description of the identified factors to the end of Sect. 2 in order to avoid the interruption of the flow of ongoing presentation. To address these factors, we proposed AWLS based adaptive predictor. In this predictor, a different weight is assigned for each neighbor pixel to reflect its relative contribution. For each neighbor pixel, the weight is assigned according to its causal intensity similarity with the causal intensity of the current pixel being predicted. The intensity similarity is better than the spatial closeness because the performance of least squares method is superior when the local area is dominated by edge pixels [10, 11].

If the local area is less likely to be dominated by edge pixels, instead of computing new predictor coefficients, previously stored coefficients that are optimized for an edge can be used repeatedly until the scanning of pixels reaches the next edge pixel. Consequently, our proposed predictor switches between two predictors which are suitable for edge areas and smooth regions. Our results show that the proposed intra-prediction mode in H.264/AVC outperforms the OLS based intra-prediction for directional images, and marginal improvements are obtained for other tested images and video sequences.

The rest of the paper is organized as follows. In Sect. 2, backgrounds for least squares method and its weaknesses are described briefly. Section 3 presents a detailed description of the proposed weighting scheme. The experimental results are then presented in Sect. 4, which is followed by the conclusion in Sect. 5.

2 Weaknesses of Ordinary Least Squares Based Method

The aim of this paper is to improve the intraprediction of H.264/AVC and H.265/HEVC video coding standards by addressing the weaknesses of ordinary OLS method. Hence, we first describe this method and its weaknesses as a background. As shown in Fig. 1, let n denote the 1-D spatial position of a pixel when the image is scanned in a raster scan order while x_n represents the intensity value of the current pixel at position *n*.



x_{n-i}^1	x_{n-i}^2	x_{n-i}^3					
x_{n-i}^4	x_{n-i}^5	x_{n-i}^6					
x_{n-i}^7	x_{n-i}^8	x_{n-i}					
					x_{n-i}^1	x_{n-i}^2	x_{n-i}^3
					x_{n-i}^4	x_{n-i}^5	x_{n-i}^6
					x_{n-i}^7	x_{n-i}^8	x_n

Fig. 1. Causal predictors of the current pixel x_n and those of its ith previously encoded pixel x_{n-i} in a local training window

Let $C_n = \left[x_n^1, \ldots .x_n^j, \ldots, x_n^N\right]^T$ be an N × 1 column vector to represent the value of the N-nearest causal neighbors of the current pixel x_n being predicted. The entire image is a 2-D non-stationary random process for the purpose of adaptive image prediction. However, the image can be modeled as a piecewise 2-D stationary autoregressive random process as shown in Eq. (1).

$$x_n = e_n + \alpha^T c_n = e_n + \sum\nolimits_{j=1}^{N} \alpha_j x_n^j \tag{1}$$

where $\alpha = \left[\alpha_1, \alpha_2, \ldots, \alpha_N\right]^T$. The term e_n is a random error which is independent of the spatial location n and the intensity value of the image signal x_n. The random error accounts for fine details of image signal and measurement noise. Since image structures such as edges and surface textures are formed by spatially coherent continuous pixels, piecewise 2-D stationary autoregressive model of the image signal is a reasonable assumption. This implies that the set of parameters $\{\alpha_j\}$ of this model remain constant or near constant in a small locality even if they often vary significantly in different segments of the entire scene. Hence, using Markov model with order N, the predicted value $x_n^\wedge$ of the current pixel x_n is given by

$$x_n^\wedge = \left(\alpha^\wedge\right)^T C_n = \sum\nolimits_{j=1}^{N} \alpha_j^\wedge x_n^j \tag{2}$$

where $\alpha^\wedge = \left[\alpha_1^\wedge, \alpha_2^\wedge, \ldots, \alpha_N^\wedge\right]^T$ is the set of estimated parameter values. These estimators are not the same as the true values of the parameters $\alpha_1, \alpha_2 \ldots \alpha_N$ and can be adjusted to local image structures such as edges and textures by fitting the model to the pixels in a local training window (see Fig. 1) using least squares method. Let $Y = [x_{n-1} \ldots x_{n-i} \ldots x_{n-M}]^T$ be an M × 1 column vector to represent M previously encoded pixels in local area. Similar to C_n, let $C_{n-i} = \left[x^1_{n-i}, x^2_{n-1} \ldots x^N_{n-i}\right]$ be 1 × N row vector to represent the N-nearest causal neighbors of the i^{th} observed pixel x_{n-1} among the M pixels in the local window of pixel x_n as shown in Fig. 1. For example, in this figure, M = 64 and N = 8. Now, we have a system of linear equations

$$A\alpha = \left[C_{n-1} \ldots C_{n-i \ldots Cn-M}\right]^T \alpha = \mathrm{Y} \tag{3}$$

Equation (3) shows that A is an M × N matrix with its rows consisting of the N-neighbors of each of the M training pixels. Since there are more equations than unknowns, this over determined type of system cannot be solved exactly. Hence, it can be solved in least squares sense by estimating the unknown set of parameters $\{\alpha_j\}$ that minimize the sum of squared errors $e_{n\text{-}i}$ between the observed value and the estimated value as follows.

$$SSE_{min} = \sum\nolimits_{i=1}^{M} \left(x_{n-i} - x^{\wedge}_{n-i}\right)^2 = \sum\nolimits_{i=1}^{M} \left(e_{n-i}\right)^2. \tag{4}$$

The minimization process is achieved by multiplying both sides of Eq. (3) by A^T:

$$(A^T A)\alpha^{\wedge} = A^T Y \tag{5}$$

where $A^T A$ is an N × N square matrix and $A^T Y$ is an N × 1 column vector. This implies that Eq. (5) becomes a system of N equations in N unknowns. If A has full rank, the square matrix $A^T A$ has inverse, and the new equation will have a unique solution using the cholesky decomposition [14] as shown in Eq. (6). On the other hand, if A is defective, the matrix $A^T A$ fails to have inverse, and the equation has no unique solution. It is solved by singular value decomposition (SVD) technique [12].

$$\alpha^{\wedge} = (A^T A)^{-1}(A^T Y) \tag{6}$$

However, we identified three major problems that can reduce the ability of OLS model in estimating the predictor coefficients accurately. The first problem is ***data over-fitting problem***. The set of prediction coefficients $\alpha^{\wedge}$ fit the training window well, but they may not fit when they are applied to a new current pixel. This is because the new pixel is not part of a training window, and it may not have similar causal context with other pixels in the training window. The second problem is ***bi-modal distribution problem***. OLS is optimal when all the data in the training window are roughly from a single data source with mono-modal distribution [10]. However, this does not holds true because of two contrasting conditions. On one hand, the shape of objects can be arbitrarily in the real scene. On the other hand, in H.264/AVC and H.265/HEVC standards, the shape of the local training window is rectangular. This shape cannot be adaptive to the local image variations like edges and texture patterns for the purpose of computational simplicity. As a result of this, the training window can contain pixels from both smooth regions and edge areas and the histogram of the pixels in the local window can bi-modal distribution. Consequently, if the pixel x_n to be predicted is in a smooth area, a sample in an edge area will provide the same contribution as the sample in the same smooth area instead of contributing much less information.

The third problem is ***constant variance problem***. The basic assumption underlying OLS estimation technique is that each estimation error in Eq. (4) has constant variance across all individuals, $\mathrm{Var}\left(e_{n-i}\right) = \sigma^2_{n-i} = \sigma^2$. In other words, each data point in the training window provides equal contribution in Eq. (1). Since this technique treats all of the data equally, it will give more influence for less precisely measured points than they should have, and it will give too little influence for highly precise points. However, this assumption does not hold true due to noise and the existence of different correlations of the causal context of each pixel in the training window with that of the causal context of the current pixel. This implies that $\mathrm{Var}\left(e_{n-i}\right) = \sigma^2_{n-i} = \sigma^2 * k_{n-i}$ for some constant k_{n-i}. To minimize the above problems, the set of model parameters $\{\alpha_j\}$ can be estimated by giving proper amount of weight W_{n-i} to the contributions of each observed pixels x_{n-i}. This implies that Eq. (4) can be modified in to Eq. (7) to minimize a weighted sum of squared residuals.

$$SSE_{Min} = \sum\nolimits_{i=1}^{M} w_{n-i}\left(x_{n-i} - x^{\wedge}_{n-i}\right)^2 = \sum\nolimits_{i=1}^{M} w_{n-i}\left(e_{n-1}\right)^2 \tag{7}$$

This also implies that Eq. (3) can be modified into Eq. (8) by scaling both the observation vector Y and the matrix A by multiplying the i^{th} element of vector Y and the i^{th} row of matrix A by W_{n-i}.

$$(\mathrm{WA})\alpha^{\wedge} = (\mathrm{WY}) \Leftrightarrow \alpha^{\wedge} = \left(\mathrm{A^TWA}\right)^{-1}\left(\mathrm{A^TWY}\right) \tag{8}$$

where W is a diagonal matrix with diagonal entries $w_{n-1}, w_{n-2}, \ldots, w_{n-i}, \ldots, w_{n-M}$.

3 The Proposed Weighting Scheme

Generally, the weight w_{n-i} of each pixel x_{n-i} in the training window decreases as its spatial distance from the current pixel x_n increases because images vary slowly over space so that near pixels are more likely to have similar values. However, the slow spatial variations fail at edges. This implies that the intensity similarity is also important in addition to spatial closeness. As a result of this, bilateral image filtering technique assigns the weight w_{n-i} to each neighbor pixel using its spatial closeness as well as its intensity similarity to the current pixel [13].

Least squares based predictor is superior when the training window is dominated by edge pixels [10, 11]. This fact motivates us to use the intensity similarity component of bilateral filtering instead of spatial closeness to assign the relative contribution of each pixel in the training window. In our proposed weighting scheme, it is impossible to use directly the intensity similarity between the current pixel x_n and each of the pixels x_{n-i}

in the window because the value of the current pixel is not known. Hence, we incorporate the intensity similarity indirectly by using the similarity between the local texture of each pixel x_{n-i} and that of the current pixel x_n as follows. Two intensity pixel values v_1 and v_2 are said be a T-approximate values if $|v_2 - v_1| \leq T$, where T is an integer threshold value. When this is generalized to the current context C_n and its i^{th} previous context C_{n-i}, with order N, they are said to be T approximate pattern matching if

$$\left|x_n^1 - x_{n-i}^1\right| \leq T \bigwedge \left|x_n^2 - x_{n-i}^2\right| \leq T \bigwedge \cdots \left|x_n^N - x_{n-i}^N\right| \leq T \tag{9}$$

Equation (9) can be generalized to L_1-distance metric, and the two contexts are said to be a T-approximate pattern matching if $L_1\left(c_n, c_{n-i}\right) = \sum_{j=1}^{N} \left|x_n^j - x_{n-i}^j\right| \leq T.$

However, instead of using a single threshold T, two empirically determined threshold values T_1 and T_2 ($T_1 < T_2$) are used to classify the level of the similarity into low, medium and high levels. To reflect the quantitative effect of these levels on the estimation of the relative weight w_{n-i} of each pixel x_{n-i} (see Eq. (10)), the similarity value is multiplied by different integer constants ($1 \leq k_1 < k_2 < k_3$). The positive constant integer $c \geq 1$ is added because the two local textures may be exact replica of each other. In this paper, empirically we found that $c = 1, k_1 = 1, k_2 = 2$ and $k_3 = 4$ works very well. This shows that a pixel which has high pattern matching will have high contribution.

Originally, we classified the M-number of training patterns into low and high similarity classes using a single threshold value T. For each test image, the optimal value of T is obtained empirically from statistical observations. Initially, we set T = 0, indicating that the previous patterns are exact matches with the current pattern. We increase the value of T in steps of 1 by observing the change in compression bit-rate and PSNR values. We observed that different values of T produce different compression results and PSNR values for the same test image as well as for different types of test images. Hence, we found that the value of T has a great effect on the performance of the proposed algorithm.

$$w_{n-i} = \begin{cases} \dfrac{1}{k_1 * L_1\left(c_n, c_{n-i}\right) + c}, & if\, 0 \leq L_1\left(c_n, c_{n-i}\right) \leq T_1 \\ \dfrac{1}{k_2 * L_1\left(c_n, c_{n-i}\right) + c} & if\, T_1 < L_1\left(c_n, c_{n-i}\right) \leq T_2 \\ \dfrac{1}{k_2 * L_1\left(c_n, c_{n-i}\right) + c}, & if\, L_1\left(c_n, c_{n-i}\right) > T_2 \end{cases} \tag{10}$$

According to our statistical analysis, we observed that the optimal value of T has direct relationship with the entropy (complexity) of test images. Since the maximum value of the entropy of 8-bit gray scale images and video sequences is 8 bits per pixel, we found that the reasonable range of T is $2 \leq T \leq 8$ for all test images. When a small value of T is used for very complex type of images, the potential number of T-approximate patterns which belongs to low dissimilarity class will be very few while a large number of patterns belongs to the high dissimilarity class. Consequently, the majority of pixels will not provide enough information for the estimation of predictor parameters. On the Other hand, if a large value of T is used for low complex type of images, a very large number of patterns belongs to the low dissimilarity class. Hence, the relative contribution of each pattern is almost similar which leads to OLS method. Finally, we found that our proposed algorithm becomes efficient when we add a medium dissimilarity class as a compromise between the two extreme types of test image classes. Hence by partitioning the range of T ($2 \leq T \leq 8$) into two sub-ranges, the reasonable ranges of T_1 and T_2 are $2 \leq T_1 \leq 4$ and $5 \leq T_2 \leq 8$ respectively for most of test images. Through extensive experiments, we found that $T_1 = 4$ and $T_2 = 8$ works very well for all test images. Thus, we used these threshold values in our experimental results.

The square matrix A^TWA shown in Eq. (8) is either singular (has no inverse) or non-singular (has inverse).This is checked in the process of cholesky decomposition [11]. If the matrix is non-singular, this equation has unique solution. If the matrix is singular, static prediction coefficients are assigned based on their spatial closeness to the current pixel, or previously stored coefficients that are optimized for an edge can be used repeatedly until the scanning of pixels reaches the next edge pixel. We used this approach because of two reasons. First, if the square matrix in Eq. (8) is non-singular, the window is more likely to be dominated by edge pixels [11]. Second, if the matrix is singular, there will be less probability for the presence of dominant edge direction in the training window. To make the description of the proposed algorithm specific and clear, its pseudo code is presented as follows.

In **step d**, the coefficients are initialized by pre-fixed values (β) according to their spatial closeness to the current pixel. In step **e.ii**, each entry of vector Y, matrix A and vector C are normalized before computing W_i. This minimizes the negative impact of outliers in the training data. In **step e.vi**, each entry of vector WY and matrix WA are normalized before computing the predictor parameters ($\alpha\hat{}$). This minimizes the problem of producing large predictor weights. As a result of this, the noise may not be amplified by the predicted value. **Step e.vii** shows that the previously stored prediction coefficients will be updated when the matrix has inverse due to the dominant number of edge pixels; Otherwise, the previously stored coefficients (β) are used repeatedly until the next edge pixel is encountered. Finally, **step e.ix** shows that the size of training window (M), vector Y and matrix A increases.

```
for all $4 \times 4$ or $8 \times 8$ blocks per $16 \times 16$ current MB do
  (a) Find $M$ and $N$ as shown in Fig. 1
  (b) Initialize the normalization factor, $NF$ : $NF = 0.0$
  (c) Compute Matrix $A$, Vector $Y$ and $NF$
      for $i = 1$ to $M$ do
        $Y[i] = X_{n-i}$ ; $X_n$ is current pixel and $X_{n-i}$ is previous
      pixel
        if $NF < Y[i]$, then $NF = Y[i]$
            for $j = 1$ to $N$ do
              $A[i][j] = x_{n-i}^{j}$
                if $NF < A[i][i]$,then $NF = A[i][j]$
  (d) Assign fixed weights $\beta = \{\beta_1, \beta_2 \dots \beta_N\}$
  (e) for all $X_n$ in $4 \times 4$ or $8 \times 8$ current block do
    (i) Find the Predictors of the current pixel $x_n$
          for $j = 1$ to $N$ do
                $C[j] = x_n^j$
                if $NF < C[j]$, then $NF = C[j]$
    (ii) Normalize Vector $Y$, Matrix $A$ and Vector $C$
          for $i = 1$ to $M$ do
            $NY[i] = Y[i]/NF$
                for $j = 1$ to $N$ do
                  $NA[i][j] = A[i][j]/NF$
                  $NC[j] = C[j]/NF$
    (iii) Compute each weight $w_i$
          for $i = 1$ to $M$ do
            $diss = 0.0$ (Pattern Dissimilarity)
              for $j = 1$ to $N$ do
                    $temp = |A[i][j] - C[j]|$
                    $diss = diss + temp$
              if ($0 \leq diss \leq T_1$), then $w[i] = [1/[(k1 * diss) + c]$
              else if ($T1 < diss \leq T2$),then $w[i] = 1/[(k2 * diss) + c]$
              else $w[i] = 1/[(k3 * diss) + c]$
    (iv) $NF_{wy} = NF_{wy} = 0.0$ (For Normalization)
    (v)  Compute $WA$, $WY$, $NF_{wy}$ and $NF_{wy}$
          for $i = 1$ to $M$ do
              $WY[i] = W[i]Y[i]$
              if $NF_{wy} < WY[i]$ then
                  $NF_{wy} = WY[i]$
              for $j = 1$ to $N$ do
                $WA[i][j] = W[i]A[i][j]$
                  if $NF_{WA} < WA[i][j]$ then, $NF_{WA} = WA[i][j]$
    (vi) Normalize Matrix $WA$ and Vector $WY$
           for $i = 1$ to $M$ do
              $WY[i] = WY[i]/NF_{wy}$
              for j = 1 to N do
                 $WA[i][j] = WA[i][j]/NF_{WA}$
    (vii) Check the existence of Inverse of Matrix $WA$
            if $WA$ has inverse, then compute $\alpha\hat{}$ and update
            $\beta$: $\beta = \alpha\hat{}$
            Otherwise, do not update $\beta$
    (viii) Compute the Prediction of $x_n$
              $x_n\hat{} = 0.0$ (Initializing predicted Value)
              for $j = 1$ to $N$ do
                    if $x_n^j$ is available, then $x_n\hat{} = x_n\hat{} + \beta_j * x_n^j$
                    Otherwise, $x_n\hat{} = x_n\hat{} + \beta_j * x_n^{\hat{}j}$
    (ix) Update $M$, matrix $A$ and Vector $Y$
           $M = M + 1$;
           $Y[M] = x_n^{\hat{}}$
          for $j = 1$ to $N$ do
            $A[M][j] = x_n^j$
```

4 Experimental Results and Discussion

We compare the performance of the proposed method with the method in [9] for improving the intra-prediction of H.264. We used standard test images (512 × 512) such as Barbara, Lena, Monarch, Peppers, and Spoke. We used CIF-resolution (352 × 288) sequences with 300 frames (such as Foreman, Paris, Mother and Daughter, and Tempete). We also used 720*p* 50-resolution sequences with 504 frames (such as Mobcal, Parkrun and shields). These video sequences are available on line at http://media.xiph.org/video/derf/. For simulation, we used the H.264 reference software JM 10.2. All the experimental results were evaluated using only the luminance component. CABAC (context-adaptive binary arithmetic coding), in-loop de-blocking filter, and rate-distortion (RD) optimization are enabled. QP values of 22, 27, 32 and 37 were used.

Table 1. PSNR gain and rate reduction for different standard test images and video sequences

Image/Video sequence	PSNR gain (dB)	Rate reduction (%)
Barbara	0.22	3.34
Lena	0.11	2.56
Pepper	0.10	2.63
Monarch	0.15	2.02
Spoke	0.23	3.95
Foreman	0.13	2.22
Paris	0.19	2.03
Tempete	0.16	1.83
Mother and daughter	0.11	1.98
Mobcal	0.12	2.04
Parkrun	0.14	2.87
Shields	0.09	1.77

Table 2. Average percentage of chosen intraprediction modes for barbara test image

Mode	4 × 4 [9]	4 × 4 Proposed	8 × 8 [9]	8 × 8 Proposed
Vertical	9.63	8.84	9.19	7.64
Horizontal	6.18	5.92	6.63	6.03
DC	17.11	17.28	14.79	15.05
Diagonal down-left	5.56	5.44	5.48	5.32
Diagonal down-left	3.99	3.67	4.01	3.91
Vertical-right	6.23	5.79	6.13	5.90
Horizontal-down	3.87	3.71	3.66	3.53
Vertical-left	8.93	8.94	8.48	7.93
Horizontal-up/LSB	**38.49**	NA	**41.63**	NA
Horizontal-up/AWLS	NA	**40.41**	NA	**44.70**

Table 1 shows the PSNR gains and the bit-rate reduction of the proposed method over the work in [9] using the BJM metric [14]. ***Horizontal-up/OLS*** **and** ***Horizontal-up/AWLS*** indicates the replacement of the original horizontal-up predictor of H.264 by OLS and AWLS predictors. It can be seen that the proposed predictor offers better results for directional images, and it yields a marginal improvement for the other tested images and video sequences. For Barbara test image, Table 2 shows the frequency of each intra prediction mode (in percentage). It can be seen that the proposed method is always ranking first among the prediction modes. This shows that the proposed method is fundamental for the PSNR gain and rate reduction for the test images.

5 Conclusion

By addressing the problems of OLS method using AWLS method, the intra-prediction of H.264/AVC was improved. The relative contribution of each pixel in a localized training window is obtained by using the local texture similarity between the current pixel and each pixel in training window. This is an effective weighting technique because the superior performance of least squares method is dominated by edge pixels in the training window. When local window is not dominated by edge pixels, the covariance matrix has no inverse. In this case, instead of computing new prediction coefficients, static coefficients are assigned for each neighing predictor according to their spatial closeness to the current pixel or previously stored coefficients that are optimized for an edge can be used repeatedly until the scanning of pixels reaches the next edge pixel. As a result of this, our proposed predictor switches between two predictors for edge areas and slowly varying regions.

Our results show that the proposed scheme outperforms the ordinary least squares for directional images and marginal improvements are obtained for other tested images and video sequences. The limitation of the proposed method comes mainly from its use of the selected threshold values T_1 and T_2 as well as the constants k_1, k_2 and k_3 in Eq. (10). Statistically selected values may not achieve the best performance for images and videos which are not evaluated in this paper. This limitation remains future work.

Acknowledgement. This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education [2016R1D1A1B03933875].

References

1. Wiegand T, Sullivan GJ, Bjontegaard G, Luthra A (2003) Overview of the H.264/AVC video coding standard. IEEE Trans Circ Syst Video Technol 13(7):560–576
2. Richardson I (2010) The H.264 advanced video compression standard. Wiley, New York
3. Sullivan GJ, Ohm J-R, Han W-J, Weigand T (2012) Overview of the high efficiency video coding (HEVC) standard. IEEE Trans Circ Syst Video Technol 22(12):1649–1668
4. Lainema J, Bossen F, Han W, Min J, Ugur K (2012) Intra coding of the HEVC standard. IEEE Trans Circ Syst Video Technol 22(12):1792–1801

5. Pan F, Lin X, Rahardja S, Lim KP, Li ZG, Wu D, Wu S (2005) Fast mode decision algorithm for intraprediction in H.264/AVC video coding. IEEE Trans Circ Syst Video Technol 15(7): 813–822
6. Wang CC, Chen TS, Tung CW (2006) Fast intra mode decision in H.264 using interblock correlation. In: IEEE international conference on image processing, pp 1345–1348
7. Lin YK, Chang TS (2005) Fast block type decision algorithm for intra prediction in H.264 FRext. In: IEEE International conference on image processing, September 2005, pp I-585–I-588
8. Silva T, Agostini L, Cruz LS (2012) Fast HEVC intraprediction mode decision based on EDGE direction information. In: Proceedings of the 20th EUSIPCO, pp 1214–1218
9. Garcia D, Queiroz R (2010) Least-squares directional intraprediction in H.264/AVC. IEEE Signal Process Lett 17(10):831–834
10. Li X, Orchard M (2001) Edge-directed prediction for lossless compression of natural images. IEEE Trans Image Process 10(6):813–817
11. Kau LJ, Lin YP (2007) Least-squares based switching structure for lossless image coding. IEEE Trans Circ Syst I 54(7):1529–1541
12. Press W, Teukolsky S, Vetterling W, Flannery B (1992) Numerical recipes in C: the art of scientific computing, 4th edn. Cambridge University Press, New York
13. Tomasi C, Manduchi R (1998) Bilateral filtering for gray and color images. In: Proceedings of the 6th international conference on computer vision, pp 839–846
14. Bjontegaard G (2001) Calculation of average PSNR differences between RD-curves. VCEG contributions, VCEG-M33

A Preference-Based Application Framework for Resource-Bounded Context-Aware Agents

Ijaz Uddin[1] and Abdur Rakib[1,2](✉)

[1] School of Computer Science, The University of Nottingham Malaysia Campus,
Semenyih, Malaysia
{khyx4iui,Abdur.Rakib}@nottingham.edu.my

[2] Department of Computer Science and Creative Technologies, The University of the West of England, Bristol, United Kingdom
Rakib.Abdur@uwe.ac.uk

Abstract. Context-awareness is an essential component of mobile and pervasive computing. It refers to the concept that an application understands its context, reason about its current situation, and provide relevant information and/or services to the users. One of the main challenges of context-aware distributed mobile computing is the dynamic adaptation to changes in the resource-bounded operating environment with user preferences. For example, a depersonalized context-aware application may exhibit behavior that is not anticipated by its user in a given situation. In this paper, we present a personalized preference model for resource-bounded context-aware applications, which provides support for the development and execution of context-aware applications using a declarative language. We implement a simple example system that demonstrates the effectiveness of the approach in a real-world scenario.

Keywords: Context-aware agents · Rule-based reasoning · Android SDK · Smart phone · User Preference

1 Introduction

In distributed mobile and pervasive computing research, context-awareness has emerged as an effective design and implementation approach for building adaptive smart-space applications. These applications rely on the use of current contextual information, and their dynamic adaptation to changes in the operating environment provides a high level of automation with very minimal or no user intervention. In developing smart-space context-aware applications, smart phones and wireless sensor technology play an important role. Smart phones have a variety of embedded sensors that can be used to automate data collection and provide a platform to infer rich contextual data about users, including location, time, and environmental condition. This is known as customized information according to the specific context. To be more precise, these sensors can be used to gather the contextual information of a user or to manipulate the context. Different notions of context have been studied across various fields of

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_20

computer science and various physical and conceptual environmental aspects can be included in the notion of context [1]. Among others, Dey et al. [2] define a context-aware system as a system which uses context to provide relevant information and/or services to its user based on the user's tasks.

There has been considerable work in the context-aware systems literature on context modeling and reasoning approach in general (see for example, [3–6]) and on context-aware reasoning based on user preferences in particular (see for example [7–10]). Much of this work aims at how can semantic (ontology-based) and/or other techniques be utilized for context-modeling, knowledge sharing and reasoning about context for pervasive computing systems. However, well developed theoretical foundations considering their resource-boundedness features are still lacking. The resources include the time, memory, and communication bandwidth required by the context-aware devices or agents to achieve a goal. In recent work [11–13], Rakib et al. have developed formal logical frameworks showing how context-aware systems can be modelled as resource-bounded multi-agent reasoning agents. In this paper, we extend and enhance our previous work [12,13] by customizing user preferences to enable the personalization of resource-bounded context-aware applications.

The rest of the paper is organized as follows. In Sect. 2, we present our context-aware system modeling framework. In Sect. 3, we present the design and implementation components of a context-aware user preference framework, which extends the existing framework [13] to allow personalized services. In Sect. 4, we present a system specification, sensor data acquisition, and discuss the results of an experiment, and conclude in Sect. 5.

2 Context-Aware System Modelling Framework

We adopt the model of multi-agent context-aware rule-based systems developed by [12]. In rule-based techniques, a context-aware system composed of a set of rule-based agents, contexts are represented using first order terms and firing of rules that infer new contexts determine context changes and represent overall behaviour of the system. In order to model contexts and rules we use ontological approach. A rule has the following format:

$$m : P_1, P_2, \ldots, P_n \rightarrow P_0 \;\; : F : CS \;\text{ where } n \geq 0.$$

where m is the rule priority. Each P_i is an atomic formula of the form $p(t_1, t_2)$, $Ask(i, j,\, p(t_1, t_2))$ or $Tell(i, j, p(t_1, t_2))$, where i and j $(i \neq j)$ represent agents, p is a predicate symbol and the t_k are terms. Where Ask and $Tell$ are special atoms used for communication between the agents [12]. Each term is either a constant symbol or a variable. Every variable occurring in a rule is universally quantified, and its scope is the clause in which the variable occurs. Every variable appearing in the head must also appear in the body of a rule. The "→" is read as *if* and "," as *and*. The atom P_0 is called consequent (or head) of the rule and the conjunction $P_1, P_2, \ldots, P_n$ is the body of the rule. If $n = 0$, then the body is equivalent to TRUE and is called a fact otherwise it is a rule. The flag F, a placeholder, associated with every rule is used to specify the type of the rule. For instance, the character 'G' is used to represent a rule containing a Goal statement, which indicates that a certain rule execution results in goal achievement. The character 'C' represents the communication rules, which can trigger a communication between agents (devices). The character 'D' represents the deduction rules. The indicator CS says which set the rule belongs to, and is mainly used for the preferences set generation, which is explained in more detail in the following section.

In our framework, we consider systems having constraint on various resources, namely time, memory, and communication. This is because many context-aware systems often run on tiny resource-bounded devices, including PDAs, smart phones, GPS system, and wireless sensor nodes. These devices usually operate under strict resource constraints, e.g., battery energy level, memory, processor, and quality of wireless connection. The logical framework developed in [12] allows us to describe a set of context-aware non-monotonic rule-based reasoning agents with bounds on computational (time and memory) and communication resources. In [13], we extended the theoretical work [12] by implementing the ontology and logic based framework using the Google Android SDK and smart phones. In this paper, we extend and enhance our previous work [12,13] by customizing user preferences to enable the personalization of resource-bounded context-aware applications. We also discuss further experimental progress of an example system. As in our previous work [13], we lacked some sensors which were then replaced by a simulated device. In the current setting, we use actual external sensors and successfully integrated them into the framework to generate experimental results considering a real world scenario.

3 Preference in Context-Aware Agents

In this section, we discuss the extended framework that allows defining components to provide personalized services. In order to implement user preferences, we add an extra preference manager layer and keep the original working inference engine [13] intact. The main idea of user preference is to select a subset of rules based on preferences, and the inference engine, instead of going through all the rules, will only process selected rules. The whole process is composed of different steps and modules which are explained in the following sections. The preference manager layer consists of Preference Set Generator(PSG), Context Monitor(CM), Context Set(CS), and Context of Interest (COI) provided by the user beforehand. Figure 1 shows how these components are related to each other.

3.1 Context Set

The context set(CS) component is basically a column added to the rule base. A literal in this column against a rule works as indicator for that particular rule. It indicates if a rule belongs to a particular set of rules that may infer a specific contextual information, e.g., *Person(?p), OfficeRoom(?o), hasLocation(?p, ?o)*

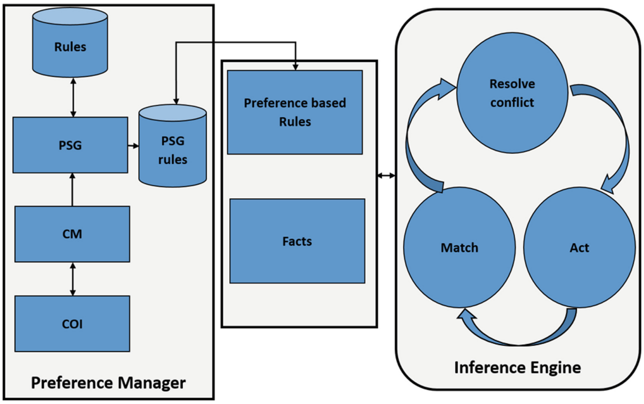


Fig. 1. Preference generation overview

→ *inOffice(?p, ?o)* with indicator "L" in CS can be attributed towards the contextual information about user's current location, which represents that the rule belongs to a group of rules that are part of location rules. The CS may contain multiple indicators, for example, if user location and his blood pressure mentioned in the same rule then CS can indicate both contexts defined with two different literals. The reason for such indication comes handy when the user preference is required. These all CS indicators can easily indicate the contexts included in a rule. For example, a user may want preference based on location only. So the preference set will add all the rules which CS indicates the location. It is pertinent to mention that any rule that does not have any CS indicator is a general rule, represented by "-" in the context set, and will be added to every sub set that is created for a preference set.

3.2 Context Monitor

The context monitor (CM) component holds the Context of Interests (COI) of a user, i.e., it holds the values provided by the user. Context monitor after reading the values passes them to the Preference Set Generator(PSG). The PSG defines a sub set of rules based on the user preferences, called preference set. This subset is then passed to the inference engine for processing. Context monitor actively monitors the contexts of interests. Any change in the context is forwarded to the PSG to derive a new set of rules to be processed according to the changed preference(s).

3.3 Preference Set Generator

The preference set generator(PSG) is the main part which gives the framework an ability to provide personalized services. Since we have added CS to the rules for indication purpose, we need a layer that can work as intermediary between the user and the rule-base. This layer provides a sub set of rules which are personalized set of rules for a current context of the user. The PSG receives instructions from the CM to derive a sub set of personalized rules. The rule base of an agent consists of a variety of rules; some rules may never get a chance to execute while some others may be actively executed. In order to generate a sub-set of the rules, the PSG has to consider the contexts that are of interest to the user.

3.4 Working Mechanism

In this section, we show how these different parts work together to give preferences to the user. We have one main repository, where all the rules are stored. As the process starts, the COI is provided by the user and the values are retrieved by the CM component. The CM forwards the values to the PSG. The PSG further communicates with the rule base and picks only those rules that are of interest to the user based on the values specified in the COI. The PSG makes use of the CS to fetch the desired rules. This CS is specified by the user as COI. When the PSG rules are ready, these rules are provided to be used as the knowledge base for further processing. Comparing to our previous work we can see, that the whole system still works as described in earlier work [13]. However, the rules in the memory are replaced with only the preference based rules. Practically addition of preference layer is the major change, which reduces the overall burden from main inference engine and making it more efficient in terms of reducing rules.

4 System Specification and Sensor Data Acquisition

We have implemented the framework using both embedded sensors i.e., GPS and external sensors which are blood pressure monitor and heart rate monitor. In our experiment we have used three different agents, namely Patient care device (an Android powered smart phone), Care giver (an Android powered smart phone), and Blood pressure and heart rate monitor (BP device) (a Bluetooth-enabled device). The patient care device uses the low-level contexts from the BP device and infers high-level contexts using the set of rules in its knowledge-base. If a patient's condition is critical or an emergency scenario is detected, it interacts with the care giver agent. Care giver can be a registered doctor or nurse. The communication between the care giver and the patient takes place via SMS messages, while blood pressure and heart rate values are sent via Bluetooth to the patient care device. Figure 2 shows the patient monitoring device that we have used in our experimental setup. In our experimental model we classify different categories of blood pressure and heart rate based on the *Blood Pressure UK* and *New Health Advisor* data charts[1]. Based on the blood pressure and heart

[1] http://www.bloodpressureuk.org/BloodPressureandyou/Thebasics/Bloodpressurechart
http://www.newhealthadvisor.com/Normal-Heart-Rate-Chart.html.

rate values, we have encoded a set of rules which are used to design and program our context-aware agents. However, due to space constraints, we have listed only few selected rules in Table 1, which are used by the patient care device.

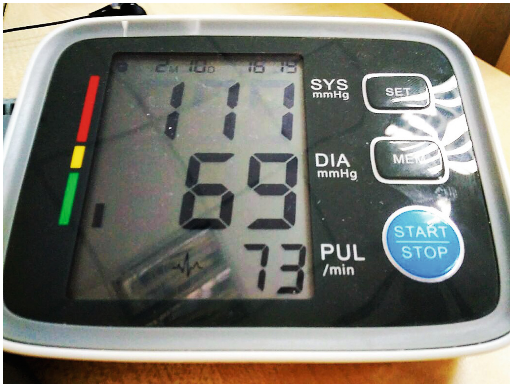


Fig. 2. Blood pressure and heart rate monitoring device

4.1 Sensor Communication

In this section, we discuss two different mechanisms of the sensor data acquisition that are used in acquiring raw data from external and embedded sensors.

External sensor: The Blood pressure and heart rate monitoring device uses the Bluetooth low energy(BLE) communication settings. The BLE is relatively new technology that is very energy efficient compared to normal Bluetooth operation [14]. The communication mechanism and blood pressure measurement procedure follow certain steps, which are discussed here. As for the prototype design, a patient has to attach the strap to the upper arm and turn the device switch ON. When the device is ON, it starts sending some signals. The structure of the signal is in the following format: [**0xFD, 0xFD, 0xXX, 0xXX, 0x0D, 0x0A**] and is adopted for all the operations that the device carry throughout the measurements. The **XX** are replaced with other values such as errors or results and vary in size. When the sensor is turned ON, it sends the following signal or we can say it broadcasts the notification of its availability by sending [**0xFD, 0xFD, 0xA5, 0x0D, 0x0A**] every half a second. Any device nearby if replies back by sending the [**0xFD, 0xFD, 0xFA, 0x05, 0x0D, 0x0A**] signal, the connection will be established and the BP monitor will start the measurement. Once the measurement is taken, the BP monitor sends the result to the connected device in the following format: [**0xFD, 0xFD, 0xFC, Systolic value, Diastolic value, Heart rate,0x0D, 0x0A**]. This format indicates that the results are accurately taken and sent to the connected smartphone device. In case of an error, which may arise due to very low heart rate or inflation taking too much time or the low battery message, the BP monitor sends its corresponding signals to the connected smart phone device. These values are written to a file and saved in the smartphone's memory. Once written to the memory, the application program can access and read the contents of the file for further processing. In our case, Patient care device receives three values from the blood pressure device and one from its embedded GPS as location.

Embedded sensor: We have also simulated the emergency case, where the location is acquired using Google Play services API. The API is recommended by Google for accurate and faster location retrieval and also consumes less energy while acquiring the location. A fine grained location can also be determined using GPS, WiFi, and Cellular network. It can also update the location on a preset interval along with the distance. For example, if a location is acquired at point A, it will recalculate the location after the preset interval or if a user moves by a preset distance e.g., say 10 meters. In that case we always get an accurate location for a user. Once the location is acquired we further make use of the reverse geo-coding technique to retrieve a user readable format from the longitude and latitude that we capture from the sensor. The end result is an accurate human readable address. These sensed values or low-level contexts make no sense at all unless they are translated into meaningful high-level contexts. For that reason, we have also followed standard blood pressure and heart rate measurement guidelines, and encoded the expert knowledge into a set of Horn-clause rules.

4.2 Experimental Results

In our previous work [13], implementation of the agents' inference engine and experimental results of a depersonalized context-aware application scenario have been provided. However, presenting detailed experimental results are out of scope of this paper. Nevertheless, we explain how the external sensor sends the raw data and how they are processed, and shows only the main rules that are most likely to be fired in the case scenario shown in Fig. 2. In Fig. 2, the blood pressure and heart rate monitoring device shows three values. The first value on top is the systolic value, the middle one is diastolic value, and the last one is the heart beats per minute. These three values appear on the screen and are forwarded to the patient care agent. Patient care agent has a variety of rules besides those presented in Table 1. For the experimental purpose, we assumed the data to be tested for the generic values of a healthy adult male. In this scenario the systolic value is between 90 and 120 and the diastolic value is between 60 and 80, which ultimately will trigger the rule resulting in the normal blood pressure category, i.e., the following rule:

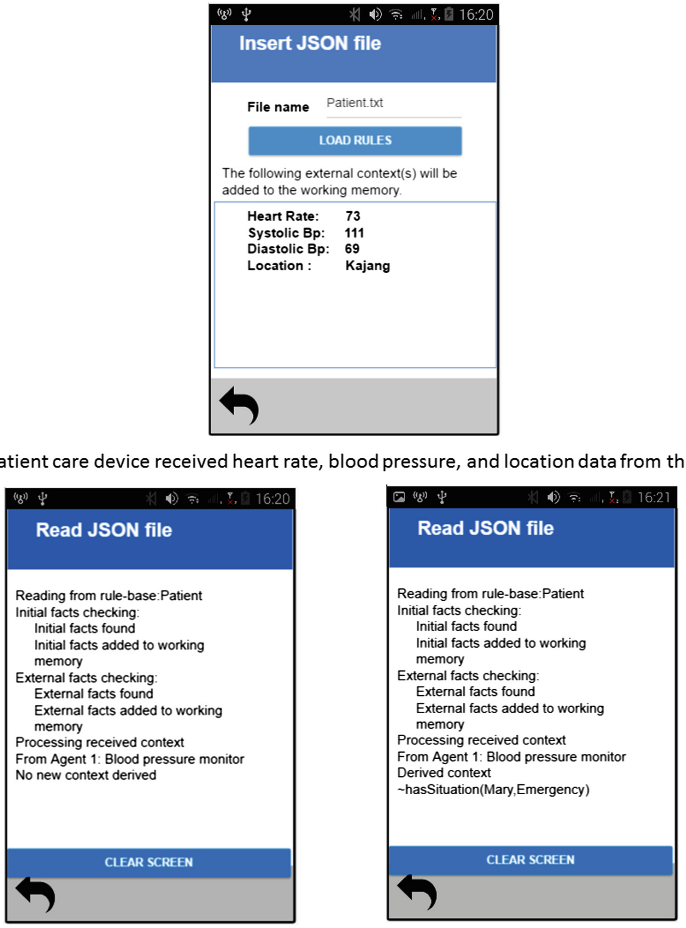
Person(?p),hasSystolicBloodPressure(?p,?sbp), hasDiastolicBloodPressure (?p, ?dbp),greaterThan(?sbp,90), greaterthan(?dbp,60), lessThan(?sbp,120), lessThan(?dbp,80)→hasBPCategory(?p,Normal)

Similarly the heart rate, as observed, falls within normal range and will trigger the following rule: *Person(?p),hasHeartRate(?p, ?hrt),greaterThan(?hrt,70), lessThan(?hrt,75)→ hasHRCategory(?p,Average)*

Table 1. Blood pressure and heart rate rules

Category	m	Corresponding rule	F	CS
Blood pressure category rules				
Low BP	1	Person(?p), hasSystolicBloodPressure(?p,?sbp), hasDiastolicBlood Pressure(?p, ?Dbp), lessThan(?sbp, '90), lessThan(?dbp,60) ⟶ hasBPCategory(?p,LowBP)	D	-
Normal	1	Person(?p),hasSystolicBloodPressure(?p,?sbp), hasDiastolicBlood Pressure(?p, ?dbp), greaterThan(?sbp,90), greaterthan(?dbp,60), lessThan(?sbp,120), less- Than(?dbp,80) ⟶ hasBPCategory (?p,Normal)	D	-
Pre high	1	Person(?p), hasSystolicBloodPressure(?p,?sbp), hasDiastolicBlood Pressure(?p, ?dbp),greaterThan(?sbp,120), greaterThan(?dbp,80), lessThan(?sbp,140), less- Than(?dbp,90)⟶ hasBPCategory(?p,PreHigh)	D	-
High	1	Person(?p), hasSystolicBloodPressure(?p,?sbp), hasDiastolicBlood Pressure(?p, ?dbp), greaterThan(?sbp,140), greaterThan(?dbp,90) ⟶ hasBPCategory(?p HighBP)	D	-
Heart rate category rules				
Athlete	1	Person(?p), hasHeartRate(?p,?hrt), greaterThan(?hrt,48), lessThan(?hrt,55) ⟶ hasHRCategory(?p, Athlete)	D	-
Excellent	1	Person(?p), hasHeartRate(?p,?hrt), greaterThan(?hrt,54), lessThan(?hrt,62) ⟶ hasHRCategory(?p,Excellent)	D	-
Good	1	Person(?p), hasHeartRate(?p,?hrt), greaterThan(?hrt,61), lessThan(?hrt,66) ⟶ hasHRCategory(?p,Good)	D	-
Above Average	1	Person(?p), hasHeartRate(?p,?hrt), greaterThan(?hrt,65), lessThan(?hrt,71) ⟶ hasHRCategory(?p,AboveAverage)	D	-
Average	1	Person(?p), hasHeartRate(?p,?hrt), greaterThan(?hrt,70), lessThan(?hrt,75) ⟶ hasHRCategory(?p,Average)	D	-
Below Average	1	Person(?p), hasHeartRate(?p,?hrt), greaterThan(?hrt,74),lessThan(?hrt,82) ⟶ hasHRCategory(?p,BelowAverage)	D	-
Poor	1	Person(?p), hasHeartRate(?p,?hrt), greaterThan(?hrt,81) ⟶ hasHRCategory(?p,Poor)	D	-
Some example rules to derive different situations				
Emergency	2	Patient(?p), hasBPCategory(?p,HighBP), hasHRCategory(?p,Poor) → hasSituation (?p,Emergency)	D	H
Emergency	2	Patient(?p), hasBPCategory(?p,PreHigh), hasHRCategory(?p,Poor) → hasSituation (?p,Emergency)	D	H
Emergency	2	Patient(?p),hasBPCategory(?p,Normal), hasHRCategory(?p,Poor) → hasSituation (?p,Emergency)	D	N
Emergency	2	Patient(?p),hasBPCategory(?p,LowBp), hasHRCategory(?p,Poor) →hasSituation (?p,Emergency)	D	L
Non emergency	1	Patient(?p),hasBPCategory(?p,Normal), hasHRCategory(?p,Average) → ∼hasSituation (?p,Emergency)	D	N
Non emergency	1	Patient(?p),hasBPCategory(?p,Normal), hasHRCategory(?p,AboveAverage) → ∼hasSituation (?p,Emergency)	D	N
Non emergency	1	Patient(?p),hasBPCategory(?p,Normal), hasHRCategory(?p,Good) → ∼hasSituation (?p,Emergency)	D	N

Hence both the blood pressure and heart rate categories fall in the normal range, which are the deciding factor in this case and the patient care agent will not interact with the care giver agent. Moreover, trying with different set of situations can produce different results, including false alarm. For example, blood pressure and heart rate readings could be high if they are measured while a person climbs stairs, and as a result the patient care agent may interact with a caregiver, which can be considered as false alarm. However, more sophisticated rules can be added to deduce about the condition of the user based on the variable such as if the person is running, climbing etc. Furthermore, preferences can be used to personalize, where the results matter more important to the user. The preference set generated is based on the COI values that reside in the CS. For example, when we provide COI as rules which deal with high blood pressure only, indicated with the symbol H in the CS, the system will not produce any new context based on the values given above, shown in Fig. 3 (b). This is because when preference is applied only rules which are of type H will be added to the preference set along with other general rules if any, and will ignore other rules having preference types L and N. However, when the COI is changed from H to H,N which includes those rules that are dealing with both high and normal blood pressure, we get different results. In this case, when we run the application again for the same input, we see that the system triggers the rule which infers ∼*hasSituation(Mary, Emergency)* as shown in Fig. 3 (c).



Patient.txt

The following external context(s) will be added to the working memory.

Heart Rate: 73
Systolic Bp: 111
Diastolic Bp: 69
Location : Kajang

(a) Patient care device received heart rate, blood pressure, and location data from the BP device

```
Reading from rule-base:Patient
Initial facts checking:
    Initial facts found
    Initial facts added to working
    memory
External facts checking:
    External facts found
    External facts added to working
    memory
Processing received context
From Agent 1: Blood pressure monitor
No new context derived
```

(b) Reasoning output with preference H

```
Reading from rule-base:Patient
Initial facts checking:
    Initial facts found
    Initial facts added to working
    memory
External facts checking:
    External facts found
    External facts added to working
    memory
Processing received context
From Agent 1: Blood pressure monitor
Derived context
~hasSituation(Mary,Emergency)
```

(c) Reasoning output with preferences H, N

Fig. 3. Sensed data and rule execution results using preference

5 Conclusion and Future Work

In this paper, we discussed and presented a preference model to personalization of resource-bounded context-aware applications. We also discussed the updated progress of the system and integration of the external sensors to the framework developed in our previous work in [13]. In future work, we would like to further narrow down the concept of preference so that it can be applied to the values of the contexts. In that case, a user will have control over the preference selection within the context rather than on rules.

References

1. Lieberman H, Selker T (2000) Out of context: computer systems that adapt to, and learn from, context. IBM Syst J 39(3–4):617–632
2. Dey AK (2001) Understanding and using context. Pers Ubiquit Comput 5(1):4–7. doi:10.1007/s007790170019
3. Ranganathan A, Campbell RH (2003) An infrastructure for context-awareness based on first order logic. Pers Ubiquit Comput 7(6):353–364
4. Korpipaa P, Mantyjarvi J, Kela J, Keranen H, Malm EJ (2003) Managing context information in mobile devices. IEEE Pervasive Comput 02(3):42–51
5. Wang XH, Zhang DQ, Gu T, Pung HK (2004) Ontology based context modeling and reasoning using OWL. In: PERCOMW 04, Washington, DC, USA. IEEE Computer Society, pp 18–22
6. Kofod-Petersen A, Mikalsen M (2005) Representing and reasoning about context in a mobile environment. Rev d'Intell Artif 19(3):479–498
7. Barkhuus L, Dey, A (2003) Is context-aware computing taking control away from the user? Three levels of interactivity examined. In: UbiComp 2003: ubiquitous computing. Lecture notes in computer science, vol 2864, pp 149–156
8. Stefanidis K, Pitoura E, Vassiliadis P (2006) Modeling and storing context-aware preferences. In: Proceedings of the 10th east european conference on advances in databases and information systems. Lecture notes in computer science, vol 4152, pp 124–140
9. Coutand O (2008) A framework for contextual personalised applications. Dissertation thesis. The University of Kassel. ISBN: 9783899587746
10. Hong J, Suh E, Kim J, Kim S (2009) Context-aware system for proactive personalized service based on context history. Expert Syst. Appl. 36(4):7448–7457
11. Rakib A, Haque HMU, Faruqui R (2014) A temporal description logic for resource-bounded rule-based context-aware agents. In: Context-aware systems and applications. Lecture notes of the institute for computer sciences, Social informatics and telecommunications engineering, vol 128. Springer, pp. 3–14
12. Rakib A, Haque HMU (2014) A logic for context-aware non-monotonic reasoning agents. In: Human-inspired computing and its applications. Lecture notes in computer science, vol 8856. Springer, pp 453–471
13. Uddin I, Rakib A, Haque HMU (2017) A framework for implementing formally verified resource-bounded smart space systems. Mobile Netw Appl 22:1–16
14. Aguilar S, Vidal R, Gomez C (2017) Opportunistic sensor data collection with bluetooth low energy. Sensors 17(1):159

B-Tree Index Layer for Multi-channel Flash Memory

Rize Jin(✉)

Department of Software, Ajou University, Worldcupro 206, Yeongtong-gu,
Suwon 16499, South Korea
rizejin@ajou.ac.kr

Abstract. Most of the recent studies on flash-aware index design focused mainly on the single channel flash memory where parallel processing of the B-tree index is not a consideration. This paper discusses efficient indexing on multi-channel flash storage and proposes a B-tree storage scheme, which not only exploits internal parallelisms of the underlying storage structure and also adjusts the node storage dynamically based on the run-time workload. Experimental results show that the proposed B-tree index layer is capable of speed up index operations near linearly while increasing degrees of internal parallelisms of flash memory.

1 Introduction

While flash memory [1] shows superiority in terms of high random access speeds, low power consumption, shock/temperature resistance, small size, etc., compared to traditional storage devices, it possesses some limitations, such as erase-before-write, discrepancy on read/write and erase units, relatively slow erase speed, and a limited number of write/erase cycles. These make most publicly available software optimizations work less well for flash based storage [2, 4]. For example, the small and frequent random accesses issued by the upper layer (e.g., buffer manager, B-tree index) can cause a substantial amount of block erases and garbage collections to flash memory. Most existing B-tree implementation assume uniform IO costs of underlying storage medium, which is not the case for flash memory, especially multi-channel flash memory which has multiple level of internal parallelism. In this paper, we exploit an adoptive B-tee index later which adjusts the node storage dynamically based on the run-time workload on a multiple channel flash memory.

Contributions of this paper are summarized as follows:

1. A channel-level block group based in-block logging approach to efficiently store and manage the B-tree on flash memory.
2. Dynamic grouping and round robin techniques to efficiently manage the number of log pages within a flash block. This contributes to reduce the number of flash erases that caused by frequently changed B-tree structure.
3. An in-block logging scheme for the index such that an index file and its log records are co-located in the same physical flash block.

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_21

4. A lazy-split LRU buffer replacement policy [6] to further classify dirty pages into clean and dirty parts. This contributes to improve the buffer hit ratio and space utilization of previous clean-first based replacement algorithms.
5. Extensive experiments to demonstrate that the proposed storage scheme markedly outperforms the related work under heavy-update workload in terms of execution time, buffer hit ratio.

The remainder of this paper is organized as follow. Section 2 discusses the background and motivation of this work. Section 3 describes the proposed solution. 4 reports the performance evaluation. Finally, the conclusion is made in Sect. 5.

2 Background

This section covers briefly some of the main characteristics of a NAND-type flash memory, such as physical structure, operations and software components. Problems of storing a B-tree in flash memory based systems are also presented.

2.1 Structure of Multi-channel Flash Devices

Flash memory is organized into a certain number of blocks, and each block consists of multiple pages. The internals of a flash memory differ in almost every aspect from a hard disk drive. Unlike traditional storages perform equalizing speed of read and write, the write and erase operations on flash memory take much longer time than a read operation. Reading and writing are performed on a unit of a page, and erasure can only be performed on the unit of a block (Table 1).

Table 1. Key features of the Samsung 2nd Generation V-NAND [1]

Features	V2 128 Gb 2bit
Page size	(16 K + 1536) Bytes
Number of pages per block	384 Pages
Block size	(4 M + 384 K) Bytes
Number of planes	2
Page read per plane	(16 K + 1536) Bytes
Random read time	35 µs
Page program per plane	2 x (16 K + 1536) Bytes1
Program time	0.39 ms
Date transfer rate	667 Mbps
Block erase time	4 ms

Note 1. Dual page program

In order to overcome these electronic limitations, flash memory is supported by an intermediate software layer called Flash Translation Layer (FTL) [2, 3]. FTL helps applications above the file system to regard flash memory as a magnetic disk-like block device. Considering the difference between the basis units of write (a page) and erase (a block, consists of a number of pages), updating the outdate data on flash memory immediately is not worthwhile. FTL adopts the wear-leveling and techniques out-of-place update [3]. The wear-leveling technique attempts to work around the erase cycle limitation by arranging data such that erasures are distributed evenly across the blocks. The out-of-place update technique writes the new contents of a page to another location instead of updating the page itself. As a result, the number of overwrites can be reduced. But, frequent out-of-place update can consume the blocks and run out them. When free blocks are not enough, it has to erase the blocks who got the outdate data inside and also merge the valid data into new blocks.

Figure 1 shows a multi-channel NAND flash memory use case, SSD drive, and its main components. Flash memory packages are organized into groups, over multiple channels.

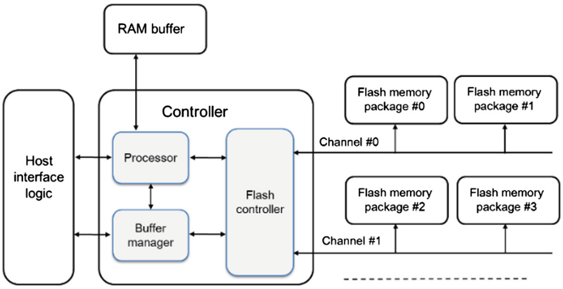


Fig. 1. Architecture of a multi-channel NAND flash memory use case.

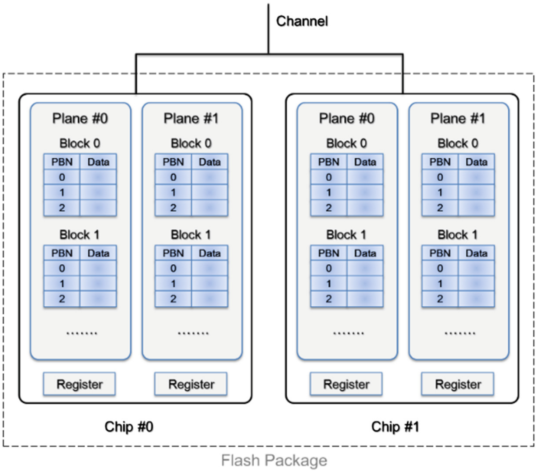


Fig. 2. The internals of a flash package

Figure 2 shows the internals of a flash package, which is also organized as a hierarchical structure. Therefore, the internal levels are channel, package, chip, plane, block, and page. Commands come from the applications through the host interface. The processor in the controller takes the commands and pass them to the flash controller. SSD drives usually contains an embedded RAM buffer, which is generally used for caching user data and mapping information. Each plane also contains a small RAM buffer, called a register, which is used for plane-level operations.

2.2 Limited I/O Bus Bandwidth

Due to physical limitations, single bus design cannot meet the bandwidth requirement of hot applications. One way to increase I/O throughput is to design these drives in such a way that multiple flash packages can be parallelized or interleaved. By utilizing all the levels of internal parallelism, multiple blocks can be accessed simultaneously in a multi-channel flash device. Different levels of internal parallelism are briefly described blow.

Channel-level parallelism: As shown in Fig. 1, the flash packages can be accessed through channels which operates independently and in parallel.

Package-level parallelism: Each channel is shared by multiple packages. Those packages can be accessed independently or simultaneously by interleaving the commands across them.

Chip-level parallelism: A package contains two or more chips, which can be accessed independently or in parallel.

Plane-level parallelism: A chip contains two or more planes. The same operation (read, write or erase) can be run simultaneously on multiple planes inside a chip.

3 Group-Based in-Block Logging Scheme

For the peculiar characteristics of flash memory, the storage scheme for B-tree index should has different properties in terms of data structure and operations. However, all of the previous works use traditional sequential logging storage scheme and oversimplified buffer for storing B-tree structured files on flash memory. In this section, we propose a group round robin based B-tree index storage scheme, which efficiently eliminates the impact that caused by frequently changed B-tree structure by using dynamic grouping and round robin techniques. In addition, the proposed scheme relies on an enhanced clean-first buffer manager [4] which not only considers imbalance of read and write speeds but also the buffer hit ratio and space utilization.

3.1 Clustered Nodes

Accesses can be referred as being “sequential” or “random”. In the host, access is said to be sequential if its starting LBA directly follows the last LBA of the previous access command. Otherwise, the operation is said to be random. However, this definition is not apply to flash based devices due to the dynamic mapping performed by the FTL, contiguous addresses in the host space may refer to addresses that are not contiguous in the flash space. This paper adopts the idea of cluster blocks to manage the B-tree nodes, clustered nodes, in the flash memory. However, in contract to clustered block, clustered nodes exploits not only internal parallelism but also interleaving across multiple packages. In detail, clustered nodes exploits several levels of internal parallelism to access to several blocks in different NAND-flash chips at once, and interleaving index operations across multiple packages in the same channel so that time-consuming operations can be executed simultaneously.

4 Performance Evaluation

To evaluate the effects of the proposed storage scheme, we compare the numbers of page copies and block erases under IPL and CPL schemes and also give the effect of the proposed concept of semi-clean state. The result is summarized here.

4.1 Varying M and K

Figure 3 compares the performances of IPL and CPL under a random trace. We also counted the buffer references that were measured by the above test. As shown in Fig. 4, CPL and IPL have similar buffer hit ratios. However, CPL always perform better than IPL. Finally, we measured the performances of CPL and IPL by varying RWR, as shown in Fig. 5. Under a heavy-read workload, the performance gaps between them increase as the write ratio increases. CPL generates fewer writes to flash memory since it adopts an enhanced clean-first replacement policy.

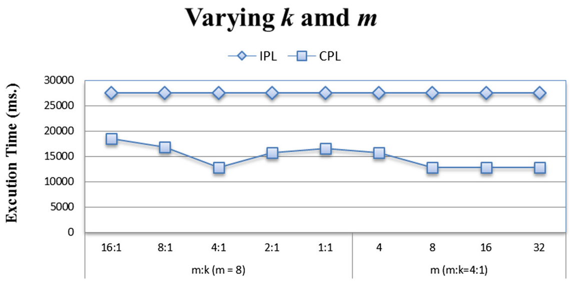


Fig. 3. Comparisons of execution times by varying *k* and *m* under random access trace

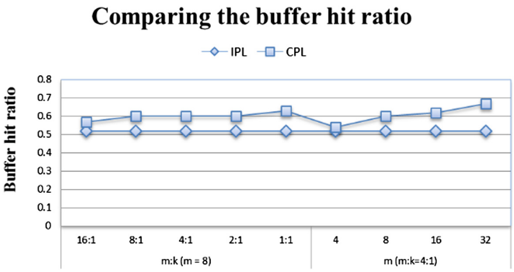


Fig. 4. Comparisons of the buffer hit ratios

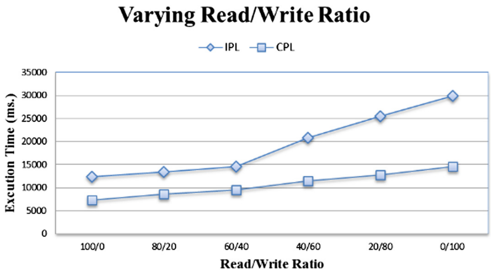


Fig. 5. Effects of increasing random write operation ratio

5 Conclusion

Despite the many advantages of flash memory, existing storage schemes perform poorly using it because of its unique electronic limitations. Especially, the B-tree index file more venerable than other data pages and generate many logs which can wear down the log area of IPL, the most-watched FTL scheme, quickly and increase the number of block erase operations. In this paper, we proposed a group round robin based B-tree index storage scheme, CPL, which uses the dynamic grouping, round robin techniques and an enhanced clean-first buffer replacement policy. The trace driven simulation showed that CPL efficiently eliminates the number of block erases that caused by frequently changed B-tree structure.

Acknowledgments. This research was supported by the MISP(Ministry of Science, ICT & Future Planning), Korea, under the National Program for Excellence in SW) supervised by the IITP(Institute for Information & communications Technology Promotion)

References

1. Park K-T et al (2015) Three-dimensional 128 Gb MLC vertical NAND flash memory with 24-WL stacked layers and 50 MB/s high-speed programming. IEEE J Solid-State Circ 50(1):204–213
2. Chung T-S et al (2009) A survey of flash translation layer. J Syst Archit Embed Syst Des 55:332–343
3. Na G et al (2012) Dynamic in-page logging for B + -tree index. IEEE Trans Knowl Data Eng 24(7):1231–1243
4. Kim J et al (2012) Parameter-aware I/O management for solid state disks (SSDs). IEEE Trans Comput 61(5):636–649
5. Roh H et al (2011) B+-tree index optimization by exploiting internal parallelism of flash-based solid state drives. In: Proceedings of VLDB'11, vol 5(4), pp 286–297
6. Jin R et al (2014) A group round robin based b-tree index storage scheme for flash memory devices. In: Proceedings of the IMCOM'14, January 9–11, 2014, Siem Reap, Cambodia

Data Security and Privacy on the Cloud: Driving to the Next Era of Technology with Confidence

Prasanna Balasooriya L N(✉), Santoso Wibowo, and Marilyn Wells

School of Engineering and Technology, Central Queensland University, Melbourne, Australia
p.balasooriya@cqu.edu.au

Abstract. Cloud services have gained popularity due to the number of advantages they provide to organizations and individuals such as reduced cost, better storage, and improved performance. However, a lot of organizations are still not willing to shift their traditional in-house services to the Cloud due to the various security implications. Many Cloud service users are worried about the security of their data and privacy being violated. There are many reported cases of Cloud service providers illegally collecting personal data of their customers, which has led to service providers being viewed with greater suspicion than before. To overcome this, Cloud service providers must ensure that they inform the users exactly about which data is being used and how it is used.

While it is the duty of the Cloud service provider to protect the data confidentiality and privacy of their customers, this should not be misunderstood or misused by customers to conduct illegal activities because Cloud service providers have to abide by the rules and regulations, including co-operating with law enforcement agencies if they need any particular customer's data. In this paper, we research the main security aspects for ensuring data confidentiality and privacy.

Keywords: Cloud security · Data privacy · Confidentiality · Challenges

1 Introduction

Cloud computing is seen as one of the emerging technologies available in the information technology domain. The National Institute of Standard and Technology (NIST) defines Cloud computing as a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources including networks, servers, storage, applications and services that can be rapidly provisioned and released with minimal management effort or service provider interaction [1, 2]. Cloud computing is found to be an important part of businesses and individuals where it helps organizations to reduce their operational costs by improving their services. In addition, the use of this technology increases the collaboration and scalability acceptance up to a non-comparable level [3].

Despite the numerous benefits of Cloud computing for businesses and individuals, there are some concerns such as security and privacy during its adoption [3]. Due to these concerns, organizations over the globe have been very slow in adopting Cloud

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_22

services, with only 10% of US organizations and 19% of European organizations are using the Cloud computing [4, 5]. It is also found that even organizations using Cloud services tend to limit their use [5]. Giannakouris and Smihily found that 49% of European organizations using Cloud services are only using the very basic services of email and data storage functionalities, and 57% of European organizations ranked the security breach as the main reason to prevent them adopting Cloud services [5]. Singh et al. [6] believe that Cloud security, availability and performance are recognized as the biggest problems for Cloud adoption [6]. Furthermore, Singh et al. [6] have further concern over the reporting structure of the incident of security and privacy violations in the Cloud computing. Therefore, ensuring data security and the privacy of the users' data on the Cloud is a critical factor that needs to be considered for the increasing use of the Cloud [7].

Nevertheless, providing secure and privacy protected Cloud services are highly challenging, as security and privacy problems could occur in different stages within the Cloud services context. Also, the success of the Cloud computing in the current information technology landscape has given a free pass for attackers to explore and target potential businesses and individuals [8]. These security and privacy issues which have been occurred due to unethical and illegal use of the user's data, could hinder the acceptance of Cloud computing [9]. It is therefore critical for businesses and individuals to address data security, data integrity and privacy issues in the use of Cloud computing [10]. Thus, the main aim of this paper is to review the issues that are related to data security and privacy of Cloud computing.

2 Literature Review

2.1 Data Security

As per ISO 27001 standards, Cloud security has been described as the preservation of confidentiality, integrity and availability of information in the use of Cloud computing [11]. However, the use of Cloud computing may face many critical issues such as competitive pressure, vendor support and third party control, performance and availability. With its growing popularity, Cloud security has become a critical factor that needs to be considered during its adoption and use [12]. Research has shown that data security, availability and performance are some of the most important elements of the quality of the service that Cloud providers need to offer to their users [13]. Table 1 presents a summary of factors that have limited the use of Cloud computing between 2013 and 2014 in Australia.

Gartner [12] indicates that more than 70% of its participants in the survey agreed that they do not intend to adopt or use Cloud services due to the fear of data security and privacy concerns. Furthermore, the number of major security breaches that occurred in the last few years has also contributed to the limited use of Cloud services by businesses and individuals. Cloud service providers have several security issues that they have to address, including (a) providing a secure connection for their users, (b) protecting data from hacker attacks, (c) ensuring that data is accessible by the customers at all times, and (d) preventing data loss during transfer [13, 14]. Mukherjee and Sahoo [15] point out that the adoption of Cloud computing lies with the security and privacy of the sensitive data of the organizations. If organizations are willing to keep their data in the Cloud, then organizations need to seek more clarification from the Cloud service provider on (a) how the Cloud provider encrypts organizational data and handle them, (b) how Cloud services providers handle the liabilities of data breaches and leakages, and (c) what is Cloud user substantiation. Figure 1 lists the most important factors that are limiting the adoption of Cloud computing services.

Table 1. Factors that are limiting the use of Cloud services in Australia

Factors	0–4 persons	5–9 persons	20–199 persons	200 or more persons	Total
Risk of a security breach	14.0	18.2	23.6	30.4	16.2
Problems accessing data or software	6.2	8.7	8.4	15.9	7.2
Difficulties with unsubscribing or changing Cloud computing service provider	3.4	4.3	5.1	7.2	3.9
Uncertainty about the location of data	9.5	11.0	14.4	19.2	10.5
Uncertainty about legal, jurisdictional or dispute resolution mechanisms	6.9	8.4	9.6	12.8	7.6
High cost of Cloud computing services	10.1	12.0	12.2	19.8	10.9
Insufficient knowledge of Cloud computing services	21.6	24.8	24.2	22.1	22.8
Other factors	4.5	4.6	4.2	5.5	4.5
No factors limited or prevented the use of paid Cloud computing	61.3	55.8	51.0	43.6	58.7

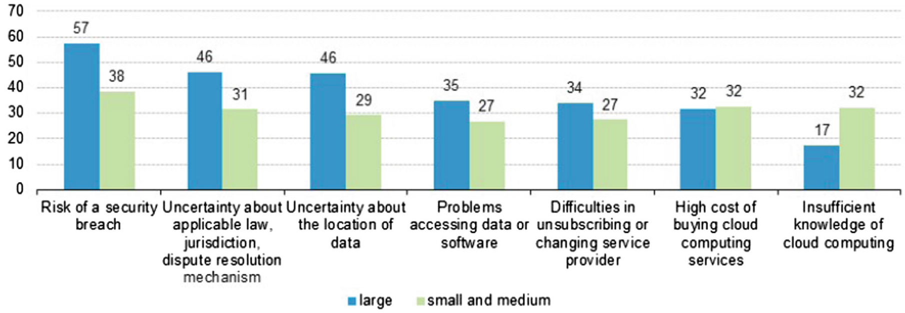


Fig. 1. Factors limiting organizations from using Cloud computing services

Data plays a very important role in Cloud services with users submitting their personal information as well as storing and transferring sensitive and confidential information. Thus, Cloud data security challenges can be broadly classified into data confidentiality issues and data integrity issues. Both of these issues arise due to failed data security measures. Data confidentiality refers to protecting the customers' data from being disclosed to illegitimate parties without their express approval while data integrity refers to protecting consumers' data from malicious modifications and ensuring the accuracy and consistency of data [16]. Table 2 presents the top security challenges that need to be considered during the adoption and use of Cloud computing.

Table 2. Top security challenges in Cloud computing

Risk/Challengers	References
Data acquisition	[17]
Confidentiality	[11, 17]
Integrity and authenticity	[17]
Multi-tenancy	[11, 18–20]
Service level agreement (SLA's)	[18]
Insider attacks	[18]

2.2 Data Confidentiality and Privacy

Data confidentiality entails preserving authorized restrictions on information access and disclosure, including means for protecting personal privacy and proprietary information [17–20]. Cloud computing has been recognized as a next generation information technology model that could help businesses and individuals fulfil their requirements. However, the operational and administration model of Cloud computing differs from traditional information computing architecture. To provide a better and reliable service at low costs, Cloud service providers have to shift their applications to data centers where the management and administration of data and services are not trustworthy [21]. This feature could contribute to a new data security and privacy challenges in adopting and using the technology [22]. Therefore, it is important for Cloud service providers to address the issue of privacy that comes along with strong and extremely sensitive data stored in the Cloud environment so that users of Cloud computing will be able to enjoy the full benefits of the Cloud computing.

In Cloud services, there are many aspects of data confidentiality. The first issue is that of unauthorized data collection by the Cloud service providers themselves. Many Cloud services are free for the customers, whereby the business model is making revenue from advertising. In order to target their advertisements better and at the same time provide personalized ads, many Cloud service providers tend to violate the privacy of their customers by collecting unauthorized personal data of their customers [17, 23]. One of the largest information technology organization, Google, which provides many Cloud services, including google drive, google play store, and Gmail, has been accused time and again of violating consumer privacy by collecting users' information surreptitiously [12, 24, 25]. In 2012, Google was fined $22.5 million by US authorities for violating privacy regulations by secretly collecting user data from the Safari browser [24]. In 2013, Google was again fined $17 million for a similar offence, and was also accused of collecting unauthorized data from every user of the google app play store and selling it to developers [24, 25]. Customers have a right to know exactly what data will be collected and how it will be used, which is why secretly collecting customer email addresses and selling it to developers without informing the users in advance, is a big violation of privacy [26].

The email Cloud service of Google, Gmail has also come under the scanner for privacy violations. In 2010, consumers filed a complaint against Google regarding the unauthorized scanning of private mails exchanged between consumers through Gmail, and using the data from the mails to target customers with personalized advertisements [27]. Google admitted to scanning emails and using the data for an advertisement generation, but stated that it had adequately informed its users of this in terms and conditions and therefore, there was no violation of privacy [27].

Encrypting user data is another key data security measure so that no one, not even employees have access to the data. [28]. A Cloud service provider, Skyhigh has conducted a survey amongst healthcare organizations regarding Cloud service usage and security risks. The research found that 33% of the organizations reported data leaks via employees in 2014, while 79% of the organizations stated data leaks as one of their topmost worries [29]. While all Cloud service providers need to ensure encryption and protection against misuse of data by employees, the healthcare industry is at the highest risk due to the high price that data mining organizations are willing to pay to obtain patient details for insurance and pharmaceutical organizations [29]. In fact, employees have resorted to selling their login credentials in order to make money, and 90% of the organizations surveyed had at least one employee credential on sale online [29]. Such data leaks by the employees constitute a violation of privacy as well as data confidentiality regulations, since the patient has no idea that his/her medical history is being sold by the Cloud service provider.

Data confidentiality may be compromised due to malicious attacks by hackers or other outside threats. As the adoption of Cloud services is growing amongst individuals and businesses, the focus of hackers is also shifting from targeting private networks to targeting the Cloud [30]. AlertLogic conducted a research on the state of Cloud security from its 2,200 customers. Their research found that in 2014, 44% of the customers experienced a brute force attack compared to 30% of customers a year earlier [31]. There was an equal percentage of vulnerability scans (44%) which have also increased from 27% a year earlier [31]. Hackers mainly carry out these attacks with the intention of stealing personal data with the ultimate aim of financial fraud or identity theft. The vulnerability of Cloud services to hacker attacks became evident with the hacking of Apple's iCloud accounts in which the privacy of several celebrities was violated and personal data was stolen [32]. This attack was caused due to weak data security measures and the absence of a two factor authentication system [32]. Retail giants Target and Home depot, both became victims of a data breach due to their private Cloud being hacked and credit card details of their clients being stolen [33]. Both attacks were caused due to loopholes in their data security such as weak data encryption systems that allowed hackers to read client data easily [33]. It can be seen that protecting data against misuse by parties with malicious intent is difficult due to the constant innovation by hackers in attacks including a new "man in the Cloud" method which hacks into the file synchronization software virtually undetected [34]. It is therefore critical for Cloud service providers to constantly evaluate their data security measures and implement the latest security measures to protect their user's data confidentiality and privacy.

Another data confidentiality issue faced by Cloud service providers is to whether or not allow access to confidential information about its users to the Government and law enforcement agencies. Hushmail is an email service that markets itself on the basis of its highly secure encryption and a network, which ensures that the data contained in the mails remains confidential and cannot be accessed by their employees or anyone else. However, this very characteristic made it a haven for nefarious activities with drug dealers who are using it to conduct their business operation [35]. When Hushmail was approached by the US law enforcement agencies regarding the matter, it chose to violate data confidentiality and disclose the relevant email data to the law enforcement officials [35]. Some may view this step as a data confidentiality violation, but Cloud service providers, including Hushmail, have stated very clearly that the encryption and data security measures are meant for protection from malicious attacks and they will always co-operate with law enforcement agencies. Therefore, while it is necessary for Cloud service providers to ensure data confidentiality and privacy of its users, their first and foremost duty is to abide by the rules and regulations [36, 37].

3 Discussion

Cloud security is a vast topic with different types of threats that have to be dealt with by having several security measures put in place. The Cloud is generally used by customers to transmit and store data, and therefore data security is one of the biggest issues in the use of Cloud, specifically data confidentiality and privacy. Customers fear losing their data, or having sensitive data leaked, which may lead to serious issues like identity theft. Data security involves protecting the customers' data all across the data life cycle starting from data input to data transfer and data destruction, with each phase requiring unique security measures. Table 3 presents the key findings of this review that can be derived in this study in relation to the protection of data confidentiality and customer privacy.

The Cloud service providers are legally bound to inform their customers about exactly which data they are collecting and how that data will be used. This helps put customers at ease since they have a clear understanding of how their personal data will be used. The major threat to data confidentiality and privacy is from the employees of the Cloud service providers who have access to customer data. In order to protect the data from being misused by the employees, it is necessary to have good data encryption, which makes it difficult for employees to access the data itself, thereby ensuring privacy and data confidentiality. The Cloud data is extremely vulnerable to threats from hackers and therefore adequate security measures need to be taken to protect Cloud services from malicious attacks. Despite the importance of data confidentiality and privacy, Cloud service providers can disclose sensitive personal data to law enforcement agencies if needed.

Protecting data on the Cloud is one of the highest priorities for Cloud providers. Thus, Cloud providers must invest more time and take robust security measures such as encryption of data during transmission and storage, limiting access to the data, continues review of security threats and implementation of system audits and accountability checks to protect data on the Cloud.

Table 3. Key findings

Key findings	References
Cloud service providers face a plethora of security issues and need to implement various security measures	[13, 14, 38]
Different data security measures need to be put in place for the various phases of the data life cycle, specially to ensure data confidentiality and privacy	[26, 39]
Cloud service providers should disclose exactly what personal data will be collected and what that data will be used for, thereby maintaining privacy and confidentiality	[26]
Data should be properly encrypted so that it cannot be accessed or misused by the employees of service provider	[28, 29]
Malicious attacks from hackers need to be prevented at every stage of the data life cycle so that there is no breach of data confidentiality	[30, 33]
Cloud services have to comply with the country's legal rules, which includes disclosing customer data to the law enforcement agencies if needed	[36, 37]

Vulnerability in the Cloud network, software applications or environment are golden opportunities for hackers who wanted to gain access and control of someone else data for their personal gain. Thus, preventing vulnerabilities and protecting data from hackers is another priority for the Cloud service provider. Here, Cloud providers could consider using some of the best known vulnerability prevention strategies such as (a) separation of infrastructure and services, (b) use of data obfuscation techniques where data can be transformed to hide the real meaning of the data, and (c) hiding or separating owners' details from the data to protect the data confidentiality further.

Protecting customers' data while complying with respective laws are also challenging for every Cloud service provider. The legal and disclosure requirements vary from country to country. Thus, respective privacy and data protection laws could be used to protect data and the privacy of the Cloud users. Based on the data protection and privacy laws, it will cover only the personal data where that is locally located. However, the fundamental rule of law is conflicting with the way that Cloud architecture is designed and developed. Cloud based e-mail can be seen as a good example of this situation. Storage of personal e-mails can be stored and located anywhere in the world. If the data goes all around the world, then it will no longer be clear which data protection laws will apply to protect users' data. Regardless of current data and privacy protection provisions, some of the countries have responded proactively to protect their citizens' data in the Cloud. The Swiss government has implemented their data protection in line with European Union (EU) law, where they have identified three key components such as (a) transfer of personal data to third parties, (b) transfer personal data abroad, and (c) data security. Thus, as per Swiss data protection provision, transfer or exporting personal data is permitted where the legislation ensures that adequate data protection measures are taken to protect personal data in accordance with the Swiss legislative requirement in the country where the data is located. Most importantly, Swiss data protection provision covers the special circumstances such as where personal data need to be transferred or exported, but there is no adequate protection is provided by the country where that data going to be stored. Thus, that provision has a mandatory requirement to mention the collection of the data and the business use of collected data in a contractual agreement.

Transferring sensitive data to a third party raises more questions than any other time. However, as per EU data protection law, service provider or data handler remains responsible for the data under their care. Furthermore, the service provider who will be looking after data is permitted to subcontract one or more third parties to process customer data on behalf of them under their instruction. However, these need to be closely monitored, and the service provider needs to ensure that contracted third parties are processing the data as per the instructions provided by them.

Sending or storing customer data could be seen as a privacy violation. There is no evidence to argue that Cloud customers know where their data is located. Furthermore, there is less evidence to indicate that Cloud service providers are providing location specific information to their customers. Thus, this issue needs to be looked at government level, and adequate data protection provisions need to be implemented to protect personal data.

Irrespective of the location of the personal data stored, Cloud service providers are responsible to safeguard the customer personal data which they collect and store. Furthermore, Cloud service providers or data collectors are required to implement additional measures to protect data from unauthorized access, illegal data destruction, thefts or misuse of data.

4 Conclusion and Future Direction

With the increasing popularity of Cloud services, Cloud security and privacy issues are gaining their importance. While there are several Cloud security issues, the one that is most worrisome for customers is data security, which includes data confidentiality and privacy protection. Stringent security measures need to be used to protect the Cloud data from hacker attacks. Hackers target user data with the intention of identity theft or financial fraud, which are very serious problems. Furthermore, Cloud service providers must consider using Service Level Agreements to provide an assurance to their customers about data protection and privacy.

In this review, applicable data and privacy protection laws have been discussed briefly, which is a really important factor in the adoption or use of this Cloud computing technology. Thus, this area needs to be explored in details in future studies.

References

1. Mell P, Grance T (2011) The NIST definition of cloud computing: recommendations of the national institute of standards and technology. NIST Special Publication 800-145
2. Shayan J, Azarnik A, Chuprat S, Karamizadeh S, Alizadeh M (2014) Identifying Benefits and risks associated with utilizing cloud computing. Int J Comput Softw Eng 3(3):1–6
3. Hashemi S (2013) Cloud computing technology: Security and trust challenges. Int J Secur Priv Trust Manage 2(5):1–7

4. Denworth J (2015) Adoption of cloud computing in the enterprise: the progress in 2015. http://betanews.com/2015/12/29/adoption-of-cloud-computing-in-the-enterprise-the-progress-in-2015. Accessed 12 Jan 2017
5. Giannakouris K, Smihily M (2016) Cloud computing - statistics on the use by enterprises. http://ec.europa.eu/eurostat/statistics-explained/index.php/Cloud_computing_-_statistics_on_the_use_by_enterprises. Accessed 18 Dec 2016
6. Singh I, Mishra KN, Alberti A, Singh D, Jara A (2015) A novel privacy and security framework for the cloud network services. In: 9th international conference on innovative mobile and internet services in ubiquitous computing, pp 301–305
7. Agrawal D, Abbadi AE, Shiyuan W (2013) Secure and privacy-preserving database services in the cloud. In: International conference on data engineering, pp 1268–1271
8. Abuhussein A, Bedi H, Shiva S (2012) Evaluating security and privacy in cloud computing services: a stakeholder's perspective. In: International conference on internet technology and secured transactions, pp 388–395
9. Tari Z (2014) Security and privacy in cloud computing. IEEE Cloud Comput 1(1):54–57
10. Al-Jaberi MF, Zainal A (2014) Data integrity and privacy model in cloud computing. In: International symposium on biometrics and security technologies, pp 280–284 (2014)
11. Alouane M, El Bakkali H (2015) Security, privacy and trust in cloud computing: A comparative study, In: International conference on cloud technologies and applications, pp 1–8 (2015)
12. Xiang T, Bo A (2011) The issues of cloud computing security in high-speed railway. Electron Mech Eng Inf Technol 8:4358–4363
13. Ashktorab V, Taghizadeh SR (2012) Security threats and countermeasures in cloud computing. Int J Appl Innov Eng Manage 1(2):234–245
14. Subashini S, Kavitha V (2011) A survey on security issues in service delivery models of cloud computing. J Netw Comput Appl 34(1):1–11
15. Mukherjee K, Sahoo G (2012) A novel methodology for security and privacy of cloud computing and its use in e-Governance. In: World congress in information and communication technologies, pp 561–566
16. Yu S, Lou W, Ren K (2012) Data security in cloud computing. In: Das SK, Kant K, Zhang N (eds) Handbook on Securing Cyber-Physical Critical Infrastructure. Morgan Kaufmann, San Francisco
17. Pant VK, Prakash J, Asthana A (2015) Three step data security model for cloud computing based on RSA and steganography. In: Green computing and internet of things, pp 490–494
18. Bamiah M, Brohi S, Chuprat S, Brohi MN (2012) Cloud implementation security challenges, cloud computing technologies, In: International conference on applications and management, pp 174–178
19. Hamouda S (2012) Security and privacy in cloud computing. In: International conference on cloud computing technologies, applications and management, pp 241–245
20. Takabi H, Joshi JBD, Gail-Joon A (2010) Security and privacy challenges in Cloud computing environments. IEEE Secur Priv 8(6):24–31
21. NICCS (2016) A glossary of common cybersecurity terminology. https://niccs.us-cert.gov/glossary#confidentiality. Accessed 13 Jan 2017
22. Jian W, Yan Z, Shuo J, Jiajin L (2010) Providing privacy preserving in Cloud computing. In: International conference on human system interactions, pp 472–475
23. Lin QT, Wang CD, Pan J, Ling L, Lai JH (2015) Local data security and privacy protection in Cloud service applications. In: 9th international conference on frontier of computer science and technology, pp 254–258

24. Acquisti, A (2017) The economics of personal data and the economics of privacy. http://www.oecd.org/sti/ieconomy/46968784.pdf. Accessed 13 Jan 2017
25. Arora H (2017) Google to pay $17 million for unauthorized tracking of safari users. http://www.techspot.com/news/54731-google-to-pay-17-million-for-unauthorized-tracking-of-safari-users.html. Accessed 14 Jan 2017
26. Firstpost (2013) Privacy violation: google play store handing out user info to developers. http://m.tech.firstpost.com/news-analysis/privacy-violation-google-play-store-handing-out-user-info-to-developers-212914.html. Accessed 14 Jan 2017
27. Chen D, Zhao H (2012) Data security and privacy protection issues in cloud computing. Int Conf Comput Electr Eng 1:647–651
28. The New York Times (2013) Claims that google violates gmail user privacy. http://www.nytimes.com/interactive/2013/10/02/technology/google-email-case.html?_r=0. Accessed 13 Dec 2016
29. Will M, Ko R (2015) A guide to homomorphic encryption. In: The cloud security ecosystem: technical, legal, business and management issues, pp 1–34
30. Bleiberg S (2016) Cloud adoption and risk in healthcare. https://www.skyhighnetworks.com/cloud-security-blog/93-of-cloud-services-in-healthcare-are-medium-to-high-risk/. Accessed 13 Dec 2016
31. Chou TS (2013) Security threats on cloud computing vulnerabilities. Int J Comput Sci Inf Technol 5(3):79
32. Ashford W (2016) Cyber attacks move to cloud with increased adoption, report shows. http://www.computerweekly.com/news/2240219265/Cyber-attacks-move-to-cloud-with-adoption-report-shows. Accessed 13 Dec 2016
33. Pachal P (2016) Apple strengthens icloud security in wake of celeb nude photo hack. http://mashable.com/2014/09/16/apple-strengthens-icloud-security/#6GyDJNTV88qn. Accessed 15 Dec 2016
34. Rando N (2016) Cloud security breaches still the stuff of IT nightmares. http://searchcloudcomputing.techtarget.com/feature/Cloud-security-breaches-still-the-stuff-of-IT-nightmares. Accessed 15 Dec 2016
35. Imperva (2016) Imperva Hacker intelligence initiative uncovers new “Man in the Cloud” attacks that use popular file synchronization services. http://globenewswire.com/news-release/2015/08/05/757974/0/en/Imperva-Hacker-Intelligence-Initiative-Uncovers-New-Man-In-the-Cloud-Attacks-That-Use-Popular-File-Synchronization-Services.html. Accessed 12 Dec 2016
36. Singel R (2016) Encrypted email company Hushmail spills to Feds. http://www.wired.com/2007/11/encrypted-e-mail. Accessed 15 Dec 2016
37. Delamore B, Ko R (2015) Security as a service (SecaaS) - an overview. In: The cloud security ecosystem: technical, legal, business and management issues, pp 1–10
38. Maxwell W, Wolf C (2016) A global reality: governmental access to data in the cloud: a comparative analysis of ten international jurisdictions. http://www-05.ibm.com/ch/services/documents/sce/br-government-access-to-cloud-data.pdf. Accessed 8 Dec 2016
39. Rautela S, Negi A, Chaudhary P (2015) Data security and updation of data lifecycle in cloud computing using key-exchange algorithm. Int J Adv Res Comput Commun Eng 8:380–386

A Decision Making Based Authentication Scheme in Cooperative Vehicular Ad-Hoc Network

Prasanna Roy[2] and Emmanuel Antwi-Boasiako[1](✉)

[1] Ghana Institute of Management and Public Administration, Accra, Ghana
abeantwi@gimpa.edu.gh
[2] Durgapur, West Bengal, India
prasanna.roy.durgapur@gmail.com

Abstract. With the growth and advancement in Intelligent Transport Systems comes the comfort and ease of driving at a rapid pace. People are also able to get both information and entertainment services when they are travelling. However, as Intelligent Transportation System (ITS) is becoming popular, the number of attacks on Vehicular Ad-Hoc Network (VANET) has also increased. One of the most important security threats is authenticity since most of the on road decisions are taken on the basis of the messages received. This paper proposes a simple decision making algorithm that helps in determining the authenticity of the occurrence of an incident and thus helping the vehicle under consideration to take decision regarding its future mobility pattern.

Keywords: VANET · Authentication

1 Background

The last two decades has seen a tremendous increase in the number of vehicles. In addition to this, passengers on board are demanding various web based services. It has therefore become necessary to bring tremendous changes in the Intelligent Transport System (ITS) that is in existence. Some of the prime services provided by ITS are ensuring on-road safety and infotainment services using internet. This has led to the introduction of a new domain known as Vehicular Ad-Hoc Network (VANET) which is one of the popular components of ITS and a famous application of Mobile Ad hoc Network (MANET). One of the distinguishing characteristics of VANET is that the pattern of the network is determined by the position of the nodes. The units that participate in a VANET change their positions very frequently and there is also no restriction in their energy consumption. One of the main objective of VANET is to support exchange of safety information among the mobile nodes that participate in the VANET. If the VANET is attacked, it can result in both financial loss and loss of lives. Thus, it is very necessary to ensure the security of VANET.

This paper discusses the various security challenges in VANET, the major attacks, brief review of some of the existing solutions and a proposal of a new security model and algorithm.

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_23

2 VANET

Vehicular Ad-Hoc Networks (VANET) is an application of MANET in which vehicles communicate with each other with the help of internet by forming an ad-hoc wireless network. Possibly, communication in VANET can occur in two modes, namely, Vehicle to Vehicle (V2V) and Vehicle to Road Side Unit (V2R) communication. Vehicles in a VANET can also indulge in Hybrid communication in which both Vehicle to Vehicle and Vehicle to Infrastructure communication is combined.

2.1 Components of VANET

Vehicles - Vehicles serve as nodes in a vehicular network. VANET wireless communication takes place between the vehicles and the access points that have been provided as a part of the VANET infrastructure and between vehicles.

Infrastructure - The main component that makes up the infrastructure of a VANET is the Road Side Base Stations. Generally these road side units are positioned at the road junctions or parking areas. The prime focus of the base station is to expand the communication area of VANET. One road side unit has the ability to transfer information to other components of the network. Safety signals regarding low bridges and accidents are issued by the road side units.

Communication Channels - To transmit signals in a VANET, radio signals are used. The wavelength of an electromagnet wave is much more compared to infrared rays. The radio waves frequency lies in the range of 190 GHz to 3 kHz. How the protocol will perform is dependent on the propagation of the radio waves. The protocol is able to determine the number of nodes that will be allowed in one collision domain based on this.

2.2 Applications of VANET

Collision Avoidance (application related to safety) - Studies have indicated that it is possible to avoid about 70% of an accident if the driver gets warning about it at least half second before the occurrence of the accident [1].

Cooperative Driving (application related to safety) - Traffic related warning signals can be also sent to the drivers. The warnings may include speed warning or lane change warnings. The drivers can enjoy uninterrupted and safe driving as there will be cooperation among them due to the exchange of these signals.

Traffic Optimization (application related to safety) - It is possible to optimize traffic if signals related to incidents such as accidents and jams are sent to the drivers on time. The driver can save time by choosing an alternative path.

Peer to Peer Application (user based application) - The vehicles that are in a network can share audio, video files with the help of these applications.

Internet Connectivity (user based application) - The participants of a VANET can remain connected with the Internet all the time.

Miscellaneous Services - Other services also provided by VANET such as paying toll taxes, locating road side restaurants and fuel stations, etc.

2.3 Characteristics of VANET

High Mobility - The nodes in a VANET change their position frequently. Thus it becomes difficult to predict its position and protect its privacy [2].

Dynamic Network Topology - The positions of the nodes keep on changing frequently because the vehicles are highly mobile due to their random speed change. This brings about continuous change in the topology of the network.

Unbound Network Size - There is no geographical limitation in the size of a VANET. Its area can span a single city, multiple cities or even countries.

Frequent Exchange of Information - Due to the ad-hoc nature of VANET it has to gather information continuously from the Road Side Units (RSU) and other vehicles.

Wireless Communication - To exchange information, the nodes in a VANET use wireless medium. Thus great emphasis has to be given to security measures while performing communication.

Time Critical - In order for a node to be able to take decisions on time and perform actions at the right time it necessary to ensure that the information from the origin node reaches the destination node on time.

Sufficient Energy - VANET does not impose any constraint on energy and computation resources. The transmission power is also not limited.

Better Physical Protection - The VANET infrastructure is safe as it is protected physically.

2.4 Challenges in VANET

Technical Challenges

Network Management - As the nodes in a VANET keep on changing their position vary frequently, there is continuous change in the condition of the channel and network topology.

Congestion and Collision Control - The unbounded nature of the network size also poses a challenge. In rural areas the traffic load is comparatively low all time of the day while in urban areas the load is less during the night only. The network suffers from high congestion during the peak hours of the day. Collision also becomes more frequent.

Environmental Impact - It is necessary to take into consideration the impact of the environmental factors on the electromagnetic waves through which communication takes place in the VANET.

MAC Design - To carry out communication the nodes participating in a VANET use shared medium. As result designing the MAC becomes a prime concern.

Security - VANET offers various road safety applications which are life critical in nature. Thus, it is very necessary to ensure that the messages that the nodes exchange are not compromised.

Security Challenges in VANET

Real time Constraint - Time is one of the most critical factors in VANET since the message must be delivered with a maximum permissible delay of 100 ms. It is necessary to use a fast cryptographic algorithm to fulfill this real time constraint. It is necessary to authenticate the entity and message within time.

Data Consistency Liability - In VANET malicious activities can be performed by a node that has already been authenticated. This might cause disturbances in the network and accident among the nodes. In order to remove this inconsistency some mechanism must be designed. One way of eliminating such inconsistency is by establishing correlation among the data that is received from various nodes.

Low tolerance for error - Probability is taken as the basis at the time of designing some protocols. A tiny error in the algorithm based on probability may cause great damage as VANET is based on information that is life critical and there is very short time available to perform the actions.

Key Distribution - VANET security mechanisms are totally based on keys. Every message is encrypted. It is necessary also to decrypt the message at the receiving end with the same key or some other key. Chances are that various manufacturers will install keys in different managers. Thus the public key infrastructure has to greatly depend on

Certificate Authority (CA). Thus, one of the prime challenges while designing security protocols is how the keys will be distributed among the vehicles.

Incentives - The major focus of the manufacturers is to build applications that are most liked by the customers. There is no driver who will accept a vehicle that will send traffic violation reports automatically. Thus, initiatives from government, manufacturers and consumers are required to overcome challenges in implementing security in VANET.

High Mobility - VANET and wired network have the same energy supply and computational capability. But in order to achieve the same throughput that is produced by wired networks, it is necessary to ensure that the security protocols execute within very less time period as the nodes in VANET are highly mobile.

Security Requirements in VANET

Authentication - Authentication ensures that whether the message has been generated or received by a legitimate user or not. It is necessary to ensure authentication due to the fact that a vehicle reacts depending on the information sent by some other vehicle.

Availability - Availability ensures that all the legitimate users get the information. The network will be brought down due to DoS attack. This prevents the information from being shared.

Non-Repudiation - Non-repudiation ensures that a node is unable to deny that it has sent a message.

Privacy - Privacy of a node guarantees that the node is kept safe from an unauthorized node. This can be used for avoiding the message delay attacks.

Data Verification - To avoid the problem of false messaging, data must be verified regularly.

Attackers on VANET

Insider and Outsider - Members who have been authenticated in the network are known as Insiders while intruders are Outsiders. The later has limited attacking capacity.

Malicious and Rational - Malicious attackers attack the network with an intention of harming the network without any personal benefit. On the other hand, rational attackers attack the network for some benefit and hence they can be predicted.

Active and Passive - While active attackers generate packets or signals the passive attackers keeps on sensing the network.

3 Previous Works on VANET Security

See Table 1.

Table 1. A comparative study of the various solutions proposed till now

S. No	Name of scheme	Attacks addressed	Solution used	Security requirements
1	ARAN [7]	• Replay attack • Impersonation • False warning	Cryptographic Certificate	• Authentication • Message integrity • Non-repudiation
2	SMT [6]	Information disclosure	Message Authentication Code	Authentication
3	SEAD [5]	• DoS • Routing attack • Resource consumption	One way hash function	• Availability • Authentication
4	NDM [4]	• Information disclosure • Location tracking	Asymmetric cryptography	Privacy
5	ARIADNE [3]	• DoS • Routing attack • Replay attack	• Symmetric cryptography • Message Authentication Code	Authentication

4 Proposed Work

As already mentioned there is a need to look at other ways of authentication that saves time and allows a driver to quickly make a decision as against current schemes that consume time because of encryption and decryption of keys that are used. In the proposed scheme an attempt has been made to ensure the authenticity of the message that is received by a vehicle from another vehicle in a cooperative driving scenario. In such a scenario vehicles send messages to one another notifying them about the various road conditions. This allows the vehicles coming from behind take decisions regarding the selection of lanes. For example, if an accident occurs on any location on a road, a driver of a vehicle should be able to know about the accident beforehand so that he can avoid the route on which the incident has occurred. However, in a cooperative driving scenario it has also been observed that sometimes, some nodes behave as selfish nodes. Selfish nodes provide misleading information about road conditions that allows to have their own ways. Such a behavior has been influenced by general human psychology. In such a situation a node coming from behind may sometimes receive wrong information about events that are occurring on roads. This will hinder the process of taking the right decision. In the proposed algorithm a simple decision making procedure at the node level is used that will allow another vehicle that is yet to arrive at the spot where the event has taken place to take an appropriate decision. It is assumed that the nodes move in the same direction. Whenever a node which is ahead observes an incident in front of it or around it, the information about the particular event is transferred to the nodes behind it.

Let the node under consideration be denoted by N_A. Let N_B be another node which is in the neighborhood of N_A, that is, N_B is in the communication range of N_A.

Here, either N_A is at a distance greater than D behind N_B or N_B is at a distance greater than D behind N_A. Here, D is considered as a threshold value. Below this particular value of D nodes are considered to be moving beside each other.

Let the neighborhood of a particular vehicle be denoted by NB(A). Let the message under consideration be denoted by M_A.

In the proposed architecture, the whole network that is under the coverage area of a single Road Side Unit (RSU) is divided into a number of sub-ranges. Number of sub-ranges that will be considered by the vehicle receiving the message about the incident is denoted by S.

Thus, number of sub-ranges = S, this is from the origin of the alert message. For example, for first-hand information S = 0, for second hand S = 1 and so on. This number is denoted by $S(M_A)$.

M_A also has a decision associated with it which is denoted by d_A. The value of d_A can be either +1 or −1.

Now, let an event occur. N is the set of nodes that witness the event.

For example if the event is an accident. The nodes report either "Accident", which is correct information and is denoted by C, or, "No Accident" which is incorrect and denoted by I.

The decision that is taken by the node N_A is denoted by d_A.

The task of node N_A is to decide whether an accident has occurred or not depending on the message it has received from its neighbors.

It considers all the neighbor nodes which are denoted by N_B. Here every N_B belongs to NB(A). The nodes under consideration are in front of the node denoted by N_A.

Now, the number of nodes which reports C is denoted by V_C, while, the number of nodes which report I are denoted by V_I.

V_{actual} on which the value of d_A depends is then calculated using the formula:

$V_{actual} = V_C - V_I$.

The decision is taken as:

If $V_{actual} >= 0$, then $d_A = 1$, N_A decides "Accident".

else

$V_{actual} < 0$, then $d_A = -1$, N_A decides "No Accident".

From the results, the concerned node decides on an action to take whether to main a route to a destination or change it.

The vehicles that report the incident take the location of the incident as one of the most important parameters. The current position of a particular vehicle is obtained from the GPS installed in it (every vehicle that is part of the proposed network must have GPS installed inside it). Thus, only nodes that are at a position which is behind the incident reports the incident. The vehicles that have passed the spot of the incident are not allowed to transmit any information about the incident. This is due to the fact that the nodes that have passed the spot before the incident have occurred, will send message that suggests that no such incident has taken place. This will greatly affect the decision making procedure. It is also assumed that the nodes are able to send notification about an incident only once. This will prevent the network from getting flooded with huge number of packets. Also, a node will not be able to confuse other nodes in the network by sending contradictory information about a particular incident.

5 Theoretical Evaluation

The proposed decision based authentication scheme has a simple, light weight and efficient algorithm compared to other authentication schemes that have been proposed in the recent time. Such schemes as Compared to our algorithm use encryption and decryption or hashing to secure the message and help destination vehicles to authenticate the message they receive. Since authentication is to be employed in vehicles which are fast moving, a simple and light weight scheme like the proposed, would help the vehicles in taking very fast decisions. The algorithm is inspired by human behavior. Thus, the scheme can also handle situations when the driver of the vehicle acts selfishly and propagate wrong information about an incident that has taken place. This is achievable due to the fact that a large number of vehicles cannot turn out to be selfish at the same moment. The proposed scheme is adaptive for any size of vehicular ad-hoc network. Thus, it is both elastic and scalable.

6 Simulation

In order to preliminarily test and evaluate our algorithm, we have simulated our algorithm using the NS-2 Simulator. We used the MAC (Media Access Control) and Physical layer parameters of IEEE 802.11a, on which DSRC is based. We simulated an urban environment that has a highway with six lanes, three in each direction with an inter-vehicle space of about 30 m. We considered the situation where vehicles are mobile and also transmit DSRC messages every 300 ms over a 300 m communication range [1]. Hence from our algorithm D = 300. We demonstrated with about 2–100 cars all moving in the same direction as shown in Fig. 1. We have assumed that it is mostly vehicles

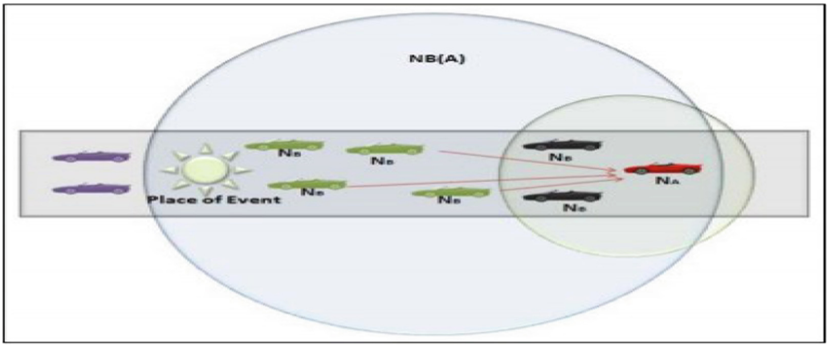


Fig. 1. Proposed system scenario

ahead of a particular vehicle that reports incidents not cars exactly beside a particular vehicle.

7 Results

Preliminary results gathered so far are as shown in the graph in Fig. 2. We have compared our algorithm with the work done in [3] since it is relatively one of the current work done in literature. The figure shows a graph of number of vehicles against execution time.

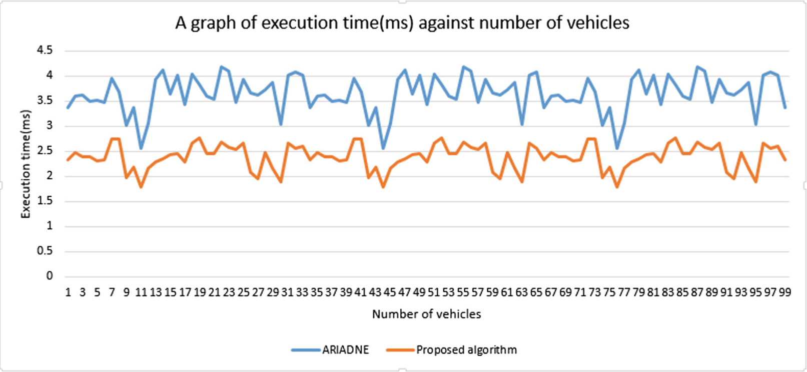


Fig. 2. Performance of proposed scheme compared to Ariadne

8 Performance Evaluation

The algorithm that has been proposed in this paper is a very simple decision making algorithm and is expected to use very less time and computational resources to execute. Hence from the results, it can be seen that our algorithm helps a vehicle under consideration to quickly make a decision as to whether to maintain a route to a destination or change it. Our algorithm uses a simple mathematical model to achieve the purposes of authentication hence low execution time. The procedure of decision making to judge the authenticity of a message is highly suitable in those scenarios that promote cooperative driving. The vehicle that is taking the decision will be able to take decision with ease and accurately depending on the number of messages favoring the incident and number of messages not favoring the incident. Also, as the nodes are allowed to report about an incidence only once, the network is not flooded with unnecessary packets. Additionally, the messages are unique which ensures authentication. A single node can only send a single message. Thus, it is unable to confuse other nodes in the network.

9 Conclusion and Future Scope of Work

Mathematically, the algorithm seems to perform well in determining the authenticity of an incident (e.g. An accident). Though the proposed work is able to carry out decision making in a very simple manner, it may fail in cases where the number of nodes that behave in a selfish manner increases. We are thinking of extending our work to incorporate the role of Road Side Units (RSUs) and network administrator in the decision making process to ensure authenticity in a much stronger way. Plans are also there to further simulate the scenario and observe how the algorithm works in various road conditions, urban and rural area and various times of the day.

References

1. Raya M, et al (2005) The security of vehicular ad hoc networks. In: SASN 2005, pp 11–21
2. Moustafa H, Zhang Y (2009) Vehicular networks: techniques, standards, and applications. Auerbach Publications, Boston
3. Hu YC, Perrig A, Johnson DB (2005) Ariadne: a secure on-demand routing protocol for ad hoc networks. Wirel Netw 11(1–2):21–38
4. Fasbender A, Kesdogan D, Kubitz O (1996) Variable and scalable security: protection of location information in mobile IP. In: IEEE 46th vehicular technology conference, 1996. Mobile technology for the human race, vol 2. IEEE, pp 963–967, May 1996
5. Hu YC, Johnson DB, Perrig A (2003) SEAD: secure efficient distance vector routing for mobile wireless ad hoc networks. Ad Hoc Netw 1(1):175–192
6. Papadimitratos P, Haas ZJ (2003) Secure data transmission in mobile ad hoc network. In: 2nd ACM workshop on wireless security (WiSe), September 2003
7. Sanzgiri K, Dahill B, Levine BN, Shields C, Belding-Royer EM (2002) A secure routing protocol for ad hoc networks. In: 10th IEEE international conference on network protocols, Proceedings. IEEE, pp 78–87, November 2002

Improved Threat Models for the Security of Encrypted and Deniable File Systems

Michal Kedziora[1(✉)], Yang-Wai Chow[2], and Willy Susilo[2]

[1] Faculty of Computer Science and Management,
Wroclaw University of Science and Technology, Wroclaw, Poland
michal.kedziora@pwr.edu.pl

[2] School of Computing and Information Technology, University of Wollongong,
Wollongong, Australia
{caseyc,wsusilo}@uow.edu.au

Abstract. This paper analyzes current widely used threat models, against which Deniable File Systems (DFSs) can potentially be secured. We contend that previously presented models are no longer adequate due to the integration of mobile and cloud computing in today's devices and operating systems, as what this implies is a shift in forensic analysis paradigms and new forensic techniques to detect and analyze Deniable File Systems. We propose improved threat models against which DFS hidden volumes and hidden operating systems can potentially be secured, this includes One-Time Access, Multiple Access and Live Response Access. We also merge currently known attack vectors and propose new ones which were previously ignored in the increasingly outdated threat models. It is vital to develop new contemporary threat models for forensic analysis that cater for the current computing environment that incorporates the increasing use of mobile and cloud technology.

1 Introduction

A Deniable File System (DFS) is one where the existence of a portion of the file system can be hidden from view [5]. An example of such a system is where a person creates a regular (non-deniable) encrypted file system, which is protected by a password. Within this file system, the person can also create a deniable file system that is protected by a second password. This inner, deniable file system is referred to as a hidden volume, which is deniable because unless the person reveals the second password to an adversary, it should be impossible for that adversary to determine whether the regular encrypted file system contains an encrypted hidden volume [5].

The current most widely used security model against which Deniable File Systems (DFSs) can potentially be secured was described in Czeskis et al. [5]. According to them, threat models against which hidden encrypted volumes can potentially be secured are based on three situations. The first is One-Time Access, when the attacker has only one copy of the disk image containing a DFS volume. This is the worst case scenario. An example of this is when the police seize a device and make a binary copy of its data. Their second model is Intermittent Access. According to Czeskis et al. [5], this is when

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_24

an attacker has several copies of the evidence volume, taken at different times. An example is when border guards make a copy of a person's device every time the person enters or leaves the country. The third model is Regular Access, this is when an attacker has several copies of the evidence data made at short intervals. For example, when the police secretly enter a person's apartment every day while the person is away and make a copy of the device's contents each time.

There are several issues with these models. The purpose of One-Time Access was to focus on a situation where there is a chance to provide information about DFS via analysis of its algorithm and implementation. In addition, it was meant to deal with a situation when a binary copy of a hard disk containing a DFS was seized and analyzed. However, this situation is altered when investigators can also take a snapshot of the device's RAM before it is shut down. Furthermore, the current forensic shift is to analyzes live running systems remotely without shutting them down. This is not captured in current threat models and therefore misses important attack vectors on DFSs, which will be discussed later. The next issue with the One-Time Access model is that current operating systems have automatic backup functions of important files. This implies that the One-Time Access model often encroaches on the Intermittent or Regular Access models depending on the number of copies and backup intervals. In common forensic investigations, the One-Time Access model is severely affected if several copies of the DFS volume exists.

As for the Intermittent and Regular Access models, the purpose of these models was based on the ability to analyze changes both in the DFS and any side channel leaks that it creates. However, the interval in which copies are seized versus the number of copies, does not play a significant role in investigations to distinguish between these models. What is primarily missing and currently ignored is the important threat vectors based on live access to a device under investigation.

In practice, the security threat models of DFSs should closely relate to the digital forensic process. There are number of guidelines and procedures used to describe this process [1, 3, 14]. The emergence of ubiquitous mobile devices and operating systems with integrated backup functions, on-the-fly encryption, mobile and cloud integration, etc. has resulted in the traditional forensic model becoming increasingly obsolete. The live forensic approach was introduced as an alternative approach by adding live analysis to forensic procedures [9]. In addition, plausibly deniable encryption on mobile devices is a growing area of research [4, 17, 21]. In light of the growing mobile and wireless environment, the previous threat models do not address the requirements of this emerging landscape and thus should be revisited with improvements.

2 Deniable File Systems

Deniable encryption was introduced in 1997 by Canetti [2]. The idea is to be able to decrypt a cipher text into two plaintexts depending on the key that is provided. An additional requirement is to guarantee that the adversary cannot detect that a hidden message is present in the cipher text. The purpose is to protect against adversaries who can force a user to provide a password to decrypt the contents. Canetti [2] proposed a shared-key encryption scenario where the sender and receiver share a random secret key to encrypt message, as well as a fake shared key. This allows the encrypted message to be decrypted into two different plaintexts depending on which key was used.

Deniable cryptography for cloud storage was also introduced by Gasti et al. [7]. The concept of deniable cloud storage includes the privacy of data even when one's communication and storage can be opened by an adversary. They designed a sender-and-receiver deniable public-key encryption scheme and provided an implementation of a deniable shared file system.

The most common DFSs which are used in practice are based on the TrueCrypt implementation [5]. TrueCrypt is an on-the-fly encryption application which also implements DFS as hidden volumes which reside in an encrypted volume. The TrueCypt project was discontinued in 2014, but alternatives exist, e.g., VeraCrypt is most popular to date. VeraCrypt is an open source software used for on-the-fly encryption [13]. Its process is user transparent so data is encrypted right before it is saved and decrypted right after it is being loaded without any user intervention [18]. Plausible deniability in VeraCrypt supports hidden volumes and hidden operating systems. Encrypted and hidden volumes can also be used in mobile devices using mobile apps like Disk Decipher [16] and Crypto Disks [20].

3 Proposed Threat Models

This paper proposes improved threat models for the security of DFSs which addresses flaws and inconsistencies in the widely used model presented in Czeskis et al. [5], in light of the changing computing environment which incorporates mobile and cloud computing. The main drawback of the current model is fusing one-time access, which is meant to be a single copy, with the current trend of multiple archive copies of DFS volumes. The next issue with the existing models is that with the increasing number of copies and automatic backups which is characteristic of the modern computing environment, this muddles the distinction between the Intermittent and Regular Access models. This therefore results in an inability to practically employ these models in the current increasingly diverse computing environment. Part of the deficiency also lies in the fact that the traditional models fail to capture the live forensic approach, which has become the commonplace when handling live access to cloud and mobile data.

The proposed approach thus amalgamates the One-Time Access model with aspects of the intermittent and regular access models as a baseline for a single system access. This is separate from multiple access which incorporates differential analysis. A third model is proposed based on the live forensic approach, which we call live response access. This not only addresses live forensics, but is also associated with new types of DFS attacks based on cloud and network integrity of today's computer systems. Figure 1 depicts the proposed threat models; it can be seen that the proposed model incorporates the One-Time Access, Multiple Access and Live Response Access models along with their associated attack vectors respectively.

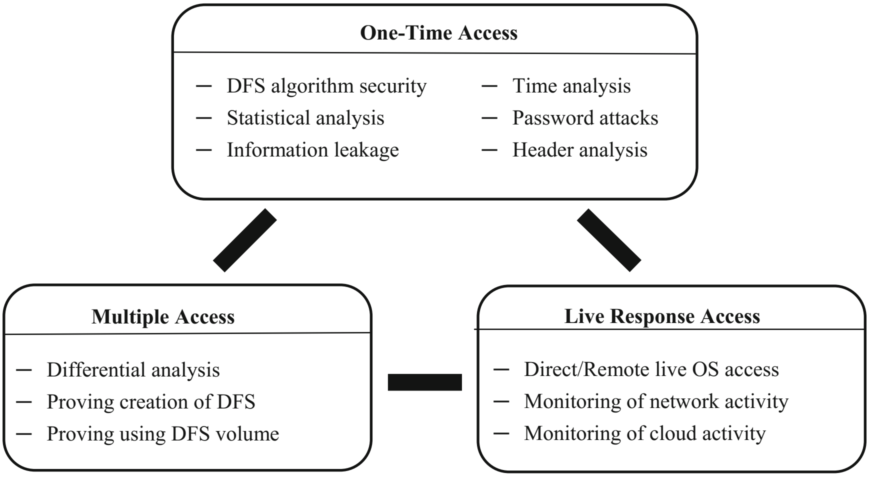


Fig. 1. Threat models and attack vectors on DFS.

4 Analysis and Discussion

This section discusses purpose of the proposed threat models and presents analysis on their significance to accommodate the current increasingly diverse computing environment, comprising of ubiquitous systems like mobile and cloud computing with their associated synchronization and auto backup features.

4.1 One-Time Access

The One-Time Access scenario is where an investigator is able to access one or more copies of a device containing only one copy of the DFS encrypted container. The most conservative variant of this model is when an investigator is able to seize and analyze forensic evidence of a binary image of an encrypted DFS volume. Two common situations are, for example, obtaining a binary copy of a hard drive encrypted with a DFS implementation like TrueCrypt/VeraCrypt, and retrieving a logical copy of the DFS encrypted container from a backup system. In either of these situations, the investigator's options are limited to analyzing the cover volume or the encrypted container itself.

The security of this is based on the cryptographic algorithm and the assumption that it can be formally and mathematically proven. However, in practice, DFSs are usually seized as a container file from a complex operating system. This results in the possibility of new attack vectors, in addition to the problem of detecting the DFS itself, as DFS implementations use encryption to hide deniable data together with encrypted cover data. Hence, all encrypted data found on an evidence device should be treated by the investigator as containing a DFS unless proven otherwise. While this problem is not commonly addressed in DFS related papers, it is very important from a forensic investigator's point of view. This was not only presented in our previous work [10], but also confirmed in Davies [6], where initial detection techniques are based on statistical detection of volumes by randomness testing. Statistical tests based on entropy, chi-square, arithmetic mean, Monte Carlo for Pi, serial correlation coefficient was used.

The main threat vector for DFS security in the One-Time Access model is information leakage, which can both compromise covert and hidden volumes. Information leakage through the operating system was first introduced in Czeskis et al. [5]. They gave an example of shortcut files that can point to data on the hidden volume or copies of hidden volume files saved in unencrypted area of disk, thus compromising its presence. The second main vector is in locating keys and password attacks against DFS. DFS systems based on TrueCrypt/VeraCrypt are only as strong as its password, which is a practical problem when many users do not comply with secure password usage policies. Furthermore, there are methods to grab passwords from memory of a running DFS volume.

Situations where an investigator can access more than one copy of a DFS volume [8], as well as the situation where an investigator can interact with a running system to find cryptographic keys should be excluded from the One-Time Access threat model. This is because the former scenario is captured in the Multiple Access model, while in the later scenario is modelled in the Live Response Access model, which are discussed in the sections to follow.

4.2 Multiple Access

A Multiple Access scenario is where an investigator has multiple device images containing multiple hidden encrypted containers. The main threat to DFSs in this case is differential analysis of hidden volumes, which can result in the ability to attack the plausible deniability attribute. This issue was first raised by Czeskis et al. [5], where they highlight that if disk snapshots can be obtained at close enough intervals, then the existence of any deniable files will be obvious, since seemingly random bytes on the hard drive will change. A practical implication of this was presented in Hargreaves and Chivers [8], where they described how hidden encrypted volumes can be detected and how their sizes can be estimated. Interesting research on detecting the creation of a DFS inside an encrypted container was presented by Jozwiak [11].

The Multiple Access model also involves the situation where more than one copy of a hidden volume can be retrieved from only one seized disk image. An example of this was presented by Hargreaves and Chivers [8], where they managed to obtain multiple copies of an encrypted container using the Shadow Copy function in the Windows Vista, Windows 7 and Windows 10 Operating Systems. Shadow Copy extends the Restore Point feature of Windows XP. The Shadow Copy feature is important for finding forensic artifacts during investigations as demonstrated by Purcell and Lang [15]. This situation is common in forensic investigations due to the standard usage of auto backup functions integrated in modern operation system including Shadow Copy and Time Machine for MacOS. The emergence of mobile and cloud computing with integrated backup also produces a source for obtaining multiple copies of DFS containers.

4.3 Live Response Access

This paper presents a new model defined as Live Response Access. Three main example scenarios for this model are where an investigator/attacker has:

1. direct/remote live access to the hosting Operating System (OS) running a DFS volume
2. direct/remote live access to DFS based hidden OS
3. access to the network environment within which a hidden OS is running, or has access to the cloud application in which the hidden OS is connected to

When Czeskis et al. [5] introduced their threat models against which a DFS could potentially be secured, forensics procedures typically involved the switching-off of computers and making a binary copy of the hard drive. Nowadays, much more effort is directed and focused toward the so called live forensics, whereby the main idea is to preserve volatile data which is mostly lost when a computer or mobile device is switched off [12]. Live response and memory analysis tools have the capabilities of collecting information from network connections, open ports and sockets, running processes, terminated processes, loaded DLLs, open files, OS kernel modules, process dumps, strings or user logs [19]. Each of these information sources can lead to compromising DFS by identifying a hidden volume disk area.

Although most of these techniques can also be used in the One-Time Access model, as volatile forensic artifacts related to hidden DFS volumes can be found in temporary system files as swap or hibernation files, it is more appropriate to extend this to the Live Response Access model. This is because it can lead to the scenario where an investigator has access to the host system, a common situation nowadays, which can generate new approaches and threats to DFS security.

A scenario that was ignored in previous models is securing a DFS when an investigator or an attacker has access to the hidden volume or the hidden operating system while it is running. The reason why this scenario was ignored was because the DFS was assumed to have a more secure encryption. However, this has changed with the hidden operating system option when it was implemented in TrueCrypt/VeraCrypt. This embraced the scenario where investigators could remotely use live response tools to have direct access to a working DFS operating system. In practice, it can be remote access via software like Team Viewer, VNC, Windows Remote Desktop or just physical access to the device. Another scenario is the running of the hidden operating system in a network environment, with the need to connect to third party mobile and cloud applications, which results in new possibilities for detecting that a system based on DFS is used.

5 Conclusion

This paper describes commonly used threat models against which Deniable File Systems can potentially be secured. With the advancements and progress in modern computer systems which include the integration of mobile and cloud solutions, the existing threat models are increasingly becoming obsolete. A new improved threat model based on One-Time Access, Multiple Access and Live Response Access is analyzed and discussed, which should supersede previous models as they have greater coverage of security issues faced by DFSs and hidden operating systems. This paper presents previously ignored scenarios and demonstrates new threat vectors for evaluating the security of DFSs and hidden operating systems. In view of the increasing probability of being able to access several copies of DFS volumes during digital investigations, this issue should be addressed by adopting new precautions or improving encryption algorithms to make it harder to perform cross data analysis, which has emerged as a major threat to the security of Deniable File Systems.

Acknowledgement. This work was undertaken with the financial support of a Thelxinoe grant in the context of the EMA2/S2 THELXINOE: Erasmus Euro-Oceanian Smart City Network project, grant reference number: 545783-EM-1-2013-1-ES-ERA MUNDUS-EMA22.

References

1. Baryamureeba MV, Tushabe F (2004) The enhanced digital investigation process. In: Digital forensic research workshop
2. Canetti R, Dwork C, Naor M, Ostrovsky R (1997) Deniable encryption. In: Kaliski BS Jr (ed) Advances in cryptology - CRYPTO 1997, Proceedings of the 17th Annual international cryptology conference, Santa Barbara, California, USA, 17–21 August 1997. Lecture notes in computer science, vol. 1294. Springer, pp 90–104
3. Carrier BD, Spafford EH (2003) Getting physical with the digital investigation process. IJDE 2(2):1–20
4. Chang B, Wang Z, Chen B, Zhang F (2015) Mobipluto: File system friendly deniable storage for mobile devices. In: Proceedings of the 31st annual computer security applications conference (ACSAC 2015). ACM, New York, pp 381–390
5. Czeskis A, Hilaire DJS, Koscher K, Gribble SD, Kohno T, Schneier B (2008) Defeating encrypted and deniable file systems: Truecrypt v5.1a and the case of the tattling OS and applications. In: Provos N (ed) Proceedings of the 3rd USENIX workshop on hot topics in security, HotSec 2008, 29 July 2008. USENIX Association, San Jose
6. Davies A (2014) A security analysis of truecrypt: Detecting hidden volumes and operating systems a security analysis of truecrypt: Detecting hidden volumes and operating systems. Information Security Group, Royal Holloway, University of London
7. Gasti P, Ateniese G, Blanton M (2010) Deniable cloud storage: Sharing files via public-key deniability. In: Al-Shaer E, Frikken KB (eds) Proceedings of the 2010 ACM workshop on privacy in the electronic society, WPES 2010, Chicago, Illinois, USA, 4 October 2010. ACM, pp 31–42
8. Hargreaves C, Chivers H (2010) Detecting hidden encrypted volumes. Springer, Heidelberg, pp 233–244
9. Hay B, Bishop M, Nance K (2009) Live analysis: Progress and challenges. IEEE Secur Priv 7(2):30–37
10. Jozwiak I, Kedziora M, Melinska (2011) Theoretical and practical aspects of encrypted containers detection - digital forensics approach. Springer, Heidelberg, pp 75–85

11. Jozwiak I, Kedziora M, Melinska A (2013) Methods for detecting and analyzing hidden FAT32 volumes created with the use of cryptographic tools. In: Zamojski W, Mazurkiewicz J, Sugier J, Walkowiak T, Kacprzyk J (eds) New results in dependability and computer systems - Proceedings of the 8th international conference on dependability and complex systems DepCoS-RELCOMEX. Advances in intelligent systems and computing, 9–13 September 2013, Brunow, Poland, vol 224. Springer, pp 237–244
12. Lessing M, von Solms B (2008) Live forensic acquisition as alternative to traditional forensic process. In: IT-incidents management & IT-forensics - IMF 2008, conference proceedings, 23–25 September 2008, Mannheim, Germany, pp 107–124
13. Loginova N, Trofimenko E, Zadereyko O, Chanyshev R (2016) Program-technical aspects of encryption protection of users' data. In: 2016 13th international conference on modern problems of radio engineering, telecommunications and computer science (TCSET), pp 443–445
14. N.I. of Justice (U.S.) (2004) Forensic examination of digital evidence: a guide for law enforcement. NIJ special report. U.S. Dept. of Justice, Office of Justice Programs, National Institute of Justice
15. Purcell DM, Lang S-D (2008) Forensic artifacts of microsoft windows vista system. Springer, Heidelberg, pp 304–319
16. Huveneers R. Disk Decipher. http://disk-decipher.hekkihek.nl/
17. Skillen A, Mannan M (2014) Mobiflage: Deniable storage encryption for mobile devices. IEEE Trans Dependable Secure Comput 11(3):224–237
18. VeraCrypt. VeraCrypt Documentation. http://veracrypt.codeplex.com/documentation
19. Waits C, Akinyele J, Nolan R, Rogers L (2008) Computer forensics: Results of live response inquiry vs. memory image analysis. Technical Report CMU/SEI-2008-TN-017. Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA
20. Zeng Y, Crypto Disks. https://itunes.apple.com/us/app/crypto-disks-disk-encryption/id889549308?mt=8
21. Yu X, Chen B, Wang Z, Chang B, Zhu WT, Jing J (2014) MobiHydra: Pragmatic and multi-level plausibly deniable encryption storage for mobile devices. Springer, Cham, pp 555–567

Performance Evaluations Power Consumption, and Heterogeneousity of WSNs in Medical Field

Reem Altaharwa(✉), Sameem Abdulkareem, and Ali Mohammed Mansoor

Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia
taharwa@siswa.um.edu.my, {sameem,ali.mansoor}@um.edu.my
www.sites.google.com/site/drreemcs/

Abstract. With the rapid development of technology, and the prevalence of the aging population, researchers are focusing on how these technology can aid medical care, especially, the care of older people. Thus, the technologist strife to develop sensors and other peripherals to address the demands in our life, and to achieve patients' satisfaction. On the other hand, clinical staff like doctors and nurses should be able to handle technology in as simple a way as possible. Nowadays we able to communicate with our peripheral environment by using different sensors and gateways. The aim of this paper is to report and survey the main applications of wireless sensors networks, which are power efficient and heterogeneous in the medical field. We attempt to show the relationship and collaboration between healthcare, engineering and the computer science fields, we will illustrate the new technologies, how they are evaluated, and what are the simulators and hardwires they use. The advantage for the development of sensors and communications and using heterogeneous at medical field make the monitoring easier, faster and efficiency.

Keywords: Wireless Body Area Network · Energy efficiency · Heterogeneous · m-health · e-health

1 Introduction

Sensors were first used for traceability and logical purposes. Subsequently, sensors were used in search and rescue operations and other emergencies, such as, monitoring operations like forest fires and volcano eruptions [1]. Sensors are also beginning to be used in medical care, especially in monitoring patients in the comfort of their own homes. In the last decade, falling fertility rates and remarkable increases in life expectancy has seen an increase in the number of people aged 65 and older. And this trend will continue and even accelerate. Wherever we go, we notice over-crowdedness, especially when it comes to the emergency units in hospitals. In some cases, the staff cannot provide enough assistance to all patients. In the urban area, it is difficult for residents to get to hospital. Under these circumstances, researchers have proposed and developed technologies that can monitor the patient at home and to take the necessary clinical tests wirelessly using wireless sensors. The Wireless Body Area Network (WBAN) idiom is mostly

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_25

widespread in the medical field. There are many idioms also known like Body Area Network (BAN), Human Body Communication (HBC), Medical Implant Communications Services (MICS), Industrial Scientific and Medical (ITU) [2].

Most devices are worn, and in some cases implanted in the body, thus, requiring surgery. In order to avoid or decrease the surgeries, researchers seek to increase batteries life time, and to enhance their functionality [3]. While other authors, concerned by the m-health [4, 5].

The paper is organized as follows. Section 2 discusses related work while Sect. 3 details the challenges, problems and solutions of this field of research highlighting some case studies and applications in Sect. 4. Summarize can be found in Sect. 5, and finally the conclusion in Sect. 6.

2 Related Work

Problems relating to the aging population in the current decade means that health organizations must provide them with health care and follow up, to ensure that, they can participate in the social community without any risk or fear. These regulations and systems will depend on many factors like Information and Communication Technologies (ICT), personal health, social network, work and economic situation. Generally, the ICT can play a role in two scenarios, namely, indoor monitoring scenario/home monitoring and the outdoor monitoring scenario. The indoor scenario is more realistic, taking advantage of the comfort and security of the home, while, the outdoor scenario provide control over the mobility of the patient. The sensors in medical devices are user friendly, portable or wearable, equipped with long range and small range wireless technology. Some health devices are used once or twice daily like glucose meter, blood pressure and weighing scales, on the other hand, some devices need real time monitoring process like electrocardiogram monitors. Authors in [4] concentrated on m-health and e-health framework of the urban and rural patient. Focusing on how to improve the sensors based on data update, size, and cost, reduction of power consumption, ease of mobility, and ease of use. Firstly, they define the e-Health as: using electronic devices and digital communication network to collect, save, retrieve, transfer, and exchange medical information between doctors and users for the achievement of the best diagnosis and intervention in emergency cases. The importance of wireless sensors e-Health, is that it is able to transfer the sensor data to mobiles, or any portable devices, where the data will be saved and sometime transfer depending on the data type. This is the benefit of IoT usage in the medical field, as stated in [6] expect the use of clouding in the near future, which will make the data available everywhere at any time. The future of healthcare will increase the quality of life for the people around the world, and will offer fast response from medical staff in emergency cases. Recently researchers, in Taiwan use clouding with their VMware workstation in their online electronic health information as a testing database, as the data is huge and needed real-time dynamic calculations [7]. At the same time, other researchers develop a new algorithm for three-dimensional ray launching (3D RL).under the android platform for wireless body area networks (WBANs) [6]. Researchers in [8], are concerned with the size of images and the delay time to transmit the images after compressed it in a heterogeneous telemedicine network. Furthermore, Athanassion and David worked on the challenges in Body Area Networks (BAN), such as how intelligence and relays can be used and to re-transmit power control and reduce outages in the MAC layer and the IEEE 802.15.6. [9]. The authors in [10] discussed Surface Electromyographic (EMG) and electrocardiographic (ECG) signals and its importance for a person's health status, body posture, fitness level, and physical performance were done by researchers in [9]. In [11, 12] the discussions were on EMG and ECG signals and how they can be used for monitoring driver's health and cognitive states for road safety improvement. The authors in [13] discussed how to calculate muscle fatigue value based on the estimation of the EMG signal root mean square, estimated during the phases of the exercise. In [14] they studied the potential of implantable Sensors in the ultrasonic network. While in [15] discussed electronic applications to reduce the documented papers. Other researchers concerned with the power consumption using NS2 [16]. Tung, Tsang, Tung, Chui, & Chi, proposed and implemented a dual radio ZigBee homecare gateway (DR-ZHG) to support remote patient monitoring. The dual radio implementation increases the transmission data rate of ZigBee and guarantees low latency and highly accurate telehealth service at home. The DR-ZHG supports seven polling service sensors but only one streaming service sensor [17].

3 Challenges and the Solutions

Nowadays technology is an essential part in all fields, and customers always demand the best performance. This optimization depends on what they need exactly, but, ultimately, most people look for the fastest and most reliable, smallest in size, clearest and easiest to use and latest update, not to mention the lowest in price, and the most secure and private gadgets. These points are considered challenges to researchers, and we illustrate in detail in these points:

3.1 The Cost

The end user, the patient, usually considers the cost and whether he can afford it. Thus, budget is considered as a one of the main priorities for the effective use of technology. The researchers, addressed this limitation in their work [5].

3.2 The Power

Reducing the energy consumption is a very difficult task, as there is no energy function to optimize. The researchers in [14], concerned with power consumption and focus their work on energy consumed by each node in the WSN. The researchers in [4] used 1.5 V AAA batteries to increase the capacity and power of their social sensors. Researchers in [7] compress medical images to reduce its size, and therefore, reduce the energy in medical image heterogeneity [8].

3.3 The Size

In social sensors, 1.5 V AAA batteries, were used by researchers in [4] to reduce the power supply size and weight. As mentioned previously, in medical image heterogeneity, researchers in [8] compressed images reducing the space in memory the overall size.

3.4 The Response Time

Also used Representational State Transfer (REST) in web-service layer to reduce the latency between the client and server in social and medical sensors [4]. While in heterogeneous telemedicine network the researchers in [7] were concerned with the transition delay and used compressive technique to reduce the size, and for that reason they increase the speed of transmission and save time.

3.5 The Mobility

Mobility is an important character in wireless sensor networks, so what is the available domain for these sensors? Aguirre, LopezIturri, et al., mentioned mobility in their study. They presented a new algorithm, the three-dimensional ray launching (3D RL). Which works under the android application. They put three sensors on the body at different areas, namely, on the chest, wrist and ankle, then collect the information using wireless techniques like ZigBee, Bluetooth, and Radio Frequency Identification. The sensors should be able to be moved around, because when it implanted in the body or worn, it will move when the patient moves. So, it should work and transfer the data during this time of movement [6].

3.6 The Environment

The environment is related to where the sensors will be placed, inside the body, outside the body, in a special place in the room, hospital, or bathroom. Therefore, the researchers in [15] focused on implanted sensors placed inside the body. The biggest challenge that they faced is how the sensors will adapt in this environment. So, they proposed new sensor which contain more than 60% water, to fit with cells within the human body.

3.7 Security and Privacy

Patient's privacy and confidentiality is an important aspect in medical records. This is a big challenge in wireless sensor networks in the medical field [18]. However the data should be easily accessed by doctors and a healthcare giver staff. Thus, the researcher added some technique to authenticate records, and protect patients' data. Also, some Viruses or malware if access to these data will change it or damage it, so, Lafayette & Jha, also proposed two techniques, Secure Execution Environment and Run-time Monitoring to secure the data and protect them against the damage that may be caused by

Viruses or malware. These techniques are: Secure Execution Environment and Run-time Monitoring [19].

3.8 Platform

Related to hardware and software, WiSE application has three tabs: Home, Training, and results. In addition, there is a menu for setting like audio and connection. Working on smartphone or tablet. And connected Wi-Fi to pc controller [5]. The details in section four.

4 Case Studies and Applications

In this section, we discussed the four most important case studies. We will start by the latest study at this time, which use NS2 simulator. Researchers, concern with the power consumption. As they mention the energy consumption in WSN include many parts: individual constituent, a global constituent, a local constituent, a sink constituent and an environmental constituent. The individual constituent contains the programmable unit and main controllable. The global constituent is interested with the maintenance in widespread network. The local constituent works in the beginning keeping the communication between the neighboring nodes. The sink constituent is responsible for the energy used by a node to communicate with the sink. While the environmental constituent is concerned with the energy collected from the environment. Researchers in [14] focus their work on energy consumed by each node.

The second study was by Aguirre, Led, et al., who used a new system and to validate it in the living lab environment supported by NASISTIC (Navarra-ASISte-TIC). This system use a Bluetooth transceiver technique, which transfers the signals through wireless channels. The new system activates the relationship between the end user (patient) and the social/medical staff. This system focused mainly on provision of tele-healthcare for citizens like people suffering mental illness or elderly. They monitor people with wireless social sensors, placed in different places in the house. These social sensors contain three main parts: first, wireless communication, which is, divided to two models; transmitter and receiver. It uses Bluetooth technology, it is responsible of establishing the connection flow control, and transmitting the data. Second, the Microcontroller, receives RSSI and BER from the wireless communication part and transfer them to a laptop or I-pad to display and/or store them. Third part: Power supply, this part depends on low voltage-low power (LV-LP) with DC-DC regulator high efficiency. They use two 1.5 V AAA batteries. The general architecture for these sensors is divided into two layers; transport layer and application layer. The transport layer establishes the communication using Bluetooth technology, and it has two protocols to information exchange. The first protocol is JAMP (JSON Agent Management Protocol). The second one is a manufacturer protocol and this type is used in medical devices and cannot be changed or updated. The application layer consists of different models like the user interface and storage module, and sensors module which collect and measure information. The cooperation between these two layers is managed by the Kernel. The main part in the software for the gateway is the core, the core manages the communication between layers. The second part is the plugins, this part includes all software modules which can be connected to devices. The plugin contains three main parts: the Agent, which, wraps the connection between the source and destination. The Transport, has many protocols and techniques like the HTTP, TCP/IP, BT. and the third part is the manager, which register the communication, allow the client to create a communication, and listen to new communication. The back-end software architecture; has four blocks. The spring framework and Spring Model-View-controller that which gives the user speed up connection with web applications. The View layer, transmits the data to the end user, it uses the AJAX framework to develop the user interface. The Controller Layer validates the data and chose the best view to show it. The last block is the data access layer, which includes two blocks: The Spring Object Relational Mapping (ORM), and Java Persistence API, which deal with object mapping and persistence. On the other hand, the web-server layer use the Representational State Transfer (REST) because it is easy to implement, and it decreases the response time between the client and server. In their study, Aguirre, Erik Led, Santiago Lopez-Iturri, Peio Azpilicueta, Leyre Serrano, Luís Falcone, Francisco, use a grid of cuboid 13 m, 7 m, 4.2 m, and utilized six sensors, placed on different places like - chairs, tables, windows, and shelves. while, some sensors were wearable on chest [4].

In the third case study of the WISE, the system acquires fitness metrics from electrocardiographic (ECG) and surface electromyography (sEMG) signals using several ultralight wireless sensing nodes. These signals are transmitted to one BS or more, through 2.4 GHz radio link through IEEE 802.15.4 protocol. Each base station is connected through USB link to a control PC running software for viewing, analyzing, and recording the data. The system is combined with a smart phone application. The most important feature for this system is: Cheap, flexible, easy to configure, easy to wear, lightweight mobile node, low cost electronic, easy to capture and process further biological signals. The application has three tabs: Home, Training, and results. Also, there is a menu for setting like audio and connection. The WiSE application on smartphone will register the metric and signal from the user and transmit it to PC controlled by Wi-Fi technology [5].

Finally, Wang, Chen, Kuo, Chen, & Shiu, focused in their study on cloud computing, wireless sensor network and communication in medical field to improve health information technology. In their study, they build a mobile web App combining two health information service functions: a collaborative recommender and a physiological indicator-based recommender. They used online electronic health information in Taiwan as a database for their study testing data base, which included very huge amount of data and needed a real-time dynamic calculations. So, they used a cloud VMware workstation to setup their system programming scripts like MySQL, PHP, HTML5, JavaScript, CSS, and jQuery. They also used a PhoneGap package to program the application. The PhoneGap supported the android and a windows phone operating system. They suggested a hybrid predictive model, which mix between Markov chain and Grey theory. These theory and algorithm dropped the cost of pursuit errors and therefore prolonged the network life time. Also, they applied a Collaborative Filtering (CF) technique to make the health information more efficient. This system can deal with a massive data, reduce the time consumption in search operation, and develop referral service. This system focus on preventive health information. In this scenario, the user will use the Fusion of Rough-Set and Average-category-rating (FRSA) to calculate the user health information and transmit physiological data signal to cloud database records using the WSN. The system makes a recommendation on the treatment/intervention required for the user based on the health information and the user's medical condition. A questionnaire and Partial Least Squares (PLS) were to analyze the data. In this study, all experiments were laboratory-based, with a future goal to expand it to the home or remote medical center. However, for the system to work in a global environment, the level of security needs to be increased. Moreover, they can observe the behaviors of user for a longer time, and save the recorded data to predict matching cases. The result of the experiment and measurement showed that, the system was satisfactory, useful, trustful, valuable, and confirmed the expectation [7].

5 Summarize

Technological development has entered all fields of science and invaded all aspects of our life, leading us to Internet of Things and allowing us to use e-medical and m-medical applications in healthcare. We use the mobile medical (m-medical) as a main device to run the software and medical applications, and a gateway to communicate between the back-end system and the sensors. Social and medical sensors are used together for the purposes of an integrated system. Figure 1 show a general architecture for social and medical sensors. As can be seen in Fig. 1 the family members make use of both the medical and social sensors to initiate a communication through the network, the communication will be transmitted to gateway by Bluetooth or ZigBee or Wi-Fi. Then it will be transferred to servers or Pcs using 3G or 4G. The staff and doctors can respond to the query/problem and diagnose the patient. However, we will need the technicians to maintain and update the hardware, devices, network and software. In the next section, we compare the recent studies done in the area and categories them depending on the

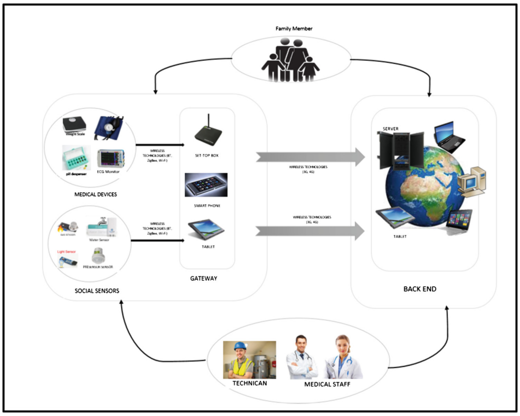


Fig. 1. General architecture for social and medical sensors

Technology used, such as, if it is wearable, and if it is indoor or outdoor, and if it uses an application on smartphones and summarized the findings in Table (1).

Table 1. compression between different techniques

Reference	Technology	Wearable	Indoor	Outdoor	Mobility	Use smartphone
[5]	IEEE 802.15.4	Yes	Yes	Yes	Yes	Yes (WiSE)
[20]	ZigBee	Yes	Yes	No	Yes	No
[17]	dual radio ZigBee	No	Yes	No	No	No
[13]	wireless electromyogr aph system	Yes	No	Yes	Yes	No
[4]	Bluetooth	No	Yes	No	No	No
[6]	ZigBee	Yes/No	Yes	No	Yes	Yes
[21]	Wi-Fi via HTTP, web server	Yes/No	Yes	Yes	No	Yes
[22]	6LoWPAN protocol	No	Yes	No	No	No
[3]	dry and wet EEG electrode Wi-Fi	Yes	Yes	No	No	No
[14]	ultrasonic waves (UsWB)	Yes	Yes	No	No	No
[15]	code blue	Yes/No	Yes	No	No	Yes
[8]	L2CAP, RFCOMM, BT, Wi-Fi	No	Yes	No	No	No
[7]	Wi-Fi and web service	Yes/No	Yes	Yes	Yes	Yes

6 Conclusions

This paper, emphasize the wireless sensors used in the medical field by studying some cases in different areas, and using different technology. We discussed energy consumption in the medical field, also, compared between different techniques according to the wearable, indoor, outdoor and the utilization of smartphones' applications.

The main goal of all these studies is to facilitate the medical service, using wireless sensors networks, and to help elder people and others who need healthcare from their residents. At the same time, these technology and progress in this field helps doctors, medical centers and technicians to save time and efforts and avoid over crowdedness in hospitals and medical centers. Also, we talked about the advantage for the development of sensors and communications and using heterogeneous at medical field which make the monitoring easier, faster and efficiency.

However, our study covers different areas in WSN in the medical field, there are a lot more that deserve further exploration in the future. These issues include system security, patients' confidentiality and privacy, and protection of the hardware and software from viruses and malware.

References

1. Tonneau AS, Mitton N, Vandaele J (2015) How to choose an experimentation platform for wireless sensor networks? A survey on static and mobile wireless sensor network experimentation facilities. Ad Hoc Netw 30:115–127
2. Cavallari R, Martelli F, Rosini R, Buratti C, Verdone R (2014) A survey on wireless body area networks: technologies and design challenges. IEEE Commun. Surv. Tutorials PP(99):1–23
3. Hasan MK, Mondal C, Al Mahmud N, Ahmad M (2016) Performance analysis of SSVEP based wireless Brain computer Interface for wet and dry electrode. In: 3rd international conference on advances in electrical engineering, ICAEE 2015, pp 64–67
4. Aguirre E, Led S, Lopez-Iturri P, Azpilicueta L, Serrano L, Falcone F (2016) Implementation of context aware e-health environments based on social sensor networks. Sensors (Switzerland) 16(3):310
5. Biagetti G, Crippa P, Falaschetti L, Member S, Orcioni S, Turchetti C (2016) Wireless surface electromyograph and electrocardiograph system on 802.15.4. IEEE Trans Consum Electron 62(3):258–266
6. Aguirre E et al (2016) Design and performance analysis of wireless body area networks in complex indoor e-Health hospital environments for patient remote monitoring. Int J Distrib Sens Netw 12(9)
7. Wang S-L, Chen YL, Kuo AM-H, Chen H-M, Shiu YS (2016) Design and evaluation of a cloud-based Mobile Health Information Recommendation system on wireless sensor networks. Comput Electr Eng 49:221–235
8. Tiwari V, Bansod PP, Kumar A (2016) Medical imaging in heterogeneous telemedicine network. In: 2015 10th international conference on information, communications signal process, ICICS 2015
9. Boulis A, Smith D, Miniutti D, Libman L, Tselishchev Y (2012) Challenges in body area networks for healthcare: The MAC. IEEE Commun Mag 50(5):100–116
10. Biagetti G, Crippa P, Falaschetti L, Orcioni S, Turchetti C (2015) A rule based framework for smart training using sEMG signal, vol 39. DII-Dipartimento di Ingegneria Dell'Informazione, Università Politecnica delle Marche: Springer Science and Business Media Deutschland GmbH
11. Oh J, Kwon M, Kim Y, Kim J, Lee S, Kim J (2013) Development and evaluation of myoelectric driving interface. In: Digest of technical papers - IEEE international conference on consumer electronics, pp 248–249
12. Son J, Kim B, Park M (2015) Lumbar cushion based real-time ECG sensing system for monitoring driver's state. In: 2015 international conference on consumer electronics, ICCE 2015, no 14, pp 261–262
13. Biagetti G, Crippa P, Orcioni S, Turchetti C (2017) Homomorphic deconvolution for MUAP estimation from surface EMG signals. IEEE J Biomed Heal Inf 1(2):2168–2194

14. Santagati GE, Melodia T, Galluccio L, Palazzo S (2013) Distributed MAC and rate adaptation for ultrasonically networked implantable sensors. In: IEEE international conference on sensing, communication and networking, SECON 2013, pp 104–112
15. Bokhari W, Patel VL, Sen A, Amresh A (2016) Development and use of a tablet-based resuscitation sheet for improving outcomes during intensive patient care. In: Proceedings of the 6th international conference on digital health conference, pp 17–21
16. Edoh TOC, Atchome A, Pawar P (2016) Simulation of energy consumption in a multi-tier heterogeneous sensor network for patient monitoring simulation using NS2-simulation-tool. IEEE xplore (2016)
17. Tung HY, Tsang KF, Tung HC, Chui KT, Chi HR (2013) The design of dual radio ZigBee homecare gateway for remote patient monitoring. IEEE Trans Consum Electron 59(4): 756–764
18. Rushanan M, Rubin AD, Kune DF, Swanson CM (2014) SoK: security and privacy in implantable medical devices and body area networks. In: 2014 IEEE Symposium on Security and Privacy, pp 524–539
19. Lafayette W, Jha NK (2013) Towards trustworthy medical devices and body area networks. ACM 978-1-4503-2071-9/13/05
20. Kobayashi H (2013) EMG/ECG acquisition system with online adjustable parameters using ZigBee wireless technology. Electron Commun Japan 96(5):1–10
21. Lomotey RK, Deters R (2012) Using a cloud-centric middleware to enable mobile hosting of web services. Procedia Comput Sci 10:634–641
22. Oliveira L, Rodrigues J, De Sousa A, Lloret J (2013) A network access control framework for 6LoWPAN networks. Sensors 13(1):1210–1230

3D Integral Imaging Based Augmented Reality with Deep Learning Implemented by Faster R-CNN

Richard Evan Sutanto, Lenny Pribadi, and Sukho Lee(✉)

Department of Software Engineering, Dongseo University,
47 Jurye-ro, Sasang-gu, Busan 47011, Korea
petrasuk@gmail.com

Abstract. In this paper, we propose a Marker-less AR method for 3D integral imaging displayed above the smart phone via an integral lenslet array. The marker-less AR is implemented by using the Deep Learning method to recognize an object detected by the smart phone device with Android platform. To obtain a real-time recognition speed we use the Faster R-CNN algorithm for the object recognition task. This system will start from the Android device that captures the image of the object the user wants to detect and sends it to the server. After the server receives the image, it starts to process the object recognition and saves the result into a database. When the database gets updated, the server sends back a feedback to the Android device. In the android system, a video file related to the content of the object it recognized begins to be played on the Android device. The video file is pre-processed so that it will appear as a 3D content when it is seen through a 3D lenslet array case which is covered above the Android device display. The integral imaging algorithm makes the pre-processed 3D content to produce a pop-up 3D/hologram. Applying this proposed method could make an application become more compatible with non-high-specs device.

Keywords: Integral imaging · Deep learning · 3D display · Augmented reality

1 Introduction

Recently, Augmented Reality (AR) and Virtual Reality (VR) become a hot topic again for research, and even more in commercial. Because of that, developer started to make more application for AR and VR. For Augmented Reality (AR) alone, there are 2 different types, Marker-based and Marker-less AR. In order to make an application with variety interaction and using unassigned object, the Marker-less method is preferred. The marker-less has to recognize an object without markers, which can be done by using a neural-network pre-trained by a deep learning algorithm. The Faster R-CNN algorithm has been developed to obtain a real-time recognition speed [1] which will be used in our implementation for object recognition task. This system will start from the android device that capture an image that the user wants to detect and will be sent to a server. After the server receives the image, it will start to process the object recognition and then save the result into a database. When the database gets updated, the server will send a feedback to the Android device. This will act as a signal, which starts an movie clip

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_26

to appear on the Android device's display. The movie clip is preprocessed such the every frame in the movie clip is an integral image which suits to the parameters to the lenslet array case covered on the smartphone display. When looking at the integral image through the lenslet array, the content in the movie clip appears as a 3D content above the phone. The lenslet array case is just made by plastic which can be reproduced for very low cost. Thus, the smartphone user can enjoy a 3D content without having to buy a costly 3D device.

In Sect. 2, we explain the related works to the proposed method, then, in Sect. 3, we introduce our implemented system for the low cost 3D integral imaging system. All the computation will be execute on the server, hence it will make the procedure in the client side faster and do not trouble their smartphone.

2 Related Works

In this section, we explain the related works of the proposed algorithm.

2.1 Integral Imaging

The 3D case is made based on the integral imaging algorithm. Figure 1 shows the concept of a smartphone device covered by an optical lenticular filter made in a plastic case form. The lenticular filter is a lenslet array which makes the underlying integral image appear as a 3D content above the smartphone.

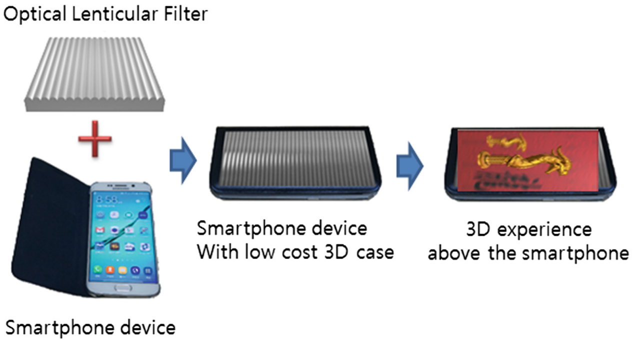


Fig. 1. Concept of the 3D augmented experience above the smartphone display by a low cost 3D case with lenslet array.

For this, the several parameters of the lenslet array should match the underlying integral image and also the parameters of the smartphone display. To understand the parameters to be adjusted we introduce the concept of the integral imaging. Integral imaging is initially proposed by Professor Gabriel Lippmann [2]. It can display a 3D image without using special glasses for the viewer. An array of lenses is placed in front of the image, where each lens looks different views depending on its viewing angle [3].

Therefore, the entire images through the lens-array are integrated, and expressed by 3D image (Fig. 2).

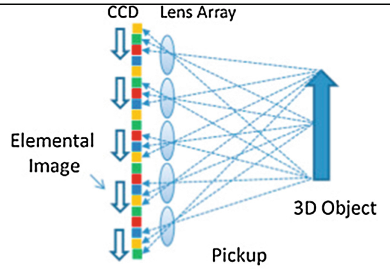
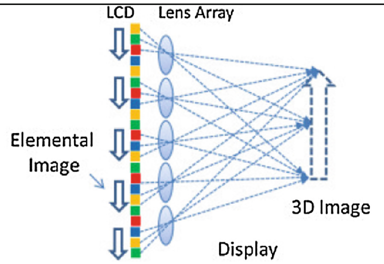
	The 3D object is captured by a lens array as individual elemental images recorded on the imaging plane which is located behind the lens array
	For display, an LCD is used to display the previously recorded elemental images to reproduce the original routes of rays irradiated from the 3D object, causing a 3D image to be formed.

Fig. 2. The formation of the elemental images into a single 3D image.

There are some characteristic equations about integral imaging. Let p be the diameter of the lens, δ be the display panel's pixel size, g the distance between the display panel and lens-array. Then, the viewing angle Ψ is defined by

$$\tan\left(\frac{\psi}{2}\right) = \frac{p}{2g} \qquad (1)$$

Here, the relationship of g with the focal distance f and distance between lens and image plane a is described as:

$$\frac{1}{f} = \frac{1}{g} + \frac{1}{a} \qquad (2)$$

Typically g is greater than f. If f and g is same, a becomes infinity. It implies that the image depth is infinity. In contrast, f is greater than g, image plane will be create at behind of the display position. This shows that the viewing angle of the 3D object is related to the parameters of the lenslets on the 3D case. In this research, the size of the 3D case that will be used is 218 mm × 127 mm × 10 mm and for lenticular sheet in the case, focal length is 2.2 mm, line per inch is 40 and the size is 212 mm × 123 mm (Fig. 3).

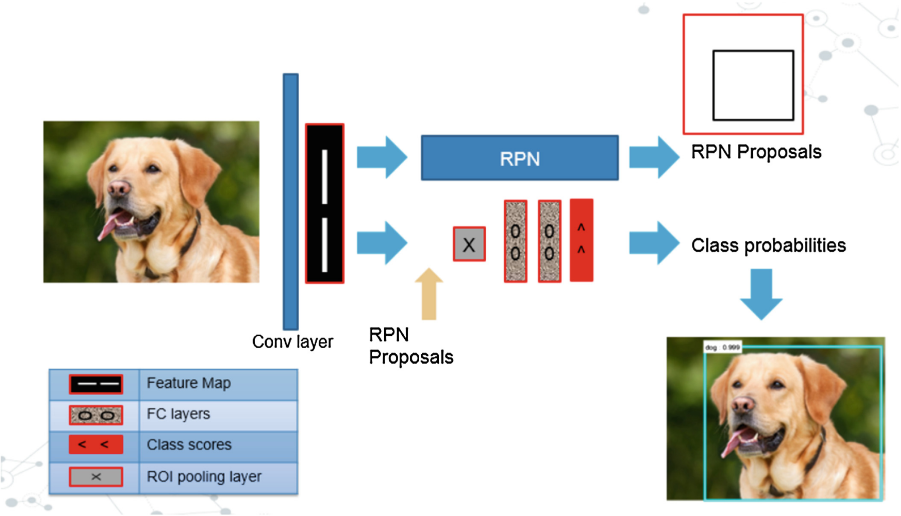


Fig. 3. Concept of the Faster R-CNN model.

2.2 Faster RCNN

In this research, Faster R-CNN pre-trained model will be used because it is already can recognized 20 objects and takes only around 0.3 s to recognized one image.

Faster R-CNN is a proposed method that gives better performance and accuracy in order to do object recognition task. By combining Region Proposal Network and Fast R-CNN into a single network and shared their convolutional features [1]. Main concept of Faster R-CNN is cutting down computing cost for proposals to be smaller by sharing convolutions. And also by putting several convolutional layers into modified RPN, it will revert to region bound and calculate the object score in the same time at each location.

3 Proposed System

The system architecture which combines all those methods is shown in Fig. 4. The application will be start by taking picture and send to server. After the server get the image, it will run object recognition by using deep learning and send result back to the user. On the user side, android will show the 3D image based on the result of object recognition. There are two main parts for this application, android side and server side. In this research, the laptop will act as a server by using xampp. The server functioning as the databases for the image results and executes Matlab to start the recognition. On Android side, application will taking the picture and sends it to server. After that, server will send back the result to smartphone, and android will display the result according to object recognition that has been save in database.

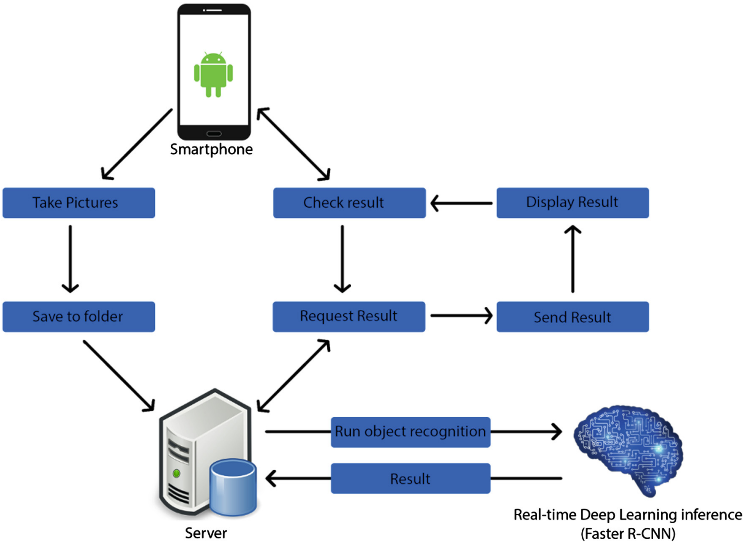


Fig. 4. System Architecture.

Figure 5 shows the algorithm flow for the image recognition method based on the client-server model explained above.

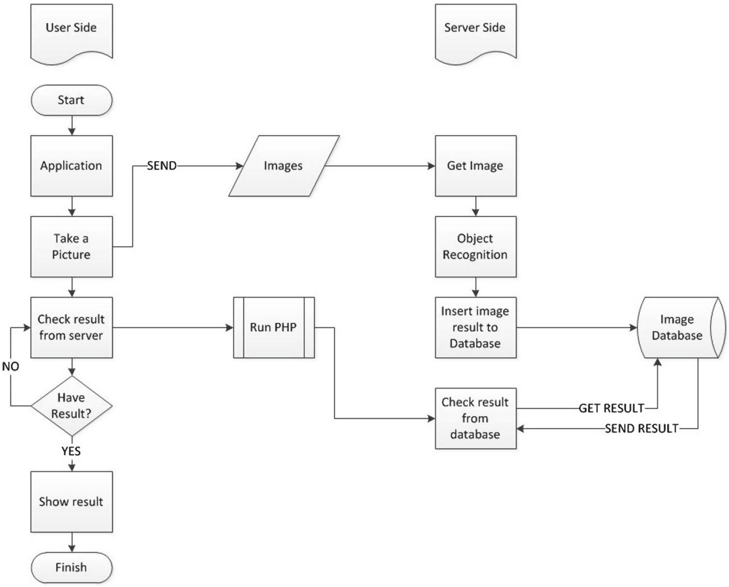
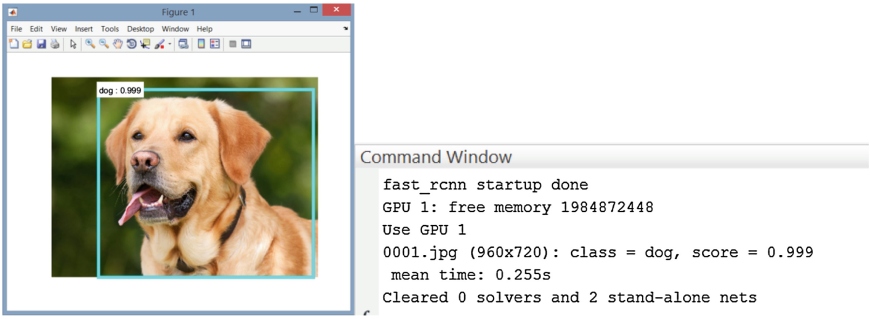


Fig. 5. Flowchart of the client-server model based image recognition implementation

To develop this research, we used a laptop with setup of Intel® Core i7-4720HQ ~ 2.6 GHz processor, 64-bit operating system, and 12.0 GB memory. The operating system is using Windows 8.1, which has the strongest compatibility now and using graphic card NVIDIA GeForce GTX 950 M 8 GB. Development environment is using android studio 2.1.2. Furthermore, this project needs to use minimal android SDK version 19, but the author compile this application using version 23 for android SDK. The reason to use that version is because of the http connection, in the new SDK version, it already change some syntax for connection and it became more stable. The server uses Xampp with PHP inside and this project use Matlab to run the Deep Learning for object recognition, and the Matlab program needs to use a MySQL-connector-java library to connect to database so it can update result of the recognition. Lastly, we used a Teclast X80 Plus Tablet PC that support both windows and Android 5.1 for testing (Fig. 6).



```
fast_rcnn startup done
GPU 1: free memory 1984872448
Use GPU 1
0001.jpg (960x720): class = dog, score = 0.999
 mean time: 0.255s
Cleared 0 solvers and 2 stand-alone nets
```

Fig. 6. Showing the recognition result on the server side.

The android device has a function to take the picture and then upload it to server and check database for result. The image sent to server will be put into one folder and then server will run Matlab to start object recognition method. If it recognizes some result, Matlab will update database according to the recognition result. And if the android detects result from server, it will start to display the object that recognized in another form.

This step will start when the server already success to get the image, the PHP will execute code to run Matlab. In Matlab, the system is using Faster R-CNN method to do object recognition task. The model that used in Faster R-CNN is pre-trained model, so it will start to detect the folder that already set from server. Images that have been accessed will be tested one by one until finish. Result of those images will be printed out in the command window of Matlab and those values will be sent to database system for update the result.

Same as in server side, this step also starts when we send the image to the server. This method will keep looping until it get the result from server, if the server already send the result of the recognition (ex. Dog, cat), Android will start to display the object that recognized in 3D form. For the 3D image, we used the jPCT library which produces the 3D object in 3ds file. Figure 7 shows the augmented object appearing in 3D above the lenslet array case after the server sends the signal for the content to appear.

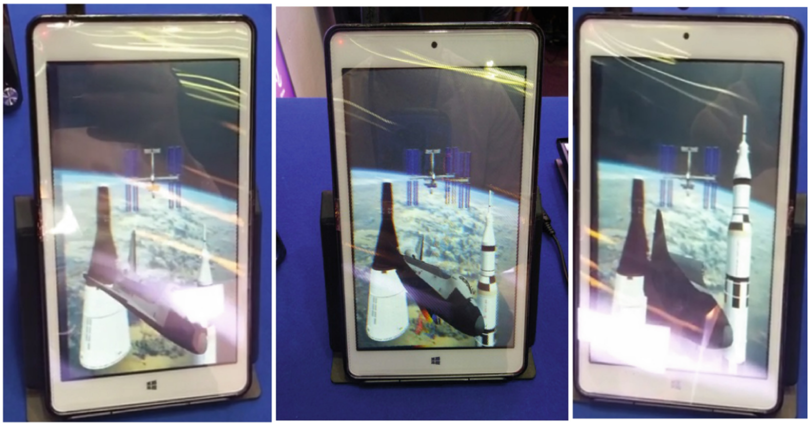


Fig. 7. Showing the 3D object displayed through the lenslet case from different viewpoints. Though not recognizable due to the 2D paper format the content appears in 3D in real environment.

4 Conclusion

In this paper, we showed an implementation of 3D augmentation with low cost lenslet case and deep learning recognition method. This maybe shows a possibility for future low cost 3D augmentation on smartphone devices. As for a further research, it could be an improvement for code efficiency. It also could be an enhancement for result display by adding 3D object generator method, so the system can generate a new 3D objects without any reference that saved in database.

Acknowledgments. This work was supported by the “Dongseo Frontier Project” Research Fund of 2016.

References

1. He K, Zhang X, Ren S, Sun J (2016) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell PP(99)
2. Stern A, Javidi B (2006) Three-dimensional image sensing, visualization, and processing using integral imaging. Proc IEEE 94:591–607
3. Park J-H, Hong K, Lee B (2009) Recent progress in three-dimensional information processing based on integral imaging. Appl Opt 48:H77–H94

Design of OpenGL SC 2.0 Shader Language Features

Nakhoon Baek[1,2,3](✉)

[1] School of Computer Science and Engineering, Kyungpook National University,
Daegu 41566, Republic of Korea
oceancru@gmail.com

[2] Software Technology Research Center, Kyungpook National University,
Daegu 41566, Republic of Korea

[3] Dassomey.com Inc., Daegu, Republic of Korea

Abstract. OpenGL is one of the most generally used 3D graphics libraries. Their new versions have special-purpose embedded high-level programming languages, named OpenGL Shading Language (OpenGL SL). The new safety critical version of OpenGL, OpenGL SC 2.0 is now available. To support the shading language features of OpenGL SC 2.0, we need an off-line compiling feature. In this paper, we present the overall design and implementation strategy for the OpenGL SC 2.0 off-line shader language specification.

Keywords: OpenGL SC · Shading language · Design

1 Introduction

OpenGL is one of the most widely used 3D graphics library [1]. It has a variation for embedded systems, named OpenGL ES [2, 3]. Originally, OpenGL and OpenGL ES was designed to be implemented into fixed function VLSI chips.

In the year of 2015, the Khronos Group, the fundamental standard management body of the famous OpenGL family, established a safety-critical working group. This “Safety Critical working group” is developing open graphics and compute acceleration standards for markets, including avionics and automotive displays.

As a result, the OpenGL SC 2.0 specification [4] defines a safety critical subset of OpenGL ES 2.0 [2]. Now the safety critical working group is working to adapt more recent Khronos standards including the new generation Vulkan API [5] for high-efficiency graphics and compute. The Safety Critical working group is also developing cross-API guidelines to aid in the development of open technology standards for safety critical systems.

The OpenGL SC 2.0 specification is different from its previous specification of OpenGL SC 1.0.1 [6], mainly in the shading language support. In this feature, programmers can provide their own low-level massively parallel executable programs to the GPU’s. This is the most efficient way of using the GPU and its massively parallel processing powers.

In this paper, we present the design for the shading language feature of the OpenGL SC 2.0 specification. Mainly, we need an offline compiler, and the compiled binary

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_27

images are downloaded to get the GPU execution codes. This is the first step to implement the OpenGL SC 2.0 shading language features into the 3D graphics rendering pipeline.

2 OpenGL SC 2.0 Shader Language Features

Since OpenGL 2.0, all the variations of the OpenGL family are equipped with the Shading Language processing. The core of this Shading Language feature is on-line compiling in the OpenGL library itself. Thus, they have on-line compiling and linking API functions, including glShaderSource, glCompileShader, glLinkProgram, and so on.

In the case of OpenGL SC 2.0, the online compiling and linking are removed from the specification, to get the safety and security of the library itself. Thus, they provide an alternative way of providing the shading language program: compile separately, and providing the compiled program to the rendering pipeline.

To provide a binary image of the pre-compiled shading language program, they can use the following OpenGL SC 2.0 API function:

void glProgramBinary(**GLuint** *program*, **GLenum** *binaryFormat*,
const void* *binary*, **GLsizei** *length*);

where

- *program* specifies the name of a program object into which to load a program binary,
- *binaryFormat* specifies the format of the binary data in binary,
- *binary* specifies the address an array containing the binary to be loaded into program, and
- *length* specifies the number of bytes contained in binary.
- The binary images read with the above glProgramBinary function are converted into the executable programs on the GPU. Internally, the binary image will be split into vertex shader and fragment shader programs. They are then downloaded on the GPU's and executed.

To create a program object, we can use the following API function:

GLuint glCreateProgram(**void**);

which returns the unique program object identifier. To install a specific program object as part of the current rendering pipeline, we can use:

void glUseProgram(**GLuint** *program*);

where *program* specifies the handle of the program object whose executables are to be used as part of current rendering state.

For a specific program object, we use the following API function to get the value of a parameter from the program object:

void glGetProgramiv(**GLuint** *program*, **GLenum** *pname*, **GLint*** *params*);

where

- *program* specifies the program object to be queried,
- *pname* specifies the object parameter, and
- *params* returns the requested object parameter.

OpenGL SC 2.0 also provides a set of shader variable related API functions. To get the location of an attribute variable, we can use the following function:

GLuint glGetAttribLocation(**GLuint** *program*, **const GLchar*** *name*);
where

- *program* specifies the program object to be queried, and
- *name* points to a null terminated string containing the name of the attribute variable whose location is to be queried.

To get the value of a uniform variable, we use the following API functions:

void glGetnUniformfv(**GLuint** program, **GLint** location,
GLsizei bufSize, **GLfloat*** params);
void glGetnUniformiv(**GLuint** program, **GLint** location,
GLsizei bufSize, **GLint*** params);
where

- *program* specifies the program object to be queried,
- *location* specifies the location of the uniform variable to be queried,
- *bufSize* specifies the size of the buffer *params*, and
- *params* returns the value of the specified uniform variable.

In the reverse way, we can also set the uniform variables, with the following API functions:

void glUniform[1,2,3,4][f,fv,i,iv](**GLint** *location*, **GLfloat**/**GLint**
v0,[*v1*,*v2*,*v3*]);
void glUniformMatrix[2,3,4]fv(**GLint** location, **GLsizei** count,
const GLfloat* value);
where

- *location* specifies the location of the uniform value to be modified,
- *count* specifies the number of elements that are to be modified. This should be 1 if the targeted uniform variable is not an array, and 1 or more if it is an array, and
- *value* specifies a pointer to an array of count values that will be used to update the specified uniform variable.

3 Implementation Design

To implement OpenGL SC 2.0 shading language features, we need an off-line shading language compiler. Fortunately, there are some open-source shading language compilers [7, 8]. Based on an existing OpenGL shading language compiler, we will generate the compiled binary images, all the uniform variable locations, and other data information, as shown in Fig. 1.

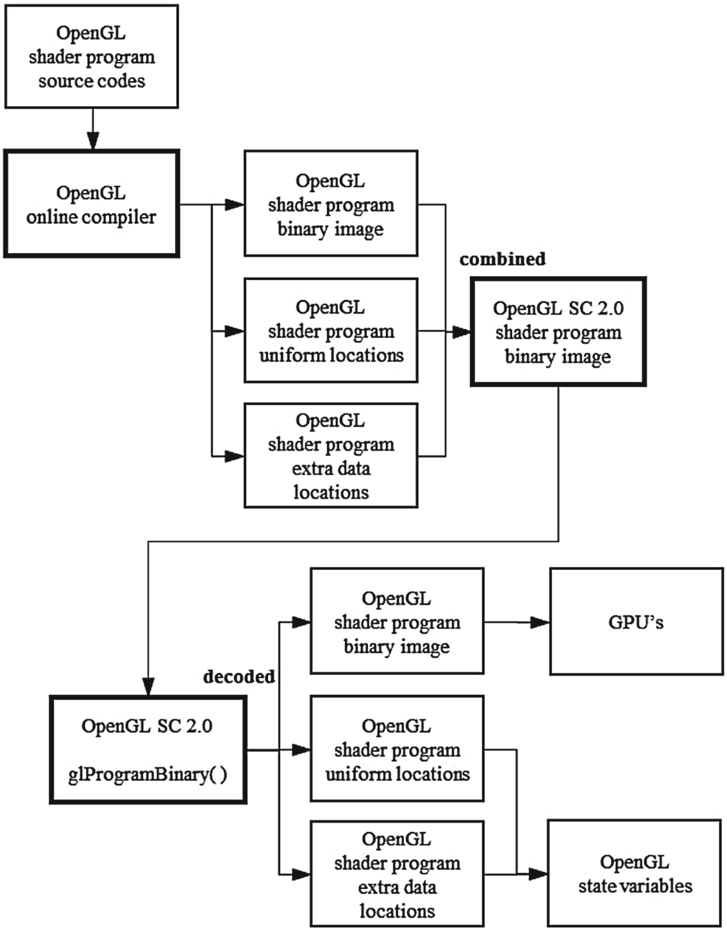

OpenGL shader program source codes
OpenGL online compiler
OpenGL shader program binary image
OpenGL shader program uniform locations
OpenGL shader program extra data locations
combined
OpenGL SC 2.0 shader program binary image
decoded
OpenGL SC 2.0 glProgramBinary( )
OpenGL shader program binary image
GPU's
OpenGL shader program uniform locations
OpenGL shader program extra data locations
OpenGL state variables

Fig. 1. Implementation strategy for the OpenGL SC 2.0 shading language feature.

Using our own OpenGL SC 2.0 specific binary image file format for the compiled result, we will transfer all these information to the OpenGL SC 2.0 execution core. Specifically, glProgramBinary function will decode all the information and spread them to the GPU's and OpenGL state variables.

Based on these implementation, we can provided the OpenGL SC 2.0 shading language features, as specified in the standard specification. Additionally, through these design scheme, we can construct the compiling system incrementally. We can minimize any potential problems, through fully testing and debugging the corresponding modules.

4 Conclusion

In this paper, we aimed to design the overall features on the OpenGL SC 2.0 Shading Language and its related API functionalities. Although we have some OpenGL on-line compilers, the OpenGL SC 2.0 features need some extra information, and thus, we provided its specific implementation design and implementation strategy. Based on these steps, we can build up the full scale OpenGL Shading Language compiling system for OpenGL SC 2.0 specification. In near future, we will show the results of our implementations.

Acknowledgement. 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 (2004) The OpenGL Graphics System: A Specification, Version 2.0
2. Munshi A, Leech J (2010) OpenGL ES Common Profile Specification, Version 2.0.25 Full Specification
3. Simpson RJ (2009) The OpenGL ES Shading Language, Version 1.00
4. Fabius A, Viggers S (2016) OpenGL SC, Version 2.0.0 Full Specification
5. Khronos Group (2016) Vulkan 1.0.35 – A Specification
6. Stockwell B (2009) OpenGL SC: Safety-Critical Profile Specification, Version 1.0.1
7. Khronos Group (2017) OpenGL / OpenGL ES Reference Compiler
8. Mesa3D.org (2017) The Mesa 3D Graphics Library

Fast Prototyping for LiDAR-Scanned Point Cloud Editing Operations

Woo-seok Shin[1] and Nakhoon Baek[1,2,3](✉)

[1] School of Computer Science and Engineering, Kyungpook National University, Daegu 41566, Republic of Korea
mel103@naver.com, oceancru@gmail.com
[2] Software Technology Research Center, Kyungpook National University, Daegu 41566, Republic of Korea
[3] dassomey.com Inc., Daegu, Republic of Korea

Abstract. Recently, we have lots of LiDAR (Light Detection And Ranging) data. They are typically a set of point cloud, for applications of high-resolution maps, geodesy, geomatics, archaeology, geography, geology, geomorphology, seismology, forestry, atmospheric physics, airborne laser swath mapping, laser altimetry, and others. We have implemented an efficient LiDAR data handling library and its GUI-style application program. In this paper, we presented an efficient way of fast prototyping some LiDAR point cloud handling operations. We also show the results of these fast prototyping.

Keywords: LiDAR data · Point clouds · Editing · Fast prototyping

1 Introduction

LiDAR (Light Detection And Ranging) is a surveying technology that measures distance by illuminating a target with a laser light [1–4]. Lidar is popularly used as a technology to make various kinds of geometric and geological data. This technique is often simply referred to as “laser scanning” or “3D scanning,” with terrestrial, airborne and mobile applications. Typical LiDAR systems produce very large data sets, with so many 3D sampling points. Sometimes, they refer these data point sets as “point clouds.”

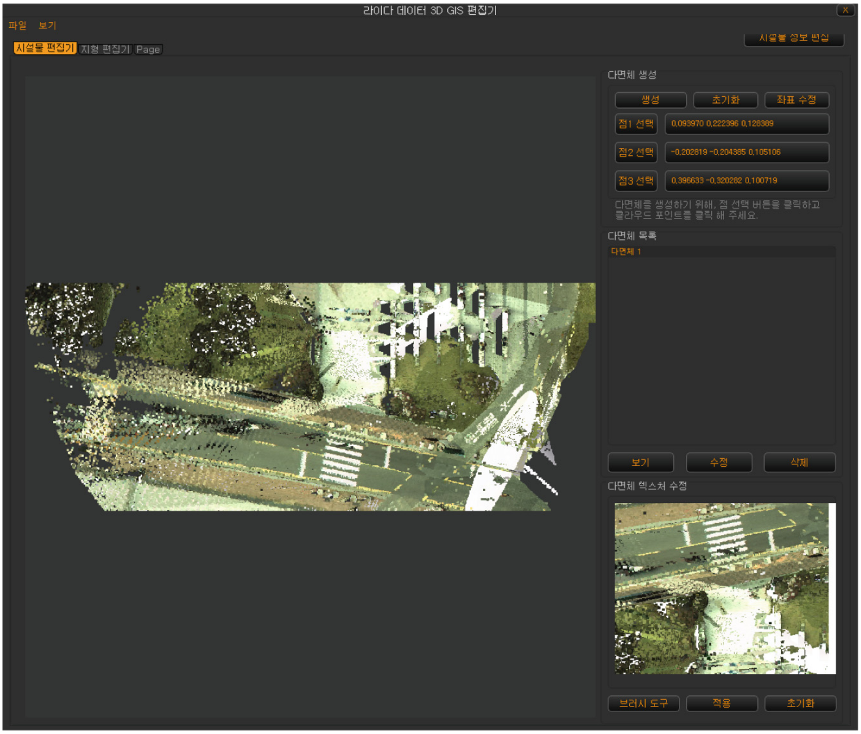
We have implemented a set of LiDAR data handling operations, and based on these operations, we also developed a simple and efficient LiDAR data handling tools. In this paper, we planned to add some basic operations for the LiDAR data handling. In the next section, we will analyze the required operations. In Sect. 3, we will show the results of fast prototyping, based on the open source projects. Our implementation will be fully tested.

2 Problems with LiDAR Point Clouds

Most of the major problems with LiDAR-scanned data are introduced by the extremely large size of the LiDAR-scanned data. In most cases, it is hard to edit the millions of

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_28

points on the single screen, and it is even hard to marge these point clouds. With our LiDAR point cloud editing system, we have implemented basic operations and graphical interfaces with them, as shown in Fig. 1.



점1 선택 0.093970 0.222396 0.128389

점2 선택 -0.202819 -0.204385 0.105106

점3 선택 0.396633 -0.320282 0.100719

다면체를 생성하기 위해, 점 선택 버튼을 클릭하고
클라우드 포인트를 클릭 해 주세요.

Fig. 1. Our LiDAR point cloud editing system.

At this time, as the advanced point cloud editing operations, the followings are immediately needed:

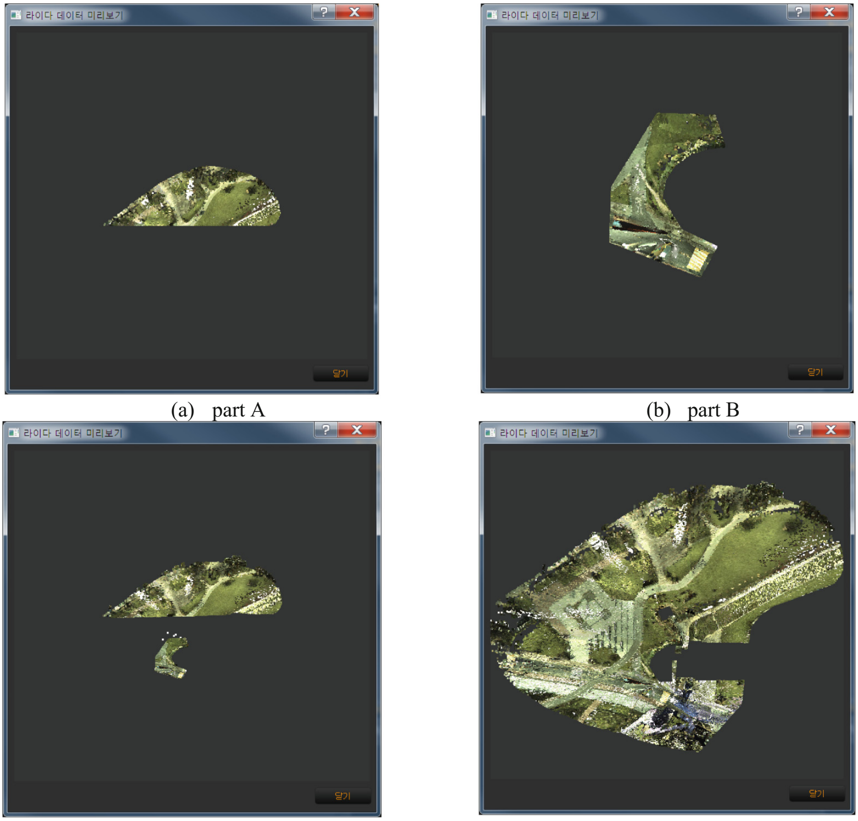
- size reduction – reduce the number of points from the given millions of LiDAR-scanned point clouds. The reduction ratio can be specified by the user.
- preview feature – fast preview images should be generated on-the-fly, from any given LiDAR-scanned point clouds.
- merging point clouds – making an integrated point cloud, from a set of point clouds.

Our final goal is to implement all these operations, and achieve the optimized implementations. As the first step, we need several reference implementations, and thus, we use an open-source library to make fast prototyping of these operations, as shown in the next section.

3 Prototyping Results

As explained in the previous section, we tested several open source projects on the point clouds data handling. We finally selected the Cloud Compare program [5] as the major tool for the fast prototyping. Actually, Cloud Compare program itself is a GUI-style application program, and we extracted some internal modules to use them as the basic operators in our fast prototyping.

Figures 2, 3, and 4 shows the implementation results with fast prototyping operations. We achieved merging, size reduction, and previewing of the large size point clouds.



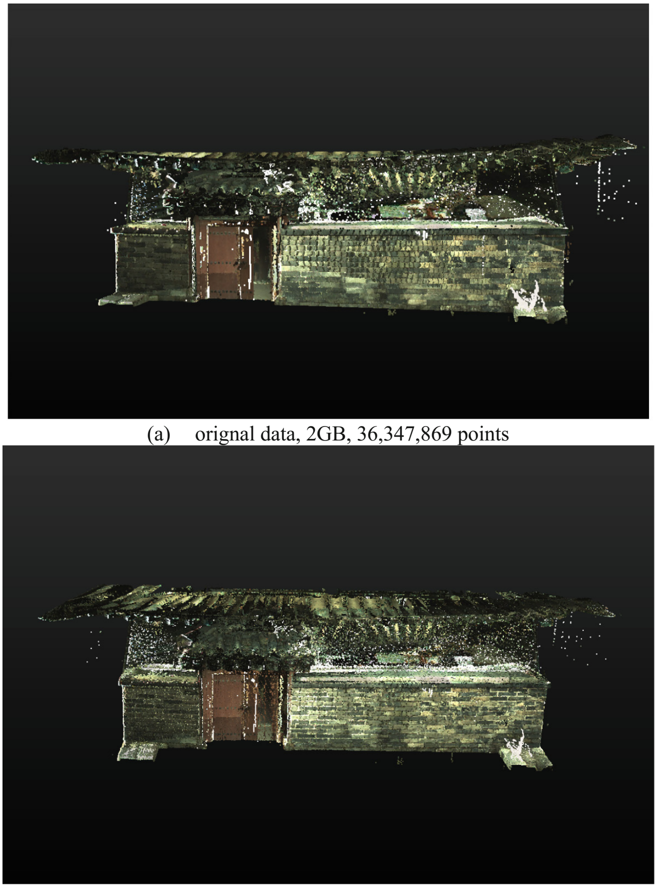
(a) part A

(b) part B

(c) combining A and B

(d) final result with 15 parts

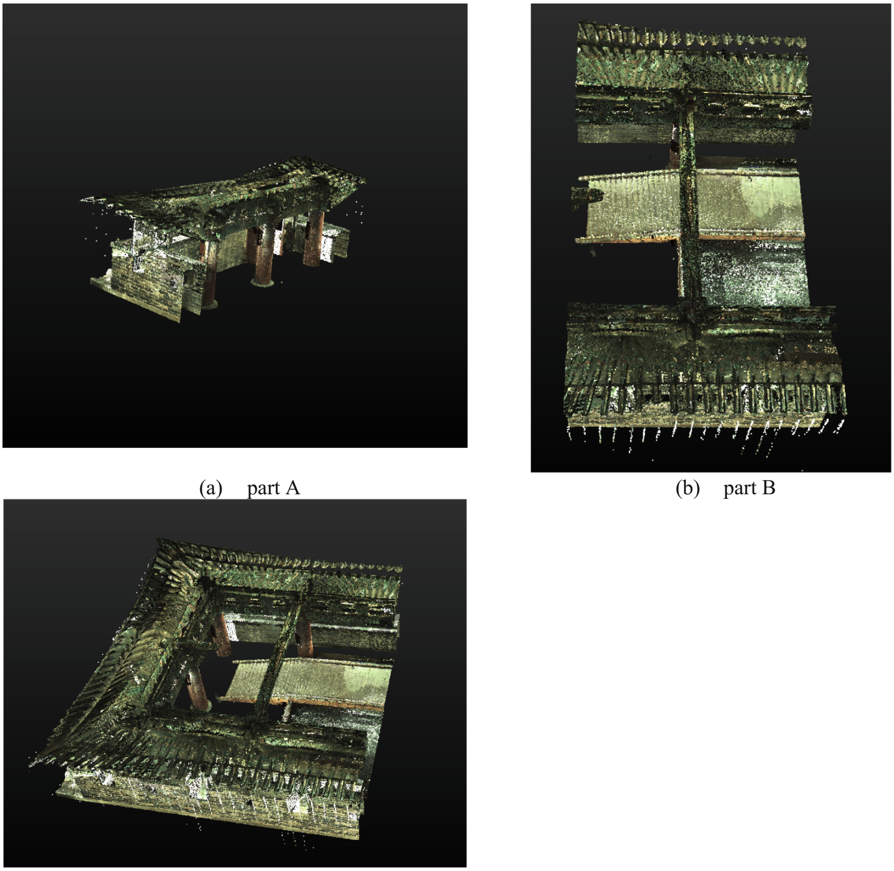
Fig. 2. Example of combining point clouds.



(a) orignal data, 2GB, 36,347,869 points

(b) reduced data. 46.2MB, 728,784 points

Fig. 3. Example of point cloud size reduction.



(a) part A

(b) part B

(c) combined result

Fig. 4. Another example of combining the point clouds.

4 Conclusion

Based on our existing implementation of point cloud editing system, we are implementing a set of more advanced large point cloud editing operations. As the first step, we acquired the reference implementation of these operations, with third part open source libraries. These fast prototyping operations are successful. We are implementing our own versions of these operations and optimize all of these operations.

Acknowledgement. This study was supported by the 2015 Technology Innovation Development program (project name: "Development of 3D GIS Platform for the LiDAR Data Utilization") funded by the Small and Medium Business Administration, Korea (project number: S2306348).

This study was 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. Heritage G, Large A (eds) (2009) Laser Scanning for The Environmental Sciences. Wiley, Chichester. ISBN 1-4051-5717-8
2. Maltamo M, Næsset E, Vauhkonen J (2014) Forestry applications of airborne laser scanning: concepts and case studies, vol 27. Springer Science & Business Media. ISBN 9-4017-8662-3
3. Shan J, Toth CK (eds) (2008) Topographic Laser Ranging and Scanning: Principles and Processing. CRC Press, Boca Raton. ISBN 1-4200-5142-3
4. Vosselman G, Maas HG (eds) (2010) Airborne and Terrestrial Laser Scanning. Whittles Publishing, Dunbeath. ISBN 1-4398-2798-2
5. Cloud Compare. http://www.cloudcompare.org/. (visited in Feb 2017)

Using Light Sensing to Acquire SpO_2 Biological Information via a Non-contact Approach

Yoshimitsu Nagao(✉), Mizuho Hatsuda, Jiang Liu, and Shigeru Shimamoto

Department of Computer Science and Communications Engineering,
Graduate School of Fundamental Science and Engineering, WASEDA University, Tokyo, Japan
ynagao@sirius.ocn.ne.jp

Abstract. Using conventional approaches to estimating arterial oxygen saturation (i.e., SpO_2) for individuals, it was impossible to measure unless the given sensor of the pulse oximeter was attached to the finger. In this study, we attempted to realize successful SpO_2 measurements using non-contact space measurements and succeeded in our experiments. Our work can be used for medical care, elder care, and other related fields. Finally, even though we observed some problems in that our approach was susceptible to other light interference, we offer our work as a first method using the laser wavelength for these purposes.

Keywords: SpO_2 · Saturation · Pulse · Non-contact · Light sensing · Biological information · PPG · LMM · Pulse oximeter · Blood oxygen saturation · Non-contact measurement · Space measurement

1 Introduction

1.1 Spatial Measurements of Blood Oxygen Saturation SpO_2

Existing probes are equipped with excited individual light-emitting elements of red and infrared LEDs. Utilizing the difference in the absorption rate of red light and infrared light due to the binding of oxygen and hemoglobin in human blood, the blood oxygen saturation degree (i.e., SpO_2) of the artery/vein, value is measured using contact-based methods. Further, only contact-based methods are used for such measurements. Aside from impractical approaches, there are no contactless approaches here. Therefore, in this study, we attempt to measure the difference of light transmitted by light sensing from fingers emitting light of different wavelengths of two types of output by non-contact methods, thus calculating the desired SpO_2 values. Measurements of these target SpO_2 values in space is a completely new approach. Here, light is measured by detecting transmitted light in the red and infrared spectra.

1.2 Measurement Principles and Types of SpO_2 Values

The light-receiving element transmits the measurement object medium and extracts the component of light that has not been absorbed. In relation to the molar extinction

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_29

coefficient and wavelength of the molecule, the artery and vein have a relationship of transmittance wavelength characteristics.

This transmittance has a wide absorption band ranging from 640 nm to 1600 nm or more. The standard deviation here is the number of received data and is in the interval between the peak of the waveform of the heartbeat and the time of the peak. It is crucial to obtain information regarding how much oxygen is being supplied to the blood to measure the state of the living body. The index used for this purpose is called the arterial blood oxygen saturation or SpO_2, where S stands for saturation, p stands for pulse, and O_2 stands for oxygen, i.e., oxygen saturation from the pulse.

At present, a device that can continuously and non-invasively measure SpO_2 is called a pulse oximeter. SpO_2 values are calculated by sandwiching the medium between irradiance devices, irradiating two kinds of lights with different wavelengths to the finger, then measuring the amount of transmitted light. Although the research method we present here in this paper is the first method that uses laser light, we face the inherent problem in that it is susceptible to interference with other light due to the light environment and is considered to be the primary difficulty in realizing contactless approaches thus far.

2 System Architecture (PPG and LMM)

2.1 SpO_2 Value Measurement Method Currently in the Blood Oxygen Saturation

Contact sensors are currently comprised of (PPG) using conventional photoelectric volume pulse waves. Such sensors are incorporated into a probe and clip-on device to be attached to the earlobe or fingertip. As shown in Fig. 1, there are two measured wavelengths here, one at approximately 650 nm in the red light range and the other at approximately 940 nm in the infrared range. As shown in Fig. 2, a pulse oximeter,

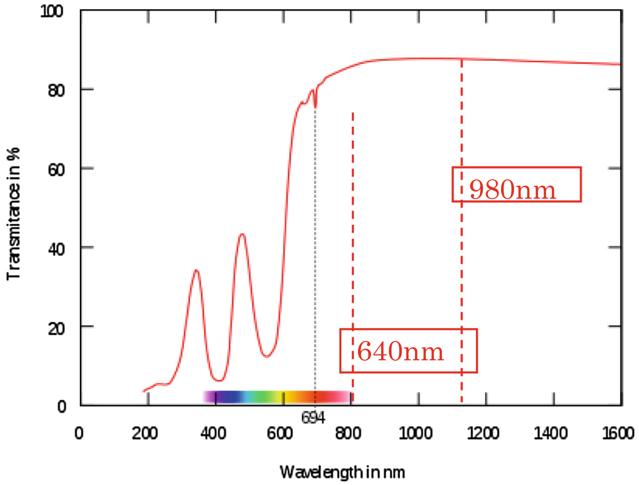


Fig. 1. Transmittance wavelength characteristics of contact

capacity pulse oximeter n contact with, such as pinching your fingers to probe measurement at the same time the current way of prototype is to compare the acquired output value in the data. The size of such devices are approximately 6.6 cm × 2.9 cm.

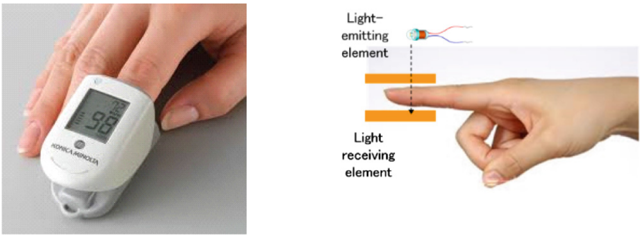


Fig. 2. (a) A p oximeter and (b) the contact SpO_2 value measurement principle

2.2 Laser Beam Measurement Method (LMM)

In this study, two different types of light wavelengths are emitted in a non-contact manner by a SpO_2 space measurement transmission method that we propose based on the laser transmission method (LMM). More specifically, as depicted in Fig. 3, we transmit light through a finger held over a human body part, which serves as the medium to measure the light difference and the target SpO_2 values without using a probe within the given space.

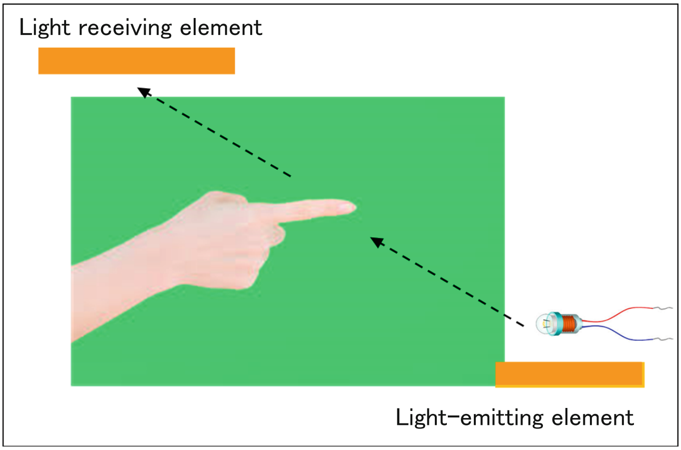


Fig. 3. Illustrating how we measure SpO_2 value in our approach

In our approach, as illustrated in Fig. 4, we measure the oxygen saturation of arterial blood and venous blood by utilizing the difference in absorption rates between red light and infrared light due to the binding of hemoglobin in oxygen to oxygen in the blood to detect the transmitted light. Given the non-contact nature of our approach, we measure the difference in light passing through the finger, which serves as the medium over which light of two different output wavelengths is radiated; by doing so, we measure the target SpO_2 value.

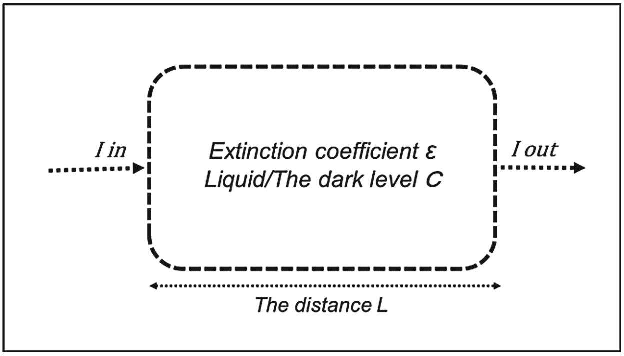


Fig. 4. Vital signs obtained from the density of the given liquid

In conventional methods with our new non-invasive approach, we use an infrared laser beam of 980 nm (5mW) and a laser light of 640 nm (1mW or less). In recent studies, researchers have confirmed that near-infrared light (i.e., 700 nm to 1,500 nm) has high permeability in living tissue; further, given that oxygen concentrations in living tissue are measured by using light in this region, advances have been expected in terms of new non-invasive measurement techniques.

3 Deriving SpO_2 Values by Irradiance Passing Through a Medium

3.1 The Principle of Deriving SpO_2 Values

We first define Apparent SpO_2 as the apparent SpO_2 value that can be derived by a calculation formula if the measured value is obtained. It is necessary to finely analyze waveforms and numerical values via measurements of laser light. The principle of deriving SpO_2 values here is that the fingertip is irradiated with two lights, i.e., Oxy-hemoglobin and Deoxy-hemoglobin, corresponding to red light and infrared light, according to the absorption characteristics of light, and from the ratio of the magnitude of the pulse wave of transmitted light to arterial blood and vein to calculate absorption characteristics. Theoretically, when the total hemoglobin contains oxygen and changes to Oxy-hemoglobin, oxygen saturation is 100%. The SpO_2 value can then be determined from the amplitude ratio of the pulse wave due to light at wavelengths of 640 nm and 980 nm, which are irradiated depending on the oxygen saturation of arterial blood and venous blood, respectively. SpO_2 is indicator of how much oxygen in the blood.

Regarding the oxygen saturation of arterial blood and venous blood, 70% of the gas in venous blood is red darkened with oxygen. Note that the numerical value of SpO_2 is expressed as a percentage. The reference value is typically 97% to 99% for a healthy person, and it is largely agreed upon that pulmonary function declines if the reference value of an individual is 90% or less.

3.2 Lambert-Beer Law

Here, we compare incident light to transmitted light in a certain concentration of solution. If the absorption coefficient of the given solution at a specific wavelength is determined beforehand, the concentration of the solution can be obtained by measuring the incident light, the transmitted light, and the distance of the solution. Arterial blood oxygen saturation of SpO$_2$ is therefore expressed as

$$SpO_2 = \frac{\Delta Coxy \cdot L^{P-P}}{\Delta Coxy \cdot L^{P-P} + \Delta Cdeoxy \cdot L^{P-P}} \tag{1}$$

Next, we obtain the maximum amplitude of each hemoglobin change within a heartbeat as

$$\Delta Coxy \cdot L^{P-P} + \Delta Cdeoxy \cdot L^{P-P} \tag{2}$$

$$\Delta Coxy \cdot L^{P-P}\text{:Red laser}$$

Further, Oxy-hemoglobin *estimation* HbO$_2$ (artery); Binding with oxygen

$$\Delta Cdeoxy \cdot L^{P-P}\text{:Infrared laser}$$

For our Deoxy-hemoglobin *estimation* Hb (vein), we check the waveform with an oscilloscope using red laser light and infrared laser light, respectively, which are not coupled with oxygen. Moreover, when a finger is held between the laser light and the light-receiving element, two changes (i.e., ΔCoxy and ΔCdeoxy) are observed in the waveform of the oscilloscope; here, these two components are transmitted normally, so detection is possible.

As illustrated in Fig. 4, using the Lambert-Beer law, the incident light Iin to a solution of constant concentration can be used to determine the transmitted light by determining absorbance A after measuring Iout, i.e.,

$$A = -log(I\,out\,/\,I\,in) = \varepsilon\,C\,L \tag{3}$$

Previously determining the extinction coefficient of the solution at a specific wavelength epsilon.

Using Iin, we determine concentration C of the solution by measuring Iout via L.

The Lambert-Beer is applied to the medium scattered by the Modified Lambert-Beer Law, i.e.,

$$-log(I\,in\,/\,I\,out) = \varepsilon\,C\,L + S \tag{4}$$

In the equations above, A represents absorbance, ε represents the extinction coefficient of the solution, C represents the concentration of the solution, L represents the distance (i.e., the average optical path length), and S represents the attenuation of the light intensity due to scattering.

When the concentration of the solution is changed from C to C + ΔC, the quantity of transmitted light is changed to Iout + ΔIout and the relation is

$$-\log\left[(\mathrm{I\,out} + \Delta\mathrm{I\,out}) / \mathrm{I\,in}\right] = \varepsilon\,(\mathrm{C} + \Delta\mathrm{C})\,\mathrm{L} + \mathrm{S} \tag{5}$$

Since Eqs. (2) and (3) as an attenuation S of the light amount due to scattering is not changed, we have

$$-\log\left[(\mathrm{I\,out} + \Delta\mathrm{I\,out}) / \mathrm{I\,out}\right] = \varepsilon\,\Delta\mathrm{C}\,\mathrm{L} \tag{6}$$

$$\begin{aligned}-\log\left[(\mathrm{I\,out}(\lambda) + \Delta\mathrm{I\,out}(\lambda)) / \mathrm{I\,out}(\lambda)\right] = (&\varepsilon\,\mathrm{oxy}(\lambda)\cdot\Delta\mathrm{C\,oxy}\\ &+\varepsilon\,\mathrm{deoxy}(\lambda)\cdot\Delta\mathrm{C\,deoxy})\cdot\mathrm{L}\end{aligned} \tag{7}$$

Of the incident light shone on living tissue at specific wavelength λ, the absorption of the amount of change in response to the absorption and scattering in the living body has been returned to the ex vivo amount of light and Iout (λ) and ΔIout (λ), oxyhemoglobin (OxyHb) the coefficient ε oxy (λ), the extinction coefficient of deoxyhemoglobin (DeoxyHb) ε deoxy (λ), ΔCoxy the change in concentration of OxyHb, changes in the concentration of deoxy Hb ΔC deoxy.

In Eq. (7) above, concentration variation ΔCoxy of OxyHb determines the concentration amount of change ΔC deoxy of DeoxyHb. Here, ΔCoxy, because two variables are used to determine ΔCdeoxy, seeks to use the red-infrared extinction coefficient at the two wavelengths of 650 nm and 980 nm, as shown in formula (8) and (9) we then obtain.

$$\begin{aligned}-log\left[(I\,out(\lambda 980) + \Delta I\,out(\lambda 980)) / I\,out(\lambda 980)\right] = (&\varepsilon\,\mathrm{oxy}(\lambda 980)\cdot\Delta\mathrm{C\,oxy}\\ &+\varepsilon\,\mathrm{deoxy}(\lambda 980)\cdot\Delta\mathrm{C\,deoxy})\cdot\mathrm{L}\end{aligned} \tag{8}$$

$$\begin{aligned}-log\left[(I\,out(\lambda 640) + \Delta I\,out(\lambda 640)) / I\,out(\lambda 640)\right] = (&\varepsilon\,\mathrm{oxy}(\lambda 640)\cdot\Delta\mathrm{C\,oxy}\\ &+\varepsilon\,\mathrm{deoxy}(\lambda 640)\cdot\Delta\mathrm{C\,deoxy})\cdot\mathrm{L}\end{aligned} \tag{9}$$

If optical path length L cannot be set, L is left as it is, i.e.,

$$\begin{aligned}&\mathrm{L}\cdot\Delta\mathrm{C\,oxy, a\ solution\ the\ L}\cdot\Delta\mathrm{C\,deoxy,\ the\ above\ formula\ from\ D}\cdot\Delta\mathrm{C\,oxy,}\\ &\mathrm{L}\cdot\Delta\mathrm{C\,deoxy\,L}\cdot\mathrm{from}\,\Delta\mathrm{C\,oxy} + \mathrm{L}\cdot\Delta\mathrm{C\,deoxy} = \mathrm{L}\cdot\Delta\mathrm{C\,total}\end{aligned} \tag{10}$$

Finally, the concentration variation of the total hemoglobin (ΔC total Hb) as Hemoglobin change (L · ΔC oxy, L · ΔC deoxy, L · ΔC total) Unit of: mM cm, mM mm; if the optical path length cannot be set, while including the optical path length, mM: millimolar).

4 Production of Non-invasive Prototype to Measure SpO_2

4.1 Sensor Module to Be Installed in Equipment

Initially, two LEDs with wavelengths of 640 nm and 980 nm were mounted on a gantry, and an attempt was made to measure SpO_2 data via non-contact light irradiation. In the LED, to ensure we did not reach a sufficient amount of light in the light-receiving element to cause light scattering within the given space, it is measured in a non-contact difficult it was found in the experiment. Initially far apart from our purpose 98–99 of SpO_2, it became possible to perform various adjustments, as shown in Fig. 5.

In our proposed measurement method, we emitted two different wavelengths of light without requiring contact to the finger, measuring the difference in the transmitted light from the finger held in between two endpoints; as such, we measured target SpO_2 values in space. Here, the shorter wavelength is transmitted through the medium using a red laser at 640 nm (i.e., light-emitting element 1) and a long wavelength using an infrared laser at 980 nm (i.e., light-emitting element 2). The Si PIN photodiode of the light-receiving element transmits through the medium to be measured and outputs the signal level of the light not absorbed by the medium.

Finally, we confirmed that the measurement range in the sensitivity wavelength was sufficient, as shown in Fig. 5.

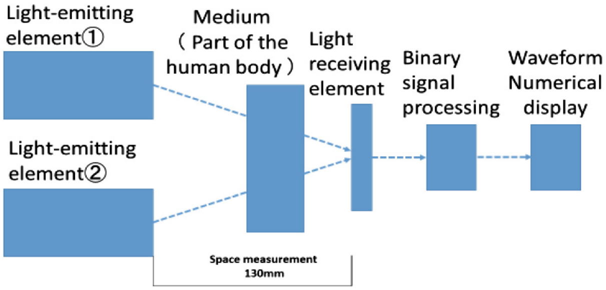


Fig. 5. Conceptual diagram of our signal processing workflow

As for our specifications, the light-receiving surface size was 5.5 mm × 4.8 mm, the effective light-receiving area was 26.4 mm^2, the reverse voltage (VR) had an absolute maximum of 35 V, the sensitivity wavelength range (λ) ranged from 320 nm–1100 nm, the maximum sensitivity wavelength (λp) receiving a good component and the signal by 960 nm doing.

4.2 Equipment

As depicted in Fig. 6, we used a device to detect SpO_2 values using red laser (i.e., artery) and infrared laser (i.e., vein) measures. Using this device, we set out to detect normal values in a stable state. Since there is a relationship between the oxidized hemoglobin concentration of the artery at 640 nm and the reduced hemoglobin concentration of the vein at 980 nm in relation to the molecule (i.e., molar extinction coefficient) and wavelength, the standard deviation will be at a peak, the time interval of this peak with the number of received data six times. We use this to identify the quality of the data and decide on the reliability by using a numerical value of 25 or less. More specifically, the value of the ratio between the AC component and the DC component is calculated as visible light and infrared light as the R-value of the SpO_2 value to be calculated.

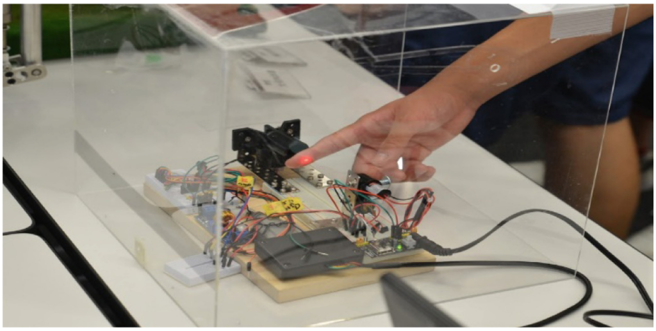


Fig. 6. Non-contact SpO_2 measuring device prototype in operation

As shown in Fig. 6, we conducted experiments to measure the effectiveness and functionality of our system by linking our created program and fabricated board to the laser beam module. From our results, we found that the SpO_2 values were effective and normal; we also confirmed that the non-invasive measurement within the space between the sensors succeeded.

5 Verification of Our Prototype for Contactless Measurement

In this section, compare the traditional contact-based and our non-contact measurement methods.

In our experiment, we measure the SpO_2 value using a commercially device and our proposed device. The result of measuring SpO_2 values using a commercially available device (CONTEC PULSE OXIMETER CMS50D+, as shown in Fig. 7) are shown in Fig. 8. Here, the basal SpO_2 value was 98.4%, while the minimum SpO_2 value was 97%. The result of using our proposed device before correction is shown as Fig. 9, In the measurement waveform of Fig. 9, the part surrounded by the dotted line is SpO_2 from 99% to 97%.

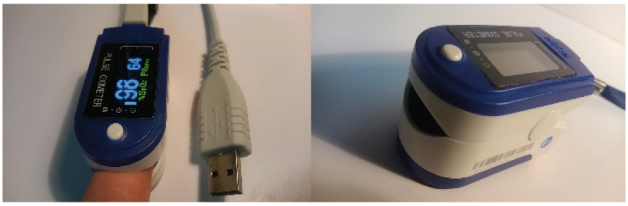


Fig. 7. CONTEC PULSE OXIMETER CMS50D+ (conventional type SpO_2 measuring instrument.)

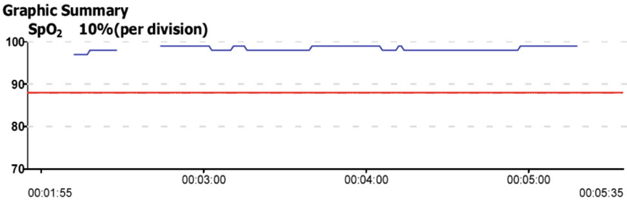


Fig. 8. Measured waveform from the commercially available pulse oximeter

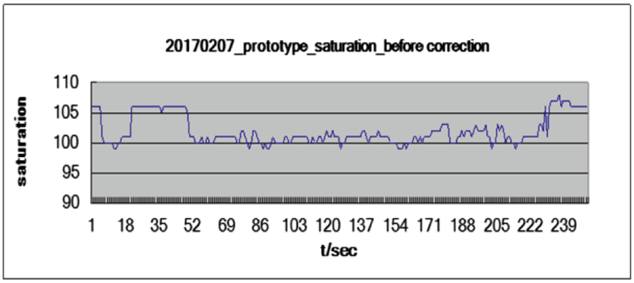


Fig. 9. Measured waveform from our p prototype SpO_2 device (before correction)

Meanwhile, our prototype SpO_2 measurement output after correction is shown in Fig. 10, cut out from the measured value of 52 s to 222 s; we analyzed its waveform and determined it to be stable in saturation from the highest 101% to the lowest 99%.

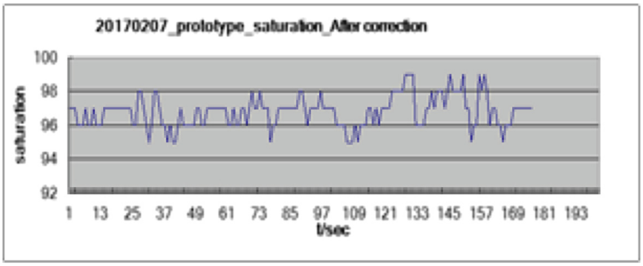


Fig. 10. Measured waveform from our prototype SpO_2 device (after correction)

Here, when we extract the interval between 52 s and 222 s, which served as the actual measurement time range, we have a time range ranging from zero to 165 s. Therefore, if we use an output parameter variable of −3, we can use the commercially available CONTEC of Fig. 8. Since we combine the output numerical values of Fig. 10, we show that the measurements in the space measurement prototype were normal.

6 Summary

In this study, we used a non-contact space measurement method to acquire biological data for medical use, nursing care, everyday life, fieldwork, and so on. Before this study, the idea of using a non-contact method for such measures has been considered impossible, such as being able to know the SpO_2 value of pets etc., which are members of the family. And there the groundbreaking knowledge was obtained.

References

1. Hamamatsu Photonics KK, HAMAMATSU SiPIN Photodiode
2. Kazuo Tsusui, Corporation Ohmsha
3. Electronic devices author. http://hooktail.maxwell.jp/bbslog/24449.html
4. Optical and color science. http://optica.cocolog-nifty.com/blog/2012/01/post-0dfa
5. Tokyo Devices. https://tokyodevices.jp/

Adaptive Resampling for Emergency Rescue Location: An Initial Concept

Wan Mohd Yaakob Wan Bejuri[1,2](✉), Mohd Murtadha Mohamad[1],
Raja Zahilah Raja Mohd Radzi[1], Mazleena Salleh[1], and Ahmad Fadhil Yusof[1]

[1] Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Malaysia
wanmohdyaakob@gmail.com,
{murtadha,zahilah,mazleena,ahmadfadhil}@utm.my
[2] Faculty of Information and Communication Technology, Universiti Teknikal Malaysia,
76100 Melaka, Malaysia

Abstract. The different of memory specification mobile devices or smart phone make it hard for developer to establish a resampling in emergency rescue location for specific smart phone. It took much time for developer to develop it. In this paper, we will introduce a good solution for developer to develop a resampling algorithm for different mobile devices or smart phone. The proposed resampling can adapt memory specification of mobile device in order to determine which suitable resampling operation or function for specific mobile device. As overall, the paper just present a concept that can be used as a guideline to develop a flexible resampling.

Keywords: Particle filter · Resampling · Sequential implementation · Memory consumption

1 Introduction

The location determination using Global Positioning System (GPS) inside building is very difficult, since the signal was blocked [1–6]. To detect location for emergency rescue purpose (usually the rescue is among firemen), the usage of GPS is useless. However, current technologies inside smart phone can be utilized as signal locator for emergency rescue purpose inside building. There are many sensor inside smart phone can utilized except GPS for that purpose such as; WiFi, Inertial sensor, Bluetooth and many more. [7–12]. Among these sensor the inertial is sensor that not depend on the building infrastructure, thus making it can be as standalone position technology and also can be used anywhere in the earth [13–15]. The usage of inertial sensor need of particle filter usage as one of emergency rescue location. However, the range of many smart phone making it hard to developer to make it specific technology solution for smart phone. In this paper, we will develop a an adaptive resampling that can be used inside particle filter which is can be installed in many smart phone that have different memory specification [16–18]. Since our paper just introduce concept only, we present as follows: Sect. 2 outlines fundamental emergency rescue location. Section 3 proposed adaptive resampling and finally Sect. 4 is a discussion about the implications of this study.

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_30

2 Fundamental of Emergency Rescue Location

Previous section discuss about introduction of the paper. This section discuss the fundamental concept of emergency rescue location (see Fig. 1 for fundamental system architecture) regards positioning determination across all environments [19–23].

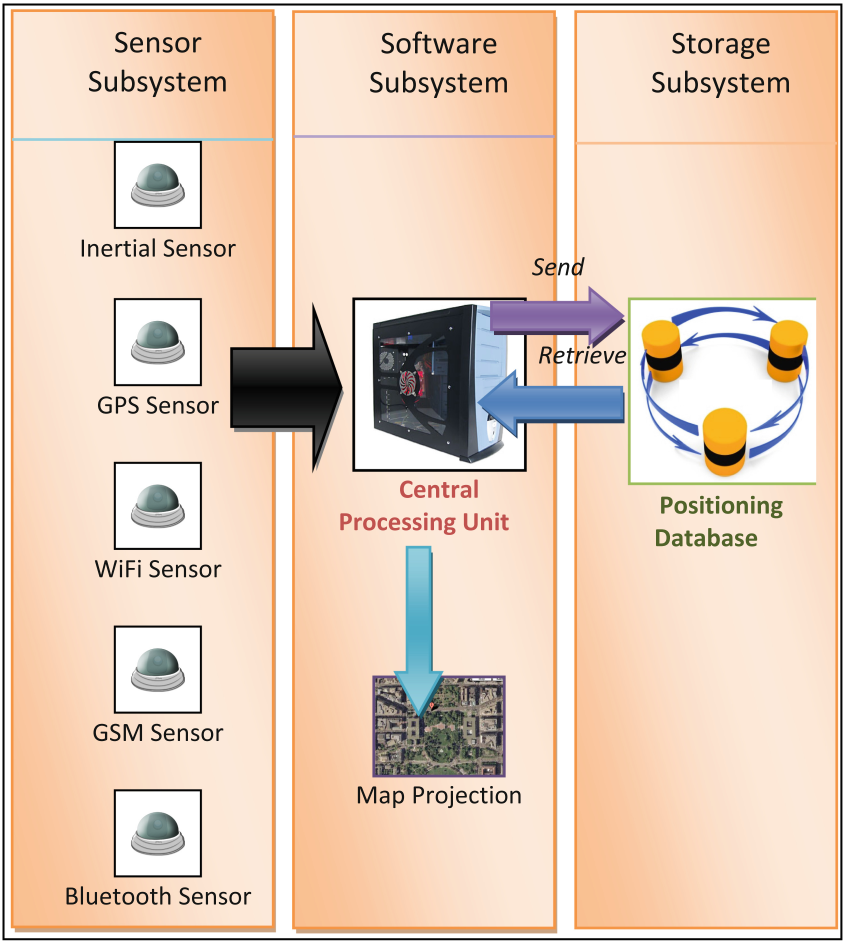


Fig. 1. Fundamental system architecture of emergency rescue location on mobile sensing platform

Usually, it requires a multi-sensor approach, augmenting standalone positioning with other signals, motion sensors, and environmental features [24–26]. According Fig. 1, it is present about fundamental system architecture of emergency rescue location on nowadays mobile phone technology. Basically, it can categories in three (3) sub system which known as sensor subsystem, software subsystem and storage subsystem. Basically, the sensor subsystem usually consist five (5) types of sensor (according to nowadays smart phone) which are, inertial sensor, Global Positioning System (GPS) sensor, WiFi sensor, Bluetooth sensor and GSM sensor. These sensor can be utilized by retrieve data from their environment. Data that retrieved will used as input for software subsystem for further processing. The subsystem of software is consists of CPU that used to processed any algorithm that written in software. The software subsystem will keep working closely by lookup information inside storage subsystem and finally display location in mobile phone screen for emergency rescue location purpose. Next section will discuss about our proposed method.

3 Proposed Adaptive Resampling

The previous section discussed the fundamental concept of emergency rescue location regards positioning determination across all environments. This section discusses the proposed design of the single distribution resampling algorithm entitled proposed adaptive resampling(see Fig. 2). Sequential important sampling is used to generate the particle and weight computation, followed by the proposed adaptive resampling (as shown in the dash box). As we can see, the proposed adaptive resampling is shown in

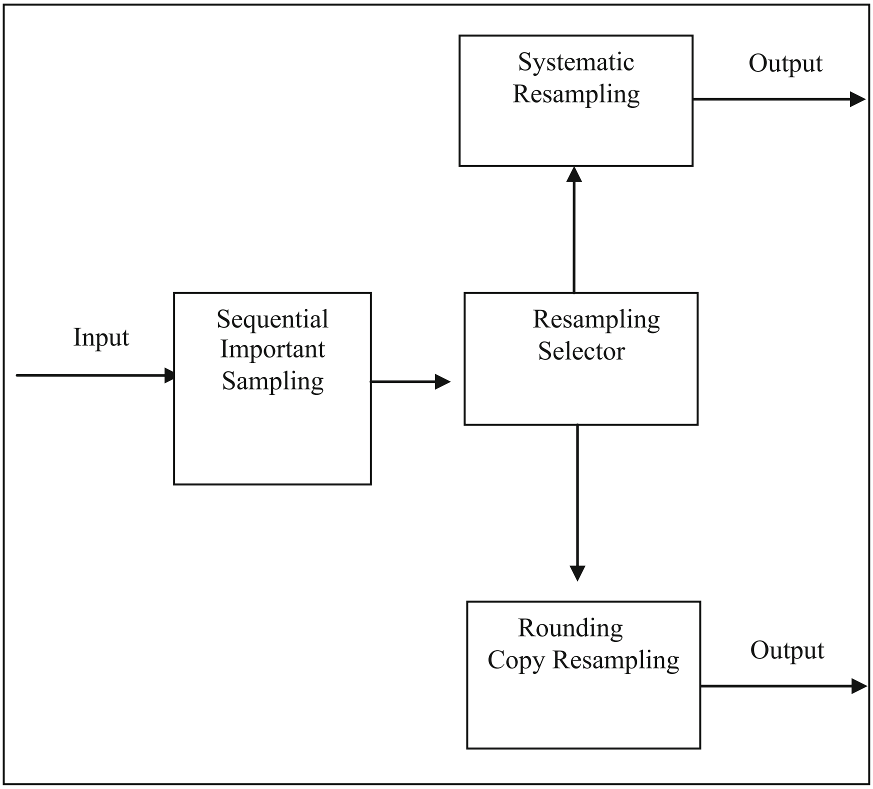


Fig. 2. Block diagram of proposed adaptive resampling in standard particle filter.

the figure which consist of resampling selector and rounding-copy resampling. The purpose of resampling selector is to adapt computer memory specification and select whether systematic resampling or rounding copy resampling as resampling function. The selection is based if the total amount of computer memory is over that 1536 MB, the resampling selector will chose resampling copy. Otherwise, it will choose resampling systematic as resampling function. The following section will conclusion and future works.

4 Conclusions and Discussions for Future Work

The different of memory specification mobile devices or smart phone make it hard for developer to establish a resampling in emergency rescue location for specific smart phone. It took much time for developer to develop it. A good solution for developer to develop a resampling algorithm for different mobile devices or smart phone has been introduced. The proposed resampling can adapt memory specification of mobile device in order to determine which suitable resampling operation or function for specific mobile device. As overall, the paper just present a concept that can be used as a guideline to develop a flexible resampling.

References

1. Chandra-Sekaran A-K, Flaig G, Kunze C, Stork W, Mueller-Glaser KD (2008) Efficient resource estimation during mass casualty emergency response based on a location aware disaster aid network. In: Verdone R (ed) Wireless sensor networks. Springer, Heidelberg, pp 205–220
2. Sitanayah L, Sreenan CJ, Brown KN (2010) ER-MAC: A hybrid MAC protocol for emergency response wireless sensor networks. In: 2010 Fourth international conference on sensor technologies and applications (SENSORCOMM), pp 244–249
3. Yang Y, May A, Yang S-H (2010) Sensor data processing for emergency response. Int J Emerg Manag 7(3):233–248
4. Rahim MSM, Shariff ARM, Mansor S, Mahmud AR, Daman D (2007) A spatiotemporal database prototype for managing volumetric surface movement data in virtual GIS. In: Gervasi O, Gavrilova ML (eds.) Computational science and its applications – ICCSA 2007. Springer, Heidelberg, pp 128–139
5. Rahim MSM, Fata AZA, Basori AH, Rosman AS, Nizar TJ, Yusof FWM (2011) Development of 3D tawaf simulation for hajj training application using virtual environment. In: Zaman HB, Robinson P, Petrou M, Olivier P, Shih TK, Velastin S, Nyström I (eds) Visual informatics: Sustaining research and innovations. Springer, Heidelberg, pp 67–76
6. Rahim MSM, Othman NZS, Daman D (2008) Visualization of surface movement data using TIN-based temporal modeling approach. In: Presented at the advances in computer science and technology
7. Bejuri WMYW, Mohamad MM, Sapri M (2011) Ubiquitous positioning: A taxonomy for location determination on mobile navigation system. Signal Image Process Int J SIPIJ 2(1): 24–34

8. Bejuri WMYW, Mohamad MM, Sapri M, Rosly MA (2012) Investigation of color constancy for ubiquitous wireless LAN/Camera positioning: An initial outcome. Int J Adv Comput Technol 4(7):269–280
9. Bejuri WMYW, Mohamad MM, Sapri M, Rosly MA (2012) Ubiquitous WLAN/camera positioning using inverse intensity chromaticity space-based feature detection and matching: A preliminary result. ArXiv Preprint arXiv:1204.2294
10. Bejuri WMYW, Mohamad MM, Sapri M, Rosly MA (2012) Performance evaluation of mobile U-navigation based on GPS/WLAN hybridization. ArXiv Preprint arXiv:1210.3091
11. Bejuri WMYW, Mohamad MM, Zahilah R (2015) Optimization of rao-blackwellized particle filter in activity pedestrian simultaneously localization and mapping (SLAM): An initial proposal. Int J Secur Appl 9(11):377–390
12. Bejuri WMYW, Mohamad MM, Radzi RZRM (2015) Offline beacon selection-based RSSI fingerprinting for location-aware shopping assistance: A preliminary result. In: New Trends in Intelligent Information and Database Systems, vol 598. Springer, pp 303–312
13. Bejuri WMYW, Saidin WMNWM, Mohamad MMB, Sapri M, Lim KS (2013) Ubiquitous positioning: Integrated GPS/Wireless LAN positioning for wheelchair navigation system. In: Asian conference on intelligent information and database systems, pp 394–403
14. Bejuri WMYW, Mohamad MM, Zahilah R, Radzi RM (2015) Emergency rescue localization (ERL) using GPS, wireless LAN and camera. Int J Softw Eng Appl 9(9):217–232
15. Bejuri WMYW, Mohamad MM, Sapri M, Rahim MSM, Chaudry JA (2014) Performance evaluation of spatial correlation-based feature detection and matching for automated wheelchair navigation system. Int J Intell Transp Syst Res 12(1):9–19
16. Bejuri WMYW, Mohamad MM, Radzi RZRM (2015) A proposal of emergency rescue location (ERL) using optimization of inertial measurement unit (IMU) based pedestrian simultaneously localization and mapping (SLAM). Int J Smart Home 9(12):9–22
17. Bejuri WMYW, Mohamad MM (2014) Performance analysis of grey-world-based feature detection and matching for mobile positioning systems. Sens Imaging 15(1):1–24
18. Bejuri WMYW, Mohamad MM (2014) Wireless LAN/FM radio-based robust mobile indoor positioning: An initial outcome. Int J Softw Eng Appl 8(2):313–324
19. Wu Y, Pan X (2013) Velocity/Position integration formula part I: Application to in-flight coarse alignment. IEEE Trans Aerosp Electron Syst 49(2):1006–1023
20. Cho SY, Choi W-S (2006) Robust positioning technique in low-cost DR/GPS for land navigation. IEEE Trans Instrum Meas 55(4):1132–1142
21. Bachrach A, Prentice S, He R, Roy N (2011) RANGE–Robust autonomous navigation in GPS-denied environments. J Field Robot 28(5):644–666
22. Fang S-H, Wang C-H, Chiou S-M, Lin P (2012) Calibration-free approaches for robust Wi-Fi positioning against device diversity: A performance comparison. In: 2012 IEEE 75th vehicular technology conference (VTC Spring), pp 1–5
23. Soleimanifar M, Lu M, Nikolaidis I, Lee S (2011) A robust positioning architecture for construction resources localization using wireless sensor networks. In: Proceedings of the winter simulation conference, Phoenix, Arizona, pp 3562–3572
24. Ren H, Chai P, Zhang Y, Xu D, Xu T, Li X (2017) Semiautomatic indoor positioning and navigation with mobile devices. Ann GIS 23(1):1–12
25. Belhajem I, Maissa YB, Tamtaoui A (2017) An improved robust low cost approach for real time vehicle positioning in a smart city. In: Industrial networks and intelligent systems. Springer, pp 77–89
26. Albert MV, Shparii I, Zhao X (2017) The applicability of inertial motion sensors for locomotion and posture. In: Locomotion and posture in older adults. Springer, pp 417–426

Using the IQRF Technology for the Internet of Things: Case Studies

Martin Pies(✉) and Radovan Hajovsky

Department of Cybernetics and Biomedical Engineering, Faculty of Electrical Engineering
and Computer Science, VSB-Technical University of Ostrava, 17. Listopadu 2172/15,
70833 Ostrava, Czech Republic
{martin.pies,radovan.hajovsky}@vsb.cz

Abstract. This article focuses on describing the IQRF technology following modern trends in the Internet of Things. There is a brief overview of the IQRF technology and then selected examples of its use in areas such as Smart Cities, Smart Parking, Smart Lighting etc. Practical examples of the use of devices developed in conjunction with the Internet of Things are provided in more detail. From the field of systems based on Inernet of Things were chosen two use cases. First of them is monitoring of concentration of carbon monoxide in old mine dumps and second one is about of monitoring of building stability using a wireless inclinometer.

Keywords: AHRS · Gas detectors · Internet of Things · IQRF · Wireless sensor networks

1 Introduction

The Internet of Things (IoT) is a very popular term today and in conjunction with another concept - Industry 4.0 - it represents a current and modern trend not only in industry but especially in communications, control, information and measurement technologies. The issues of IoT stretch practically across all fields of human activity and in the future they will form an integral part of everyday life for everyone.

The Internet of Things interconnects devices via the Internet without active human participation. These devices may be automobiles, home appliances, wearable accessories and various sensors and detectors that mutually exchange information and cooperate. The Internet of Things is made possible, among other things, thanks to miniaturization, reduced consumption and prices of chips and wireless technology that enables communication with other technologies with very low power consumption.

The basic principles are the functional characteristics of the Internet of Things: the ability to provide connectivity, security, interoperability, information analysis and projection of all this into cash on the side of both costs and revenues.

It is assumed that in 2020 there will be more than 50 billion devices connected to the Internet of Things. IoT is a very dynamically developing area in particular with regard to Smart Cities and represents great potential in search of new opportunities.

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_31

A typical application of the Internet of Things is the currently already widely implemented area of Smart homes. The Smart home contains a range of electronic devices and equipment that can communicate through a variety of technologies. One application is the control of the orientation of photovoltaic panels towards the sun, described in [5]. Other areas of deployment in a Smart home is the control of electronic equipment via smartphones, see [11, 13].

The notion of the Internet of Things is also beginning to affect the automotive industry, where cars can communicate with each other for example through headlights and cameras or through other communication technologies operating at a relatively short distance. This relatively new term is called Car2Car communication. In terms of linking the car and the Smart home, the article [2] describes the mutual exchange of electric power between the electric vehicle and smart home.

This article presents an overview of the use of the IQRF technology in the field of the Internet of Things in connection with Smart Cities and the monitoring of environmental variables. For this purpose, the following structure of the measuring chain is suitable:

- At the lowest level are sensors and actuators, operating with very low power consumption. Publication [1] is dedicated to this area.
- The second area is the collection of measured data via an IQRF gateway that sends data to cloud storage. These gateways also allow control of the wireless nodes. Commercial solutions from different manufacturers exist, but an open platform for the communication protocol of the IQRF technology enables implementation of one’s own IQRF gateway, with Raspberry Pi as the basic hardware, see [3]. Another alternative is to send the collected data to an FTP server as described in [10].
- The third area is the visualization and assessment of measurement data. This area includes not only the various trends and statistics, but also decision algorithms which provide feedback to control the monitored technology.

2 The IQRF Technology

IQRF is usable with almost any equipment and in almost any application where low speed and low data volume wireless is needed, ranging tens or hundreds of metres. It allows either simple peer-to-peer communication or huge complex MESH networks, which can count up more than 1000 nodes [7].

IQRF domain

- Packet-oriented communication. Maximum payload is 64 B data packet, which can be sent in about 40 ms to another device in range. Hopping to the end point out of direct range via other network nodes is possible with a corresponding delay. Maximum hops are up to 239.
- Low power consumption enables battery-operated devices with the lifetime of a number of years. It is possible to reach down to 15 μA while receiving when the transceiver module operates in XLP mode.

- Another way to decrease the consumption is the LP mode, in which power consumption is 330 μA while receiving combined with 380 nA while precisely timed sleeping.
- Most suitable for MESH networks.
- Semistatic topology. The fully automated networking algorithm requires static routing devices. End devices are fully relocatable. However, topologies with all devices moving are also possible.

IQRF is best suited for:

- Control applications where end devices are managed and monitored from a central point.
- Telemetry - data collection from end devices (sensors etc.). Acquired data can transparently be forwarded to the Internet and stored in Cloud depository accessible from anywhere. End devices in the network can also be managed remotely via the cloud.
- Automation of processes.
- Intelligent buildings and cities.
- Internet of Things.

The above-described technology can be advantageously used for the realization of extensive monitoring systems, which serve for the measurement of variables, and provide wireless transmission to the dispatching centre [6, 8]. The first impulse for applied research in the area of use of IQRF was an idea to create a fully autonomous measuring and monitoring system for the long-term monitoring of temperature, concentration of hazardous gases, air quality and stability of buildings in the area affected by previous mining activities. These areas mainly consist of old mining dumps, but these systems can generally be applied to all storage locations of extractive waste (industrial landfills, etc.).

3 Case Studies

For the Internet of Things based on the IQRF technology, there is a large variety of commercial, but also prototype solutions. The commercial and deployed solutions include the following areas:

- Wireless CO_2 sensor with data visualization on the IQRF Cloud
- Automatic blinds
- Public parking

Among prototype solutions developed by the Department of Cybernetics and Biomedical Engineering team are the following areas:

- Monitoring of environmental variables in mine dumps
- Monitoring of the stability of slopes and building structures

The following subsections describe some of these solutions in detail.

3.1 Wireless CO_2 Sensor with Data Visualization on the IQRF Cloud

Although carbon dioxide is invisible and odourless, its increased level is obvious due to fatigue and decrease in the ability to concentrate. Particularly in areas with large numbers of people such as schools, offices or medical facilities, the negative impact of elevated CO_2 concentration in the air is very significant.

CO2 concentrations up to 5000 ppm do not pose a serious risk to human health. However, according to research, elevated CO_2 concentration leads to drowsiness, lethargy, fatigue and a decrease in the ability to concentrate and an unpleasant feeling of stale air. Some studies have examined the link between increased concentration of CO_2 in the air and drop in productivity and performance. The recommended concentration of CO_2 in the air should ideally be kept below 1000 ppm.

A CO_2 sensor from an unnamed Czech company monitors the air quality in the building and indicates the level of CO_2 using three colour LEDs, so that the user can instantly check the current state of the atmosphere. The measured data is sent to the gateway that sends the data to cloud storage. Above this cloud storage layer operates visualization with advanced options including the ability to control the air conditioning unit.

3.2 Automatic Blinds

Automatic blinds are wirelessly operated automatically or manually, based on the information from the sensors for lighting, humidity and temperature. As in the previous case, the sensor data is sent to the gateway which sends the data to the cloud above which the visualization and evaluation applications run.

3.3 Public Parking

Parking cars in parking spaces is monitored by a Smart Parking System from another Czech company. Smart Parking System is a modular system that allows users to detect occupied parking spaces anonymously. The system in its basic form is used for installation of the detection technology of vehicle presence in selected parking spaces, allowing online monitoring (availability of a particular parking space, the percentage utilization of parking etc.). To manage the parking system on the basis of the continuous collection of data on the occupancy of parking spaces, the system automatically offers the possibility of long-term evaluation of the parking for the adjustment of tariff policy, or for example for comparing and quantifying the difference in the choice of parking charges between the current system (data from the payment terminals) and smartphone system parking (data about the real use of the parking area). In the advanced forms it enables guidance of passenger vehicle drivers to vacant slots before driving using a special web application, or when driving through the installed directional LED signs, or through a developed mobile application. The occupancy data can also be offered to providers of navigation or traffic data for the subsequent development of services for the travelling public.

The presence of parked vehicles is determined by a magnetic detector, which is hidden in the tarmac. The detector sends a signal to the solar-powered GSM gateway located on a street light pole. The gateway forwards the data to the cloud, where it is visualized and processed.

3.4 Monitoring Environmental Values in Mining Dumps

The area of the Moravian-Silesian region, particularly the districts of Ostrava and Karviná, was and partially still is the most important locality of black coal mining. A very negative accompanying effect connected to mining dump administration is the autoxidation processes of coal substance remains present within the dumps and stockpiles. In many cases these effects have been cumulated, resulting in subsurface fire, on a minor or major scale. An extensive part of the mining dumps have been affected by thermal processes in the past and some of them are still active today. Another monitored and highly risky factor at these dumps is the presence of toxic gases, especially poisonous lethal carbon monoxide (CO). Since the beginning of 2014 there have been two measurement locations installed at the Ema dump for the monitoring of dangerous CO. Both particular locations were picked on the basis of long-term manual measurement of the effects of thermal processes. The highest concentration of dangerous gases was found at the installation of the so-called grave-pile, see Fig. 1.

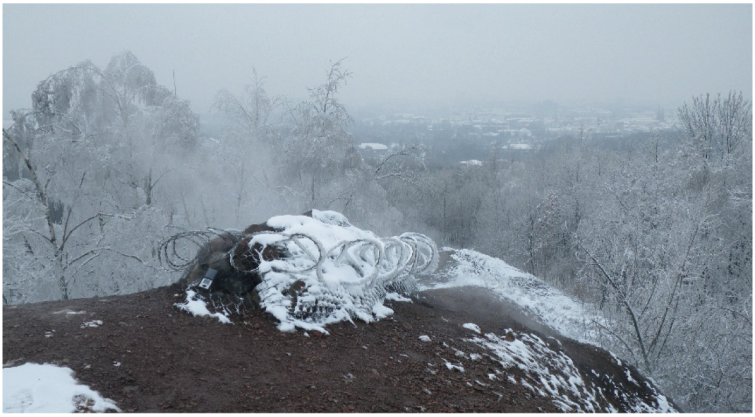


Fig. 1. Dispersion location at the Ema dump in Ostrava, Czech Republic.

The given locations are fitted with special measurement probes containing CO sensors. The primary cell of the CO sensor comprises an IQRF transceiver capable of communication with an IQRF receiver connected to the analogue inputs of the telemetry unit. The data from this unit is then sent by means of GPRS technology to a central database and visualized at the website [10].

This version of the concentration measurement system consists of two modules – transceivers, capable of sending the measured CO concentration by wireless ISM band 868 MHz. The schematic layout is show in Fig. 2.

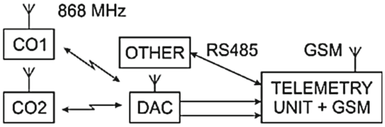


Fig. 2. Measurement of CO concentration at the Ema mining dump.

The “OTHER” block stands for converters for other non-electric signals such as temperatures, light exposure, heat flow and so on [10].

The system for CO concentration measurement basically consists of two parts (nodes), while each node includes a transceiver module, power supply circuit and circuit for analogue signal processing. Detailed structures of the CO and DAC nodes are given in Fig. 3.

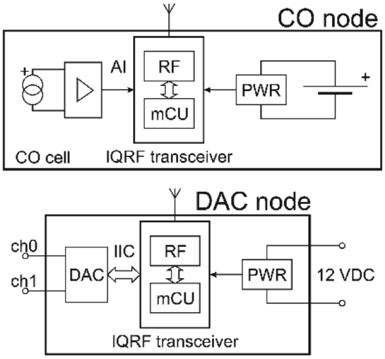


Fig. 3. Inner structures of the CO and DAC nodes.

Measurement of carbon monoxide is performed by a TGS5042 sensor made by the Figaro company. It operates in the range from 10 to 10000 ppm, but a range from 10 to 1000 ppm is applicable. The most important advantage is that there is no need of the power supply as its output is a current in nA, proportional to the measured value of ppm. The output current then depends linearly on the CO concentration.

The DAC node is activated by the external telemetry unit, therefore it is fitted with a supply module with LDO, capable of connecting to 12 V level. The IQRF module in this DAC node does not use an internal digital-analogue converter due to its low bit resolution – 6 bits only. The intended function was achieved by use of an external double channel 10-bit DAC converter connected to the IQRF module through the I2C bus. The output voltage of the DAC converter changes between 0 V to 2.5 V and is connected to the analogue voltage inputs of the telemetry unit [12].

3.5 Monitoring Stability of Slopes and Building Structures

The objective of this use case is the non-traditional measurement of warehouse tilt, using a wireless inclinometer, see Fig. 4. This measurement is conducted by measuring the physical quantity of a three-axis accelerometer, gyroscope and magnetometer. The data given by these devices serves to calculate the object rotation and the trajectory of the measurement module motion in space.

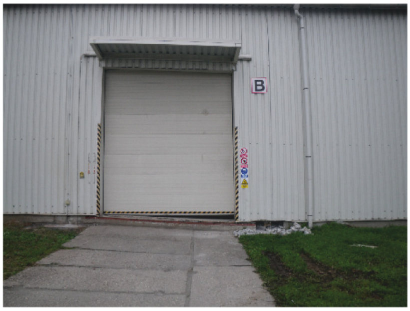


Fig. 4. The consequence of ground descent to the warehouse.

The Measurement Chain

The measurement chain consists of a sensor part, data controlled transceiver modules (DCTR) and a personal computer, which processes and interprets the data measured. Figure 5 shows the block scheme of the measurement chain. The sensory part is given by an AltIMU-10v4 development board; apart from the above-mentioned features, the AltIMU-10 v4 development board also contains a barometer. However, the barometer data regarding the warehouse tilt is not discussed in the present paper. The data reading of these devices is carried out by the IQRF DCTR modules.

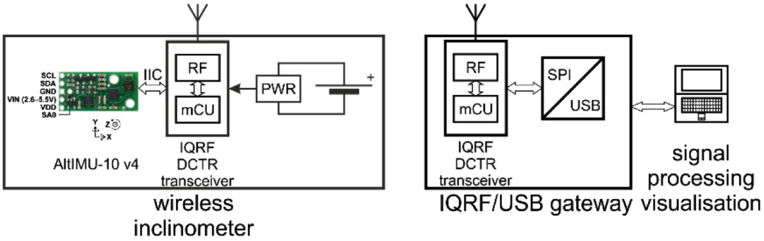


Fig. 5. Measurement chain – peer-to-peer communication.

The measurement chain consists of two independent parts:

- Transmitting part (measurement module/wireless inclinometer) – it measures by current acceleration value sensors, an Earth magnetic field and gyroscope rotation. The DCTR sends the data to the receiving part. The measuring part only reads the data and sends the raw data format to the receiving part; doing this we can avoid any useless DCTR overloading caused by converting the data to physical units as these processes would include floating point operations, which in the case of a 8 bit microcontroller tend to take a considerable amount of computing time.

- Receiving part - the raw data received is forwarded using the IQRF/USB gateway to a PC. This gateway is in CDC mode meaning that the PC identifies it as another serial port. The PC runs NI LabVIEW runtime, which processes the measured raw data by VISA protocol. The data is further filtered by a Kalman filter. After that the data is processed using the AHRS algorithm. The AHRS algorithm calculates quaternions; based on these results it then calculates the measurement module rotation as well as its trajectory.

The Tilt and Motion Trajectory Visualisation

The tilt and motion trajectory of the wireless inclinometer is resolved by the AHRS algorithm [9]. Figure 6 on the left shows the rotation using the AHRS algorithm. The real measurement clearly illustrates that even a slight hand tremor affects the rotation motion and the object is moving also in the other axes. Because the motion was slow, there is a medium size error. In the case of a motion of medium speed the Kalman filter setting is optimal and therefore there was just a minor error. However, in the case of a quicker motion the error is bigger because of the fixed Kalman filter setting. An extremely quick motion cannot be recorded due to the data transmission speed restriction, which causes major errors. In this case it is necessary to reset positions before any new measurement, bringing the object position back to the original value.

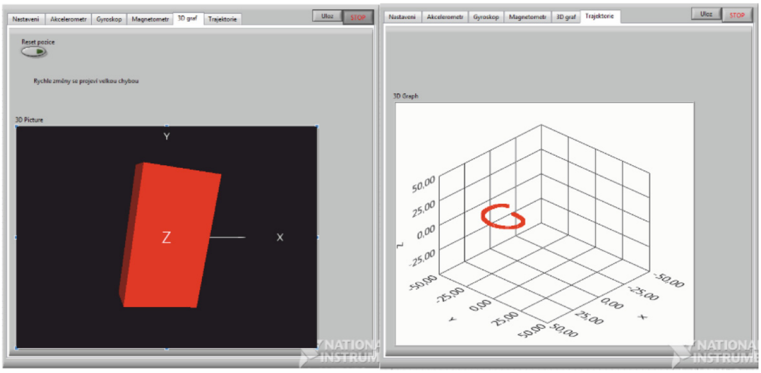

Rychle změny se projeví velkou chybou

Fig. 6. 3D graph of an object tilt and its 3D trajectory.

Figure 6 on the right shows the trajectory of a measured object, the rotation around the *z* axis in particular [4].

4 Conclusion

The paper focused on describing the connection of IQRF technology with the modern and ever-expanding area of IoT. This technology, thanks to its parameters, in particular very low power consumption and simplicity of implementation, can be used advantageously in various application areas of IoT. The Department of Cybernetics and Biomedical Engineering has been engaged with IQRF technology for several years, and the result is several applications, particularly with regard to the monitoring of environmental variables on old mine dumps. Another area in which IQRF is used is the monitoring of the condition of safety nets and barriers. Regarding IoT its development and implementation focuses on modern trends in the field of Smart Cities, Smart Parking, Street Lighting, etc. Currently, the research team at the Department is developing low power gateway for the IQRF network and low power sensors for monitoring the condition of safety nets, barriers and stability of riparian zones.

Acknowledgements. This work was supported by the project of the Ministry of Industry and Trade - MPO FV10706 “Research and development of intelligent sensor network to monitor the safety barriers for geotechnical use”, further by project No. CZ.01.1.02/0.0/0.0/15_019/0004694 of the European structural funds and by project SP2017/158, “Development of algorithms and systems for control, measurement and safety applications III” of the Student Grant System, VSB-TU Ostrava.

References

1. Prauzek M, Krömer P, Rodway J, Musilek P (2016) Differential evolution of fuzzy controller for environmentally-powered wireless sensors. Appl Soft Comput J 48:193–206. doi:10.1016/j.asoc.2016.06.040 Elsevier Science BV, Amsterdam
2. Slanina Z, Docekal T (2016) Energy monitoring and managing for electromobility purposes. In: Proceedings of SPIE - the international society for optical engineering, vol 100311P. doi: 10.1117/12.2247799
3. Calvo I, Gil-García JM, Recio I, López A, Quesada J (2016) Building IoT applications with raspberry Pi and low power IQRF communication modules. Electronics 5(3):54. doi:10.3390/electronics5030054 (Switzerland)
4. Pies M, Hajovsky R: Use of IQRF technology for detection of construction inclination. In: AIP conference proceedings, vol 1738, P 480121 (2016). doi:10.1063/1.4952357
5. Vit J, Krejcar O (2016) Smart solution of alternative energy source for smart houses. Lecture notes in computer science, vol 9799, pp 830–840. doi:10.1007/978-3-319-42007-3_70
6. Hajovsky R, Pies M (2015) Use of IQRF technology for large monitoring systems. IFAC-PapersOnLine 28(4):486–491. doi:10.1016/j.ifacol.2015.07.082
7. Bazydło P, Dąbrowski S, Szewczyk R (2015) Wireless temperature measurement system based on the IQRF platform. In: Advances in intelligent systems and computing, vol 317, pp 281–288. doi:10.1007/978-3-319-10990-9_25
8. Bazydło P, Dąbrowski S, Szewczyk R (2015) Distributed temperature and humidity measurement system utilizing iqmesh wireless routing algorithms. In: Advances in intelligent systems and computing, vol 352, pp 1–9. doi:10.1007/978-3-319-15835-8_1
9. Machaj J, Brida P (2015) Wireless positioning as a cloud based service. Lecture notes in computer science, vol 9012, pp 430–439. doi:10.1007/978-3-319-15705-4_42
10. Pies M, Hajovsky R, Latocha M, Ozana S (2014) Radio telemetry unit for online monitoring system at mining dumps. Appl Mech Mater 548-549:736–743. doi:10.4028/www.scientific.net/AMM.548-549.736
11. Dvorak J, Berger O, Krejcar O (2014) Universal central control of home appliances as an expanding element of the smart home concepts—Case study on low cost smart solution. Lecture notes in computer science, vol 8838, pp 479–488

12. Pies M, Hajovsky R, Ozana S (2014) Wireless measurement of carbon monoxide concentration. In: International conference on control, automation and systems, vol 6987843, pp 567–571, doi:10.1109/ICCAS.2014.6987843
13. Behan M, Krejcar O (2013) Vision of smart home point solution as sustainable intelligent house concept. In: IFAC-PapersOnline, vol 12 (PART 1), pp 383–387. doi: 10.3182/20130925-3-CZ-3023.00057

C-HYFLUR: Recovery for Power-off Failure in Flash Memory Storage Systems Using Compression Scheme for HYbrid FLUsh Recovery

Ji-Hwan Chung, Sungsoo Kim, and Tae-Sun Chung(✉)

Department of Computer Engineering, Ajou University, Suwon, Korea
{batoski10,sskim,tschung}@ajou.ac.kr

Abstract. Flash memory has a variety of advantages such as the better performance compare to hard disk, shock resistance, small size, and so on. Therefore the use ratio of flash memory is being increased. But if the power-off failure is occurred, flash memory storage systems may loss mapping information. So in this paper, we propose power-off recovery scheme, called C-HYFLUR. We have applied C-HYFLUR to page mapping FTL and implemented on a platform board, and compared with the existing recovery schemes through several evaluations. Compare to existing recovery schemes, the overhead of C-HYFLUR is very low.

Keywords: Power-off failure recovery · Compression · FTL · Flash memory

1 Introduction

Nowadays, the use ratio of flash memory is very high in many storage systems. This is because that flash memory has a variety of advantages such as better performance compared to hard disk, light weight, shock resistance, and low power consumption. Thus, flash memory is used in various fields and is used as an essential element of many embedded systems. However, flash memory has some drawbacks because of the constraints on the hardware level. For example, when write operations are performed to the location that some data is already written, the erase operation should be performed before the write operation. This is called erase-before-write architecture [1]. Because of this problem, flash memory has a performance degradation.

To reduce the performance degradation, flash memory uses an FTL (Flash Translation Layer) [2] algorithms. The FTL algorithm performs translating logical addresses from the file system to physical addresses to flash memory. And it can be classified in three schemes, page mapping scheme, block mapping scheme and hybrid mapping scheme [3]. Shortly, block mapping scheme is to translate lbn (logical block number) to pbn (physical block number). Similarly, page mapping scheme is to translate lpn (logical page number) to ppn (physical page number) and the hybrid mapping scheme is mixture of block and page mapping scheme.

Another disadvantage of flash memory is that it is more expensive than other storage devices and that there exist hardware errors due to sudden power-off. First, flash memory is very expensive compared to hard disk, so it is very difficult to apply it to a system even

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_32

though it has better performance than hard disk. To overcome these drawbacks, various algorithms have been developed to compress and store data in flash memory [4–6].

Next discussion point of the drawbacks of flash memory is hardware level faults. Various hardware level faults exist in flash memory, and this paper focuses on the sudden power-off failure. Power-off failure means that a sudden loss of power that is supplied to flash memory. The sudden power-off failure is still occurred [7], and it is a critical problem in data and metadata reliability [8]. The data reliability problem is very important point to the storage device based on flash memory. And metadata such as mapping information is required in the FTL operation and this data is important. So metadata reliability is also very important point.

To prevent this problem, many recovery schemes are proposed [8–11]. Usually it can be divided into three schemes, In-Block Backup, In-Page Backup, and Hybrid Backup.

Briefly, a certain blocks in flash memory is reserved to save metadata in In-Block Backup, metadata is always stored in the spare area of the page in In-Page Backup, and finally Hybrid Backup is the mixture of In-Block and In-Page Backup.

In this paper, we focus on Hybrid Backup scheme and the page mapping FTL algorithm. We apply the compression algorithm to existing HYFLUR [12] algorithm to solve the capacity problem which is a disadvantage of the page mapping algorithm. So in this paper, we propose a C-HYFLUR (compression scheme for hybrid flash recovery) scheme on the page mapping FTL.

The rest of this paper is organized as follows. Section 2 is the background of flash memory, power-off failure and compression algorithm. Section 3 gives the related work of recovery scheme of power-off failure. In Sect. 4, we propose our scheme C-HYFLUR. And Sect. 5 presents evaluation results. Finally, Sect. 6 is the conclusion.

2 Background

2.1 Power Source for Flash Memory

All hardware systems require power to operate. Because flash memory is also a hardware system, power can be seen as a very important source of flash memory. However, the power supply of flash memory may become unstable in a real environment, and if the power is lost, a serious error occurs in flash memory [7, 8]. Therefore, a recovery algorithm is required to help flash memory to operate normally even if an unexpected power-off occurs.

There are two types of power-off in flash memory: normal power-off and sudden power-off. Normal power-off is literally normal power interruption and so does not cause serious damage to flash memory. However, in case of sudden power-off, an unexpected power-off occurs, which can cause serious damage to flash memory. Therefore, this paper focuses on the flash memory recovery algorithm for sudden power-off.

2.2 FTL Mapping Algorithm

FTL is a firmware that manages the host operating system and flash memory [2]. As mentioned earlier, FTL algorithm can be classified into three schemes: block map-ping, page mapping and hybrid mapping [3]. And each of the schemes is divided based on the mapping unit. Block mapping is a mapping in block units, page map-ping is a mapping in page units and hybrid mapping is mixture of block and page mapping. The existing HYFLUR algorithm [12] focuses on the page mapping FTL. Therefore, in this paper, we also propose C-HYFLUR, an improved version of HYFLUR, focusing on the page mapping.

Basically, FTL has metadata for managing flash memory. Metadata refers to da-ta for managing flash memory, and mapping information also corresponds to metadata. Most of the mapping information is composed of mapping table, and it plays a role of matching lpn of operating system and ppn of flash memory. There-fore, if the operating system requests lpn to read or write specific data, the FTL will find the ppn corresponding to lpn and help to access the flash memory.

2.3 Flush Scheme for Power-off Failure

When the mapping table changes in the page mapping FTL, existing data is updated or erased from flash memory. First, if existing data is updated, flash memory must write data to another address because of the feature that it cannot be overwritten. Accordingly, the mapping table must also be modified to the changed address, and if not modified, the updated data cannot be accessed. Next, when the mapping table is modified, the data is deleted from flash memory. If the data is erased by the host or the garbage collection of the FTL, the mapping table must also remove address information for the erased data.

When the mapping table is updated as described above, it is necessary to store the changed information in flash memory to recover a power-off failure occurring later. However, storing the mapping information in flash memory every time data is updated causes the performance of flash memory to be degraded, and the storage of excessive mapping information may lead to a capacity shortage. There-fore, in order to solve the above problem, the FTL performs a data backup operation called ‘flush’. FTL generates flush processes periodically and usually stores mapping information in the flash memory’s data area or spare area.

2.4 Compression Algorithm for Page Mapping Algorithm

In the case of page mapping FTL, unlike other mapping FTLs, the mapping information occupies a large amount in flash memory. However, it is inefficient to store large amounts of mapping tables in expensive flash memory compared to hard disk, as it causes space overhead in flash memory. Therefore, this paper focuses on the compression algorithm to reduce page mapping information.

In order to solve these problems, we apply Lempel-Ziv-Storer-Szymanski (LZSS) algorithm, which is an improved version of the LZ77 algorithm [13]. By applying LZSS algorithm to HYFLUR, we can overcome space overhead which is a disadvantage of page mapping FTL and recover large amount of mapping table with small amount of mapping information. Detailed description of this process is in shown Sect. 4.

3 Related Work

3.1 Compression Algorithm for FTL

Many FTL algorithms with compression techniques have been addressed [4–6]. A compression layer for smart media cards [4], a real-time compression scheme for devices using flash memory [5], and zFTL [6] containing compression algorithms are proposed. Additionally, the MEW (Metadata Embedded Write) scheme [14] deals a compression algorithm with power-off recovery.

The main motivation for the above paper is to reduce the amount of data written to flash memory, and authors proposed a scheme applying different compression algorithms. In this paper, we apply LZSS algorithm to HYFLUR algorithm to reduce space overhead of flash memory.

3.2 In-Block Backup

The In-Block Backup scheme stores the backup data at a set of reserved blocks in flash memory, and this set of blocks is called map-blocks. At first, this scheme prepares map blocks for storing mapping information in flash memory. So when the mapping information is generated, this scheme stores it in map-blocks [10]. As this scheme also generates the additional write operation, it will create additional overhead.

When there is a power-off failure in flash memory, the recovery module will scan map-blocks. And based on the scanned mapping information, it constitutes the mapping tables. Compared to the In-page Backup recovery scheme, the recovery delay of In-Block Backup recovery scheme is small.

3.3 In-Page Backup

This scheme stores the backup data in spare area of pages. Generally, spare area is used to store FTL management information and ECC bits. However, as the page size of flash memory is increased with the development of technology, spare area is also increased. So this scheme can write the backup data on the unused area [10, 11]. Also when this scheme is executed in flash memory, it does not create additional overhead.

And when there is power-off failure in flash memory, the recovery scheme will be executed. In this case, the recovery scheme is very simple. First, it scans all spare areas of pages in flash memory. And with this scanned information, it constitutes the mapping tables. This scheme is very simple, but the long recovery delay is possible. In this paper, we compare the recovery delay of this scheme and our proposed scheme.

3.4 A-PLR

As we mentioned above, the main drawback of In-Page Backup is the recovery delay. Because of this drawback, Jung et al. [11] proposed enhanced scheme of the In-Page Backup, called A-PLR (Accumulation based Power Loss Recovery). Also this scheme stores the mapping information in spare area in a manner similar to the In-Page Backup. However unlike the In-Page Backup, the mapping information to be stored in the spare area different way.

By using other methods in this scheme, it was able to solve the drawbacks of the In-Page Backup. Briefly, the process of the method is as follows. At first, A-PLR reserves the special buffer in RAM called MIB (Mapping Information Buffer). This buffer size is fixed, and it stores the accumulation of mapping information. When A-PLR stores the mapping information in the spare area, it stores the MIB. As the A-PLR stores more mapping information, the size of MIB is increased. When the size of MIB is full, it is also stored in spare area and this page is called the intermediate page. Also this page is used in the recovery process.

Recovery process of the A-PLR is very simple. When the recovery process is running because of a power-off failure, A-PLR uses intermediate pages. For example, A-PLR scans intermediate pages instead of scanning all the pages. Therefore, recovery delay is lower than the In-Page Backup. So A-PLR enables the mapping tables to be reconstructed through scanning the intermediate pages. Also in this paper, we com-pare the recovery delay of the scheme to our proposed scheme.

4 C-HYFLUR (Compression Scheme for HYbrid FLUsh Recovery)

4.1 Overview

We propose C-HYFLUR with the compression algorithm applied to the recovery algorithm HYFLUR [12] proposed in the previous paper. The compression algorithm we apply is the LZSS algorithm, which is an enhancement of LZ77 [13], and it is applied to the existing HYFLUR algorithm to solve the space overhead which is a disadvantage of the page mapping scheme.

C-HYFLUR is implemented in the Open SSD Project Board [15] and the hardware specification is as follows. The platform uses NAND flash [16] of the 64 GB, and this memory is composed of 8 KB pages. Also blocks of the memory is composed of 128 pages. In summary, NAND memory consists of total 66,432 blocks. Also this plat-form uses mobile SDRAM [17] of the 64 MB.

Also our scheme reserves the particular block of flash memory as well as certain memory of RAM to store all mapping information. And the blocks don't store data. For storing mapping information and tables, we reserve the 15 blocks and 25 blocks, respectively. Also we call each block MSB (mapping information store block) and TSB (table store block).

Table 1 below shows the data storage ratio of HYFLUR, which is the conventional recovery algorithm, and C-HYFLUR, which applied the compression algorithm. The compression ratio of LZSS is about 31%, and data of 8 K can be compressed to about

2.6 K. Mapping information stored in the MSB of the existing HYFLUR can store update data of about 31% of flash memory, but the update data of C-HYFLUR can be stored about three times as much as the HYFLUR. Also, TSB requires 25 blocks to store all mapping tables, but C-HYFLUR can store all mapping tables with only 9 blocks. This can solve the space overhead which is a disadvantage of the page mapping FTL.

Table 1. Data storage ratio

	C-HYFLUR	HYFLUR
MSB	Store 92% of update data	Store 31% of update data
TSB	9 blocks	25 blocks

4.2 Flush Process

As we mentioned earlier, the flush process is to transfer information from RAM into flash memory. In our scheme, the flush is divided into three sub flush: SWF (spare write flush), URF (update RAM flush) and MTF (mapping table flush). And Fig. 1 shows that how the three flush operations are performed with timeline. In our proposed scheme, MTF, URF and SWF are operated by the appropriate policies.

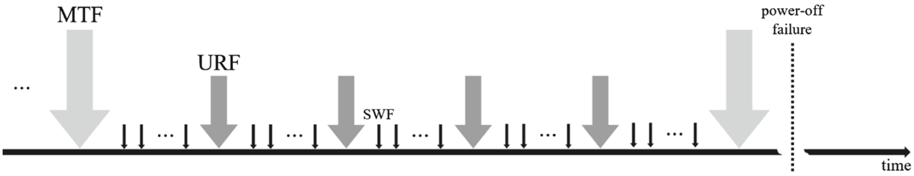


Fig. 1. Flush with timeline

The SWF operation is very simple operation. For example, if the page is to be updated, data should be write to another page. This situation evokes changing mapping information of ppn (physical page number). In our scheme, the change between old and new ppn is written to RAM. In this situation, old and new ppn should be stored in the spare area. And this process is called SWF.

The MTF and URF operation will be discussed in more detail in the following sections.

4.2.1 URF Operation

As mentioned earlier, when a page is updated, information of ppn is written in the 1st RAM. For storing ppn, our scheme assigned 6 byte of 8 KB. Therefore, it is possible to write 1365 ppn in 8 KB memory. This means that it can store mapping information of 1365 pages. And, as shown in Fig. 2, when the 1st RAM is full, the compression process is performed and the compressed data is stored again in 2nd RAM. This process is repeated three times to make 2nd RAM full. Then URF operation as shown in Fig. 3 is started.

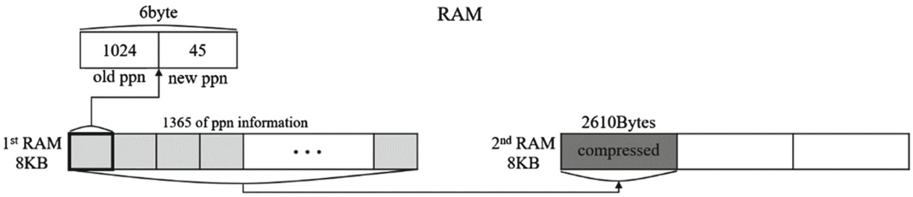


Fig. 2. RAM configuration

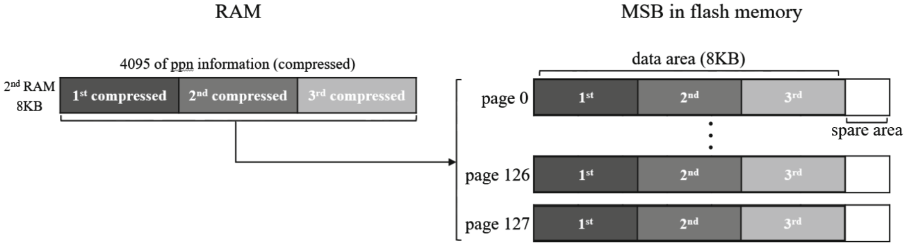


Fig. 3. URF operation

Figure 3 shows a brief URF (update RAM flush) operation. If the 8 KB of 2nd RAM is full, URF operation is performed. Also this means that approximately 32 MB of flash memory has been updated. And URF operation is performed and all information is written to the pages of the MSB. And MSB can store compressed mapping information about 92.4% of flash memory. This provides the space to store most of the update data.

4.2.2 MTF Operation

The MTF (mapping table flush) operation is the process of writing all mapping information in flash memory. For storing mapping table, our scheme assigned 3byte for writing ppn, compresses the three 8 KB pages with 2730 ppn, and writes to one page of TSB. This operation is occurred when MSB is full because of the URF operation.

So when MTF operation is performed, it creates a new compressed mapping table by using the existing mapping table and MSB. Additionally, it stores new compressed mapping table information, as shown in Fig. 4 and writes the ppn in pages of TSB. For example, when MTF operation is executed, new ppn is written to pages and these page index is become lpn. It compresses three 8 KB pages and writes them to one page of TSB, so that total 8190 ppn can be stored in on one page. Compared with the existing algorithm HYFLUR, HYFLUR requires 25 blocks in total to store all mapping tables, but C-HYFLUR requires only 9 blocks. Therefore, C-HYFLUR can reduce space overhead, which is a disadvantage of page mapping FTL.

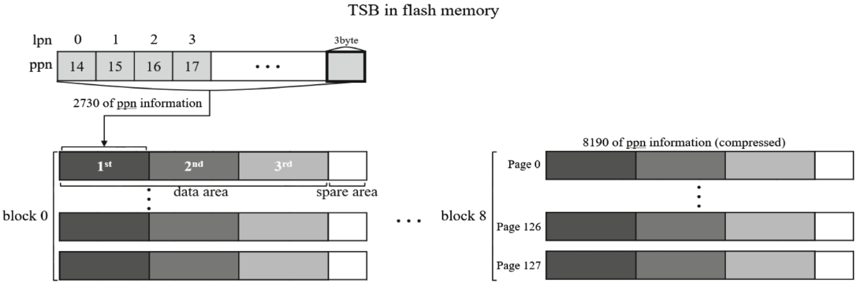


Fig. 4. The structure of TSB

4.3 Power-off Failure Recovery Process

Our power-off recovery process is divided into 3 cases. To recover mapping in-formation, (1) reading the pages of MSB created by URF operations & decompression processes, (2) reading only TSB created by MTF operations & decompression processes and (3) reading MSB and TSB & decompression processes.

(1) At the first case, when power-off failure occurs, TSB is not generated. There-fore, our technique reads the existing mapping table and pages of MSB for recovering mapping table. However, since the pages storing the mapping information are compressed, the decompression process should be performed. Finally, it reads the spare area of pages to find out mapping information created by the SWF operation.
(2) Second case is power-off failure occurs when TSB is made. In this situation, our recovery scheme reads only TSB for creating mapping table. This is because that all mapping information is stored in TSB. Therefore, no need to reading pages of MSB. This process also requires decompression the compressed pages to complete the mapping table. And our proposed scheme has the best efficiency in this situation.
(3) The last case is power-off failure occurs when TSB is made and before mak-ing the new TSB. In this case, our recovery scheme read existing TSB and pages of MSB. It also involves a decompression process. And finally read the spare area of pages to find out mapping information created by the SWF operation.

So our proposed scheme is possible to recover all mapping information in all three cases.

5 Evaluation

5.1 Evaluation Setup

In this section, the proposed scheme C-HYFLUR is implemented on Open SSD Project Board [15]. Figure 5 shows the real board snapshot and hardware architecture, also this board using an SSD [16], SDRAM [17] and so on. Also this board employs an ARM7TDMI-S core (running at 87.5 MHz) and a SATA 2.0 host interface (3Gbps). The objective of the evaluation is comparing the overhead of recovery to the existing schemes (In-Page Backup, A-PLR, HYFLUR) and C-HYFLUR. And we use benchmark Iometer [18] to evaluation.

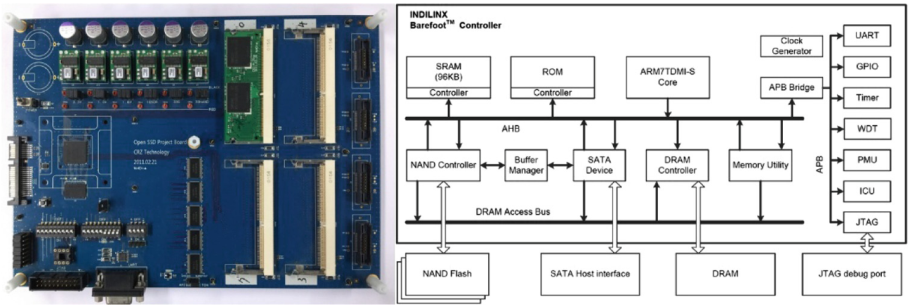


Fig. 5. Open SSD project board and platform H/W architecture [9]

5.2 Evaluation Results and Discussion

In this section, the first evaluation is to estimate the response time of the platform board. Figure 6 above shows the data storage ratio of the HYFLUR algorithm and C-HYFLUR and the time it takes to read the blocks to recovery. The compression ratio of the LZSS is about 31%, and since the data of 8 K can be compressed to about 2.6 K, there exists the difference as in Fig. 6.

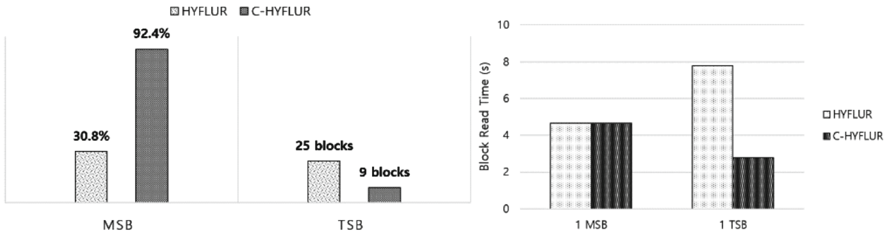


Fig. 6. Compression efficiency and block read time

The mapping information stored in the MSB of the existing HYFLUR can store the update data in about 30.8% of flash memory, but the update data of C-HYFLUR can be stored in 92.4% of flash memory. Also, TSB requires 25 blocks to store all mapping tables, but C-HYFLUR can store all mapping tables with only 9 blocks. This can solve the space overhead which is a disadvantage of the page mapping FTL. Also, as shown in the graph on the right, C-HYFLUR has a significantly lower time to read TSB for recovery than HYFLUR. This means that the compression algorithm can reduce the time it takes for C-HYFLUR to recover (Fig. 7).

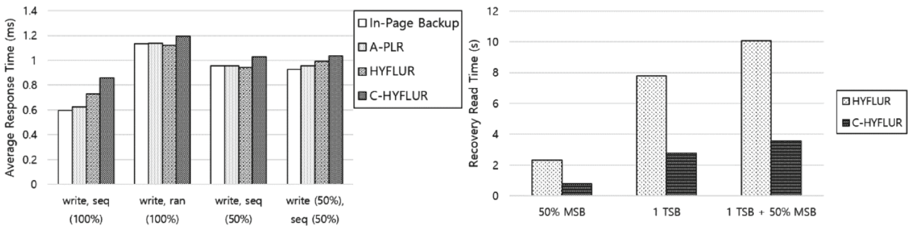


Fig. 7. Average response time recovery read time

The next evaluation is to evaluate the average response time of the platform board and the time it takes to read the mapping information to recover. And in this evaluation, the benchmark Iometer [18] is used to measure the response time. Before the left graph evaluation, we divided into four cases to measure the results. First, only write operations are issued and all requests are sequential (100%). Second, only write operations are issued and all requests are random (100%). Third, only write operations are issued and request are sequential and random (50%). The last case is write (50%) and read (50%) operation and request is sequential (50%) and random (50%). Overall, C-HYFLUR is similar to the other scheme in response time. However C-HYFLUR shows low average response time in the left graph, these results are negligible because all of time is very narrow margin.

The graph on the right is a graph of the three recovery process cases described ear-lier in Sect. 4.3. The time taken to read the mapping information to recover the existing HYFLUR algorithm and the C-HYFLUR algorithm is compared for the three recovery processes. The three recovery processes are: (1) updated information due to URF operation is stored in 50% of MSB, (2) one TSB is created due to MTF operation, and (3) MSB and TSB coexist. Overall, the recov-ery read performance of C-HYFLUR is much faster than that of HYFLUR. This is because C-HYFLUR applies the compression algorithm, and it requires fewer page reads than HYFLUR. Therefore, C-HYFLUR recovery read time overhead is very lower than HYFLUR.

6 Conclusion

This paper proposes the recovery scheme for power-off failure, called C-HYFLUR. C-HYFLUR applies the page mapping FTL and adds compression algorithm to existing HYFLUR. For recovery, C-HYFLUR using URF and MTF operation, and reserves special blocks MSB and TSB. And URF and MTF operation reduces the number of the write operations for storing mapping information. In addition, compression algorithm reduces the number of read operation, and resolves the space overhead of the page mapping FTL. Therefore C-HYFLUR recovery time overhead is very lower than other recovery schemes. However, the C-HYFLUR response time is little longer than other schemes. Because additional page write operations and compression process is required. But these overheads are negligible.

Acknowledgement. This research was supported by the MISP (Ministry of Science, ICT & Future Planning), Korea, under the National Program for Excellence in SW supervised by the IITP (Institute for Information & communications Technology Promotion) (2015-0-00908)

References

1. Samsung Electronics (2007) NAND Flash Memory &Smart media Data book
2. Ban A (1995) Flash file system, United States Patent No. 5,404,485
3. Chung T-S, Park D-J, Park S-W, Lee D-H, Lee S-W, Song H-J (2009) A survey of flash translation layer. J Syst Archit 55:332–343
4. Yim KS, Bahn H, Koh K (2004) A flash compression layer for SmartMedia card systems. IEEE Trans Consum Electr 50:192–197
5. Huang WT, Chen CT, Chen CH (2007) The real-time compression layer for flash memory in mobile multimedia devices. In: International conference on multimedia and ubiquitous engineering (MUE 2007)
6. Park Y, Kim J-S (2011) zFTL"power-efficient data compression support for NAND flash-based consumer electronics devices". IEEE Trans Consum Electr 57:1148–1156
7. Zheng M, Tucek J, Qin F, Lillibridge M (2013) Understanding the robustness of SSDs under power fault. In: Proceedings of the 11th USENIX conference on file and storage technologies (FAST 2013)
8. Tseng H-W, Grupp L, Swanson S (2011) Understanding the impact of power loss on flash memory. In: Proceedings of the 48th design automation conference (DAC 2011)
9. Zhang C, Wang Y, Wang T, Chen R, Liu D, Shao Z (2014) Deterministic crash recovery for NAND flash based storage systems. In: Proceedings of the 51th design automation conference (DAC 2014)
10. Chung T-S, Lee M, Ryu Y, Lee K (2008) PORCE: an efficient power off recovery scheme for flash memory. J Syst Archit 54(10):935–943
11. Jung S, Song YH (2015) Data loss recovery for power failure in flash memory storage systems. J Syst Archit 61(1):12–27
12. Chung J-H, Chung T-S (2016) HYFLUR: recovery for power-off failure in flash memory storage systems using HYbrid FLUsh recovery. In: Information science and applications (ICISA 2016)
13. Ziv J, Lempel A (1977) A universal algorithm for sequential data compression. IEEE Trans Inf Theor 23:337–343
14. Kim D, Won Y, Cha J, Yoon S, Choi J, Kang S (2016) Exploiting compression-induced internal fragmentation for power-off recovery in SSD. IEEE Trans Comput 65:1720–1733
15. Indilinx Jasmine Platform Specification. http://www.openssd-project.org/wiki/The_OpenSSD_Project/
16. Samsung Electronics, K9LCG08U1 M Information. http://www.elnec.com/device/Samsung/K9LCG08U1M+%5BLGA52%5D/
17. Samsung mobile SDRAM chip, K4M51323PG-HG75, Datasheet
18. Iometer benchmark. http://www.iometer.org/

ERF: Efficient Cache Eviction Strategy for E-commerce Applications

Jung Hwa Lee[1], Se Jin Kwon[2], and Tae-Sun Chung[3](✉)

[1] Software Engineering, Ajou University, Suwon, Republic of Korea
paychecks@ajou.ac.kr

[2] Department of Computer Engineering, Kangwon National University,
Samcheok, Republic of Korea
sjkwon@kangwon.ac.kr

[3] Computer Engineering, Ajou University, Suwon, Republic of Korea
tschung@ajou.ac.kr

Abstract. Most vendors of e-commerce applications deploy the cache memory to deliver the web objects to clients faster. However, they face many problems in dealing with the cache memory due to limited resources and dynamic access patterns. As a result, we need to efficiently manage the cache memory by evicting the unused data. The performance of cache manager depends upon the efficiency of delete determination. In this paper, we propose ERF, a cache eviction policy using natural exponential function on time with frequency in order to cope with dynamic nature of e-commerce business with limited memory. It sorts the caches in the order of result value which come from coordination between frequency and recency and evicts the caches according to it. We evaluate the performance of ERF by using the workload which reflects the real-world applications and compare it with conventional algorithms. By increasing the cache hit ratio with ERF, we can expect the decrease of copy and delete operations of cache with improving the overall system performance.

Keywords: Multimedia databases and file systems

1 Introduction

In e-commerce business, there are lots of information about clients and products and the information is updated continuously. Due to a large amount of information, it is necessary to manage the information in efficient manner. So, most of E-commerce Web sites especially, online shopping malls are based on database to manage the information easily. Thus, the importance of utilizing database is on the rise. Furthermore, response time to clients' request via database processing is a critical one for e-commerce Web sites. Website process is expected to be loaded within 2 s [1]. Moreover 40% of clients tend to leave the website if the time to load the web objects takes more than 3 s. Clients are not apt to keep waiting for the site to response [1]. To reduce the response time, a

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_33

large number of e-commerce vendors use caching strategies to speed up the transmission. The architecture of database-driven commercial Web sites incorporates the association between cache and DBMS.

In case of online malls, there are many problems in dealing with the caches due to dynamically changed demands for items. Lots of items are created and discarded from day to day. However, conventional algorithms for managing the caches do not take into account fluctuation of demands for items.

In this paper, we bring up the cache management for e-commerce malls in terms of thumbnail delivery and propose to consider the change of demands in e-commerce environment for efficient caching management algorithm. Our technique minimizes the copy and the delete operations for thumbnail cache. Also, the experiment result shows high hit ratio, the probability of cache hit on the total number of data requests [2].

The remainder of this paper is organized as follows. We examine thumbnail delivery solutions and address the performance problem when improper cache eviction strategy is applied in Sect. 2. We examine the conventional cache management algorithms in Sect. 3. Then, we propose ERF which is a cache eviction policy using natural exponential function on time with frequency in Sect. 4. Finally, we compare ERF to previous algorithms in Sect. 5 and conclude our paper in Sect. 6.

2 Motivation

Thumbnail images are deployed for improvement in recognizing the original images in many e-commerce applications. By deployment of thumbnail images, it solves the latency of downloading web objects which include images and pages. There are two alternative ways to provide thumbnail images to clients considering response time to clients' requests. One possible solution is resizing the whole set of images to thumbnails in advance which is generally used. Whenever user uploads original images to NAS (Network Attached Storage), system resizes the images to thumbnail images. It can improve the clients' response time, but this solution can drastically increase the memory consumption for cache so that it is hard to be implemented due to limited resources in real world.

Given the limited resources for cache, there is a solution called dynamic cache (as shown in Fig. 1), which resizes images whenever there is a request to a thumbnail. The storage for dynamic cache can be divided into main storage and cache which stores the original images and the thumbnails respectively. The system with the dynamic cache resizes the original images to thumbnail images and inserts resizing result to main DB. Then, it delivers the objects to cache in NAS.

Due to limited resources for cache, dynamic cache needs cache manager which evicts the caches when the cache memory exceeds the threshold. Resizing the images on the fly by resizing server, in fact, requires more CPU and RAM than dealing with one general request. Therefore, the importance of leaving the objects which will be frequently used in the near future in cache is stressed.

Whenever there is a request to thumbnails, dynamic cache searches the relevant cache and returns the cached images to clients. If there is no cache for request, dynamic

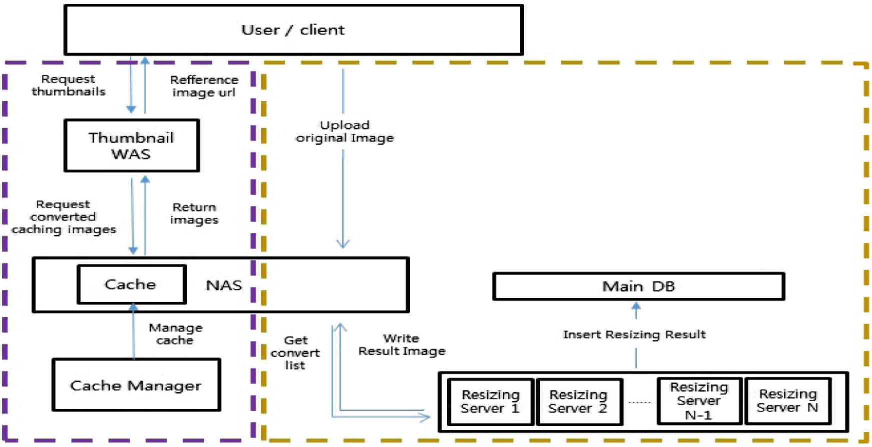


Fig. 1. The structure of dynamic cache

cache should perform the image resizing operations. And it sends the thumbnail images to NAS and saves it. Then, it updates database according to it, and returns the cached images to clients (as shown in Fig. 2).

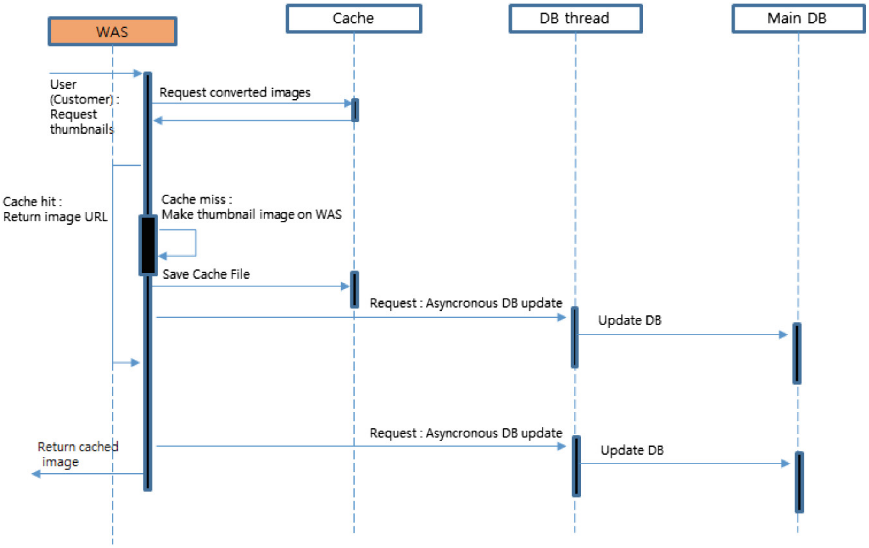

WAS
Cache
DB thread
Main DB
User (Customer) : Request thumbnails
Request converted images
Cache hit : Return image URL
Cache miss : Make thumbnail image on WAS
Save Cache File
Request : Asyncronous DB update
Update DB
Request : Asyncronous DB update
Update DB
Return cached image

Fig. 2. Thumbnail cache request flow in dynamic cache

As the resizing operation contains a series of request, insert and convert operation, it is crucial overhead to system performance. In the worst case, cache manager can delete the frequently used caches. Then, there has no choice but to impose a large number of cache miss in the future in a given condition and it can be a major obstacle for performance in terms of response time to clients' request. It motivates us to come up with the cache eviction policy considering the characteristics of the real-world workloads. We have conducted experiments to evaluate our cache eviction algorithms and compare its performance with LRU and LFU algorithms.

3 Related Works

In this section, we examine some popular approaches for cache replacements strategies. First of all, when a requested object does not appear in cache, the cache gets the requested object and stores it in cache. Given the above condition, when the cache memory exceeds the fixed threshold *T*, a portion of cached objects are evicted until the cache memory decreases to given size *t, with* $T > t$ [3].

Frequency and recency are crucial factors for cache eviction processes. They have association with two types of locality, temporal locality and spatial locality. Temporal locality is an access pattern to the identical objects within recent period whereas spatial locality is based on the number of occurrence of accesses to objects within whole time [5]. Namely, we can expect the pattern of request streams in near future based on each factors.

3.1 Recency-Based Strategy

This strategy utilizes the temporal locality for cache eviction processes. Temporal locality is preference for the identical objects within recent period. The initial concept of recency-based strategy is evicting the least recently used objects.

Recent researches have shown that the LRU replacement strategy has potential for performance improvement. As a result, most of researches have devoted to improve LLC (Last Level Cache) replacement [7, 8]. Even if recency-based strategy does not take into account the frequency information, it is widely used for different areas. But it can be the major obstacles to react dynamically changed access patterns.

LRU

The last accessed object is always acquired. If the cache memory exceeds T, it evicts the least recently used objects [4].

LRU-Thold

This strategy does keep the cache memory from receiving an object if the object causes eviction process which removes a lot of remaining cache. If not, it operates as LRU [5].

Size

It sorts caches based on the size of files. Largest file is removed first whereas same sizes of files are under LRU strategy. As most of references are interrelated with small files, it makes high hit ratio [6].

3.2 Frequency-Based Strategy

Frequency-based strategies utilize the frequency as a main factor to judge whether the object will be a target of an eviction process. It shows strength in static environments. But it can be confused because objects can have same values. Furthermore, it needs aging technique due to static value of frequency [3].

LFU
This strategy keeps the most frequently accessed objects by far in cache [4].

Perfect LFU (PLFU)
This strategy keeps the histogram of frequency about all objects ever accessed before. As it has to keep the all stats, it can cause space overhead [4].

In-Memory LFU
It only keeps the access frequency in cache and acquires the most recently accessed objects in cache. It evicts the least frequently used objects [4].

3.3 Other Strategy

DIP (Dynamic Insertion Policy) [9] is a hybrid caching strategy that dynamically calculates the number of misses incurred by LRU and LFU policies. It selects a strategy that makes comparatively lower misses. But it incurs overhead for verifying the cache misses from two different caching strategies.

4 Our Approach

We observed that due to dynamic nature of E-commerce business, the access patterns to items of online malls are various as day goes by, as hour goes by, as minutes goes by. As web requests show a certain level of temporal locality, information about recency is a significant part for the victim selection process.

With strengthening the strength and making up for the weakness of LRU and LFU, we used two eviction processes at once at the expense of overhead caused by bringing two eviction processes at once. First, to combine spatial and temporal locality together maintaining their characteristics, we use $1/e^x$ equation and use the cache hit count and recent accessed time of items which is remaining in cache.

$$Eviction\,Value = \frac{\text{CacheHitCount}}{exp\left(\frac{Current\,Time - Recent\,Accessed\,Time}{600}\right)} \tag{1}$$

In fact, it is hard to reconcile the time value with cache hit count due to drastically increased value of cache hit count and difficulty in controlling time value. Notwithstanding the elapse of time and drastically changed cache hit count, we can adequately cope with dynamically changed demands for items in E-commerce business with $1/e^x$ equation with $x = \textit{difference between current time and recent accessed time}$. As shown in Fig. 3, if x exceeds certain value, y converges to zero. Value of (1) converges to zero likewise. Thus, the objects which remain in cache without any access over certain time cannot help having high priority of target to be evicted. With this approach, we can solve the situation that an item which is very popular in the past, rests in the cache even if there is no access to the item for a long time.

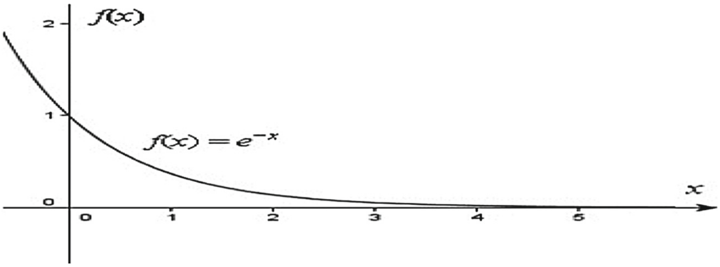


Fig. 3. f(x) = e^{-x} graph

However, in a given equation, the last accessed items can be evicted due to low frequency count even if they have high value of time and can be accessed in near future. To prevent this situation, we adopt one more process to evict items which remain in cache over a long time without any access. With these processes, we can expect conformity to dynamic nature of e-commerce business. Also it is just based on the query result set, it does not cause changes to conventional DBMS systems.

5 Experiments

We have conducted experiments to evaluate our cache eviction algorithm (*ERF: natural Exponential function on Recent access time with Frequency*) and compare its performance with LRU and LFU algorithms in perspective of hit ratio.

The general experimental settings are as follows. The total number of the original images in the system is 5000 whereas the condition when the cache eviction operation is implemented is when the number of tuples in cache exceeds over 1000 threshold (T). Cache eviction operation makes the number of objects remaining in cache be 800 (t). This ratio between the values of threshold and remaining caches reflects the cache managers deployed in real-world applications.

We made a trace which reflects the real-world workloads in efficient manner. First, we divided all of the original images to three groups, assuming that one (i.e. group A) consists of items frequently used in general and another (i.e. group B) contains items which are used less than former one. Items which are scarcely used are classified as group C.

Given the above conditions, the numbers of images are 350, 750 and 3900 in each group A, B and C respectively. Then, the access rate to thumbnail cache is fifty per a second. All of fifty images are picked out from one of three groups. And the probability of choosing a group among three groups is 70%, 20% and 10% for each group A, B and C as shown in Table 1 to reflect the access locality of real-world.

Table 1. The environments of trace build

Group	The number of images	Selection probability
A	350	70%
B	750	20%
C	3900	10%

It updates fifty objects in the selected group in every second and changes the constitution of the each group in every ten minutes. Since we take into account the change of preference of clients' accesses, we change the composition of each group periodically. When we change the composition of each group, twenty percent of objects in group A are ejected to group C and fifty percent of objects in group B are ejected to group C. Then, the equivalent percent of objects in group C are put into each group. Finally, experiments are implemented on a 2-hour basis. The cache eviction operation is implemented whenever the number of objects meets the condition. If the condition of the cache memory is met, it evicts items which remain in cache over a long time without any access. After that, it sorts the caches into the order of result of above algorithm and evicts the caches according to it until the number of objects remaining in cache be 800 (t).

Given the above conditions, we compared five eviction processes to each other. First, we can notice that the delete counts jagged before the cache counts drastically increased in LRU (as shown in Fig. 4). It shows that some items are evicted even if they have higher value in cache counts than others. In contrast, delete counts show even form before the sudden increase of cache counts but it almost never appear after drastic increase of cache counts in LFU as shown in Fig. 5. We can realize that items which

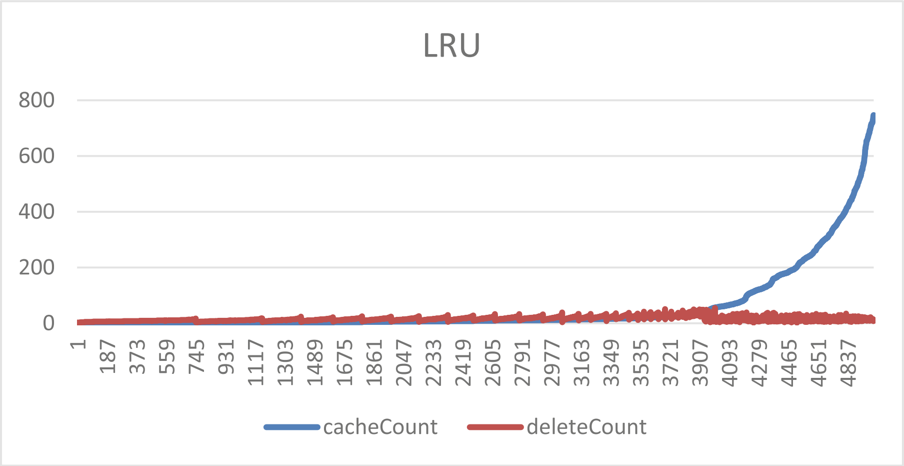


Fig. 4. Relationship between cache count and delete count in LRU

have much higher frequency value than others are scarcely deleted even if there is no access to them for a long time.

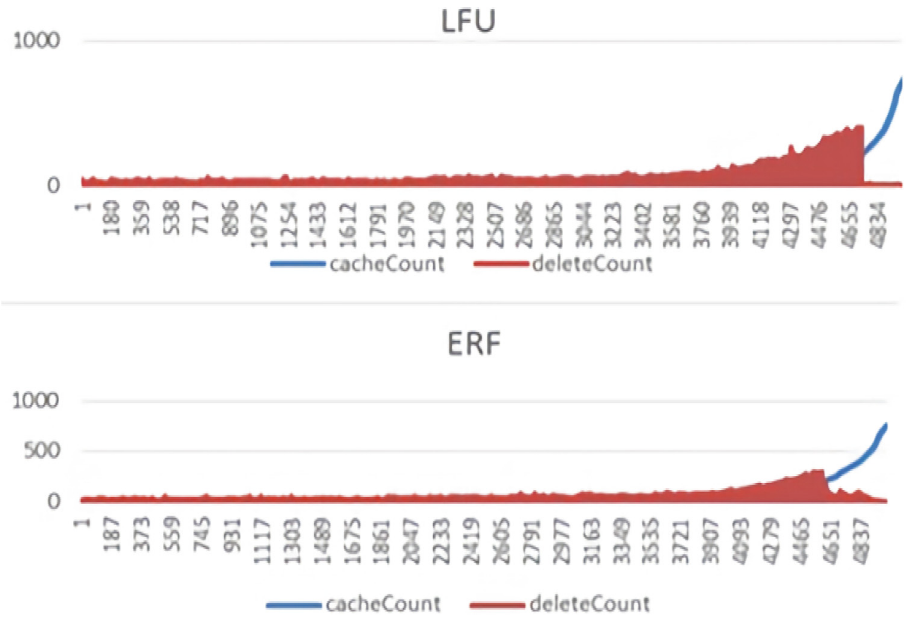


Fig. 5. Relationship between cache count and delete count in each algorithms

Meanwhile, ERF shows intermediate form compared to LFU and LRU as shown in Fig. 5. The form of graph in ERF looks optimal, but it also needs certain level of protection to last accessed items due to their low cache count.

As shown in Table 2, bringing out two eviction processes outperform LRU and LFU. In this context, *Overtime* is a simple eviction process which evicts objects which were not requested for a certain time $\bar{t}$. Also, eviction operations are applied evenly to all objects regardless of cache counts as shown in Fig. 6. As request counts increases, the gap between delete count of conventional algorithms and our approach increases. As a result, high hit ratio derives the decrease of the copy and the delete operations of thumbnail cache. It also drives improvement of the response time to clients' requests.

Table 2. The results of each algorithms

	Request count	Cache count	Delete count	Hit ratio (%)
LFU	360000	191033	168084	53.06
LRU	360000	279886	79313	77.74
ERF	360000	229717	129356	63.81
Overtime - > LFU	360000	288622	70421	80.17
Overtime - > ERF	360000	289615	69398	80.45

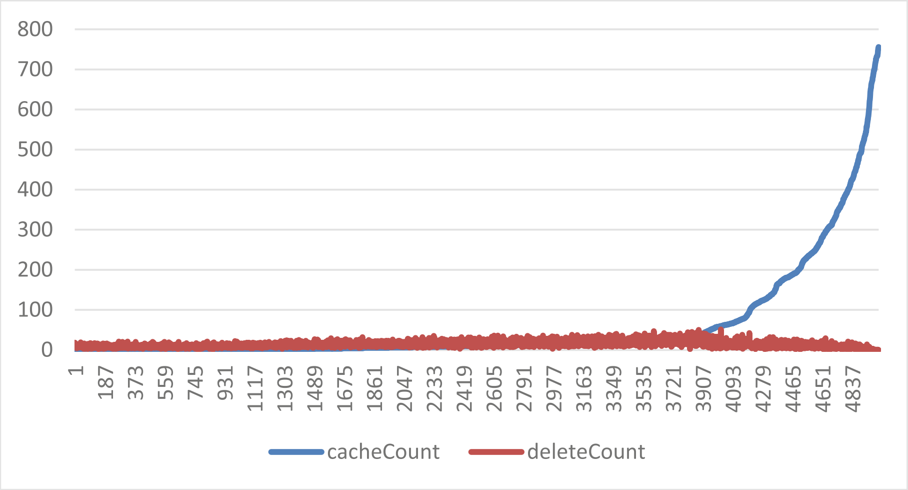


Fig. 6. Relationship between cache count and delete count in ERF with *overtime*

6 Conclusion and Future Work

This article introduces brief overview about conventional caching replacement strategies and introduces ERF. It shows that ERF outperforms conventional replacement algorithms. In dynamic nature of e-commerce business, the importance of making good balance between frequency and recency come to the fore. Thus, ERF seems a good solution for e-commerce business.

Future works will have comparisons in terms of memory consumption and time with conventional algorithms. Moreover, we can apply the *Markov chain theory* [10] in production of traces for verification of cache replacement algorithm's efficiency.

Acknowledgement. This research was supported by the MISP(Ministry of Science, ICT & Future Planning), Korea, under the National Program for Excellence in SW) supervised by the IITP(Institute for Information & communications Technology Promotion) (2015-0-00908)

References

1. Website Magazine (n.d.). https://www.websitemagazine.com/blog/5-reasons-visitors-leave-your-website. Accessed 20 Feb 2017
2. Mano MM (2008) Computer System Architecture. Prentice-Hall of India, New Delhi
3. Podlipnig S, Böszörmenyi L (2003) A survey of Web cache replacement strategies. ACM Comput Surv 35(4):374–398
4. Einziger G, Friedman R (2014) TinyLFU: a highly efficient cache admission policy. In: 2014 22nd euromicro international conference on parallel, distributed, and network-based processing

5. Abrams M (1995) Caching proxies: limitations and potentials. Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA
6. Abrams M, Standridge CR (1996) Removal policies in network caches for World-Wide Web documents. In: Conference proceedings on applications, technologies, architectures, and protocols for computer communications - SIGCOMM 1996
7. Kharbutli M, Solihin Y (2008) Counter-based cache replacement and bypassing algorithms. IEEE Trans Comput 57:433–447
8. Wu C-J, Martonosi M (2011) Adaptive timekeeping replacement: fine-grained capacity management for shared CMP caches. ACM Trans Archit Code Optim 8:3
9. Qureshi MK, Jaleel A, Patt YN, Steely SC Jr, Emer J (2007) Adaptive insertion policies for high performance caching. In: Proceedings of the 35th international sympsium on computer architecture
10. Markov chain (04 March 2017). https://en.wikipedia.org/wiki/Markov_chain. Accessed 05 Mar 2017

A Recursive Analysis Approach for Retrial Mobile Networks with Two Customers Classes and Non-preemptive Priority

Nawel Gharbi[1(✉)] and Leila Charabi[2]

[1] Department of Computer Science, University of Sciences and Technology, USTHB, Algiers, Algeria
ngharbi@usthb.dz

[2] National Computer Science Engineering School, ESI, Algiers, Algeria
l_charabi@esi.dz

Abstract. Retrial queueing models with two classes of customers arise in various practical mobile networks and telecommunication systems. The consideration of retrials (or repeated attempts) introduces analytical difficulties and most of works consider either models with preemptive priority or non-preemptive priority in the single server case. This paper aims to propose a recursive algorithmic approach for the performance analysis of a multiserver retrial queueing model with non-preemptive priority and two customers classes: ordinary customers whose access to the service depends on the number of available servers and who join the orbit when blocked; and impatient priority customers who have access to all servers and are lost when no server is available. In addition, we develop the formulae of the main stationary performance measures. Through numerical examples, we study the effect of the system parameters on the blocking probability for ordinary customers and the loss probability for priority customers.

Keywords: Retrial multiserver queueing model · Mobile networks · Two customers classes · Impatient customers · Non-preemptive priority · Recursive algorithm · Performance measures

1 Introduction

Retrial queueing models are characterized by the feature of retrial phenomenon that an arriving customer who finds all servers (or resources) occupied, joins the virtual group of blocked customers, called *orbit* and retry again for service after a random amount of time. Models with retrials have been widely used to analyze several practical problems in telecommunication systems as the call centers, the cellular mobile networks [1–3] and the wireless sensor networks [4]. Significant surveys on this topic [5,6] reveal the non-negligible impact of retrials, which arise due to the lack of available resources and which can negatively affect the system performances, because they generate more load. However, the consideration of retrial phenomenon introduces great analytical complications to obtain

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_34

most important performance measures. In particular, for multiserver models, no explicit closed-form solution exist for the performance measures [6]. These analytical difficulties are due to the simultaneous presence of the repeated requests stream from the orbit and the normal stream of primary requests arrivals. Therefore, lots of attention have been paid to approximation methods, computational algorithms and tail asymptotics to estimate the performance measures [6].

On the other hand, the heterogeneity of customers from the point of view of customers characteristics such as the arrivals, the service and/or the retrial process distributions, is an another important problem in retrial queueing area, because models with different types of customers arise in various practical systems. For example, in cellular mobile networks, the base station channels are used by a class of fresh calls initiated in the same cell and a second class of handoff calls incoming from adjacent cells. Similarly, in modern call centers, multiple types of calls arrive at service station over different communication channels such as telephone, internet, e-mail, mobile device, etc.

However, retrial queues with multiple classes of customers (called also multiclass retrial queues) have been known to be far more difficult for mathematical analysis than models with a single class of customers (or homogeneous customers). So, explicit results for this subject are limited to some particular cases [7] and recently, sufficient stability conditions were defined for a multiserver multiclass retrial queue [8].

For the single server retrial queues with two classes of customers, a number of analytic results have been obtained [5,6,9]. As regards to multiserver case with two customers classes, as far as we know, there are no explicit formulae and only a few algorithmic methods are proposed using matrix geometric methods [10], matrix analytic methods [11] or computational approaches as the one we have proposed using the Colored Generalized Stochastic Petri nets formalism [12]. Recently, Kim et al. [13] studied the stability of a two-class two-server retrial queue.

The objective of this paper is to propose a new recursive algorithmic approach for the performance analysis of a multiserver retrial queue with two classes of customers: ordinary customers whose access to the service depends on the number of available servers and who join the retrial group with a certain degree of impatience when blocked; and priority customers who have access to all servers and leave definitively the system when no server is available. Hence, and in order to minimize the loss probability of priority customers, they should be given a higher priority over ordinary customers in access to the system resources. To this end, we give them the possibility to use all servers, unlike ordinary customers whose access to the service depends on a threshold on the number of available servers. Further, we assume that all servers follow the non-preemptive priority rule, which means that if one or more priority customers arrive during the service time of an ordinary customer, the current service of this non-priority customer continues and is not stopped.

Some papers considered retrial models with two customers classes and preemptive priority [9,11,14] or a non-preemptive priority in the single server case

[15,16]. However, there is no work that deals with multiserver retrial queueing systems with two customers classes and non-preemptive priority. That motivates us to investigate such queueing model in this work.

The layout of the paper is given as follows: After the introduction, a detailed mathematical description of the model under study is given in Sect. 2. Then, we present our analysis approach and the details of the recursive algorithm we propose to calculate the stationary states probabilities in Sect. 3. Next, we give the formulae of the main performance measures. In Sect. 5, we discuss through numerical examples, the effect of the dedicated servers number and retrial rate on the system performances, namely the blocking probability for ordinary customers and loss probability for priority customers. Finally, we give a conclusion.

2 Mathematical Description of the Model

We consider a retrial multi-server queueing system with two classes of customers; ordinary and priority ones. The service area consists of C, $(C \geq 1)$ homogeneous servers with the same exponential service rate μ. The ordinary (priority) customers arrive in the system following a Poisson process with a mean arrival rate λ_1 (λ_2 respectively). The global arrival rate is then given by $\lambda = \lambda_1 + \lambda_2$. In order to ensure that priority customers are served prior to ordinary (non-priority) ones, our strategy consists of reserving a certain number of servers d $(1 \leq d \leq C)$, called *dedicated servers* only for priority class of customers. Thus, on the arrival of a priority customer, if at least one server of the C servers of the service station is idle, it will be served immediately, otherwise, it will be lost definitively, whereas an arriving ordinary customer must find at least $(d+1)$ available servers to get service, otherwise, it joins the orbit and retry for the service later. A blocked customer in the orbit decides to retry with probability θ or give up and returns to the free state with probability $(1-\theta)$. Note that θ is used to represent the degree of *impatience of customers*. The retrial time is exponentially distributed with rate α. All involved random variables are independent and identically distributed.

3 Recursive Analysis Algorithm

From the stochastic behavior of both two classes of customers and the servers allocation policy, the retrial system described above can be modeled by means of a two-dimensional Continuous-Time Markov Chain (CTMC) where each state is described by means of two random variables $(X(t), Y(t); t \geq 0)$. Let $X(t)$ be the number of customers being in service (which equals the number of busy servers), and $Y(t)$ the number of customers waiting in the orbit at time t. Hence, the steady state probabilities are defined by the probabilities $\pi_{i,j} = Pr\left\{X = i, Y = j\right\}, i = 0, 1, ..., C \; j = 0, 1, ..., ...$ of having i customers in service and j customers in the orbit.

The state space $S = \{0, ..., C\} \times Z_+$ of this CTMC is infinite because the population size and the orbit capacity are supposed to be infinite. In order to obtain a finite CMTC model, we propose the truncation of the state space to $S' = \{0, ..., C\} \times \{0, ..., q\}$ with q large enough. In other terms, the probability of being in states with a number of customers in orbit greater than q is neglected.

The *balance equation* describing the probability flux in and out of state (i, j) is defined by:

$$E(i,j) : \sum_{(k,l)\in S'\setminus(i,j)} \pi_{i,j}.R_{(i,j),(k,l)} = \sum_{(k,l)\in S'\setminus(i,j)} \pi_{k,l}.R_{(k,l),(i,j)}$$

where $R_{(i,j),(k,l)}$ is the transition rate from state (i, j) to state (k, l).

We put $K_0 = \pi_{0,q}$, $K_1 = \pi_{0,(q-1)}$, ..., $K_d = \pi_{(0,q-d)}$. We first should express all probabilities as a function of $K_i,\ i = 0, ..., d$.

$$\pi_{i,j} = K_0.u_0(i,j) + K_1.u_1(i,j) + ... + K_d.u_d(i,j)$$

Then, we express coefficients $K_i,\ i = 0, ..., d$ as a function of K_0, and finally, we use the normalization equation, where the unique unknown is K_0,

$$\sum_{i=0}^{C}\sum_{j=0}^{q} \pi_{i,j} = 1 \tag{1}$$

to find its value.

We now proceed to explain the details of the algorithm:

Step1. Expressing all probabilities in K_i

1. Columns q down to $q - d$

Starting with column q, it's obvious that $\pi_{(0,q)} = K_0$, such as $u_0(0, q) = 1$, $u_1(0, q) = 0$, . . ., $u_d(0, q) = 0$.

We calculate recursively, for $i = 1, ..., C - (d + 1)$, $\pi_{i,q}$ using the balance equation $E(i - 1, q)$. We get:

$$\begin{aligned} \pi_{1,q} &= \frac{q.\alpha + \lambda}{\mu}.\pi_{0,q} \\ \pi_{i+1,q} &= \frac{q.\alpha + \lambda + i.\mu}{(i+1)\mu}\pi_{i,q} - \frac{\lambda}{(i+1)\mu}.\pi_{i-1,q} \end{aligned}$$

In the same way, we calculate for columns $(q - j)$, $j = 1, ..., d$, the value of $\pi_{1,q-j}$ using $E(0, q - j)$ first, then $\pi_{i+1,q-j}$ using $E(i, q - j)$, $i = 0, ..., C - d - 1$. We get:

$$\begin{aligned} \pi_{1,q-j} &= \frac{(q-j).\alpha + \lambda}{\mu}.K_j \\ \pi_{i+1,q-j} &= \frac{(q-j).\alpha + \lambda + i.\mu}{(i+1)\mu}\pi_{i,q-j} - \frac{\lambda}{(i+1)\mu}\pi_{i-1,q-j} - \frac{(q-j+1)\alpha}{(i+1)\mu}\pi_{i-1,q-j+1} \end{aligned}$$

Then, inside the line $(C-d+1)$, columns from $j=(q-d+1)$ to $j=(q-1)$ can be calculated. Actually, for each probability $\pi_{C-d+1,j}$, we use the balance equation $E(C-d,j)$:

$$\pi_{C-d+1,j}.[(C-d+1)\mu] = [(C-d)\mu+\theta\lambda_1+\lambda_2+j(1-\theta)\alpha].\pi_{C-d,j}-\lambda.\pi_{C-d-1,j}$$
$$-[(1-\theta)(j+1)\alpha].\pi_{C-d,j+1}-\theta\lambda_1.\pi_{C-d,j-1}-[(j+1)\alpha].\pi_{C-d-1,j+1}$$

And for $\pi_{C-d+1,q}$, we use $E(C-d,q)$, we have:

$$\pi_{C-d+1,q}.[(C-d+1)\mu] = [(C-d)\mu+\lambda_2+(1-\theta)q\alpha].\pi_{C-d,q}$$
$$-\lambda.\pi_{C-d-1,q}-\theta\lambda_1.\pi_{C-d,q-1}$$

For the rest of lines, i.e. from $i=C-d+2$ to $i=C$, only columns $j=i+q-C,...,q$ can be deduced for the moment, they are calculated from balance equations $E(i-1,j)$:

We have for $j=i+q-C$ to $j=q-1$:

$$\pi_{i,j}.(i.\mu) = [(i-1).\mu+(1-\theta).j.\alpha+\lambda_2+\theta.\lambda_1].\pi_{(i-1),j}$$
$$-\lambda_2.\pi_{(i-2),j}-(1-\theta).(j+1).\alpha.\pi_{(i-1),(j+1)}-\theta.\lambda_1.\pi_{(i-1),(j-1)}$$

And when $j=q$:

$$\pi_{i,q}.(i.\mu) = [(i-1).\mu+(1-\theta).q.\alpha+\lambda_2].\pi_{(i-1),q}-\lambda_2.\pi_{(i-2),q}-\theta.\lambda_1.\pi_{(i-1),(q-1)}$$

From $E(C,q)$, we obtain the value of $\pi_{C,q-1}$ as follows:

$$\pi_{C,q-1}.\theta.\lambda_1 = [C.\mu+(1-\theta).q.\alpha].\pi_{C,q}-\lambda_2.\pi_{(C-1),q}$$

Now, in order to have the rest of columns, we proceed like this. For each $j=q-2$ to $q-d$, we obtain first $\pi_{C,j}$ using $E(C,j+1)$:

$$\pi_{C,j}.\theta.\lambda_1 = [(1-\theta).(j+1).\alpha+C.\mu+\theta.\lambda_1].\pi_{C,(j+1)}$$
$$-(1-\theta).(j+2).\alpha.\pi_{C,(j+2)}-\lambda_2.\pi_{(C-1),(j+1)}$$

Then, we obtain the other lines $i=C-1$ to $i=C+1+j-q$, thanks to $E(i,j+1)$:

$$\pi_{i,j}.\theta.\lambda_1 = [(1-\theta).(j+1).\alpha+i.\mu+\theta.\lambda_1+\lambda_2].\pi_{i,(j+1)}$$
$$-(1-\theta).(j+2).\alpha.\pi_{i,(j+2)}-\lambda_2.\pi_{(i-1),(j+1)}-(i+1).\mu.\pi_{(i+1),(j+1)}$$

Up to now, we have expressed all probabilities from column $j=q$ to $j=q-d$, in $K_0,...,K_d$.

1. Column $(q-d-1)$ down to 1

In a similar way as explained above, by invoking $E(i-1,j)$, for each $\pi_{i,j}$, we find:

$$\begin{aligned}\pi_{1,j} &= \frac{j.\alpha+\lambda}{\mu}.\pi_{0,j}\\ \pi_{i+1,j} &= \frac{j.\alpha+\lambda+i.\mu}{(i+1)\mu}\pi_{i,j} - \frac{\lambda}{(i+1)\mu}\pi_{i-1,j} - \frac{(j+1)\alpha}{(i+1)\mu}\pi_{i-1,j+1}\end{aligned}$$

In particular, when $i = C-d-1$, we can find numbers $v_0(C-d,j)$, $v_1(C-d,j),...,v_{d+1}(C-d,j)$, such that:

$$\pi_{(C-d),j} = v_0(C-d,j).K_0 + ... + v_d(C-d,j).K_d + v_{(d+1)}(C-d,j).\pi_{0,j}$$

On the other hand, by invoking $E(C-d,j+1)$, we get:

$$\begin{aligned}&\pi_{(C-d),j}.\theta.\lambda_1 = [(C-d).\mu+\theta.\lambda_1+\lambda_2+(1-\theta).(j+1).\alpha].\pi_{(C-d),(j+1)} - \lambda.\pi_{(C-d-1),(j+1)}\\ &-(j+2).\alpha.\pi_{(C-d-1,j+2)} - (1-\theta).(j+2).\alpha.\pi_{(C-d),(j+2)} - (C-d+1).\mu.\pi_{(C-d+1),(j+1)}\end{aligned}$$

which implies that we can find explicitly numbers $u_0(C-d,j), u_1(C-d,j),...,$ $u_d(C-d,j)$, such that:

$$\pi_{(C-d),j} = u_0(C-d,j).K_0 + u_1(C-d,j).K_1 + ... + u_d(C-d,j).K_d$$

From equation (2) and (2), we can deduce $\pi_{0,j}$ value,

$$\pi_{0,j} = \frac{\sum_{k=0}^{d}[u_k(C-d,j) - v_k(C-d,j)].K_k}{v_{(d+1)}(C-d,j)}$$

Thus, we calculate again $\pi_{i,j}$, $i = 1,...,C-d$, in $K_0, K_1,...,K_d$ only, we get for $k = 0,...,d$:

$$u_k(i,j) = v_k(i,j) + \frac{u_k(C-d,j) - v_k(C-d,j)}{v_{(d+1)}(C-d,j)}.v_{(d+1)}(i,j)$$

After that, equations for $\pi_{i,j}$, such that $i = C-d+1,...,C$ can be easily derived from $E(i,j+1)$.

Step2. Expressing coefficients $K_i, i = 1,...,d$ in K_0

Let's consider the balance equation $E(C,0)$:

$$\pi_{C,0}.(C.\mu+\theta.\lambda_1) = \pi_{(C-1),0}.\lambda_2 + \pi_{C,1}.(1-\theta).\alpha$$

Keeping in mind that both $\pi_{C,0}$, $\pi_{(C-1),0}$ and $\pi_{C,1}$ can be written as a linear combination of $K_0,...,K_d$, equation (3) is equivalent to:

$$\sum_{k=0}^{d} u_k(C,0).K_k.(C.\mu+\theta.\lambda_1) = \sum_{k=0}^{d} u_k(C-1,0).K_k.\lambda_2 + \sum_{k=0}^{d} u_k(C,1).K_k.(1-\theta).\alpha$$

It's a question of a simple algebra to extract K_d in $K_{(d-1)},...,K_0$. In the same way, we consider $E(i,0)$, $i = C-1,...,C-d+1$, to have K_x in $K_{x-1},...,K_0$, $x = d-1,...,1$

Step3. Finding K_0

Finally, we solve the normalization equation (1), in order to extract the value of K_0 which is the unique unknown.

4 Performance Measures

Once the stationary probabilities are determined thanks to the above algorithm, several performance measures can be calculated applying the following formulas. The most significant performance indices are as follows:

- Mean number of busy servers:: $N_{Busy} = \sum_{i=0}^{C} \sum_{j=0}^{q} i.\pi_{i,j}$
- Mean rate of ordinary customers served at the first attempt:

$$\bar{\lambda}_{FS} = \lambda_1. \sum_{i=0}^{C-(d+1)} \sum_{j=0}^{q} .\pi_{i,j}$$

- Mean rate of blocked ordinary customers:

$$\bar{\lambda}_{FU} = \lambda_1.\theta. \sum_{i=C-d}^{C} \sum_{j=0}^{q} .\pi_{i,j}$$

- Mean rate of blocked ordinary customers leaving the system without being served:

$$\bar{\lambda}_{FB} = \lambda_1.(1-\theta). \sum_{i=C-d}^{C} \sum_{j=0}^{q} \pi_{i,j}$$

- Effective mean ordinary customers arrival rate:

$$\bar{\lambda}_F = \bar{\lambda}_{FS} + \bar{\lambda}_{FU} + \bar{\lambda}_{FB}$$

- Mean rate of retrials served at the first attempt:

$$\bar{\alpha}_{RS} = \alpha. \sum_{i=0}^{C-(d+1)} \sum_{j=0}^{q} j.\pi_{i,j}$$

- Mean rate of blocked retrials: $\bar{\alpha}_{RU} = \alpha.\theta. \sum_{i=C-d}^{C} \sum_{j=0}^{q} j.\pi_{i,j}$
- Mean rate of blocked retrials leaving the system without being served:

$$\bar{\alpha}_{RB} = \alpha.(1-\theta). \sum_{i=C-d}^{C} \sum_{j=0}^{q} j.\pi_{i,j}$$

- Effective mean retrial rate: $\bar{\alpha} = \bar{\alpha}_{RS} + \bar{\alpha}_{RU} + \bar{\alpha}_{RB}$
- Mean rate of priority customers being served: $\bar{\lambda}_{HS} = \lambda_2. \sum_{i=0}^{C-1} \sum_{j=0}^{q} \pi_{i,j}$
- Mean rate of lost priority customers: $\bar{\lambda}_{HB} = \lambda_2.\ \sum_{j=0}^{q} \pi_{C,j}$
- Effective mean priority customers arrival rate: $\bar{\lambda}_H = \bar{\lambda}_{HS} + \bar{\lambda}_{HB}$
- Blocking probability of ordinary customers: $P_{BF} = \frac{\bar{\lambda}_{FU}+\bar{\lambda}_{FB}}{\bar{\lambda}_F}$
- Blocking probability of retrial customers: $P_{BR} = \frac{\bar{\alpha}_{RU}+\bar{\alpha}_{RB}}{\bar{\alpha}}$
- Loss probability (of priority customers): $P_{BH} = \frac{\bar{\lambda}_{HB}}{\bar{\lambda}_H}$

5 Numerical Results

In this section, we examine the impact that have some system parameters like the number of dedicated servers, degree of persistence and the arrival and service rates on the system performance, namely the blocking and loss probability. In that follows, unless otherwise stated, we assume that $C = 15$, $1/\mu = 120s$, $\alpha = \mu = 20$, $\lambda_1 = \lambda_2 = 24$ and $\theta = 0.6$. The offered load is defined by $\rho = \lambda/(C.\mu)$.

The effect of the traffic load ρ and the number of dedicated servers on the blocking probability and the loss probability is shown in Figs. 1 and 2

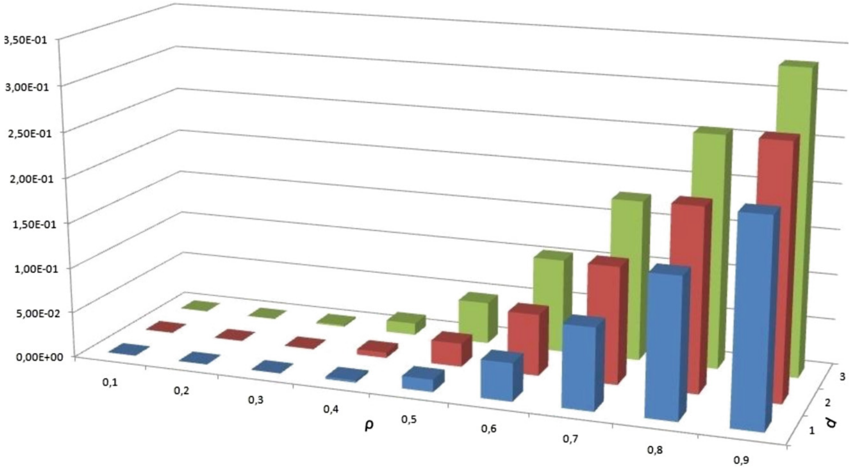


Fig. 1. Influence of the offered traffic and the number of dedicated servers on the blocking probability.

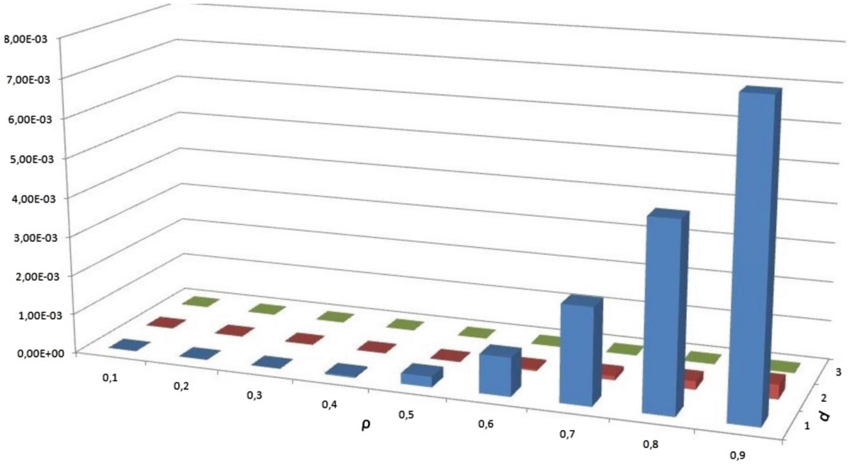


Fig. 2. Influence of the offered traffic and the number of dedicated servers on the loss probability.

respectively. We can note that the increase in parameters ρ affects negatively both of the two probabilities. On the other hand, as expected, increasing the number of dedicated servers can significantly improve the loss probability of priority customers. It is just perfect ($\simeq 0$) when $d = 3$. We observe the opposite effect on the blocking probability, when more servers are reserved to priority customers, more ordinary customers are blocked at their arrival.

6 Conclusion

A recursive algorithmic approach for the performance analysis of a mobile network with repeated attempts, two classes of customers: ordinary (impatient) and priority customers and non-preemptive priority was investigated in this paper. In order to minimize the loss probability of priority customers, they should be given a higher priority over ordinary customers in access to the network servers. Our proposition was to reserve some servers to be used only by priority customers. The analysis of the model was performed using a bi-dimensional Time Continuous Markov chain, and an efficient recursive algorithm was proposed and implemented in order to calculate the steady state probability distribution. Moreover, the formulae of several performance measures were developed. We showed via numerical examples that dedicated servers technique improves the system performance, mainly the loss probability, but at the expense of ordinary customers.

References

1. Kim C, Klimenok VI, Dudin AN (2014) Analysis and optimization of guard channel policy in cellular mobile networks with account of retrials. Comput Oper Res 43:181–190
2. Charabi L, Gharbi N, Ben-Othman J, Mokdad L (2016) Call admission control in small cell networks with retrials and guard channels. In: Proceedings of The IEEE global communications conference 2016 (GLOBECOM 2016), USA
3. Gharbi N (2016) Using GSPNs for performance analysis of a new admission control strategy with retrials and guard channels. In: The 3rd international conference on mobile and wireless technology (ICMWT 2016), Korea
4. Wuechner P, Sztrik J, De Meer H (2011) Modeling wireless sensor networks using finite-source retrial queues with unreliable orbit. In: Proceedings of the Workshop on Performance Evaluation of Computer and Communication Systems (PERFORM 2010). LNCS, vol 6821. Springer-Verlag (2011)
5. Artalejo JR, Gómez-Corral A (2008) Retrial queueing systems: a computational approach. Springer, Berlin
6. Kim J, Kim B (2015) A survey of retrial queueing systems. Ann Oper Res 247:1–34
7. Shin YW, Moon DH (2014) M/M/c retrial queue with multiclass of customers. Method Comput Appl Probab 16:931–949
8. Avrachenkov K, Morozov E, Steyaert B (2016) Sufficient stability conditions for multi-class constant retrial rate systems. Queu Syst 82(1):149–171
9. Gao S (2015) A preemptive priority retrial queue with two classes of customers and general retrial times. Oper Res 15(2):233–251

10. Choi BD, Chang Y (1999) MAP1, MAP2/M/c retrial queue with the retrial group of finite capacity and geometric loss. Math Comput Model 30:99–113
11. Kumar MS, Chakravarthy SR, Arumuganathan R (2013) Preemptive resume priority retrial queue with two classes of MAP arrivals. Appl Math Sci 7(52):2569–2589
12. Gharbi N, Dutheillet C, Ioualalen M (2009) Colored stochastic petri nets for modelling and analysis of multiclass retrial systems. Math Comput Model 49:1436–1448
13. Kim B, Kim J (2015) Stability of a two-class two-server retrial queueing system. Perform Eval 88–89:1–17
14. Boutarfa L, Djellab N (2015) On the performance of the M1, M2/G1, G2/1 retrial queue with pre-emptive resume policy. Yugoslav J Oper Res 25(1):153–164
15. Ayyapan G, Muthu Ganapathi Subramanian A, Sekar G (2010) M/M/1 retrial queueing system with loss and feedback under non-pre-emptive priority service by matrix geometric method. Appl Math Sci 4(48):2379–2389
16. Madan KC (2011) A non-preemptive priority queueing system with a single server serving two queues M/G/1 and M/D/1 with optional server vacations based on exhaustive service of the priority units. Appl Math 2:791–799

A Survey of Vehicular Ad-Hoc Network Security

MinSu Kim(✉)

Convergence Security Department, Kyonggi University, Iui-dong,
Yeongtong-gu, Suwon-si, Gyeonggi, South Korea
fortcom@hanmail.net

Abstract. A variety of functional control of vehicles has developed along with information and communication technology. In particular, with the application of wireless network for real-time information offering, it has been possible to establish Vehicular Ad-hoc Network (VANET), an intelligent vehicle service for convenience and safety, which makes possible collision accident warning and avoidance, warning of dangerous factors on road, traffic information offering, and other kinds of service offering. However, the VANET service environment has physical and technical vulnerabilities caused by the vehicular internal/external communication based on wireless network. Therefore, Vehicular Security has emerged as an essential factor to prevent malicious threats and privacy violation from vehicles, drivers, and traffic network. This study tries to find the main security components of the VANET environment, analyze the latest research trend to overcome security vulnerabilities in each area, and propose the future research direction of Vehicular Security.

Keywords: VANET (Vehicular Ad-hoc Network) · Routing algorithm · Vehicle security · ITS (Intelligent transportation System) · V2V (Vehicle to Vehicle) · V2I (Vehicle to Infrastructure) · V2X (Vehicle to Everything)

1 Introduction

Vehicular ad hoc network (VANET) is an application of mobile ad hoc network (MANET) which is a key component of intelligent transportation systems (ITS). VANET provides communication between vehicle-to-vehicle (V2 V) using vehicle OBU (On Board Unit) and Vehicle-to-Infrastructure (V2I) [1].

VANET uses multi-hop typed broadcast with the WAVE standard developed in the combination of IEEE 802.11p and IEEE 1609.x in order to provide emergent warning messages for dangerous situations and traffic information service [2]. However, VANET with high mobility has such problems as frequent change in network topology and communication disconnection [3].

As shown in Fig. 1 'VANET SYSTEM', a vehicle can make V2I communication with the use of RSU serving as a base station, or can directly communicate with other vehicles without any help of RSU.

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_35

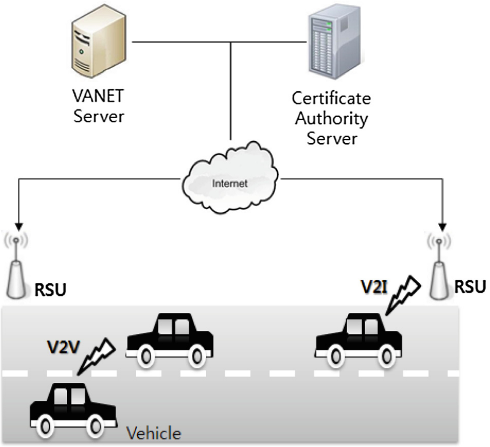


Fig. 1. VANET system

VANET Server manages all information on RUSs through wired or wireless internet. CA (Certificate Authority) Server in a reliable organization serves as the roles of registering a vehicle and permitting communication between vehicles through authentication.

VANET communication means that the information collected from vehicles and infrastructure is delivered accurately to relevant vehicles and infrastructure timely over wireless network. VANET communication technology is specialized to the vehicles running at high speed as shown in Fig. 2. It uses WAVE(Wireless Access in Vehicular Environment) developed in combination of IEEE 802.11p [4] and IEEE 1609.x [5–8]. WAVE is standardized with the IEEE 1609 standard documents, among which IEEE 1609.1 describes resource manager; IEEE 1609.2 describes application and management message security service; IEEE 1609.3 describes networking service; IEEE 1609.4 describes multi-channel operation.

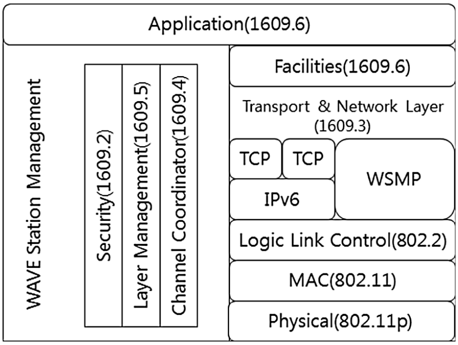


Fig. 2. WAVE architecture and protocol stack [5]

2 Security Requirements for VANET

Security requirements for VANET are authentication, confidentiality, integrity, availability, access control, and non-repudiation as the criteria to define a level of network security [9–11]. More details are described as follows:

2.1 Authentication

Through authentication, it is possible to guarantee message reliability and recognize a sender rightly. In the vehicular communication standard WAVE reference model, IEEE 1609.2 is the standard for authentication and security layer.

2.2 Integrity

In order to guarantee the accuracy of a transmitted message, it is required to limit data changes and prevent the obstacles to stability. As such, integrity means keeping data states correct.

2.3 Availability

An authorized vehicle needs to access the event to require a warning to other vehicles in a wireless channel.

2.4 Confidentiality

This function is aimed at preventing the access to a message sent by an illegal or non-permitted node.

2.5 Non-repudiation

This is the security mechanism of proving non-repudiation of a transaction that occurs between a sender and a receiver.

3 Vulnerabilities in VANET

Of a lot of information transmitted through vehicular network, some information is able to use vulnerabilities of the vehicular environment so as to cause a fatal accident. Security vulnerabilities of vehicle include vulnerability of internal communication of vehicle, vulnerability of communication between vehicles, and vulnerability of traffic infrastructure, which are described as follows [12, 13].

3.1 Bogus Information

By spreading incorrect information to network, this attack influences a different driver's behavior. If an attack vehicle transmits bogus information on traffic situation to a neighboring victim vehicle, the victim travels to the intended path of the attacker on the basis of the bogus information.

3.2 Fake Position Information

This attack is used to induce the changes in a vehicle's known position, speed, and direction in order to avoid responsibilities from an accident.

3.3 ID Exposure

For position-tracking, a different vehicle's ID is exposed. In the logic of Big Brother, it is possible to monitor the travelling path of a vehicle targeted by an observer and use the information for other purposes. A passive attacker is able to reveal an identification of a target through VANET system, rather than a physical tool. Aside from an ID, a time, a position, travel information, and other kinds of personal information can be exposed.

3.4 Denial of Service

This attack is aimed at paralyzing communication or causing confusion in VANET, or triggering an accident. Jamming is a sort of DoS attack, which generates signals of a different vehicle's communication in a particular network area of VANET in order to paralyze communication.

3.5 Impersonation

An attacker uses an ID of an authorized vehicle in order to confuse neighboring vehicles.

3.6 Forgery

An attacker generates fake information in order to confuse other vehicles in a certain network area. A case in point is spreading false warnings like icy road warning in order to slow overall traffic.

3.7 In-transit Traffic Tampering

This attack is aimed at impeding normal communication through the deletion or change of a message to transmit in vehicle travelling.

3.8 Vehicle Information Tampering

This attack is aimed at tampering with internal information of vehicle (e.g., speed, position, sensing information). In the attack, a speed, a position, and other kinds of information are provided incorrectly. Like forgery, this attack also forges information. What is difference is that this attack uses sensors or internal devices, rather than transmitted information, in order to change internal information and trigger malfunction of a vehicle.

3.9 Information Block

This attack uses the features of the protocol used in VANET. If the protocol transmits the information of a vehicle in transmission to its neighboring vehicle favorably positioned for transmission, the information stops being transmitted. At this time, an attacker cheats the vehicle in transmission as if it is the vehicle favorably positioned for transmission, and consequently the victim vehicle stops transmitting information. As a result, by stop transmitting information which should be sent to other vehicles, it is possible to cause confusion.

3.10 Tunnel

By sending fake information to a vehicle that enters and exits in and from places with temporary no service of GPS service, such as a tunnel, this attack makes it possible for the vehicle to update incorrect information.

3.11 Wormhole

This is a sort of disturbance attack. By sending meaningless information, though authorized, it is possible to disturb network.

3.12 Jungle Communication

This is the expansion of Bogus Information attack. By sending information to each vehicle and continuing to change it, it is possible to change the initial information to different information.

4 Related Work

Ram [14] proposed security mechanisms for symmetric and asymmetric algorithms to secure the safety of routing protocol and protecting personal information from malicious insertion and modification of data in the open access environment of VANET. Nirbhay [19] analyzed the security requirements and proposed solutions to security problems in VANET environment. To ensure security in VANET, he proposed to consider certain attributes which includes Authentication, Availability, Non-Repudiation, Access control, Privacy, Con-fidentiality, Data Verification, Integrity, Real time guarantees. Chen [16] observed that there is a security problem in the routing protocol information due to the characteristics of large-scale networks, fast mobile nodes, and frequently changing topology structures in VANET. He proposed a security mechanism for data encryption, security authentication, and intrusion detection to protect the integrity and consistency of network information.

Tomar [17] considered the problem of optimizing traffic flow in a vehicular network where some vehicle interferes with each others. Then, he allocated the time slots by the RSU using the SINR model to maximize the time slot utilization in the vehicular network. In this paper, he presented the model for the interference range messages to prevent the potentially interfering nodes from initiating new transmissions.

Tomar proposed Spatial Division Multiple Access (SDMA) to optimize channel allocation and throughput for secure transmission of messages. Vijayakarthika [18] presented CAN (Controller Area Network) DELIVER, which is part of a complete system for providing car drivers and passengers pervasive access to needed data while on the road. The proposed method reduced a delay time of data communication and increase communication efficiency by using the security mechanism of designating RSU as a proxy server and providing reliable data communication between vehicles. Vijayalakshmi [19] pointed out that in order to provide security services to all users in the VANET environment, the problem of security and scalability should be solved. Nam [20] proposed a VANET performance analysis method that uses the AODV (Ad-hoc On Demand Distance Vector) and DSDV (Destination Sequenced Distance Vector) throughput, packet loss rate, and average delay time as parameters. Pham [21] designed the secured linkability protocol using pseudonym-based encryption and Bloom filter Private Set intersection technique and a context-aware trust management scheme working compatibly with the linkability protocol.

Donato [22] proposed Desync mechanism to improve transmission performance through the recalculation of transmission delay. The proposed mechanism uses ABSM and AID protocols to reduce a collision and maximize a transmission speed. Kaur [23] proposed Decision Packet. All nodes that create a path from a departure node to a destination node make a check with the hash value of Decision Packet so that an attacker is less likely to change a hop count. One attacker is able to use multiple IDs to attack VANET. To solve the Sybil attack detection problem, Rahvari [24] proposed the mechanism to detect an attack and secure authentication, non-repudiation, privacy protection and data integrity through encryption.

In the communication of VANET, DoS attacks such as Sybil Attack, and selfish driver attack can occur. Gandhi [25] proposed RRDA (Request Response Detection

Algorithm) to detect DoS attacks. The proposed model uses a hash table to reduce DoS attacks of a forgery vehicle, transmits packets to all vehicles in between a departure and a destination, and updates a hop count. The model was found to reduce packet delay and request retransmission in the way of evaluating packets with a hop count and updating through the limitation of a counter capacity. RoselinMary [26] proposed APDA(Attacked Packet Detection Algorithm) to transmit a message safely against security threats like DoS(Denial of Service) in the VANET environment. The proposed APDA minimizes a delay overhead at the beginning in order to improve security of a VANET system. Nitish [27] proposed Multi Operating Channels Model to protect vehicle network against the attacks that can trigger malfunction of network and data confidentiality loss in the VANET environment. To examine the proposed model, the researcher analyzed Message Suppression Attack, Denial of Service Attack, SYN flooding Attack, Alteration Attack, Link spoofing Attack and Link withholding Attack, and Fabrication Attack. As a result, the model improved security.

Mohammed [28] proposed IDS (Intrusion Detection System) application technology in the VANET environment by classifying detection technology into Signature based system [29], Anomaly detection system [30], and Specifications based system [31]. Amarpreet [32] proposed EAPDA (Enhanced Attacked Packet Detection Algorithm) to prevent network performance deterioration of vehicles and RUS from Denial of Service attacks such as Sybil attack, Alteration attack, and Selfish Driver attack in the VANET environment. DoS attacks are detected with time slot. Compared to conventional algorithms, the EAPDA had higher response, less delay and more throughput.

Grzybek [33] provided stable community detection in the dynamic mobile network in consideration of vehicles' high mobility in the VANET. Therefore, the researcher expanded LPA (Label Propagation Algorithms) for community detection [34, 35] and SandSHARC(Stability And Network Dynamics over a Sharper Heuristic for Assignment of Robust Communities) [36] in the dynamic mobile network, and proposed the evaluation framework for examining the stability of a detected community.

Hussain [37] proposed a technique that was found to protect privacy through multiple anonymity, track a path by saving a Beacon message in Cloud, guarantee safe and conditional anonymity through the application of anonymity withdrawal, and have less operations than conventional techniques.

Hussain [38] proposed CaaS (Cooperation as a Service) architecture which has three sub objects- TIaaS (Traffic Information as a Service), WaaS (Warning as a Service), and IfaaS (Infotainment as a Service)-for for VANET Clouds. In the proposed architecture, communication between cloud infrastructure and vehicles is accomplished by GT (Gateway Terminals), and positioning based encryption is applied to keep privacy for a vehicle's position and identification. Park [39] designed the framework based on vehicular security requirements, including RSU (Road Side Unit) authentication, message integrity, confidentiality, privacy protection, non-repudiation, and availability in order for safe communication in the vehicular cloud environment.

As an encryption standard of a communication message between vehicles, BSM (Basic Safety Message) defined in 'SAE J2735' was used in order to design an authentication and message protocol. In this way, the designed model was found to secure stability and efficiency from the security threats such as forgery attack, data tampering and MITM, repudiation, and information leak in the vehicular cloud environment with the combination of the VANET and internet based cloud environment.

Park [40] pointed out that in 802.11 MAC protocol as a 802.11p based technology, if a node with low transmission rate holds a channel long, a node with high transmission rate is standardized downward to the low transmission rate; and a rise in nodes leads to a high probability of collision. Therefore, the researcher proposed the algorithm that makes a CWmin value low to reduce the backoff time of a node of holding a channel and sets CWmin value large to lower a collision probability in order for a node with good channel status to have a high probability of holding a channel.

Fengzhong [41] proposed the solution to the problem of security and privacy protection in the VANET open access environment. The proposed method was efficient for reducing much time and calculation cost in the process of examination and withdrawal. It improved the process of certificate revocation.

In order to improve a conventional warning message transmission type in the urban areas with poor radio environment, Lee [42] designed the method in which all nodes use neighboring nodes' information to calculate Forwarding Priority of themselves and neighboring nodes; and a node with the highest forwarding priority becomes a transmission node to send a warning message. Also, for the blind spots, the researcher proposed the algorithm to select the next transmission node and send a warning message to a blind spot.

Park [43] proposed ICRC algorithm, an efficient placement algorithm of RSU (Roadside Unit) which is an essential factor for transmitting, collecting, and analyzing traffic information in the VANET environment. The ICRC algorithm determines initial RUS candidate positions on the basis of IP (Intersection Priority) and ED(Even Distribution) approaches, and then removes a RUS with strong connection of RSU candidate positions in order to minimize the number of RSUs. According to the performance comparison, in the roads with good connection of intersections, both IP and ED based ICRA had excellent performance; in the complex roads with bad connection of intersections, ED based ICRA had better performance than IP.

In order to provide vehicle authentication and conditional privacy protection for safe communication of V2 V, Kim [44] proposed the batch verification technique to prevent unnecessary group subscription in previous studies using group signature technique. The proposed method met various security requirements on the basis of group signature. And, the Bloom Filter based batch verification method was found to improve node-by-node calculation efficiency more than conventional methods.

According to the research of Nazmul [45], when vehicles communicate through a wireless channel, it is essential to guarantee safety vehicle communication against various attacks, such as injection of wrong information and change and reproduction of a distributed message. Using PKI(Public Key Infrastructure) is able to meet such requirements as entity authentication, message integrity, non-repudiation, and personal information protection. If a vehicle cancels communication by going out of a region, efficient certificate revocation is required. Also, if a vehicle enters in a new region, it is required to update a certificate of the region efficiently. Therefore, the researcher proposed the security mechanism to reduce a message loss rate caused by message check delay in the way of shortening the time of message authentication and.

5 Research Trends of VANET Security

According to the analysis on previous studies on Vehicle Security, there were studies to overcome the attacks on vulnerabilities in the VANET environment.

The studies have actively been conducted since the mid 2000 s along with the development of wireless communication technology. Up to now, many studies are being performed on routing protocols for safe communication and privacy protection against various attacks.

The suggested security requirements in the VANET environment are authentication, availability, non-repudiation, access control, privacy protection, confidentiality, data verification, and data integrity in consideration of high mobility, dynamic topology, and other VANET features.

They focused on reliable communication for routing protocols and minimized communication delay time in order for safe communication.

Regarding algorithm design in the VANET environment, there are studies on the algorithms for attack detection and prevention, security in the cloud environment, and efficient communication improvement. In particular, since 2014, VuC (VANET using Clouds) has continued to be studied in order to increase the structural efficiency of data processing in the VANET environment, predict a situation far away to go out of spatial restrictions caused by a vehicle's position, and solve the problems of security and privacy.

The future VANET environment will be changed to V2X (Vehicle to Everything) as shown in Fig. 3 and its areas will be widened.

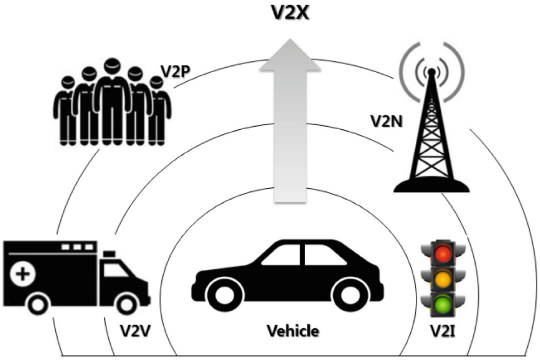


Fig. 3. Research trends of vehicle security

6 Conclusions

In this paper, we investigate the VANET system structure and previous studies on the security requirements in the VANET environment. The objects to communicate with vehicles are not limited to vehicles and infrastructure, but all objects supporting communication can be connected with others. It is requited to study security requirements widely, efficient encryption techniques in the VANET environment, and the application of routing algorithms in the VANET environment in order for efficient communication to respond to a wide range of communication objects. In addition, with the development of IoT (Internet of Things) and the V2X environment to share information through access to all objects, security requirements in the VANET environment are expected to expand to wider areas. Therefore, it is required to continue to research how to handle V2X security threats and how to reduce network loads in large network.

References

1. Caballero-Gil P (2011) Security issues in vehicular ad hoc network. In: Mobile ad-hoc networks: applications, pp 67–88
2. Li F, Wang Yu (2007) Routing in vehicular ad hoc networks: a survey. IEEE Veh Technol Mag 2(2):12–22
3. Martinez FJ, Toh C-K, Cano J-C, Calafate CT, Manzoni P (2010) Emergency services in future intelligent transportation systems based on vehicular communication networks. IEEE Intel Transp Syst Mag 2(2):6–20
4. IEEE Std 802.11p, Part 11 (2010) Wireless LAN Medium Access Control(MAC) and Physical Layer(PHY) Specifications, Amendment: Wireless Access in Vehicular Environments
5. IEEE Guide for Wireless Access in Vehicular Environments(WAVE)-Architecture, IEEE Std 1609.0, 2013
6. IEEE 1609.1,2 (2006) Trial Use Standards for Wireless Access in Vehicular Environments (WAVE)-Resource Manager, Security Services for Applications and Management Messages, Networking Service
7. IEEE 1609.3 (2007) Trial Use Standards for Wireless Access in Vehicular Environments (WAVE)-Networking Service
8. IEEE 1609.4 (2006) Trial Use Standards for Wireless Access in Vehicular Environments (WAVE)-Multi Channel Operation
9. Delgrossi L, Zhang T (2012) Vehicle Safety Communications: Protocols Security, and Privacy. Wiley, Hoboken
10. Al-Qutayri M, Yeun C, Al-Hawi F (2009) Security and privacy of intelligent VANETs. In: Computational intelligence & modern heuristics. IN—TECH Publisher
11. Yeun C, Al-Qutayri M, AlHawi F (2009) Efficient security implementation for emerging VANETs. Spec Issue Appl Comput Ubiquit Comput Commun J 4(4)
12. Raya M, Hubaux JP (2007) Security vehicular ad hoc NETworks. J Comput Secur 15(1):3–38
13. Cho Y, Lee H, Park N, Choi D, Won D, Kim S (2009) Security technology trend in VANET. Korea Inst Inf Secur Cryptol 19(1)
14. Raw RS, Kumar M, Singh N (2013) Security challenges, issues and their solutions for VANET. Int J Netw Secur Appl 5(5):95–105
15. Chaubey NK (2016) Security analysis of vehicular ad hoc networks (VANETs): a comprehensive study. Int J Secur Appl 10(5):261–274
16. Chen L, Tang H, Wang J (2013) Analysis of VANET security based on routing protocol information. In: 2013 fourth international conference on intelligent control and information processing

17. Tomar RS, Verma S (2012) Enhanced SDMA for VANET communication. In: 26th international conference on advanced information networking and applications
18. Vijayakarthika R, Banumathi V (2014) Efficient data dissemination for secured communication in vanet. In: International conference on current trends in engineering and technology
19. Vijayalakshmi V, Saranya S, Sathya M, Selvaroopini C (2014) Survey on various mechanisms for Secure and Efficient VANET communication. In: ICICE
20. Nam J (2016) Implementation of VANET simulator using Matlab. J Korea Inst Inf Commun Eng 20(6):1171–1176
21. Diep PTN, Yeo CK (2016) A trust-privacy framework in vehicular ad hoc networks (VANET). In: Wireless telecommunications symposium
22. Donato EA, Maia G, Madeira ERM, Villas LA (2015) Impact of 802.11p channel hopping on VANET communication Protocols. IEEE Latin Am Trans 13(1):315–320
23. Kaur H, Batish S, Kakaria A (2012) An approach to detect the wormhole attack in vehicular ad-hoc networks. IJSSAN **1**(4)
24. Rahvari M, Jamali MAJ (2011) Efficient detection of sybil attack based on crytography in VANET. IJNSA 3(6)
25. Gandhi UD, Keerthana RVSM (2014) Request response detection algorithm for detecting DoS attack in VANET. In: International conference on reliability, optimization and information technology
26. RoselinMary S, Maheshwari M, Thamaraiselvan M (2013) Early detection of DoS attacks in VANET using attacked packet detection algorithm (APDA). In: 2013 international conference on information communication and embedded systems
27. Shukla N, Dinker AG, Srivastava N, Singh A (2016) Security in vehicular ad hoc network by using multiple operating channels. IEEE
28. Eritali M, El Ouahidi B (2013) A review and classification of various VANET instrusion detection systems. IEEE
29. Anjum F, Subhadrabandhu D, Sarkar S (2003) Signature intrusion detection for wireless ad hoc networks: a comparative study of various routing protocols
30. Kishore Raja PC, Suganthi M, Sunder MR (2006) Wireless node behavior based Intrusion detection using genetic algorithm. Ubiquit Comput Commun 7:143–148
31. Ping Y, Yichuan, Yiping Z, Shinyong Z (2005) Distributed instruction detection for mobile ad hoc networks. In: Processing of the 2005 IEEE symposium on application and the internet workshops AINT-W05
32. Singh A, Sharma P (2015) A novel mechanism for detecting DOS Attack in VANET using enhanced attacked packet detectioin algorithm (EAPDA). In: RAECS UIET
33. Grzybek A, Seredynski M, Danoy G, Bouvry P (2014) Detection of stable mobile communities in vehicular ad hoc networks. In: IEEE international conference on intelligent transportation systems
34. Raghavan UN, Albert R, Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E 76(3):036106
35. Leung IX, Hui P, Lio P, Crowcroft J (2009) Towards real-time community detection in large networks. Phys Rev E 79(6):1–10
36. Herbiet G, Bouvry P, Guinand F (2011) Social relevance of topological communities in ad hoc communication networks. In: IEEE computational aspects of social networks
37. Hussain R, Oh H (2014) A secure and privacy-aware route tracing and revocation mechanism in VANET-based clouds. J Korea Inst Inf Secur Cryptol 24(5):795–807
38. Hussain R, Oh H (2014) Cooperation-aware VANET clouds: providing secure cloud services to vehicular ad hoc networks. In: Korea information processing society
39. Park J, Choi D (2015) A design of framework for secure communication in vehicular cloud environment. J Korea Inst Inf Commun Eng 19(9):2114–2120

40. Park S, Kim N (2012) Design of MAC algorithm supporting adaptive transmission rate on VANET. J Korea Inst Electron Inf Eng 49(11):132–138
41. Qu F, Wu Z, Wang F, Cho W (2015) A security and privacy review of VANETs. IEEE Trans Intell Transp Syst 16(6):1–12
42. Lee WY (2014) A performance enhancement of VANET warning message propagation on electric wave blind area problem in the urban environment. J Korea Multimed Soc 17 (10):1220–1228
43. Park H, Hwang T, Jo Y, chi J (2014) An intersection connectivity-based RSU allocation algorithm in VANET. Korea Inf Sci Soc 10(1)
44. Kim S-H, Lee I-Y (2012) A study on message authentication scheme based on efficient group signature in VANET. J Korea Inst Inf Secur Cryptol 22(2):239–248
45. Islam N (2016) Certificate revocation in vehicular ad hoc networks: a novel approach. In: International conference on networking
46. Perer C, Georgakopoulos D (2014) Context aware computing for the internet of things: a survey. IEEE Commun Surv Tutor 16(1):414–454

Software Networking

Detecting Negative Deceptive Opinion from Tweets

Alemu Molla, Yenewondim Biadgie, and Kyung-Ah Sohn(✉)

Department of Software and Computer Engineering, Ajou University, Suwon, South Korea
{alemu,wondim,kasohn}@ajou.ac.kr

Abstract. Nowadays, a huge amount of opinions about specific brands of a company are shared on the Web. Such opinions are an important source of information for customers and companies. Unfortunately, there is an increasing number of deceptive opinions in order to deceive consumers by promoting a low quality product (positive deceptive) or by criticizing a potentially better quality product (negative deceptive). This paper focuses on the detection of negative deceptive opinions from tweets on specific brands of a company. We developed a classifier that detects negative deceptive opinions by combining lexical features of a tweet and personal profile and behavioural features of the writer. One of the challenges to develop this system is the lack of labeled dataset for training and testing. To resolve this issue, we collect our own dataset and label each tweet by multiple experts. Our experimental results show that the proposed system is a promising approach for detecting negative deceptive opinions. Our approach can help to identify defamers by analyzing personal profiles and writing style of each writer.

Keywords: Opinion mining · Negative deceptive opinion · Positive deceptive opinion · Tweet · Lexical features · Personal profile and behavioral features

1 Introduction

Social media such as Twitter has become extremely popular these days. Millions of users use this service to share their opinions, thoughts, and emotions. These user-generated comments represent a potential complementary source of essential information about company's brand. The comments can be classified as positive, neutral and negative [1]. Using this opportunity, users can disseminate intentionally the negative buzz to the rest of the customers. These negative opinions may come from an unsatisfied customer, a disgruntled employee, and a competitor company. Those involved parties can create a negative impact on a company's brand by writing their minds continuously when they are unhappy or disappointed on the reputation of the company. This situation can be spread to a large audience by a simple click of a mouse [2], and have the potential to create big impact on the company's entire brand.

It has been shown that most people believe that buying decisions are influenced by negative and positive reviews. However, people tend to share their negative experience than their positive experience. There are also some people involved in spreading fabricated rumors and misrepresenting the truth by hiding themselves. This situation causes the reputation of an individual, company or brand of a company to be defamed on the

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_36

twitter almost instantaneously worldwide with a little or no cost. These opinions are called spam opinion. It can be any deceptive texts like, fake comments, fake reviews, or fake social network postings. Studying and investigating such types of opinions can play a big role for companies, government institutions and policies. However, such opinions may come with short text, like a tweet using different writing styles from anonymous people. Detecting such opinions is still a challenging task.

In this study, we focus on negative deceptive tweets that are purposely written to defame a product of a company. We propose a method that can detect negative deceptives from general tweets. Various types of features are investigated, such as lexical features of tweets, personal profiles and behavioral analysis of writers. The proposed method is composed of two steps. In the first step, the sentiment of the given tweet is classified into positive, negative or neutral class. In the second step, the negatively classified tweets are further classified into deceptive negative or non-deceptive negative classes using lexical features of tweets as well as personal profile and behavioral features of writers. To develop this classifier, there is no labeled public dataset. Hence, we prepared our own labeled dataset to train and test the proposed classifier.

The rest of the paper is organized as follows. In Sect. 2, the review of related work is presented. In Sect. 3, the proposed approach is described in detail. Experimental results come in Sect. 4. Conclusions and future work are given in Sect. 5.

2 Related Work

Most researches on opinion mining and spam detection using the web mainly focus on classifying opinions into positive, negative and neutral based on the sentiments of the polarity [3–7]. The focus of our work is identifying the negative deceptive opinion spams from negative tweets. In [8], different types of spams are described such as web spam, email spam, and opinion spam. Web spam has the objective of making an attraction for target pages so as to draw individuals to go to these pages. There are two kinds of web spams: content spam and link spam. Content spam tries to include remotely important or insignificant words to the target pages to attract individuals to go to these pages. Numerous studies considered this issue [9–15]. Link spams do not consider the content of spams, only focused on the hyperlink of spams. Spammers can give a positive or negative opinion to their target object. Spams in the reviews are very distinctive. In [16], a comparative work is presented for the issue of review spam, and sorted the distinctive types of spam reviews. The well-known spam categorization issue (spam versus non-spam) is traditionally studied in [17, 18] in the domain of web and email. Automatic detection of opinion spam that aims to deceive readers is merely another difficult face of comparative problems studied in the sentiment analysis domain [19].

In [8, 16], the authors defined three categories of features in the context of product review sites. The primary category is review-driven features, the second is reviewer-driven features and the last category is item-driven features. In [20], behavioral attributes between harmful and normal users are examined, and three behaviors of spammers were proposed, namely, similarity of comments across time, the consideration of the target in the comment, and the time stamp of the comment. If a user gives similar and repeated comments, if a user does not consider the target of the comment, and if the time stamp of the comments is almost similar, authors conclude that a user is a fake user. Based these three criteria, they tried to differentiate harmful and normal users.

Deceivers are getting smarter these days, and they attempt to conceal data by having extra fanciful exertion [21, 22]. These progressions impact writing style of spammers. The stylometric features are considered as a system to discover deceptive reviews. In [19, 23], four types of features in light of the writing style are extricated, namely, syntactic, lexical, content-specific and structural features. In [23], three distinctive clustering algorithms are used. Among them bisecting k-means and k-means algorithms beat Expectation Maximization with 90% of F-measure. In a recent work [24], lexico-syntactic patterns are applied to detect deceptive opinion reviews. Similar work in [21] tried to detect deceptive reviews by using lexical and syntactical features. Considering these previous results, our technique is relying upon lexical as well as personal profile and behavioral features to distinguish deceivers. To extract personal profile and behavioral features of the spammer, we consider the following aspects.

- What is the personal tweet score distribution?
- Does the spammer tweets mostly negative or not?
- Does the spammer hiding spatial information in his/her profile information?
- Does the spammer tend to use more negative emoticons in the tweet and inadequate comments with offensive words?
- What is the age of the account? Is it new or old?

In addition to the above clues, we also consider the lexical analysis of the tweet as well. The details of the proposed method will be described in the following sections.

3 System Architecture of the Proposed Deceptive Detector

The general architecture of the proposed systems is illustrated in Fig. 1. It is composed of two main components: General Sentiment Classifier and Negative Deception Detector. Each component of the proposed detector is depicted below in detail.

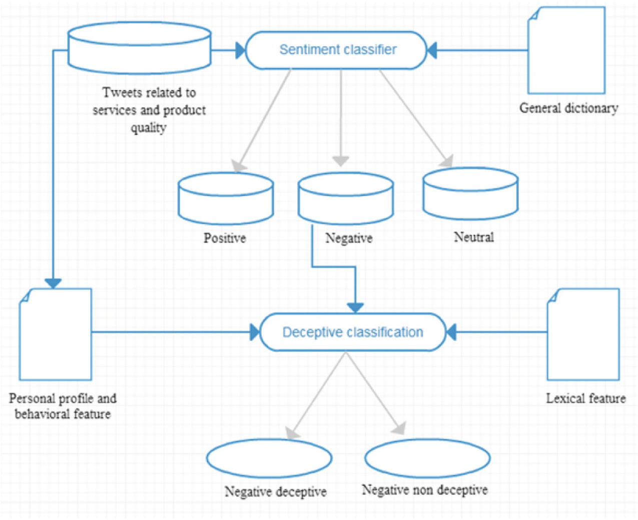


Fig. 1. The architecture of the proposed deceptive detector

3.1 General Sentiment Classifier Module

The general sentiment classifier assigns tweets into one of the three classes of positive, negative or neutral. We collected around 10,000 tweets related to Samsung products and services using different hash tags. After data collection, several data preprocessing tasks are performed like removing of URLs, excluding retweeted tweets and minimizing repeated letters. After pre-processing the data, we score each tweet by using a general negative and positive word dictionary from [4] which is publicly available at http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html.

A Naïve Bayes classification algorithm is used to develop this classifier using the "Sentiment" package in R. Among 10,000 tweets, the numbers of negative, positive and neutral tweets are found to be 2,315, 2,808, and 4, 877, respectively. We used Naïve Bayes classification algorithm because of its computational and memory efficiency. It is also effective with small training data size compared to the other methods [25, 26]. The sentiment score distributions of the collected tweets are illustrated in Fig. 2.

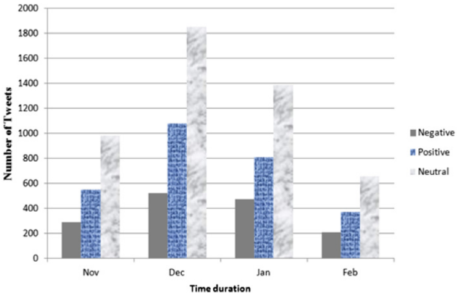


Fig. 2. The distribution of negative, positive and neural tweets in time series.

3.2 Negative Deceptive Detection Module

There are some users who criticize purposely the products and services of a company by expressing a negative deceptive opinion. Investigating the behavior of writers and analyzing their writing style play a vital role to identify spammers. In this module, we utilized lexical features of the tweet as well as personal profile and behavioral features of writers to recognize negative deceptive opinions from the negative tweets. The generation of these features is explained in detail in the following sections.

3.3 Feature Generation

Syntactic, lexical, content specific and structural features are the common features used for general text analysis. Among these features, lexical features are effective in detecting deceptions because features focus on the writing styles of reviews and reviewers [21, 23, 27, 28]. In this paper, in addition to lexical features, personal profiles and behavioral features are used to develop the negative deceptive detector. We implemented the feature extraction module based on the characteristics of spam comments given in [20, 21]. Totally, 45 features are extracted, which are described in detail as follows.

Lexical features. Lexical features include attributes such as part-of-speech and n-grams. These features help to identify the writer's preferred usage of words and characters. There are two categories of lexical features, word-based and character-based. Character-based features include mostly numbers of capital letters, small letters, digits, white-spaces, tabs, and frequency of other ASCII characters. The second category of lexical feature depends on word orders and its occurrence such as average word length, sentence length, frequency of words etc. [19, 23, 28]. Based on this principle, we extracted the following 36 lexical feature types.

Total Number of White Spaces in the Tweet (Twhite). Spam comments tend to have a large number of white spaces so that it has more impact on the user who reads it.

Total Number of Sentences in the Tweet (Tsente). The number of sentences in a spam comment is typically smaller than the one in a legitimate comment; legitimate comments usually have a coherent use of words and sentences are clearly delimited.

Punctuation Marks in the Tweet. Spam comments tend to have a larger number of punctuation marks, especially exclamation marks and dots, to draw attention of the readers. This feature includes 7 punctuation marks, namely, period (.), question mark (?), exclamation mark (!), semi-colon (;), colon (:), single quote (‘) and double quote (“).

Special Characters in the Tweet. Spam comments have a tendency to have a higher number of non-ASCII Characters in comparison to legitimate comments. This feature includes 20 special characters: < , > , ~ , ^, |, {,}, [,], \/, #, $, +, –, *, %, -,., &, and@.

Stop Word Ratio. The stop word ratio is calculated as the ratio between the range of stop words and the aggregate number of words in the comment. True comments tend to have a proportional number of stop word ratio in contrast with spam comments.

Number of Capital Letters from the Tweet. Capital letters are used in spam comments to attract more attention than legitimate comments. Spam comments are often written entirely using capital letters.

Number of New Lines in the Tweet. Spam comments tend to have a high number of new lines in order to draw attention upon certain facts by creating a visual effect.

Total Number of Words in the Tweet(Twords). An increased and decreased number of total letters in the tweet can indicate legitimate user comment and spam user comment, respectively. Most of the time spammers use less words than legitimate users.

Total Number of Characters in the Tweet (Tchar). It includes all letters upper case A-Z, and lower case a-Z letters, all digits, all punctuation marks. Spam comments have less count number of characters than legitimate users.

Number of Alphabets in the Tweet(Talphabets). This represents the total number of upper case A-Z and lower case a-z alphabets. Spammers would like to use a less number of alphabets in their tweets than legitimate user opinions.

Number of Digits(Tdigits). It has been observed that there is a certain class of spam comments that have a high digit number (0–9) and sometimes use symbols in place of letters related to the length of the comment.

Personal Profile and Behavioral Features. This feature can have an important role in representing the identity of a person. Authors in [20] tried to identify spammers based on three malicious user's behaviors. If the timestamp of a user's comment is frequent and unique than other normal users, if a user posts a redundant comment and has no relation to the domain of the target, a user must be a spammer. It was an effective approach to detect spammers. Our feature extraction technique is based on similar idea, but besides behavioral features, we also include personal profile information. To extract the profile information of a user, we use a web API from http://foller.me/ and http://tffratio.com. We extract 9 personal profile and behavioral features: Total number of followers, Total number of following, Ratio of total number followers to total number of following, Total number of negative tweets, Total number of positive tweets, Total number of neutral tweets, Ratio of total number of negative tweets to total number of positive tweets, Attitude of the writer (ratio of the total number of happy feelings to sad feelings), and Number of years of the user account created.

3.4 Preparation of Negative Deceptive and Negative Non-deceptive Corpus

One of the main challenges to develop negative deceptive detector is the lack of labeled data set. Therefore, most initial works regarding the detection of opinion spam considered unsupervised approaches using meta information from reviews and reviewers. Recently, authors in [29] release the gold-standard datasets which contain examples of positive and negative deceptive opinion spam about hotel reviews. The nature of this datasets is different from the tweet dataset because the number of characters in review is usually much longer than 140 in a tweet. As a result, we cannot use this data to develop a system that identifies negative deceptive users of a tweeter application. To the best of our knowledge, there is no publicly available labeled tweet data set. To address this problem, we create our own dataset with negative deceptive and negative non-deceptive labels tweets about Samsung's products and services. Five graduate students took part in labeling the dataset. Since the resulting labels can disagree between participants, we use the following two strategies to create labeled datasets. If the given tweet is labeled by five students and if all of them give the same label, we include the tweet in our dataset. We called this consensus vote labeling.

Initially, among 10,000 tweets, the general sentiment classifier generates 2,315, 2,808, and 4, 877 number of negative, positive and neutral tweets respectively. Among 2,315 negative tweets, we got 53 negative deceptive and 51 negative non-deceptive tweets using consensus vote labeling method. Since the resulting dataset is small, we use our second strategy to expand the dataset: if a given tweet is marked by five students and if the marked labels of three or four of them agree, we include this as our training dataset. We call this majority vote labeling. Using this technique, we increase the number of samples from 53 to 98 and from 51 to 98 in each type of negative tweets. Therefore, we have 104 and 196 training samples from the consensus vote and majority vote labeling methods, respectively. We train six different classification algorithms, namely, Support Vector Machine (SVM), Naïve Bayes, Decision Tree, Maximum Entropy, BAGGING and Random Forest to identify negative deceptive tweets among negative tweets.

4 Experimental Results

To validate our method, we perform three types of experiments. We first examine the performance of the proposed detector using 45 features. In the second experiment, we further investigate the exactness of the proposed detector by selecting best 9 features using the Akaike information criterion (AIC) method. The third experiment compares the frequency distribution of the major lexical feature between negative deceptive and negative non-deceptive tweets..

Experiment 1. We trained six different negative deceptive detectors using 45 features and check the 10-fold cross validation accuracy. Table 1 shows the result using the data set labeled by consensus labeling method. It shows that Maximum Entropy classifier performs the best with 100% accuracy and Naïve Bayes with 98% using combined features. The high accuracy seems to be due to the labeling strategy. If a tweet is labeled unanimously by all the participants, then it is more likely that the classification task is rather easy. On the other hand, the training of classifiers using the data set labeled by majority voting method faces more challenging shown in Table 2. Still, maximum Entropy classifier outperformed all the others while the overall accuracy using the majority voting labeling method is lower than consensuses voting method as expected. In terms of the comparison among features, lexical features dominate the overall performance across all the classification algorithms. Personal profile and behaviors features tend to also improve the performance of classifiers, especially Naïve Bayes and Decision Tree classifiers.

Table 1. Classification accuracy on the dataset labeled by consensus labeling method

Classifier	Lexical features	Personal profile and behavioral features	Combined features
Naïve Bayes	0.9552239	0.9701493	0.9850746
Maximum entropy	1.0000000	0.9359376	1.0000000
Decision tree	0.8527458	0.5428986	0.8853164
SVM	0.97725	0.7560703	0.9776535
Random forest	0.9793651	0.7578879	0.9609921
Bagging	0.9669001	0.7327058	0.9643797

Table 2. Classification accuracy on the dataset labeled by majority vote labeling method

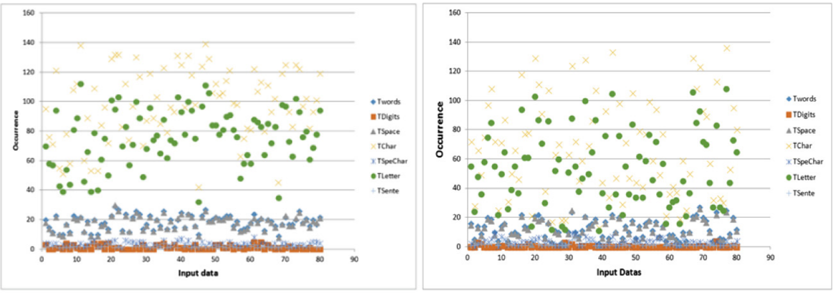
Classifier	Lexical features	Personal profile and behavioral features	Combined features
Naïve Bayes	0.7777778	0.8055556	0.9166667
Maximum entropy	0.9518532	0.9441787	0.9465301
Decision tree	0.7588071	0.7440828	0.7390722
SVM	0.8061051	0.8037717	0.8100775
Random forest	0.8403794	0.8392034	0.8215295
Bagging	0.7651701	0.7747905	0.7900969

Experiment 2. To select the best features and evaluate their impact on the result, we perform a stepwise variable selection, which produced 9 best features among 45. As shown in Table 3, Maximum Entropy performed the best on both data. Naïve Bayes performed well on the data labeled by consensus labeling method. Considering each variable has been normalized to have zero mean and standard deviation of one, lexical features seem to have more discriminative power than others profiles in general.

Table 3. Performance after variable selection on the dataset labeled by majority vote labeling

Classifier	Consensus labeling	Majority vote labeling
Naïve Bayes	1.0000000	0.7777778
Maximum entropy	1.0000000	0.9548633
Decision tree	0.8664477	0.7296332
SVM	0.9646252	0.8194008
Random forest	0.9733333	0.8237779
Bagging	0.9462189	0.7789863

Experiment 3. We compare the frequency of major features to see their similarities and differences between two classes using majority vote labeling data. There is a clear difference in the feature occurrences between the two categories. For instance, the natures of negative non-deceptive tweets are longer than deceptive negatives tweets (Fig. 3).



(a) (b)

Fig. 3. Frequency of lexical features in negative (a) non-deceptive and (b) deceptive tweets.

5 Conclusion and Future Work

We conducted an in-depth analysis on negative tweets to identify negative deceptive comments. We found that lexical features as well as personal profile and behavioral features play a vital role to detect negative deceptive opinions. Experimentally, we found that the negative deceptive detector developed by the Maximum Entropy algorithm is the best detector. We tried to prepare negative deceptive and negative non-deceptive tweet datasets from collected negative tweets about Samsung's products and services.

In our future work, we want to introduce social network module to the architecture of the proposed negative deception detector. This module will be used in order to integrate opinion of users with social networks. Although our dataset size is not that much large, the results using our proposed methods seem promising.

Acknowledgments. This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education [NRF-2016R1D1A1B03933875].

References

1. Tsytsarau M, Palpanas T (2012) Survey on mining subjective data on the web. Data Min Knowl Disc 24(3):478–514
2. Brown JS, Duguid P (2000) The Social Life of Information. Harvard Business Press, Boston
3. Dave K, Lawrence S, Pennoc DM (2003) Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In: Proceedings of the 12th international conference on world wide web, pp 519–528. ACM, Budapest, Hungary
4. Hu M, Liu B (2004) Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp 168–177. ACM, Seattle, Washington
5. Li W, Zhong N, Liu C (2006) Combining multiple email filters based on multivariate statistical analysis. In: International symposium on methodologies for intelligent systems, pp 729–738. Springer, Berlin, Heidelberg
6. Ntoulas A, Najork M, Manasse M, Fetterly D (2006) Detecting spam web pages through content analysis. In: Proceedings of the 15th international conference on world wide web, pp 83–92. ACM, Edinburgh, Scotland
7. Sahami M, Dumais S, Heckerman D, Horvitz E (1998) A Bayesian approach to filtering junk e-mail. In: Learning for text categorization: papers from the 1998 workshop, vol 62, pp 98–105, Madison, Wiscon
8. Jindal N, Liu B (2007) Analyzing and detecting review spam. In: Seventh IEEE international conference on data mining, pp 547–552. IEEE, Omaha
9. Fetterly D, Manasse M, Najork M (2005) Detecting phrase-level duplication on the world wide web. In: Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval, pp 170–177. ACM
10. Castillo C, Donato D, Becchetti L, Boldi P, Leonardi S, Santini M, Vigna S (2006) A reference collection for web spam. ACM SIGIR Forum 40(2):11–24
11. Gyongyi Z, Gartia-Molina H, Pedersen J (2004) Combating web spam with TrustRank. In: Proceedings of the 30th VLDB conference, pp 576–587, Toronto, Canada
12. Henzinger M (2006) Finding near-duplicate web pages: a large-scale evaluation of algorithms. In: Proceedings of the 29th annual international SIGIR conference, pp 284–291, ACM, Seattle, Washington, USA
13. Liu B (2007) Web Data Mining: Exploring Hyperlinks, Contents and Usage Data. Springer, New York
14. Turney PD (2002) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th annual meeting on association for computational linguistics, pp 417–424, Philadelphia

15. Wang Y-M, Ma M, Niu Y, Chen H (2007) Spam double-funnel: connecting web spammers with advertisers. In: Proceedings of the 16th international conference on world wide web, pp 291–300. ACM, Banff, Alberta, Canada
16. Jindal N, Liu B (2008) Opinion spam and analysis. In: Proceedings of the 2008 international conference on web search and web data mining, pp 219–230. ACM, Palo Alto, California, USA
17. Gyongyi Z, Gartia-Molina H (2005) Web spam taxonomy. In: First international workshop on adversarial information retrieval on the web, Chiba, Japan
18. Drucker H, Wu D, Vapnik VN (1999) Support vector machines for spam categorization. IEEE Trans Neural Netw 10(5):1048–1054
19. Zheng R, Li J, Chen H, Huang Z (2006) A framework for authorship identification of online messages: writing-style features and classification techniques. J Am Soc Inf Sci Technol 57(3):378–393
20. Wang Q, Liang B, Shi W, Liang Z, Sun W (2010) Detecting spam comments with malicious users' behavioral characteristics. In: International conference on information theory and information security, pp 563–567. IEEE, Beijing, China
21. Shojaee S, Azrifah M, Murad A, Azman A, Sharef NM, Nadali S (2013) Detecting deceptive reviews using lexical and syntactic features. In: 13th international conference on intelligent systems design and applications, pp 53–58. IEEE, Bangi, Malaysia
22. Frank MG, Menasco MA (2009) Human behavior and deception detection. In: Handbook of science and technology for homeland security, Wiley, New York
23. Iqbal F, Binsalleeh H, Fung BCM, Debbabi M (2010) Mining writeprints from anonymous e-mails for forensic investigation. Digital Invest. Int J Digital Forensics 7(1–2):56–64
24. Ott M, Choi Y, Cardie C, Hancock JT (2011) Finding deceptive opinion spam by any stretch of the imagination. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, vol 1, pp 309–319 Portland, Oregon
25. Manning CD, Raghavan P, Schutze H (2009) Introduction to Information Retrieval. Cambridge University Press, Cambridge
26. Huang J, Lu J, Ling CX (2003) Comparing naive bayes, decision trees, and SVM with AUC and accuracy. In: Proceedings of the third international conference on data mining, pp 553–556. IEEE, Melbourne, Florida, USA
27. Feng S, Banerjee R, Choi Y (2012) Syntactic stylometry for deception detection. In: Proceedings of the 50th annual meeting of the association for computational linguistics, pp 171–175, Jeju, Republic of Korea
28. Pearl L, Steyvers M (2012) Detecting authorship deception: a supervised machine learning approach using author writeprints. Literary Linguist Comput 27(2):183–196
29. Ott M (2011) Deceptive Opinion Spam Corpus v1.4. http://my1eott.com/op_spam/

A Study on Effectiveness of Network Attack Using Analysis of Eigenvalue

Ayumi Ishimaru(✉) and Hidema Tanaka

National Defense Academy of Japan, 1-10-20 Hashirimizu, Yokosuka, Kanagawa, Japan
{em55037,hidema}@nda.ac.jp

Abstract. Network communication is based on IP packets which are standardized by international organization. Therefore, network attack does not work without following the standardized manner. Hence, network attack also leaks adversaries' information in their IP packets. In previous works, malicious topology maps are derived using these IP packet informations. And the effectiveness of network attack is evaluated by increment in eigenvalue of topology matrix (adjacency matrix and Laplacian matrix). However, previous method does not consider security level of each node. In this paper, we propose the solution to this problem and the improvement of previous method using total accessibility matrix. As a result, we confirm that the effectiveness of attack can be increased by attacking some nodes with low accessibility. We also found that the shape of topology map has big influence to the effectiveness.

Keywords: Network attack · Topology map · Total accessibility matrix

1 Introduction

1.1 Background

Recently, we can find some previous works studying topology map and network attack [1, 2]. Since network attacks use TCP/IP, attack packets also include information of adversaries. Based on this fact, previous work [1] shows network counter attack strategy using darknet monitoring. The attack scenarios are as follows.

Scenario-1: Spread of malware and disinformation
Scenario-2: Concentration and confusion of information sharing

And the following three tactics are proposed.

Tactics-1: Down of server
Tactics-2: Construction of agent server
Tactics-3: Combination of Tactics-1 and Tactics-2

By combining of two kinds of scenarios and three types of tactics, we can deduce six patterns of strategies. In previous work, they adopt exhaust search for the optimum strategy using following procedure.

Step-1. Collect IP addresses from the target area (target IP group).

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_37

Step-2. Execute traceroute command for target IP group.
Step-3. Estimate the topology of target area.
Step-4. Execute simulation of Tactics-1 ~ -3.
Step-5. Choice the scenario and tactics (strategy).

They used darknet log data in Step-1 and derived network topology map of adversary area in Step-2 and -3 (Country B and C [1]) [3–5]. They executed exhaust simulation assuming all node as target, in Step-4.

The effectiveness of these six patterns of strategies are estimated by eigenvalue analysis of the network topology map [1]. The effectiveness of Scenario-1 is evaluated using adjacency matrix and one of Scenario-2 using Laplacian matrix [6, 7].

1.2 Adjacency Matrix and Laplacian Matrix

Let G be a network with n nodes. Let $A(n \times n)$ be the adjacency matrix of G, and $A_{i,j}(1 \leq i,j \leq n)$ be its elements.

$$A_{i,j} = \begin{cases} 1 & (i\,\text{link to}\,j) \\ 0 & (otherwise) \end{cases} \tag{1}$$

Let λ(A) is the eigenvalue of the adjacency matrix A obtained from the following characteristic equation. Also, eigenvalue λ(A) has eigenvector $I(L)$ corresponding to each.

$$\det(\lambda I - A) = 0 \tag{2}$$

Since this is an n-th order characteristic equation, there are $m(1 \leq m \leq n)$ eigenvalues. Let $\lambda_{max}(A)$ be the maximum value of λ. Since this indicates the degree of connectivity between hub nodes, $\lambda_{max}(A)$ represents "spread speed" in the network.

The network G is also represented by the Laplacian matrix L. Let $L_{i,j}(1 \leq i,j \leq n)$ be an element of the Laplacian matrix L.

$$L_{i,j} = \begin{cases} d_i & (i = j) \\ -1 & (i\,link\,to\,j) \\ 0 & (otherwise) \end{cases} \tag{3}$$

Where d_i denote the degree of the i-th node. The eigenvalues of Laplacian matrix L are obtained in the same way as the eigenvalues of the adjacency matrix shown in the Eq. (2). Therefore, there are different m $(1 \leq m \leq n)$ eigenvalues L, and the minimum eigenvalue $\lambda_{min}(L)$ always equals to 0. There is following relationship among eigenvalues.

$$0 = \lambda_1(L) \leq \lambda_2(L) \leq \ldots \leq \lambda_{max}(L) \tag{4}$$

The network is divided into some independent areas and the number of such areas is equals to the number of eigenvalues $\lambda(L) = 0$. For example, when $\lambda_2(L) = 0$, network G is unlinked and divided into two areas. In order to maintain network connectivity, it is necessary to keep the second minimum eigenvalue $\lambda_2(L)$ with a positive value. Because of this property, the second minimum eigenvalue $\lambda_2(L)$ is called algebraic connectivity of network G. When $\lambda_2(L)$ takes a large value, the network has high connectivity. In addition, the maximum eigenvalue $\lambda_{max}(L)$ represents the difficulty in delay of communication. The ease of synchronization in the network can be evaluated by the ratio $R = \lambda_2(L)/\lambda_{max}(L)$. Therefore, the value of R represents "convergence" in the network.

1.3 Problem of Previous Work and Motivation

In previous work [1], they set security levels of all nodes zero. Actually, such condition is not adequate. Therefore, the attack may be infeasible against the optimal node derived by the method. Or, even if we attack many weak nodes successfully, the purpose of strategy may not be archived. In the first place, we cannot judge the security level of which node is low or high from the topology map or matrixes derived from Step-3 and -4. To solve this problem is our motivation.

2 Proposal Method

To solve the problem, we assume that the node with high degree has high security level and vice versa. There are some estimation methods for degree of node, we focus on total accessibility matrix T [8]. In particular, total accessibility matrix shows not only degree but also accessibility according to degrees of neighbor nodes. Therefore, even if two nodes have same degree, their accessibility may be different each other. This characteristic is adequate for our purpose.

For network G, let d be a diameter and A $(n \times n)$ be adjacency matrix, total accessibility matrix T is calculated as follows,

$$T = A + A^2 + A^3 + \ldots + A^d. \tag{5}$$

Let $T_{i,j}$ be an element of T and V_i be accessibility of node i, we can calculate accessibility V_a of node a as follows,

$$V_a = \sum\nolimits_{i=1}^{n} T_{i,a}. \tag{6}$$

In this paper, we set target node group whose accessibility is less than the threshold $\mathbb{V}$ which is determined by the attacker. The difference of attack condition between previous method and proposal is summarized in Table 1. The procedure of our proposal method is same as previous shown in Sect. 1.1.

Table 1. Attack condition

	Previous	Proposal
Down of server	1	2, 3 with $V_i < V$
Agent server	1	1
Generated link	2	2, 3, 4 with $V_i < V$

3 Experiments

3.1 Darknet Log Data

To compare the previous results [1], we use the same darknet log data (Table 2). As same as previous work, we restricted the network with 100 nodes around the capital area (Fig. 1). Actual attacks are often executed through some springboard PC. So, it is difficult to know the true IP address of the attacker. However, even if the springboard PC did not participate in the attack intentionally, we regard them as adversaries. Note that some detection methods of springboard PC are proposed [9–11].

Table 2. Number of accesses to darknet (2013/3/1~21)

	Number of access	IP address	Traceroute IP	Traceroute link
Country B	75,785	53,390	17,684	24,163
Country C	8,728	3,674	2,119	3,819

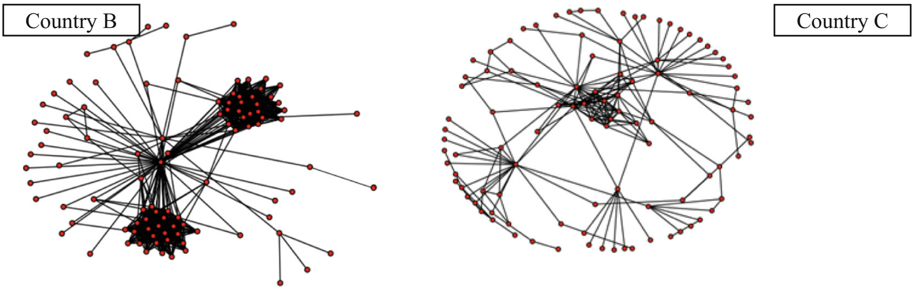


Fig. 1. Malicious topology map in Country B and C

3.2 Total Accessibility Matrix and Threshold $\mathbb{V}$

We derive network topology map from Step-1 ~ -3 and calculate adjacency matrix A. Because of same reason described in [1], since darknet log data has many sensitive information, we cannot show any details in the followings. From adjacency matrix A, we calculate diameter d. In our experiment, we use diameter command provided Python language, and obtain $\boldsymbol{d} = \mathbf{4}$ for Country B, and $\boldsymbol{d} = \mathbf{8}$ for Country C. The distribution of accessibility in each country shown in Fig. 2. The vertical denotes V_i and the horizontal denotes index number i for the node.

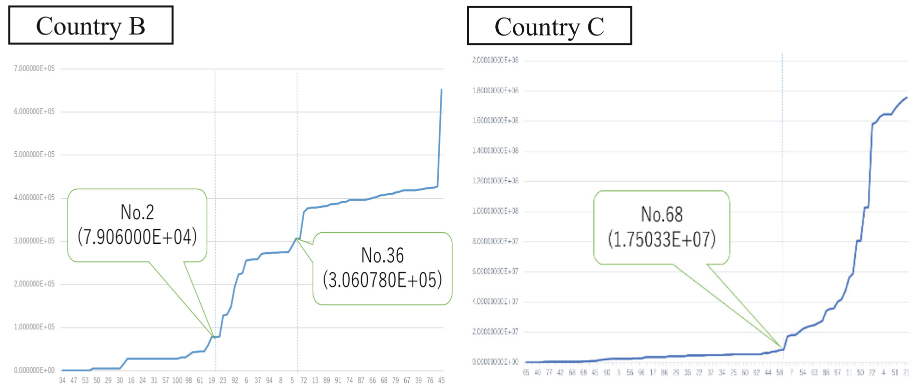


Fig. 2. Distribution of accessibility in Country B and C

We focus on the significant change in increment of accessibility in Fig. 2. In Country B, there are two significant points; No. 2 node and No. 36 node. In this paper, we judge $\mathbb{V} = \mathbf{3.06 \times 10^5}$ from the view point of reasonable ratio of number of backbone nodes with high security and one of endpoint nodes with weak security. In the same view point, we set $\mathbb{V} = \mathbf{1.75 \times 10^7}$ for Country C.

3.3 Computer Simulation and Results

We execute Step-4 by computer simulation (Table 3). Table 3 shows computational cost for each simulation. The cost is estimated by the number of calculation of eigenvalues (Table 4). As described above, we choose target nodes under the threshold $\mathbb{V}$ with attack condition (see Table 1). Due to the limitation of computational cost, we execute only Tactics-1 and-2 in this paper. For Tactics-1, we took 1.4 s for one node down, about one hour for 2 nodes down and almost one day for 3 nodes down. And for Tactics-2, we took two minutes for 2 nodes link generation, about one hour for 3 nodes link generation, and almost one day for 4 nodes link generation (Table 5). The best results of Tactics-1 is shown in Table 6. For Tactics-2, we also executed additional experiments to determine the necessary number of links for $\lambda_{max}(A)$ which becomes larger than the initial value or previous result [1] (Table 7).

Table 3. Specification of our computer environment

OS	Windows 10 home 64-bit
Compiler	Python 3.5.1 (Anaconda 4.0.0)
CPU	Intel (R) Core (TM) i7-6700 CPU @ 3.40 GHz
Memory	8.00 GB

Table 4. Computational cost

	Country B	Country C
1 node down	100	100
2 nodes down or 2 nodes link	1,953 ($_{63}C_2$)	2,278 ($_{68}C_2$)
3 nodes down or 3 nodes link	39,711 ($_{63}C_3$)	50,116 ($_{68}C_3$)
4 nodes link	595,665 ($_{63}C_4$)	814,385 ($_{68}C_4$)

Table 5. Necessary time for calculation (sec)

	Tactics-1	Tactics-2
1 node	1.4	–
2 nodes	3241.5	122.9
3 nodes	81,037.5	3904
4 nodes	–	85,905

Table 6. Best Results of Tactics-1

	Country B		Country C	
	$\lambda_{max}(A)$	R	$\lambda_{max}(A)$	R
Initial value	24.2098	0.002853	10.0785	0.005487
1node down [1]	24.2098	0.012527	10.0785	0.005950
1node down	24.2098	0.012527	10.0785	0.005950
2nodes down	24.2098	0.014056	10.0785	0.006012
3nodes down	24.2098	0.014390	10.0785	0.006455

Table 7. Best Results of Tactics-2

	Country B		Country C	
	$\lambda_{max}(A)$	R	$\lambda_{max}(A)$	R
Initial value	24.2098	0.002853	10.0785	0.005487
2nodes link [1]	24.2165	0.003553	10.1152	0.006329
2nodes link	24.2098	0.003549	10.0785	0.006247
3nodes link	24.2098	0.003728	10.0785	0.006805
4nodes link	24.2098	0.004153	10.0785	0.007254
Over Initial	18links	–	23links	–
Over result [1]	43links	–	64links	–

4 Consideration

4.1 Effectiveness of Scenario-1

4.1.1 Tactics-1

As shown in previous work [1], Tactics-1 has no influence to Scenario-1. The value of $\lambda_{max}(A)$ is determined by the number of nodes n and the number of links l, and the maximum value is $n - 1$. Since Tactics-1 decreases n and l, the resultant eigenvalue also decreases. From this fact, previous work expected that Tactic-1 is invalid for Scenario-1. We can confirm it from the results shown in Table 6. In particular, the results of multiple attacks against nodes with low accessibility maintain the same as the initial value and do not change the properties of network.

On the other hand, paper [12] shows that attack of top 5% of high degree nodes, the average path length of topology map increases to about twice. Then, the contrary to our expectation, we can predict that $\lambda_{max}(A)$ will be decreased greatly when attacking the nodes with high accessibility. We confirm this expectation by an additional experiment. As the same as paper [12], we attack top 5% of accessibility node. As a result, for Country B, we attacked two nodes and obtained $\lambda_{max}(A) = 22.6036$. In the same way, the result for Country C shows that $\lambda_{max}(A) = 9.1832$ with one node attack. From these results, we can conclude that miss attacking to the high accessibility node will cause greatly down of effectiveness of attack.

4.1.2 Tactics-2

From Table 7, 2 node links generation, which is the minimum attack of Tactics-2, has the best effectiveness with the target nodes of previous result [1], and ones of our attacks are same as the initial value. This is the same situation of Tactics-1. However, previous result attacks nodes with high accessibility (Country B: $V_{45} = 6.52 \times 10^5$ and $V_{77} = 4.28 \times 10^5$, Country C: $V_{21} = 1.76 \times 10^8$ and $V_{24} = 1.74 \times 10^8$), the strategy will be very hard to execute. In addition, we simulate the necessary number of nodes to

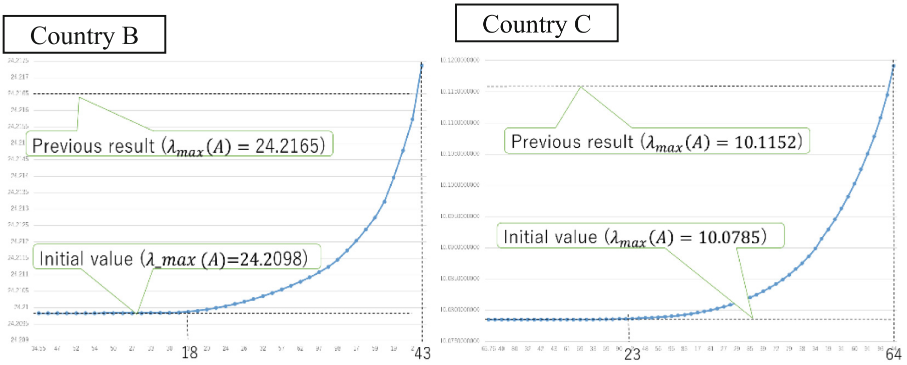


Fig. 3. Country B and C increasing link ($\lambda_{max}(A)$)

achieve more than the value of initial and result [1], adding to connect nodes from the lowest accessibility (Fig. 3).

From the view point of comparing with initial value, necessary number for Country B is 18 and one for Country C is 23. And from comparing with result [1], for Country B is 43 and for Country C is 64. Note that the number of nodes in target topology maps are same in each country (see Sect. 3.1). Therefore, we assume that these difference is caused by the shape of topology map and number of links. Obviously, big hub nodes are founded in Country B from Fig. 1. Therefore, it will be easy to construct another hub node. On the other hand, in Country C, all nodes are almost uniformly distributed. As a result, it will be hard to construct another hub node, we need more number of attack nodes than Country B. The detailed analysis of our assumption is our future work.

4.2 Effectiveness of Scenario-2

4.2.1 Tactics-1

The main purpose of Scenario-2 is to concentrate and confuse their information sharing, in other word, it is to divide the network topology map into some isolate areas. From the result shown in Table 6, we can confirm the effectiveness of our tactics. And the result of previous work [1] and our 1 node down have same attack target. In the case of Country B, R increases greatly when attacking the highest accessibility ($V_{45} = 6.52 \times 10^5$). This is the inverse situation of the case of Scenario-1 (see Sect. 4.1.1). On the other hand, in the case of Country C, we cannot find such situation by attacking high accessibility node. From the shape of topology map (Fig. 1), in particular Country B, we consider that the down of huge hub node will generate many isolate small hub nodes, then R is increased. However, the topology map which all nodes are almost uniformly distributed is considered to be difficult to generate smaller hub nodes by the down of hub node. The detailed analysis of correlation between effectiveness of Scenario-2 with Tactics-1 and shape of topology map is our future work.

4.2.2 Tactics-2

The results of Tactics-2 for Scenario-2 is shown in Table 7. The previous result [1], which is the minimum attack condition, connects $V_{30} = 5.53 \times 10^3$ and $V_{70} = 3.91 \times 10^5$ for Country B. And for Country C, it connects $V_{38} = 4.96 \times 10^6$ and $V_{57} = 2.43 \times 10^7$. These results are better than our attack generating 2 nodes link, however, since the accessibility of target nodes are high, the strategy derived by previous result is difficult to execute. On the other hand, our results derive targets which are feasible to attack and effectiveness of attack are almost same as previous. And we can confirm that larger number of generated links will derive more effective strategy. From the experimental results, we found that the agent server which connects nodes of low accessibility and one of middle range, increase the value of R effectively. This fact will point out that connecting end points of topology map which belong to different cluster from each other, is effective. Since the effectiveness of method based on above, our prediction will depend on the shape of topology map, the detailed analysis is our future work.

The main purpose of Scenario-2 is to concentrate and confuse their information sharing. Therefore, when a target node for converging information is determined, it is necessary to search for an input node of disinformation using eigenvectors. Therefore, it is insufficient for this purpose only to evaluate using the value of R. The evaluation method based on R and eigenvector is also our future work.

5 Conclusions

Previous work [1] proposed network counter attack strategy using darknet monitoring, however, the security level of all nodes is set zero. In this paper, we focus on total accessibility matrix to solve this problem. We assume that the node with high accessibility has high security level and vice versa. As a result, Tactics-1 has no influence to Scenario-1, and miss attacking to the high accessibility node will cause greatly down of effectiveness of strategy. Tactics-2 require many end point nodes to increase the effectiveness, so it is expected that the attack cost will increase. Therefore, we conclude that the method of attacking nodes with low accessibility is not effective for both Tactics-1 and-2. We also confirmed that Tactics-1 and-2 for Scenario-2 are effective for improving convergence R.

In addition, we found that the shape of topology map has big influence to the effectiveness of attack. Comparing the results of Country B and C in Tactics-2 for Scenario-1, we need more number of attack nodes in Country C. And in the case of Tactics-1 for Scenario-2, the down of huge hub node will generate many isolate small hub nodes in Country B. But this situation is not found in Country C. The detailed analysis of the correlation between the shape of topology map and attack effectiveness is our future work. In particular, Tactics-2 for Scenario-1, we found that the hub node can be constructed more easily in Country B. However, the detailed reason is not clear. Also, we should derive the threshold of $\lambda_{max}(A)$ and R. In the previous works, the effectiveness is estimated from comparison with the initial value. However, it does not yet clarify how much the value increases from the initial value can contribute to the attack result. In addition, the evaluation method based on R and eigenvalue is necessary for Tactics-2 for Scenario-2. These are also our future work.

References

1. Komoriya K, Iwai K, Tanaka H, Kurokawa K (2015) Network attack strategy by topological analysis. In: The second international conference on information security and digital forensics
2. Gallos LK, Cohen R, Argyrakis P, Havlin S (2005) Stability and topology of scale-free networks under attack and strategies. Phys Rev Lett 94(18):188701.1–188701.4
3. Dall' Asta L, Alvarez-Hamelin L, Barrat A, Vazquez A, Vespignani A (2006) Exploring networks with traceroute-like probes: theory and simulations. Theoret Comput Sci 355:6–24
4. Bilo D, Guala L, Leucci S, Proietti G (2014) Network creation games with traceroute-based strategies. Lecture Notes in Computer Science, vol 8576, pp 210–223
5. Tomita Y, Nakao A (2009) Inferring an AS path from an incomplete traceroute. J Inst Electron Inf Commun Eng 109(273):17–22

6. Rojo O, Soto R (2005) The spectra of the adjacency matrix and Laplacian matrix for some balanced trees. Linear Algebra Appl 401(1–3):97–117
7. Wu CW (2005) On rayleigh-ritz ratios of a generalized laplacian matrix of directed graphs. Linear Algebra Appl 402(1–3):207–227
8. Taaffee EJ, Gauthier HL (1973) Geography of Transportation. Prentice-Hall, Upper Saddle River Ch. 5
9. Takeo D, Ito M, Suzuki H, Okazaki N, Watanabe A (2007) Proposal of a detection technique on stepping-stone at-tacks using, connection-based method. IPSJ J 48(2):644–655
10. Kisamori K, Shimoda A, Mori T, Goto S (2011) Analysis of malicious traffic based on TCP fingerprinting. IPSJ J 52(6):2009–2018
11. Yokota R, Okubo R, Sone N, Morii M (2013) The affect of the honeypot on the darknet observation, part 2, IE-ICE technical report, vol 2013-GN-88, No 16, pp 1–4
12. Albert R et al (2000) Error and attack tolerance of complex networks. Nature 406:378–382

Analysis of Slow Read DoS Attack and Communication Environment

Shunsuke Tayama(✉) and Hidema Tanaka

National Defense Academy of Japan, 1-10-20 Hashirimizu, Yokosuka,
Kanagawa 239-8686, Japan
{em55035,hidema}@nda.ac.jp

Abstract. Slow Read DoS attack is a technique which interferes Web server by exhausting resources. There are no effective countermeasures against from this attack nowadays. In this paper, we analyze Slow Read DoS attack, we found that the efficient attack can be realized when the bandwidth is over 500[Kbps]. In addition, we found that attacker can more effective attack by setting the connection rate to be equal to the process capability of Web server. At the same time, we can derive the secure setting of Web server against Slow Read DoS attack.

Keywords: Slow Read DoS attack · Slowhttptest · Bandwidth · Process capability

1 Introduction

Since 2015, the number of detections of DoS attacks worldwide exceeds 12,000 cases per week, and its method is also increasingly diversified and sophisticated [3]. In this paper, we focus on Slow Read DoS attack developed by Shekyan [9]. The attacker sends multiple requests to the Web server, and deliberately slows down the speed of reading the response with keeping many connections connected. As a result, by using up resources of Web server, it is impossible to connect from other clients [6, 7]. In addition, since an attacker transmits apparently normal request, it is necessary to monitor the transport layer in order to detect packet for attack. As a prominent case, there is a speculation for which Slow Read DoS Attack was used by Anonymous [1] to attack the site of JASRAC (Japanese Society for Rights of Authors, Composers and Publishers) for a protest against copyright protection in September, 2012 [5]. Currently, various studies [13–15, 17, 18, 20] have been reported, but no way to completely protect Slow Read DoS attack has been established. ModSecurity, which is one of WAF (Web Application Firewall) [4], has been implemented in many Web servers, however, Park has proven that it is vulnerable to Slow Read DoS attack [16].

In our previous study [19], in order to analyze the relationship between communication environment and effectiveness of Slow Read DoS attack, we attack in the virtual environment while changing the attacker's bandwidth and RTT. As the result, we conclude that RTT does not affect the effectiveness of attack, while bandwidth affects the speed of generating the connection for successful attack. In this paper, we have two

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_38

purposes. The first one is to analyze the effectiveness of bandwidth. And, the second one is to analyze the effectiveness of server setting, especially in timeout of Web server.

2 Slow Read DoS Attack

Slow Read DoS attack is a technique that an attacker occupies the maximum number of concurrent connections of Web server and interferes with connections from other clients (Fig. 1). The attacker sends legitimate requests and receives data after TCP 3-way handshake. Then, by setting the window size to be small when returning ACK, the attacker decreases the data size from Web server and slow down the response speed. When the window size is extremely set to 0, the Web server stops sending data and keeps connections. By connecting such connections up to the maximum number of concurrent connections of Web server, the service becomes disabled state.

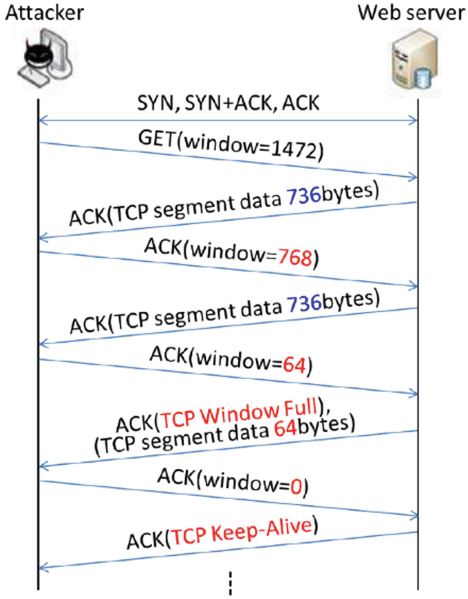


Fig. 1. Outline of Slow Read DoS attack

Since Slow Read DoS attack is different from traditional DoS attack strategy, monitoring of the transport layer by security vendors is an only effective defense method, but it requires higher cost. Therefore, it is necessary to establish a new counter measure.

3 Experiment

3.1 Purpose

Slow Read DoS attack is succeeded when the number of connection kept from the attacker (P_{max}) achieves the maximum number of concurrent connections of Web server (MC). Then, the condition of successful attack is $P_{max} \geq MC$. On the contrary, when $P_{max} < MC$, the attack fails.

We consider the condition that satisfies successful attack changing environment of network. In this experiment, we set environment using bandwidth X, timeout T and MC. These parameters are not open and impossible to manipulate for attacker. On the other hand, the attacker can optimize the connection rate R. Therefore, we search for the best connection rate R_0 under the environment with X, T and MC.

3.2 Experiment Environment

As experiment environment, we use a virtual environment (Table 1). We set the target Web server using Apache which is most popular one [2, 12]. The attacker and Web server are connected by one virtual switch (Fig. 2).

Table 1. Experiment environment

Parameter	Version or value
Host OS	Windows 10 home
Application for virtual environment	VMware workstation 12.1.0 player [11]
Guest OS	CentOS 6.7
Memory capacity of guest	1 GB
Program for attack	Slowhttptest-1.6 [10]
Application for web server	Apache (httpd-2.2.15) with 100 KB page

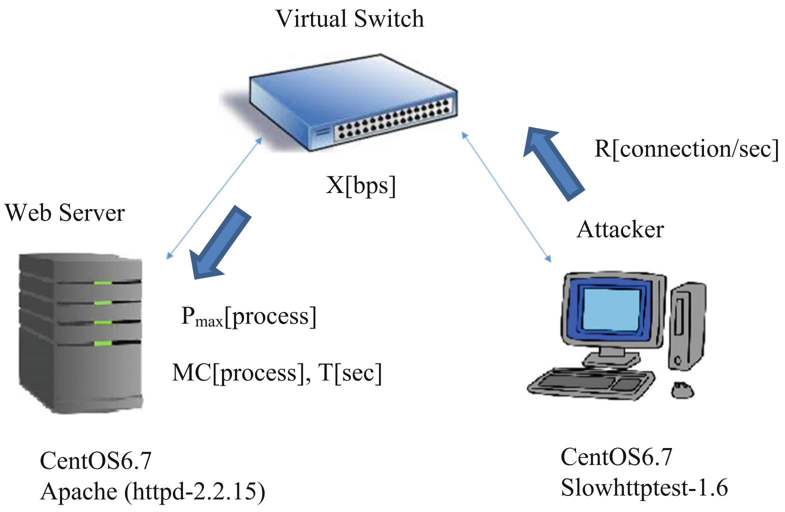


Fig. 2. Virtual environment

Table 2 shows the parameters of the Web server (httpd.conf). "Directive" controls the connection with the client, "prefork MPM" controls the operation of the process.

Table 2. Parameter of http.conf

	Parameter	Value
Directive	Timeout	T
	KeepAlive	Off
prefork MPM	StartServer	8
	MinSpareServer	5
	MaxSpareServer	20
	ServerLimit	1200
	MaxClients	1200
	MaxRequestChild	4000

Table 3 shows the parameters of the slowhttptest tool [10]. In this experiment, T (Table 2) and R (Table 3) denote variables. Timeout T is the time until the connection is forcibly disconnected, and connection rate R is the number of connections generated by attacker per second. In addition, the value of X is given to the communication line of attacker by virtual switch.

Table 3. Parameter of slowhttptest

Parameter	Value
Number of attack connections	1200
Connection rate	R
Window size	0
Pipeline factor	1
Read rate from receive buffer	5
Timeout for prove connection	10

3.3 Setting of Parameters and Variables

For the value of T, the default setting of Apache is T = 60[sec]. In addition, we set 30[sec] and 90[sec] for comparison. For the value of X, we set 50[Kbps], 100[Kbps], 200[Kbps], 300[Kbps], 400[Kbps], 500[Kbps], 1[Mbps], 10[Mbps], 100[Mbps] and 1[Gbps]. From the specification of VMware settings, we set connection rate $0 < R \leq 150$.

4 Experimental Result

Due to the limited space, in the followings, we show only the results of 100[Kbps], 500[Kbps], 1[Mbps] and 10[Mbps].

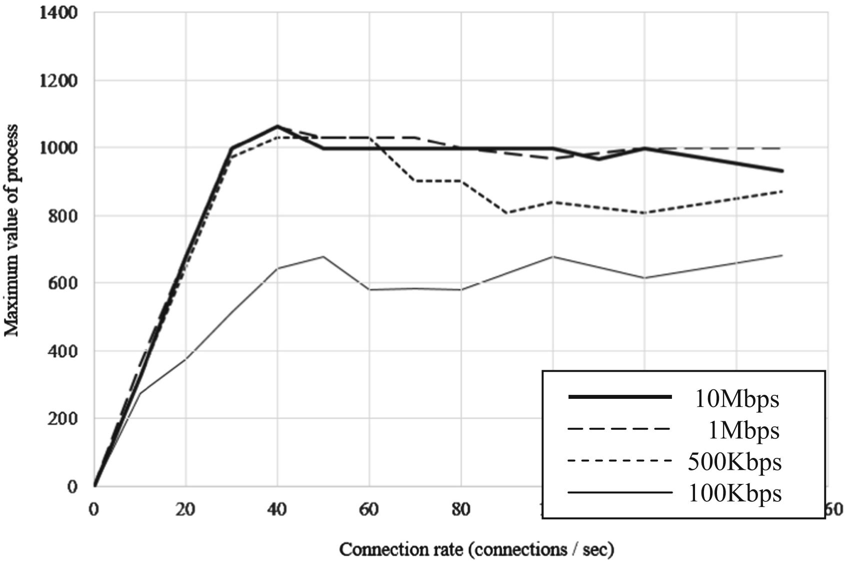


Fig. 3. Experiment 1 (Timeout 30)

4.1 Timeout 30[Sec]

Figure 3 shows the results of T = 30[sec] (Experiment 1). Comparing for each bandwidth, we can find that P_{max} of X = 100[Kbps] is obviously smaller than other. And, when X = 100[Kbps], the attack fails because P_{max} achieves only 600 ~ 700 against MC = 1200, and it is randomly with respect to the value of R. On the other hand, although attacks also fail when X ≥ 500[Kbps], P_{max} has constant tendency.

In the case of X ≥ 500[Kbps], each P_{max} takes the maximum value around R = 40. We also find that P_{max} of 500[Kbps] is smaller than one of 1[Mbps] and 10[Mbps], with R in the range greater than 60. Furthermore, in the case of X ≥ 500[Kbps], we can confirm that P_{max} decreases as the value of R increases. Therefore, we conclude that huge value of R will degrade the effectiveness of attack.

4.2 Timeout 60[Sec]

Figure 4 shows the results of T = 60[sec] (Experiment 2). Comparing for each bandwidth, we can confirm that the same characteristics in Experiment 1 are found. As new discovered features, P_{max} achieves 1200 which is the maximum value when X ≥ 500[Kbps]. From these results, we can find that P_{max} achieves 1200 at R = 20, and the maximum value is kept until R = 30 with X = 500[Kbps]. On the other hand, when X = 1[Mbps] and 10[Mbps], the maximum value is kept until R = 100, so we can confirm that the difference in the effectiveness of attack is caused by the bandwidth. In the case of X = 100[Kbps], P_{max} is larger than the value of Experiment 1, but P_{max} takes the maximum value at R = 30 and its value is at most 800 (attack fails).

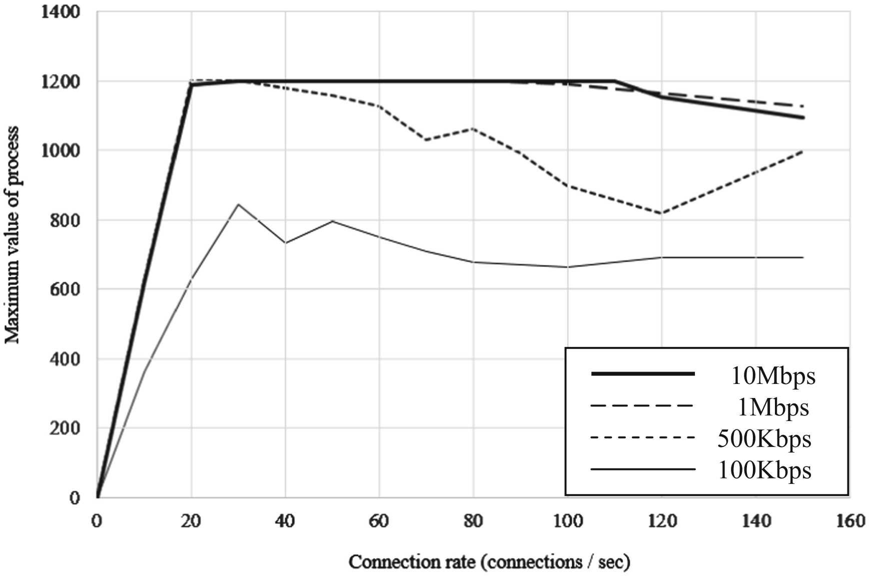


Fig. 4. Experiment 2 (Timeout 60)

4.3 Timeout 90[Sec]

Figure 5 shows the results of T = 90[sec] (Experiment 3). We can also see the same characteristics shown in Experiment 1 and 2. We can confirm that P_{max} = 1200 is kept until R = 40 with X = 500[Kbps]. On the other hand, since there is no significant change in the results when X = 1[Mbps] and 10[Mbps] through Experiments 1 to 3, we conclude that when the communication speed is more than or equal to 1[Mbps], the effectiveness of attack does not depend on timeout T.

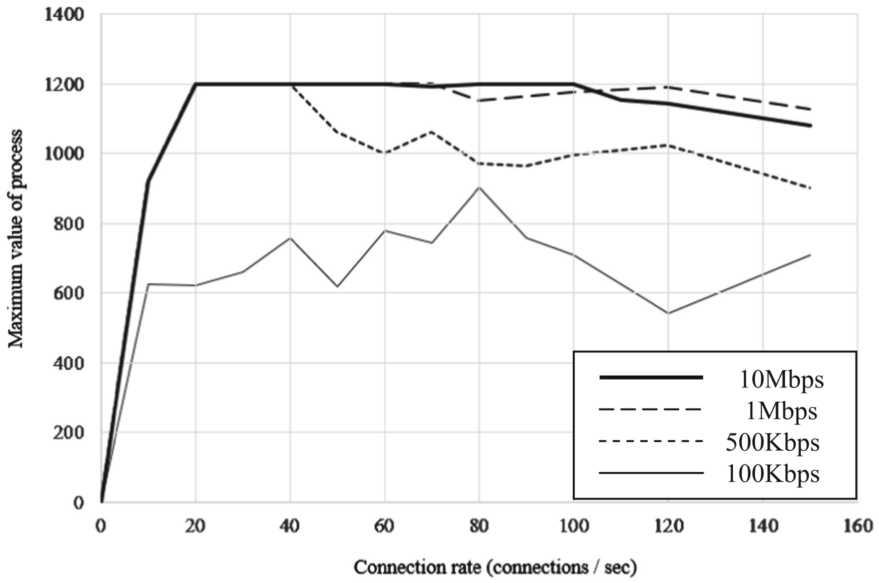


Fig. 5. Experiment 3 (Timeout 90)

5 Consideration

5.1 Effectiveness of Communication Environment for Attack

From the results of Experiments 1 to 3, we can summarize as follows.

(a) When X = 100[Kbps], P_{max} takes the smallest value among other bandwidth.
(b) When X ≥ 500[Kbps], the maximum value of P_{max} is generally obtained in the range of 20 ≤ R ≤ 40.
(c) When X ≥ 1[Mbps], the effectiveness of attack does not depend on T.
(d) Huge value of R will degrade the effectiveness of attack.

Experimental results of 200 ~ 400[Kbps], 100[Mbps], and 1[Gbps] are omitted for the limited space, however, we can find that P_{max} of X ≤ 400[Kbps] cannot succeed attack and their results have same tendency of 100[Kbps]. Therefore, from result (a) and (b), we can conclude that the necessary communication speed for attack is faster than

500[Kbps]. In addition, from result (c), if it is possible to hold $X \geq 1$[Mbps], stable effectiveness of attack can be expected. Our previous result [19] shows that the communication speed will affect the process capability of Web server because of its adaptive algorithm. Since there is relationship between result (d) and process capability of Web server, we show the detailed analysis in the next section.

5.2 Analysis of Result (D) and Condition for Optimum Attack

Let Q[process/sec] be process capability of Web server. When observing the data in Figs. 3, 4 and 5, we measured the number of process increments per second for each X and R, and the results are shown in Table 4. From the table, we can find that Q takes the maximum value of 36[process/sec] when $X \geq 1$[Mbps], and the Web server used in this experiment can generate processes of up to 36[process/sec]. Since this value of Q can be the upper bound, we consider that the results (b) and (d) are derived from this fact. As the result (d), huge value of R generates a large number of connections that are disconnected without being processed by Web server. Thereby, we conclude that huge value of R degrades effectiveness of attack. On the other hand, excluding $X = 100$ [Kbps], we consider that more effective attack can be performed by selecting the optimum connection rate R_0 for Q that depend on X. In order to derive R_0, we analyzed Experiment 1 ~ 3 in detail. From Fig. 3, P_{max} is maximized when $R = 40$. Also, from Figs. 4 and 5, we confirm that P_{max} is maximized in the range of $20 \leq R \leq 40$. Comparing these results with the process capability shown in Table 4, we can conclude that P_{max} is maximized when R takes the value close to Q. Therefore, we expect that the optimum R_0 will be equals to Q.

Table 4. Process capability of Web server

X [bps]	Q [process/sec]
100 K	20
500 K	32
1 M	36
10 M	36

5.3 Secure Setting of Web Server Against Slow Read DoS Attack

The unit [process/sec] denotes the number of communication requests received by Web server per unit time, and the unit [connection/sec] denotes one generated by attacker. So, both of them have same physical meaning, but they are from different view point each other. Therefore, in fact, we can see R[connection/sec] = Q[process/sec] in the followings. In our experiments, we measure P_{max} using *ps* command under condition of T, R and $MC = 1200$. The relationship among them are summarized in Table 5. From the result, we have following.

$$T \times R = P_{max}(Q \geq R) \text{ and } T \times Q = P_{max}(Q < R)$$

Table 5. P_{max} for each T and R when $X \geq 500$[Kbps]

T[sec]	30		60		90	
R[connection/sec]	≥ 30	20	≥ 20	10	≥ 20	10
P_{max}[process]	1000	600	1200	600	1200	900

If P_{max} < MC, the attack fails. Thus, we can derive the condition for secure Web server against Slow Read DoS attack as $T \times Q < MC$. From the above section, the optimum condition for attack is derived as $R_0 \approx Q$. Thus even if $Q < R_0$, the condition of $T \times Q < MC$ holds. Therefore, when the Web server is set $T \times Q < MC$, it is secure against Slow Read DoS attack. We confirmed the validity of this condition by additional experiments.

6 Conclusion

In this paper, we analyze Slow Read DoS attack in virtual environment and derive the optimum connection rate for the attacker's bandwidth (X) and Web server's timeout (T). From these results, we conclude that more effective attack can be executed by holding the communication speed above 500[Kbps] setting the connection rate (R) is equal to the process capability of Web server (Q). Using this result, we also derive the secure setting of Web server against Slow Read DoS attack, $T \times Q < MC$. We confirmed the validity of this condition by additional experiments. In this paper, we clarify the influence of the process capability of Web server on effectiveness of attack. However, it will depend on the specification and performance of Web server. Our future work is to analyze under other Web server and communication environment.

In our experiments, we set a single attacker using one attack PC, however, distributed attack schemes such as DDoS are mainstream [8]. Analysis of effectiveness of Distributed Slow Read DoS attack [16] under the real network environment and derivation of secure settings for Web server are also our future works.

References

1. Anonymous Operation Japan, twitter account "@OP-japan."
2. Apache: http://httpd.apache.org. Last Accessed 12 Feb 2017
3. Anstee D, Bowen P, Chui CF, Sockrider G (2016) Worldwide infrastructure security report, vol 11. ARBOR Networks
4. Muscat I (2017) How to mitigate slow HTTP DoS attacks in apache HTTP Server. acunetix. https://www.acunetix.com/blog/articles/slow-http-dos-attacks-mitigate-apache-http-server/. Last Accessed 12 Feb 2017
5. Japanese bulletin board, 2ch.net
6. Higgins KJ (2017) New denial-of-service attack cripples web server by reading slowly. InformationsWeek DarkReading. http://www.darkreading.com/attacks-breaches/new-denial-of-service-attack-cripples-we/232301367. Last Accessed 12 Feb 2017
7. Constantin L (2017) Researcher devises hard-to-detect DoS attack against HTTP servers. InfoWorld. http://www.infoworld.com/article/2618359/security-management/researcher-devises-hard-to-detect-dos-attack-against-http-servers.html. Last Accessed 12 Feb 2017

8. Zelasko M: DDoS attacks rose significantly in 2016.", COLOCROSSING Dec 21. 2016. https://blog.colocrossing.com/ddos-attacks-rising-2016/. Last Accessed 12 Feb 2017
9. Shekyan S (2017) Are you ready for slow reading?. https://blog.qualys.com/securitylabs/2012/01/05/slowread. Last Accessed 12 Feb 2017
10. Shekyan S (2017) Application layer DoS attack simulator. https://github.com/shekyan/slowhttptest. Last Accessed 12 Feb 2017
11. VMware. https://my.vmware.com/jp/web/vmware/details?productId=524&downloadGroup=WKST-1210-WIN. Last Accessed 12 Feb 2017
12. W3Techs: Most popular web servers. http://w3techs.com. Last Accessed 12 Feb 2017
13. Enrico C, Gianluca P, Giovanni C, Maurizio A (2013) Slow DoS attacks: definition and categorisation. Int. J. Trust Manage Comput Commun 1(3/4):300–319
14. Li JJ, Savor T (2014) Detecting DoS attacks on notification services. In: Software security and reliability-companion, pp 192–198
15. Park J, Iwai K, Tanaka H, Kurokawa T (2014) Analysis of slow read DoS attack. In: Information Theory and its Applications, pp 60–64
16. Park Junhan, Iwai Keisuke, Tanaka Hidema, Kurokawa Takakazu (2015) Analysis of slow read DoS attack and countermeasures on web servers. Int J Cyber-Secur Digital Forensics (IJCSDF) 4(2):339–353
17. Tripathi N, Hubballi N, Singh Y (2016) How secure are web servers? An empirical study of slow HTTP DoS attacks and detection. In: Availability, reliability and security, pp. 454–463
18. Oshima S, Nakashima T, Sueyoshi T (2010) Early DoS/DDoS detection method using short-term statistics. In: Complex, intelligent and software intensive systems, pp. 168–174
19. Tayama S, Tanaka H (2016) A study on the relationship between communication environment and effectiveness of slow read DoS attack. In: Proceeding of the computer security symposium, vol 2016(2), pp 749–755
20. Hirakawa T, Ogura K, Bista BB, Taketa T A defense method against distributed slow HTTP DoS attack. In: Network-based information system, pp 152–158 (2016)

A SFC Network Management System in SDN

Li-Der Chou(✉), Chia-Wei Tseng, Hsin-Yao Chou, and Yao-Tsung Yang

Department of Computer Science and Information Engineering,
National Central University, Taoyuan, Taiwan
cld@csie.ncu.edu.tw, cwtseng@g.ncu.edu.tw,
qooqoonick@networklab.csie.ncu.edu.tw, yao.vct@gmail.com

Abstract. With the rapid growth of the Internet, more demand for network resources allocation and management of value-added services are appeared. The new concept of network service architecture, Service Function Chain (SFC), has been proposed. The SFC is a capability that uses Software-defined networking (SDN) technology to create a service chain of connected network services and connect them in a virtual chain. This paper designed a VLAN and OpenFlow based Service Function Chain (VOFSFC) system to improve the flexibility of the network service deployment and increase the utilization of service functions.

Keywords: SDN · NFV · SFC · Network management

1 Introduction

With the rapid growth of the Internet, the network communication architecture becomes relative complicated and large. Traditional network infrastructure combines complex network functions with routers and switches for support more network services. This make the network becomes more complicated and decrease the performance of network severely. The purpose of SDN is to change the physical, complex network into a virtual, programmable and open network architecture. The SDN architecture decouples the network control plane from the data plane, and control the network by SDN controller. The OpenFlow protocol is a foundational element for building SDN solutions [1, 2].

As the Internet becomes more and more sophisticated, different traffic needs to apply different network service functions. Such as the video streaming service needs to apply the firewall, Video Optimization Server, etc. To satisfy the demands on network design and deployment as well as service application, Internet Engineering Task Force (IETF) is developing a new technology called Service Function Chain [3]. The main purpose of SFC is to lead the traffic flow of users to their required services according to users' demands. The SFC approach promises high flexibility and lower costs for network operator at the same time. The paper proposed a VLAN and OpenFlow based Service Function Chain (VOFSFC) system in the SDN environment. The VOFSFC system can identify different requirements of network traffic, steer/deliver packets to service functions and record information of the traffic by the VLAN tag. The VOFSFC system

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_39

provides ability to manage the service function paths, and makes creation, deletion and modification of service function to be easier.

The main idea of this research is adopting the VLAN ID tag to identify different service function paths, performed services chain and services decision-making positions which reduce the load on these service functions in the chain. Our major contributions are summarized as follows:

- We proposed a VOFSFC system to manage service function paths in the SDN environment.
- We implemented a SDN experiment environment for development and validation of SFC deployment and experiments.

The remainder of this paper is organized as follows: in Sect. 2, the background and related works are addressed. Section 3 is to describe the design and implement of the VOFSFC system. Section 4 presents the experimental results. The last section concludes this paper and addresses potential future works.

2 Background and Related Works

SDN attracted much attention these years. Many researchers are working on this research area [4–6]. SDN architecture decouples the network control plane from the data plane, so that the network administrator can manage the network by the programmable operators [7, 8]. The OpenFlow protocol is the first SDN language to be productized and implemented [9]. The OpenFlow Controller is the brains of the SDN network that can manage the Flow table on the OpenFlow switch by the OpenFlow protocol. Flow table is composed of many flow-entry, and the flow-entry contains three parts, Rule, Action and Stats. Rule is used to define and identify a flow; Action defines the behavior of packets which was matched by the Flow table; Stats gathers the flow information statistic. OpenFlow controller as the control center can be more intelligent, automated and efficient to manage network. There are many open-source project of OpenFlow controller for network operators and researchers, like Ryu [10], Floodlight [11], OpenDaylight [12] and ONOS [13]. Ryu is an open-sourced Network Operating System (NOS) developed and maintained by NTT (Nippon Telegraph and Telephone) in Japan. Ryu provides software components with well-defined API that make it easy for developers to create new network management and control applications. Therefore this research will use the Ryu to develop the VOSFC system.

SFC is a kind of network technology currently under research and development by IETF which is formulated by the work groups of IETF Service Function Chain Working Group (SFC WG) and Source Packet Routing in Network Working Group (SPRING WG) [14]. SFC is being include in many SDN and network functions virtualization (NFV) use cases and deployments, including data centers and carrier networks [15, 16]. Many research have propose various implementation and application for SFC. Network Service Header (NSH) [17] is an IETF draft pro-posed to address the Service Function Chaining based on the SFC encapsulation to support the SFC architecture as outlined in the RFC7665 [18]. In the framework of NGSON (Next-Generation Service Overlay

Network), Service Overlay Network (SON) is a new network service framework which is formed by the combination of original service-oriented architecture and service delivery platform to realize service functions [19–21]. described a use case of SFC, the authors proposed a security service on-demand system that offers a security service function chain that enables ICT service providers to provision a dynamic and flexible secure service on the SDN network. To achieve the service function deployment, [22] proposed a greedy algorithm for service function placement. It consider the order of service function to decision the place of service function.

This paper addressed a SFC management system to manage service function paths in the SDN environment. In order to realize the SFC deployment in our experiment, the NFV and SDN technologies are both taken into consideration in the proposed VOFSFC system.

3 VLAN and OpenFlow Based Service Function Chain System

Figure 1 shows the system architecture of the VOFSFC system. The system is divided into three parts, the first part is OpenFlow Switch (Pica8 Pronto Open Switch) that act as the service function forwarder role in the network. The second part is the VOFSFC

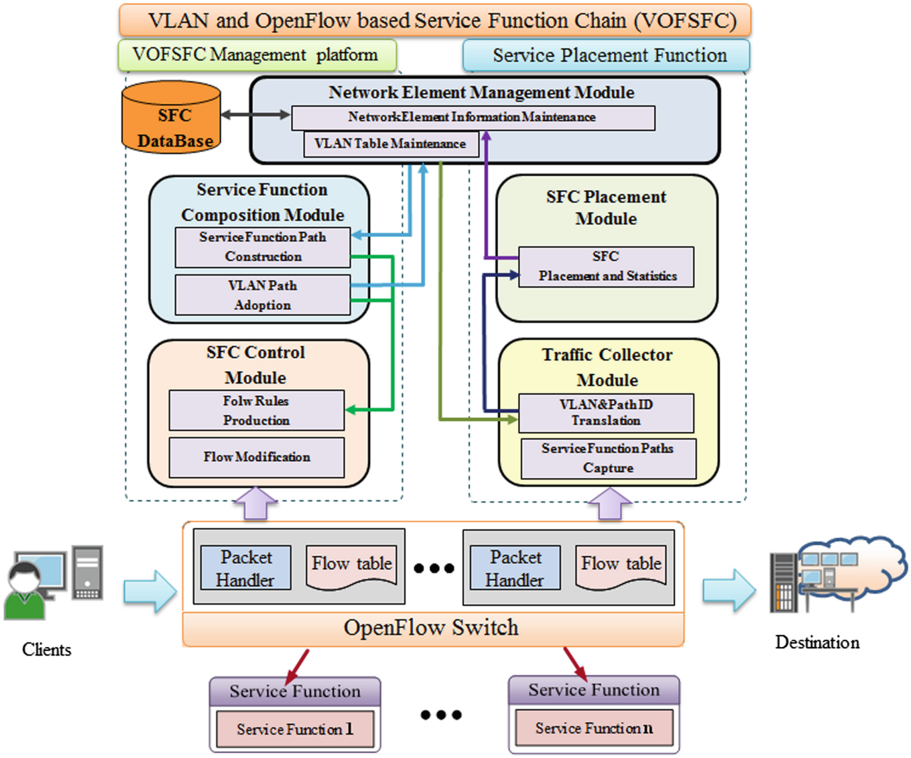


Fig. 1. VOFSFC system architecture

management platform. The third part is the service placement function. There are five different modules in the VOFSFC system.

3.1 Network Element Management Module

The Network Element Management Module provides the network element information maintain function and VLAN table maintain function in the VOFSFC system. The network element information maintenance function will regularly access and confirm the *SF_Infor* and the *Dest_Server_Infor* information in the SFC Database. These information will transmit to the service function path construction function and the VLAN&Path ID function to establish the service function paths. The VLAN table maintenance function will communicate with the service function path construction function and the VLAN path adoption function to maintain the VLAN table.

3.2 Service Function Composition Module

The Service Function Composition Module provides the service function path construction function and VLAN path adoption function in the VOFSFC system. The service function path construction function will get the service function paths information from the SFC Database and combined with the information obtained by the Network Element Information Maintenance Module to construct the path of the service function to the destination. And request the VLAN Path Adoption function to provide the VLAN_ID for the service function path. The VLAN path adoption function can receive the SF_Path and SF_List from the service function path construction function. In accordance with the desired SFC path and the demand service functions, calculated the Path_ID for each SFC. The VLAN path adoption function will use this Path_ID to register with the VLAN Table Maintenance function. When the VLAN table maintenance function received queries from the VLAN path adoption function, it will check the VLAN table that the Path_ID is registered in the SFC database or not. If the system have not register before, the system will assign a VLAN_ID to this service function path, and use the Path_ID to register the VLAN table. The VLAN table information will return to the VLAN Path Adoption function and update the SFC Database.

3.3 Service Function Chain Control Module

The Service Function Chain Control Module provides flow rules production function and flow modification function in the VOFSFC system. After receiving the information service function path from Service Function Composition Module. The flow rules production function will divide into three cases. One for the packet into the VOFSFC system, add the VLAN_ID, turn to the next service function path; the second from the service function server back to the system packet, through the packet ID on the packet identification, modify the VLAN_ID to send To the next service function; Finally, the service flow will untag the VLAN through the OpenFlow switch and forward to the destination server.

A flow modification function is use to modify the flow rules in the OpenFlow Switch. After receiving the information from the flow rules production function, the flow modification function will add OFPT_FLOW_MOD message by OFPFC_ADD command and remove the old OpenFlow rule through OFPFC_DELETE. The OpenFlow rule will be configured in the OpenFlow switch. If the Flow table receives the OpenFlow rule from the Flow Modification function, it performs the modification of the Flow table, adds the new flows to the Flow table, and removes the old not applicable flow.

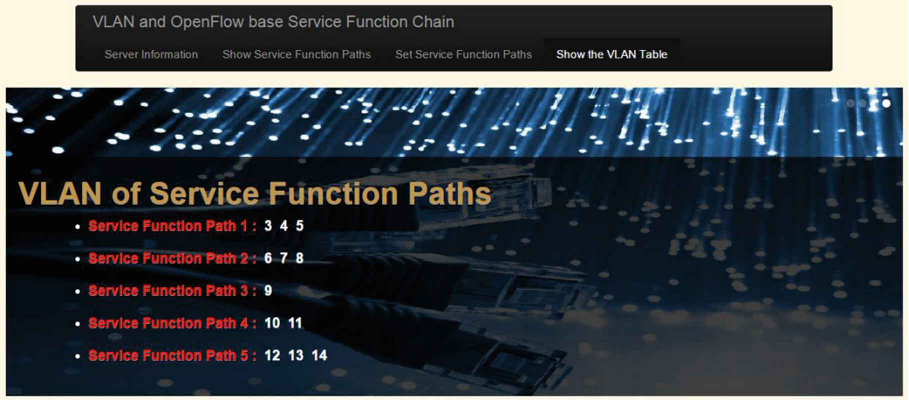
3.4 Traffic Collector Module

The Traffic Collector Module provides the VLAN&Path ID Translation function and service function paths capture function in the VOFSFC system. When the packet enter the OpenFlow switch, the Packet Handler on the OpenFlow switch will compare the OpenFlow rule with the VOFSFC system and tag the VLAN_ID forward to the Service Function Placement Module. When the service function paths capture function received the packets from OpenFlow switch, it will calculate the *statistic_VLAN* and *statistic_Packet* information and send back to the OpenFlow switch. After received this packet, the OpenFlow switch will forward it to the next service function based on the VLAN_ID information. The VLAN&Path ID Translation function will regularly receive the *statistic_VLAN* and *statistic_Packet* information from service function paths capture function. Combined with the *SF_Infor* and the *Dest_Server_Infor* provided by the network element information maintenance function, the VLAN&Path ID Translation function can identify the relationship between service function and Path_ID. Afterwards, the VLAN&Path ID Translation function will generate the *Statistic_Loading* information to the path statistic and prediction function.

3.5 SFC Placement Module

The Service Function Placement Module provides the service function placement and statistics functions in the VOFSFC system. After receiving the *Statistic_Loading* information from Traffic Collector Module, the service function placement and statistics functions will count the number of each SFC paths in the system and evaluate the available service functions when new service flows arrive.

In our design, the VOFSFC system provides a Web GUI as user interface. Figure 2 shows the designed VOFSFC system. The graphical user interface allows users to interact with our system where the user can designate the service function path in the experiments or find additional information about SFC status.


VLAN and OpenFlow base Service Function Chain
Server Information
Show Service Function Paths
Set Service Function Paths
Show the VLAN Table

VLAN of Service Function Paths

- Service Function Path 1 : 3 4 5
- Service Function Path 2 : 6 7 8
- Service Function Path 3 : 9
- Service Function Path 4 : 10 11
- Service Function Path 5 : 12 13 14

Fig. 2. VOFSFC system's user interface

4 Experiment

The experiment environment is shown in Fig. 3. The OpenFlow Controller is the central manager for this system. The Pica8 OpenFlow switch is act as the service function forwarder in our design. There are two All-in-one Hypervisors worked as the SFC servers to provide service functions. Table 1 summarizes our experimental environment.

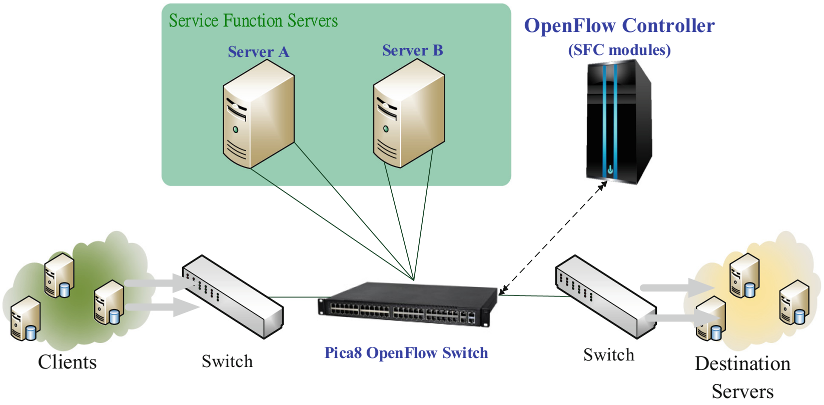


Fig. 3. Experiment environment

When the system is initialized, the OpenFlow controller adds the flow rules of the service function path to the Flow table of the OpenFlow switch. When the client's traffic flow enter the system, the OpenFlow switch will compare the Flow table in the OpenFlow switch and forward to the service function servers based on the flow rules. Then the Flow will start designated SFC routing, since VOFSFC system has determined the corresponding SFC routing to the decision, every corresponding service function can lead flows to next service function after its processing to complete required service chains. After the Flow complete all demand service chains, OpenFlow switch will lead Flow to the destination according its original destination location.

Table 1. Hardware and software resource

Hardware/software	SFC server	OpenFlow controller (Ryu 3.6)
Manufacturer	ASUSTeK computer INC.	Dell INC.
CPU	Intel(R) Core(TM) i3-3220 CPU @ 3.30 GHz	Intel(R) Core(TM) i3-4130 CPU @ 3.40 GHz
Memory	DDR-3 1600 4 GB	DDR-3 1600 4 GB
Network interface	TP-LINK TG-3468 Gigabit PCI Express * 2	TP-LINK TG-3468 Gigabit PCI Express
Operating system	Ubuntu 12.04.1 LTS	Ubuntu 12.04.4 LTS
OS Kernel	3.2.0-29-generic-pae	3.11.0-15-generic

To evaluate the functionality of the proposed VOFSFC system, the experiment designed five different SFC for testing. Table 2 shows the details of different service function path. There are 8 different service functions in this experiment. The existing service functions such as an Service Placement Function (SPF), Monitoring (Mon), Firewall (FW), intrusion detection system (IDS), Web Optimizer (Web), DPI, Video Optimizer (Video) and NAT. In order to achieve the service function placement and traffic statistics, SPF function is necessary for each SFC in our system.

Table 2. Service function path table

SFP	SFs	Path ID
SF_Path_1	SPF, FW, IDS, NAT	23
SF_Path_2	SPF, Mon, Web, NAT	147
SF_Path_3	SPF, FW, DPI, Video	141
SF_Path_4	SPF, FW, IDS, DPI, Video, NAT	237
SF_Path_5	SPF, Mon, FW, Web	101

In the experiment, the VOFSFC system utilize the service function placement and statistics functions to monitor the status of traffic flow through different SFCs. Figure 4 shows the experiment result between the SF_Path1 and SF_Path2. The combination of different service functions is the main factor affects the experiment result. The utilization of the FW and IDS is higher than the Mon and Web service function. Figure 5 shows the experiment result among the SF_Path3, SF_Path4 and SF_Path5. As shown in Fig. 5, the traffic statistics in the SF_Path4 is higher than SF_Path3 and SF_Path5. The result indicates that if a new traffic flow is entered into the system, the SF_Path3 is the better choice than other paths.

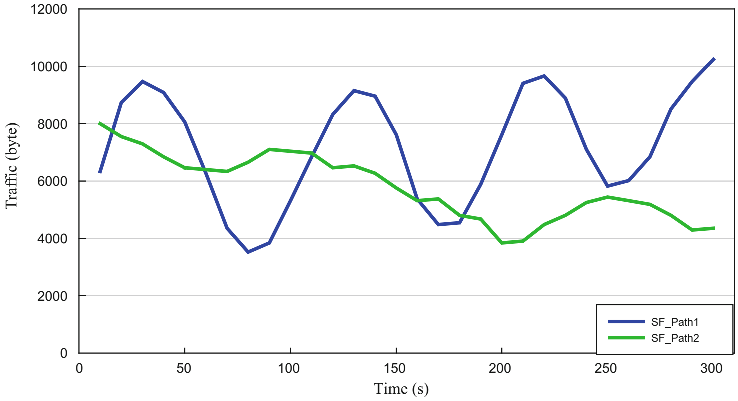


Fig. 4. Traffic statistics of *SF_Path$_1$* and *SF_Path$_2$*

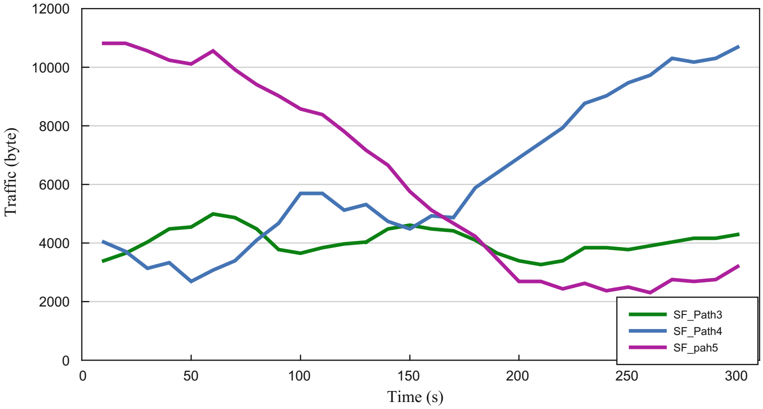


Fig. 5. Traffic statistics of *SF_Path$_3$*, *SF_Path$_4$* and *SF_Path$_5$*

5 Conclusion and Future Works

This paper proposes a VLAN based Service Function Chain management system in SDN network. The VOFSFC system can deploy service functions in the SDN network environment. The system achieves the redirection of network flows and the addition of the needed service functions. The SFC can make a flexible order of the comprehensive services, also practicing SDN/NFV will bring benefits on the operation and deployment of overall service resources.

Our future work will involve more functions and analyses of the VOFSFC system. The impact of the VNF movement technology over SDN and study the capacity influenced by different network architecture will be also taken into consideration.

Acknowledgment. The work described in this paper was supported in part by Ministry of Science and Technology of the Republic of China. (Project No. 104-2221-E-008-039-MY3 and 105-2221-E-008-071-MY3)

References

1. Mcdysan D (2013) Software defined networking opportunities for transport. IEEE Commun Mag 51(3):28–31
2. OpenFlow. http://archive.openflow.org/
3. Service Function Chaining (sfc). https://datatracker.ietf.org/wg/sfc/documents/
4. Chou L-D, Tseng C-W, Lai P-H, Hsieh S-Y, Wu M-C (2016) SDN/NFV virtualization testbed with automatic deployment and management functions. In: Proceedings of The second international conference on electronics and software science (ICESS2016), Takamatsu, Japan, November, pp. 112–121
5. Sieber C, Basta A, Blenk A, Kellerer W (2016) Online resource mapping for SDN network hypervisors using machine learning. In: IEEE NetSoft conference and workshops (NetSoft), Seoul, Korea, pp. 78–82, June 2016
6. Nikbazm R, Dashtbani M, Ahmadi M (2015) Enabling SDN on a special deployment of OpenStack. In: 2015 5th international conference on computer and knowledge engineering (ICCKE), pp. 337–342
7. Casado M, Foster N, Guha A (2014) Abstractions for software-defined networks. Commun ACM 57(10):86–95
8. ONF, "SDN architecture," Open Networking Foundation, Technical report, June 2014. https://www.opennetworking.org/images/stories/downloads/sdn-resources/technical-reports/TRSDNARCH1.006062014.pdf
9. McKeown N, Anderson T, Balakrishnan H, Parulkar G, Peterson L, Rexford J, Shenker S, Turner J (2008) OpenFlow. SIGCOMM Comput Commun Rev 38(2):69
10. Osrg.github.io, Ryu SDN Framework (2015). http://osrg.github.io/ryu/
11. Project Floodlight (2015): Floodlight OpenFlow Controller. http://www.projectfloodlight.org/floodlight/
12. Opendaylight.org (2015) OpenDaylight | A Linux Foundation Collaborative Project. http://www.opendaylight.org/
13. Onosproject.org (2015) Open Network Operating System. http://onosproject.org/
14. Source Packet Routing in Networking (spring). https://datatracker.ietf.org/wg/spring/documents/
15. King D, Farrel A, Georgalas N (2015) The role of SDN and NFV for flexible optical networks: current status, challenges and opportunities. In: 2015 17th international conference on transparent optical networks (ICTON), pp 1–6
16. Akhtar N, Matta I, Wang Y (2016) Managing NFV using SDN and control theory. In: NOMS 2016 - 2016 IEEE/IFIP network operations and management symposium, pp 1005–1006
17. Network Service Header. https://tools.ietf.org/html/draft-ietf-sfc-nsh-11
18. Service Function Chaining (SFC) Architecture. https://tools.ietf.org/html/rfc7665
19. Duan Z, Zhang Z, Hou YT (2003) Service overlay networks: SLAs, QoS, and bandwidth provisioning. IEEE/ACM Trans Net 11:870–883 Erl T (2007) SOA: Principles of Service Design, 1st edn., Prentice Hall
20. Pavlovski C (2007) Service delivery platforms in practice. IEEE Commun Mag 45(3):114–121

21. Chou L-D, Tseng C-W, Huang Y-K, Chen K-C, Ou T-F, Yen C-K (2016) A security service on-demand architecture in SDN. In: The 7th international conference on information and communication technology convergence (ICTC 2016), Jeju Island, Korea, October
22. Quinn P, Guichard J (2014) Service function chaining: creating a service plane via network service headers. Computer 47(11):38–44

A Performance Comparison of Deterministic and Adaptive Routing for an x-Folded TM Topology

Mehrnaz Moudi[1](✉), Mohamed Othman[1,2](✉), Amir Rizaan Abdul Rahiman[1], and Kweh Yeah Lun[1]

[1] Department of Communication Technology and Network, Universiti Putra Malaysia, UPM,
43400 Serdang, Selangor D.E., Malaysia
mehrnazmoudi@gmail.com, mothman@upm.edu.my

[2] Computational Science and Mathematical Physics Lab., Institute for Mathematical Research,
Universiti Putra Malaysia, Serdang, Malaysia

Abstract. This article focuses on a novel topology in interconnection networks which was called x-Folded TM. We explored the dynamic communication performance of this topology under various traffic patterns by designing new scheme for routing algorithms. In this scheme, the topology is partitioned to make the routing without deadlock in x-Folded TM topology. Using deterministic and adaptive routing, we have evaluated the performance of x-Folded TM topology in compare with Torus. A comparison of results reveals that the average delay grows steadily with the increase of injection rate until reaching saturation, which is the smallest rate of traffic for channel saturation in the network. Therefore, the achievements introduce x-Folded TM topology as efficient one in interconnection networks.

Keywords: x-Folded TM topology · Partitioning scheme · Deadlock · Deterministic · Adaptive · Delay · Throughput

1 Introduction

With the advance of the technology, smart interconnect architectures integrate the huge number of nodes or processing elements (PE) in the forms of processors available on a network. For a considerable period of time, the number of processors simply increases to get a considerable performance improvement. By adding more processors to the network, their communication performance is degraded in sending and receiving packets. Deadlock is an anomalous state among the network resources without a careful design for a network when a network congested. Deadlock problem plays a vital role in the communication performance and it can be solved using new topology in interconnection networks and congestion-aware routing scheme [1–5]. The proposed partitioning scheme in this article used to find the shortest path from a source to a destination without deadlock. We validate the contributions of this article by simulation with considering the routing algorithms under different traffic patterns. This validation provides a detailed comparative evaluation between the x-Folded TM and Torus topologies. The achievements show the superiority of the x-Folded TM topology through a

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_40

comprehensive study of different two algorithms to minimize the network congestion as a severe problem in interconnection networks. In interconnection networks, the topology and the employed routing within it influence on the network performance.

2 x-Folded TM Topology

The evolutionary advances in interconnection networks are suitable and improve the majority of network performance. In previous studies, many interconnect topologies have been performed while solved the dependency of the network performance to the network topology properties. One of the major topologies is k-ary n-cube. They are very popular topologies because of their desirable properties such as low diameter, small node degree, symmetry plus a straightforward implementation [6, 7]. The k-ary n-cube topologies are introduced with $N = k^n$ nodes where n is the number of dimensions and k is the number of nodes per dimension in the interconnection networks. In this research, k-ary n-cube topologies are utilized for the scope. Among most popular topologies in k-ary n-cube, Torus (Definition 1) is the most popular due to provide low-latency and high-bandwidth in the network communication.

Definition 1. A Torus contains $2n$ additional wraparound links, connecting the leftmost node to the rightmost node in the same row or connecting the uppermost node to the lowermost node in the same column. That is, two nodes $(x, 0)$ and $(x, k - 1)$ are connected by a wraparound link for $0 \leq x \leq (k - 1)$, and two nodes $(0, y)$ and $(k - 1, y)$ are connected by a wraparound link for $0 \leq y \leq (k - 1)$.

Since, the key element in the interconnection network performance is having an attractive topology; x-Folded TM topology has been introduced as a novel topology in k-ary n-cube. In [8], x-Folded TM is derived from a TM topology by folding the topology based on the imaginary x-axis, removing several links and sharing several nodes. x-Folded TM is defined according to Definition 2.

Definition 2. In x-Folded TM topology, node (x, y) is a valid node if $0 \leq x \leq (k - 1)$ and $0 \leq y \leq (k - 1)$. Along x-axis, the nodes connecting to node (x, y) are: $(x + 1, y)$ if $x < (k - 1)$ and $(x - 1, y)$ if $x > 0$. Along y-axis, nodes $(x, y + 1)$ if $y < (k - 1)$ and $(x, y - 1)$ if $y > 0$ are connected to this node. Then, node (x, y) is removed from the x-Folded TM if $(x + y) \bmod k = 0$ or 1 or… or $(k - 3)$, where $x > y$ and $(k - 3) \leq x \leq (k - 1)$ and $1 \leq y \leq (k - 2)$. In addition, there is no link between two nodes (x, y) and $(x + 1, y)$ if $x = i$ and $y = i + 1$, when k is even and $i = \left(\frac{k}{2}\right) - 1$.

In Fig. 1, x-Folded TM topology has been illustrated with size 8 × 8. Followed by folding, some of the nodes are shared in the x-Folded TM topology and have been displayed with purple color. The color code for all nodes is applied in this topology to show the deadlock-free partitions in x-Folded TM topology. Since we defined the color codes, all nodes in x-Folded TM topology were found in three different color triangles; purple, blue and red. The concept of the shared nodes is applied in x-Folded TM topology when we concern the routing in this topology, which is presented with details in Sect. 3.

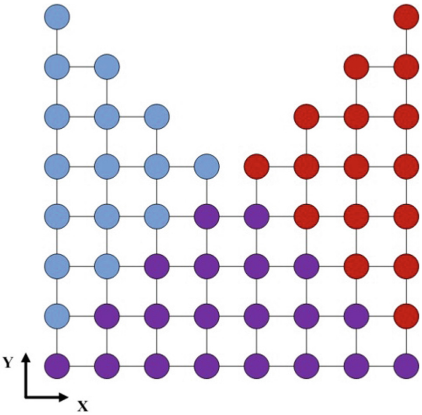


Fig. 1. x-folded TM topology with size 8 × 8

In [8], x-Folded TM has been introduced and simplified to prove it is a topology with low diameter and average distance which influence on the network performance. The main properties for the x-Folded TM topologies are diameter, degree, cost, number of links and average distance. The most important among these properties are diameter and average distance, which are maximum and average of the minimum distances between any two nodes in this topology. Table 1 presents the various topological properties of Torus and x-Folded TM topologies where n is the number of dimensions and k is the number of nodes per dimension in each topology.

Table 1. Topological properties

Topology	Torus	x-Folded TM
Number of links	$2k^2$	$2k^2-2k$
Degree	$2n$	$n-1, n, n+1, 2n$
Diameter	k	$k-1$
Average distance (AD_k)	$\frac{k}{2}$	$\frac{AD_{k-1}+k^2-k}{k^2-1}$
Cost (*degree* × *diameter*)	$(2n \times k)$	$(n-1 \times k-1)$ or $(n \times k-1)$ or $(n-1 \times k-1)$ or $(2n \times k-1)$

3 Routing Algorithm

One of the most important challenges in these networks is how to apply the routing algorithms between the processors connection without limiting the network performance. Recently, many approaches have been used to improve the performance of the interconnection networks. A significant step to achieve high network performance in the interconnection networks is having efficient communication. Since the packet delivery rate affects on the dynamic communication performance, the routing algorithm is used to find the shortest path without deadlock between each source and destination [9–11].

To solve deadlock problem, new partitioning scheme is developed to improve the topological performance. The new scheme is to partition x-Folded TM topology into some specific directions and then to impose the routing algorithm to traverse the packets without deadlock. The topology is split into three deadlock free partitions (1) X + Y + or X−Y−, (2) X + Y−, and (3) X−Y +. Following these partitions can be assigned into purple, blue, and red triangles in x-Folded TM topology. Algorithm 1 presents how to assign each partition to purple, blue and red triangles.

Algorithm 1. Topology Partitioning Scheme

```
Input: $(x_D, y_D)$                                  // It is the destination node coordinate.
if Destination node $\epsilon$ Purple Triangle then   // It is located at the purple triangle.
    if $0 \le x_D \le (\frac{k}{2}) - 1$ then              // $x_D \epsilon\ [0, (\frac{k}{2}) - 1]$
        Partition = $X+ Y+$ ;
    else
        Partition = $X- Y-$ ;
    end if
end if
if Destination node $\epsilon$ Blue Triangle then     // It is located at the blue triangle.
    Partition = $X+ Y-$ ;
else
    Destination node $\epsilon$ Red Triangle;         // It is located at the red nodes.
    Partition = $X- Y+$ ;
end if
Output: Assigned partitions; X+Y+, X-Y-, X+Y- and X-Y+
```

The applied routing algorithms in this article are deterministic and adaptive routing. Both algorithms are applied for the x-Folded TM topology. The deterministic routing scheme determines about packets direction in each dimension only for the routing between the source and destination pair. A successful route in each dimension when the packet arrives at the proper coordinates and the distance is zero, then the route will proceeds to the next dimension. [12, 13]. Second algorithm is adaptive routing algorithm which by adding adaptive features to the routing can improve the deterministic routing algorithm. An adaptive routing algorithm is more suitable for dynamic ne works to choose the routing paths. The chosen path is used if the links are not busy when the packets arrive. A number of pairs of nodes transmit packets simultaneously using alternate paths without blocking. The algorithm improves the performance potentially and its delay is near to deterministic algorithm, however it increases the cost of preventing deadlock which overwhelms the adaptive routing advantages [14, 15]. Accordingly, the partitioning scheme considers these two routing algorithms in x-Folded TM topology. Then, the x-Folded TM topology performance is compared with Torus under both routing algorithms for verification purpose in this paper.

4 Simulation Environment

Simulation is carried out by Booksim2.0 simulator [16] to evaluate fairly the performance of the x-Folded TM topology along with two routing algorithms. The current simulator comprises a collection of routers and channels with topologies which defines the interconnection between routers and channels. By using this simulator, our simulation is accompanied by a set of parameters which have been presented in Table 2.

Table 2. Simulation setup

Parameter	Value
Network size	8 × 8 ($k = 8$ and $n = 2$)
Virtual channels	2 VCs
Buffer depth	2 flits
Packet size	8 flits
No. of cycles	10000 cycles

The simulation parameters have significant performance implications to present the superiority of the x-Folded TM topology under different routing algorithms. In the following, the applied evaluation metrics are average delay and network throughput. They are highly dependent on the packet injected rate. The maximum amount of accepted information is introduced as network throughput (*Th*) for a particular traffic pattern. The elapsed time for traversing a packet from any source to any destination is presented as average network delay (*Di*). They measure in flits/node/cycle according to Eqs. (1) and (2).

$$Th = \frac{Total\,Recieved\,Flits}{Number\,of\,Nodes \times Total\,Cycles} \tag{1}$$

$$D = \frac{1}{N_p} . \sum\nolimits_{i=1}^{N_p} D_i \left(N_p\text{: } Total\ Number\ of\ Packets \right) \tag{2}$$

5 Performance Evaluation

For performance evaluation purpose, the deterministic and adaptive routing algorithms considering uniform, transpose and hotspot traffic patterns are used for router design in x-Folded TM topology. These routing algorithms are used to find a short and single path from source to destination between several paths. The simulation results for x-Folded TM topology compared with Torus are presented in this section.

Under uniform traffic pattern in Fig. 2, the average delay using deterministic and adaptive routing algorithms for Torus and x-Folded TM topologies are approximately similar. However x-Folded TM topology saturated earlier than Torus and its saturation point is 0.009 (flits/node/cycle). In Fig. 3, it is seen that the network throughput for Torus with deterministic routing is higher than that with using adaptive algorithm and it is better than x-Folded TM topology under both routing as well.

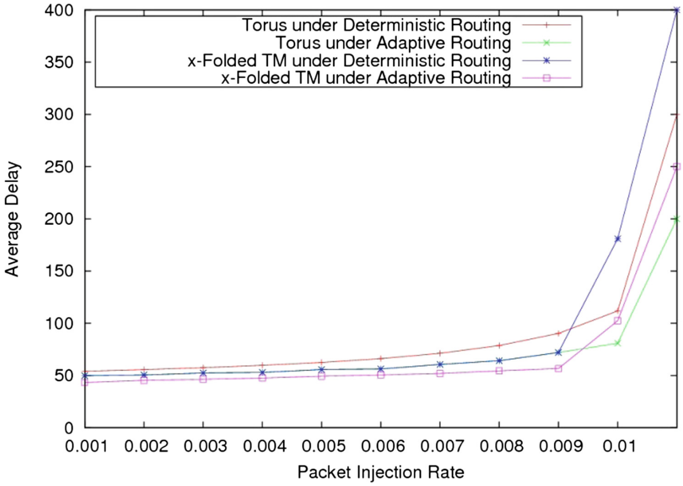


Fig. 2. Average delay curves uniform traffic pattern

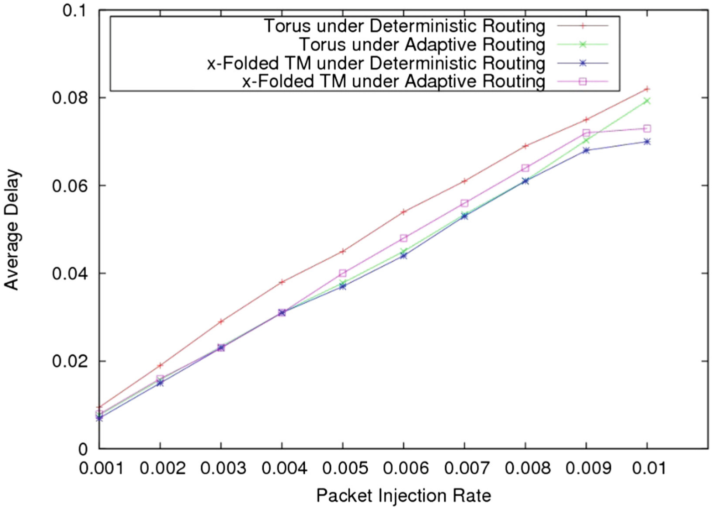


Fig. 3. Network throughput curves under uniform traffic pattern

Figure 4 depicts a considerable improvement for x-Folded TM topology using deterministic and adaptive routing with the increase of injection rate under transpose traffic pattern. Obviously, the network throughput for x-Folded TM is better than the Torus with both routing in Fig. 5, as there is a direct trade-off between network throughput and average delay.

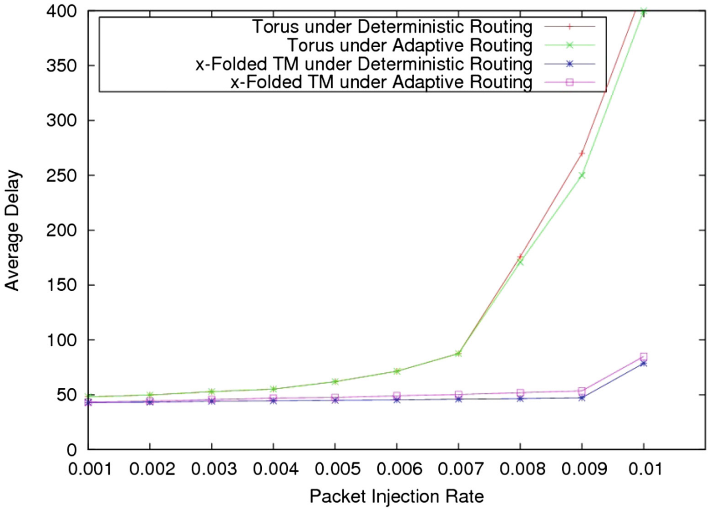


Fig. 4. Average delay curves under transpose traffic pattern

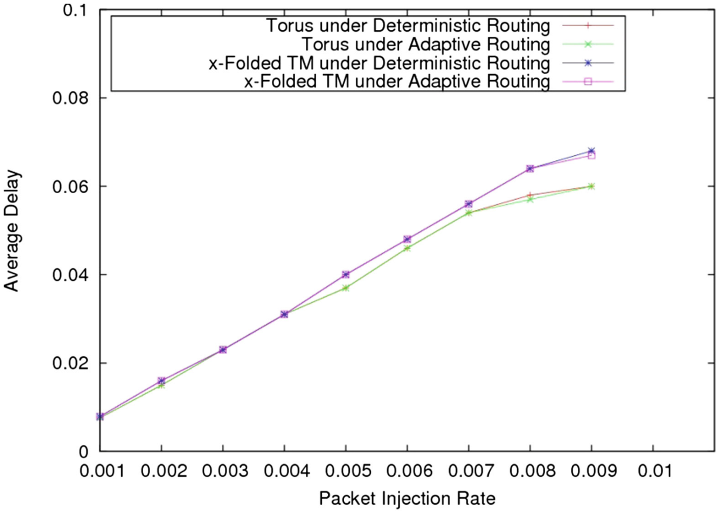


Fig. 5. Network throughput curves under transpose traffic pattern

In the presence of hotspot traffic pattern, there is a considerable reduction in the average delay of x-Folded TM topology compared with Torus by using adaptive routing algorithm. This reduction is provided using deterministic routing algorithm as well. Although, it does not have big difference compared with Torus. The topologies performance using both routing algorithms is presented in Fig. 6. For network throughput, as shown in Fig. 7, x-Folded TM topology has more improvement using adaptive routing algorithm in compare with deterministic. The improvement of the x-Folded TM is presented with increasing the packet injection rate.

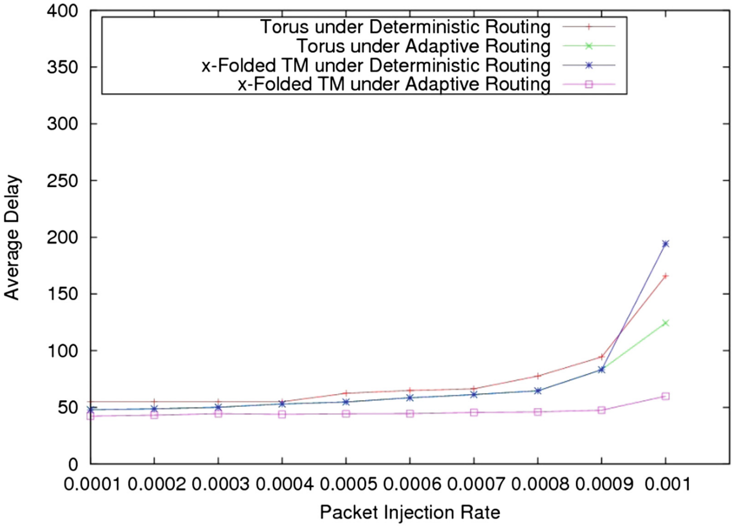


Fig. 6. Average delay curves under hotspot traffic pattern

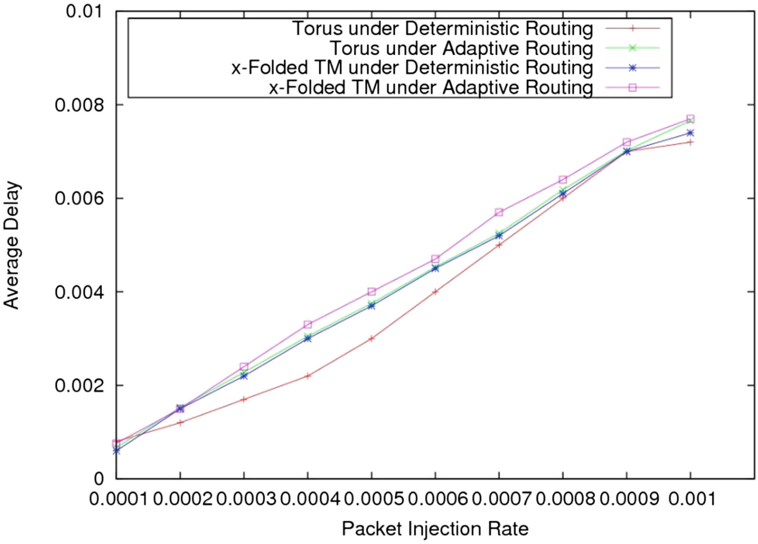


Fig. 7. Network throughput curves under hotspot traffic pattern

6 Conclusion

Over the years, several research efforts have aimed to propose new topology with applied routing as essential components to solve deadlock problems in interconnection networks. In this article, we described x-Folded TM topology as an efficient one to avoid deadlock while we presented the partitioning scheme on applied routing algorithms as well. The conclusion is positive effects on the average delay in the presence of various traffic patterns. In the field of the performance evaluation, it is carried out using simulation as the main research tool. Torus and x-Folded TM topologies have been evaluated to show the improvement in the performance of x-Folded TM. The most obvious advantage of this topology over the other compared topology is a reduction in the average delay which causes the better performance. In addition, the difference between deterministic and adaptive routing algorithms used to prove the superiority of adaptive routing algorithm over the deterministic algorithm in x-Folded TM topology.

Acknowledgement. This research was supported by Malaysian Ministry of Education, Fundamental Research Grant Scheme (FRGS, Ref: FRGS/1/2014/ICT03/UPM/01/1).

References

1. Feero BS, Pande PP (2009) Networks-on-chip in a three-dimensional environment: a performance evaluation. IEEE Trans Comput 58(1):32–45
2. Das S, Lee D, Kim DH, Pande PP (2015) Small-world network enabled energy efficient and robust 3D NoC architectures. In: Proceedings of the 25th edition on great lakes symposium on VLSI, pp 133–138, New York, NY, USA
3. Chen Y, Hu J, Ling X, Huang T (2012) A novel 3D NoC architecture based on De Bruijn graph. Comput Electr Eng 38(3):801–810
4. Camara JM, Moreto M, Vallejo R, Beivide R, Miguel-Alonso J, Martinez C, Navaridas J (2010) Twisted torus topologies for enhanced interconnection networks. IEEE Trans Parallel Distrib Syst 21(12):1765–1778
5. Moudi M, Othman M (2011) A challenge for routing algorithms in optical multistage interconnection networks. J Comput Sci 7(11):1685–1690
6. Liu Y, Han J, Du H (2008) A hypercube-based scalable interconnection network for massively parallel computing. J Comput 3(10):58–65
7. Hafizur Rahman MM, Shah A, Inoguchi Y (2012) On dynamic communication performance of a hierarchical 3D-Mesh network. In: Park JJ, Zomaya A, Yeo S-S, Sahni S (eds) Network and parallel computing, pp 180–187. Springer, Berlin, Heidelberg
8. Moudi M, Othman M, Lun KY, Abdul Rahiman AR (2016) x-Folded TM: an efficient topology for interconnection networks. J Network Comput Appl 73:27–34
9. Valinataj M (2013) Reliability and performance evaluation of fault-aware routing methods for network-onchip architectures (research note). Int J Eng – Trans A: Basics 27(4):509–516
10. Chiu G-M (2000) The odd-even turn model for adaptive routing. IEEE Trans Parallel Distrib Syst 11(7):729–738
11. Moudi M, Othman M, Lun KY, Abdul Rahiman AR (2016) Adaptive routing algorithm in x-Folded TM topology. In: IEEE symposium on computer applications & industrial electronics (ISCAIE), pp 63–66

12. Borhani AH, Movaghar A, Cole RG (2010) A new deterministic fault tolerant wormhole routing strategy for k-ary 2-cubes. In: IEEE International conference on computational intelligence and computing research (ICCIC), pp 1–7
13. Bohm C, Krebs F, Kriegel H-P (2002) Optimal dimension order: a generic technique for the similarity join. In: Kambayashi Y, Winiwarter W, Arikawa M (eds) Data warehousing and knowledge discovery, vol 2454. LNCS. Springer, Berlin, Heidelberg
14. Su KM, Yum KH (2008) Simple and effective adaptive routing algorithms in multi-layer wormhole networks. In: IEEE international performance, computing and communications conference (IPCCC), pp 176–184
15. Safaei F, Khonsari A, Fathy M, Ould-Khaoua M (2007) Performance evaluation of fully adaptive routing for the torus interconnect networks. In: Shi Y, Albada GDV, Dongarra J, Sloot PMA (eds) Computational science ICCS 2007, vol 4490 in LNCS, pp 606–613. Springer, Berlin, Heidelberg
16. Jiang N, Becker DU, Michelogiannakis G, Balfour J, Towles B, Shaw DE, Kim J, Dally WJ (2013) A detailed and flexible cycle accurate Network-on-Chip simulator. In: IEEE international symposium on performance analysis of systems and software (ISPASS), pp 86–96

Temporal Citation Network-Based Feature Extraction for Cited Count Prediction

Ho-Min Park, Yenewondim Biadgie Sinshaw, and Kyung-Ah Sohn(✉)

Department of Software and Computer Engineering, Ajou University, Suwon, South Korea
{simmani91,wondim,kasohn}@ajou.ac.kr

Abstract. The academic data that keeps the advanced knowledge of mankind continues to increase. Accordingly, researchers have been actively conducted research to find important ones among the academic data. This study presents new features for citation count prediction problem. The new features are derived from the network centrality analysis with time transition variance and are compared with the existing author, venue, and content features to verify their excellence. We use coefficient of determination $\left(\mathrm{R}^2\right)$ as a performance measure, and it has been confirmed that our proposed features are more useful for predicting the citation count than the existing features. Along with presenting new features, we have also attempted time-series analysis to observe whether the features used in the prediction change their influence with time. Thus, we have found that the influence of features does not change much over time.

1 Introduction

We are now living in the big bang era of data. In the world, a lot of data is stored and circulated. Academic data is one of the most important data that stores the latest technology and information. As the amount of academic data grows, we need to classify the papers which have a more scientific impact. These demands have been around for a long time and various metrics have been proposed. Typical examples are Eugene Garfield's impact factor [1] that measures the influence of the publisher, the H-index [2] that measures the academic capacity of the author, and the citation count which is the measure of the academic impact of the paper.

The citation count method of measurement is very useful for automatically finding highly impact papers from a bunch of academic data. However, this does not mean that it will still have a high impact for 'future' because the papers that are considered "highly cited" at the present were lowly cited at the beginning. In addition, forecasting the future potential of scholarly data is something many people want. For example, in the case of a government agency or corporation, the set of predictive information about citation count can be used as a guide to determine the future policy. It can also help researchers determine future research directions.

In recent years, research has been actively conducted to extract important features from academic data and to use them to view the future performance of scholarly papers or researchers [1]. Yan et al. have been conducting research to extract and quantify

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_41

information from academic data to predict citation counts [2]. They created a predictive model that learned the extracted information and applied it to the actual academic search system [3], and it was a remarkable achievement. Their information extraction studies have been applied in future studies and used in the model to predict the authors' H-index [4], and have shown such changes using graph pattern mining [5, 6].

Bibliographic network analysis, which models the relationship between papers, or the relationship between authors, is also one of the leading research areas for measuring academic promise. Some studies have been conducted on engineering, medicine, and patent areas based on academic data such as PubMed or web of science [7–10].

Iwami et al. calculated centrality features from evolving citation network [11]. They analyzed the correlation of this centrality with the famous papers in several academic fields. However, those features were not compared to the features of other research works. On the contrary, Yan's research has found many good features but needs an approach to temporal problems [2]. For example, their study assumes that the model for predicting the number of citations in 2009 by learning the data for the year 2000, and it will be able to predict the year 2019 by learning the data for the year 2010. However, the academic field changes from time to time and we need to make sure that past models can be used to predict the future.

In this paper, we address two questions raised in the above-mentioned works. First, how useful is the network feature extraction to find promising papers when compared to other features? Second, is the point of view fixed for prediction to be compatible with other data? We run some experiments and analyze the results to find answers to these two questions.

2 Problem Definition and Data Description

The academic paper corpus of total time span $\mathrm{D}_{1:\mathrm{T}} = \left\{\mathrm{D}_1, \ldots, \mathrm{D}_\mathrm{t}, \ldots \mathrm{D}_\mathrm{T}\right\}$ is a set of document dump D_t which denotes the documents published on time span t and each document $\mathrm{d}_\mathrm{t} \in \mathrm{D}_\mathrm{t}$ represents a scholastic paper which published in time span t. Each of document has citing-cited relations to other documents. The cited count of time t $\left(\mathrm{C}_\mathrm{t}(.)\right)$ is defined as

$$\mathrm{cited}\left(\mathrm{d}_\mathrm{t}\right) = \left\{d'_s \in D_{1:T}{:}d'_s cites\ d_t\ and\ t \le s\right\} \tag{1}$$

$$\mathrm{C}\left(d_t\right) = \left|\mathrm{cited}\left(\mathrm{d}_\mathrm{t}\right)\right| \tag{2}$$

Given a feature matrix X_t which is extracted features from D_t and $\mathrm{C}_{t+\Delta t}$ that is a cited count vector at time span t + Δ t, our problem objective is learning an approximation function $\mathbb{F}(.)$ that can predict true cited count $\mathrm{C}_{t+\Delta t}$ from X_t .

The authors in [2–4] used arnetminor citation dataset [12] to train their models. The arnetminor citation dataset consists of the metadata (title, author, year, abstract, citation information, etc.) of the thesis. We extracted new features using temporal citation network as described in Sect. 3. Authors in [2–4] extracted almost 33 features to predict future citations or author's H-index as shown Table 1. These features can be divided into three categories: author, venue and content. However, we included only 13 features to our network based extracted features by dropping the remaining 20 features. The selected 15 features are known to be associated with cited counts.

Table 1. Types and descriptions of existing features [2–4]

Type	Feature	Description
Author (Added)	*H-index*	The relationship between the number of citations of an author and the number of paper of an author [13]
	Author rank	Rank based on the number of citations received by the author
	MPIA	The most cited count among authors' published papers
	Productivity	Number of papers published by authors
	Sociality	Number of co-authors working together
	Authority	Author's Topic distribution times cited count which he/she get
	Versatility	Subject range of papers published by author
Venue (Added)	*Venue rank*	A ranking based on the number of citations in the venue
	MPIV	Cited count from Venue's most cited paper
Content (Added)	*Popularity*	The prospect of the paper based on the number of citations times topic probability
	Recency	Number of years since the paper was published
	Novelty	Subject similarity with papers cited by a given article
	Diversity	The diversity of topic distribution that the article includes
Author (Dropped)	*TPIA*	Total past influence of author (total number of citation count)
	NOCA	*Number of co-authors in a document*
	R.Author Rank	Author Rank using only recent N years of data
	R.H-index	H-index using only recent N years of data
	R.Productivity	*Productivity* using only recent N years of data
	R.MPIA	*MPIA* using only recent N years of data
	R.TPIA	*TPIA* using only recent N years of data
	R.NOCA	*NOCA* using only recent N years of data
	R.Sociality	*Sociality* using only recent N years of data
	R.Authority	*Authority* using only recent N years of data
	R.Versatility	*Versatility* using only recent N years of data
Venue (Dropped)	*V.Centality*	Centrality values of each venues in citation-cited network
	TPIV	Total past influence of venue (total number of citation count)
	R.Venue Rank	*Venue Rank* using only recent N years of data
	R.V.Centrality	*Venue Centrality* using only recent N years of data
	R.MPIV	*MPIV* using only recent N years of data
	R.TPIV	*TPIV* using only recent N years of data
Content (Dropped)	*Topic Rank*	The rank of topic that a document has
	R.Novelty	Novelty using only recent N years of data
	R.Topic Rank	Topic Rank using only recent N years of data

3 Feature Extraction Based on Temporal Citation Network

In this section, we introduce feature extraction based on the temporal citation network, a new feature extraction method for citation count prediction, and summarizes the necessary prior knowledge. The Iwami et al. [11] argued that the feature extracted by the citation network of the paper can be used to find the paper with the greatest academic influence in the future. However, two examples of centrality, betweenness and closeness, do not significantly affect the detection of articles with higher citation counts in previous papers [14]. Therefore, we use the degree of centrality which is considered to have the highest correlation, the eigenvector centrality that extends it, and the PageRank as features. We used NetworkX [15], an open source graph library, to compute Network features.

Citation network $\mathrm{N}_{1:t} = \left(\mathrm{V}_{\mathrm{D}_{1:t}}, \mathrm{E}_{\mathrm{D}_{1:t}}\right)$ is a directed graph, where each node $\mathrm{v} \in \mathrm{V}_{\mathrm{D}_{1:t}}$ represent a published paper and $\mathrm{edge}(\mathrm{v}, \mathrm{u}) \in \mathrm{E}_{D_{1:t}}$ exists if paper u and v have a cited relation of Eq. (1).

Centricity is a measure of the relative importance of individual nodes in a graph. The higher the centroid of a node, the more it acts as a hub or intermediary within the network, which in turn has a great influence within the network. The centroid can be computed in various ways, but the disadvantage is that the computational complexity increases exponentially as the network grows. In this study, degree, PageRank centrality which is an extended version of degree, in-degree, and eigenvector centrality are calculated and used for experiments.

The citation network has an increasing shape over time. Since the centrality value is a relative importance in the network, so it changes over time. Thus, we can obtain the graph shown in Fig. 1. For example, from this graph, we can extract five kinds of features, the value itself, slope, height, span, and area. Value is a calculated centrality of the time span t. Slope means the highest slope value. Height is the largest centrality value of the whole section. Span refers to the time it takes to reach the maximum value, and area refers to the cumulative value of the centrality of the year.

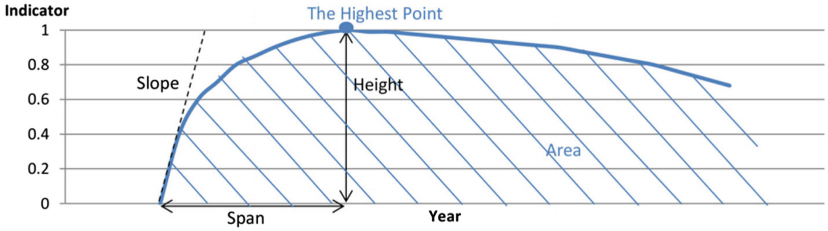


Fig. 1. Centrality value graph and features though the time span

4 Experimental Setup

We calculated 33 features in each year, including 13 Author, Venue, Content features as shown in Table 1 and 20 network features. We sampled 10,000 training and testing data from the papers generated from 2000 to 2005. With $y_{t+\Delta t}$ values, we used citation counts after 1, 5, and 10 years of each data. For citation count prediction, we used three representative regression algorithms: Linear Regression, K-Nearest Neighbor, and Classification and Regression Tree. In this study, we used the code provided by scikit-learn [16], an open source library implemented in the python language.

To evaluate our proposed method, we used Coefficient of determination. This is a measure of the degree to which the estimated linear model is appropriate for the given data. It refers to the percentage of the variable that can be explained by the applied model among the variation of the response variable. The usual sign of the coefficient of determination is R^2. Normally, R^2 has a value between 0 and 1, and it is said that the closer to 1 it describes the variable. But in the case of the nonlinear models, such as KNN, a value less than 0 may be obtained due to the nature of the equation. The formula for obtaining R^2 is as follows.

$$R^2 = 1 - \frac{\sum_i^N (y_i - f_i)^2}{\sum_i^N (y_i - \bar{y})} \tag{3}$$

where y_i is the true value of the test data and f_i is the predicted value through the model. $\bar{y}$ is the mean value of the test data set y.

5 Result

5.1 Performance Analysis with Adding/Dropping Features

Table 2 shows the value of R^2 when each feature is divided into groups and then only groups are excluded (Drop) or models are tested using only that group (add). In general, CART model has the best performance among all models using all features. It is also found that the probability of predicting a small Δt is high and that it decreases with decreasing Δt. One of the most noticeable points in the drop section is that the performance is improved when all data features are used except for the content feature group. Therefore, it can be interpreted that the content feature may not contribute to model learning more than other features. In the left part, it was difficult to predict the model with fewer features, but it was confirmed that the model using the network centrality features showed the same or better performance than the full model (Table 3).

Table 2. Types and descriptions of temporal citation network-based features

Type	Feature	Description
Degree centrality	*Area*	Sum of degree centrality of document until sampling year
	Height	Highest degree centrality of document until sampling year
	Slope	Max difference of degree centrality until sampling year
	Span	The time required to reach the highest point
	Value	Degree centrality of the sampling year
In-degree centrality	*Area*	Sum of in-degree centrality of document until sampling year
	Height	Highest in-degree centrality of document until sampling year
	Slope	Max difference of in-degree centrality until sampling year
	Span	The time required to reach the highest point
	Value	In-degree centrality of the sampling year
Eigenvector centrality	*Area*	Sum of eigenvector centrality of document until sampling year
	Height	Highest eigenvector centrality of document until sampling year
	Slope	Max difference of eigenvector centrality until sampling year
	Span	The time required to reach the highest point
	Value	eigenvector centrality of the sampling year
PageRank	*Area*	Sum of PageRank of document until sampling year
	Height	Highest PageRank of document until sampling year
	Slope	Max difference of PageRank until sampling year
	Span	The time required to reach the highest point
	Value	PageRank of the sampling year

Table 3. Performance comparison for Add/Dropping features by prediction model and Δt. Bold-faced ones are the best R^2 values in the corresponding year and condition.

		Δt = 1			Δt = 5			Δt = 10		
		KNN	LR	CART	KNN	LR	CART	KNN	LR	CART
Drop (−)	Author	0.486	0.963	0.942	0.539	0.877	0.881	0.446	0.729	0.801
	Venue	0.479	0.963	0.942	0.445	0.875	0.881	0.358	0.728	0.715
	Content	0.431	0.963	0.943	0.423	0.876	**0.882**	0.342	0.728	**0.808**
	Net_deg	0.505	0.963	0.942	0.685	0.873	0.814	0.580	0.720	0.799
	Net_indeg	0.510	0.891	0.906	0.642	0.860	0.832	0.526	0.727	0.761
	Net_eigen	0.475	0.963	0.942	0.511	0.877	0.833	0.418	0.732	0.713
	Net_page	0.481	0.962	**0.946**	0.521	0.877	0.878	0.428	0.720	0.583
Add (+)	Author	−0.073	0.479	−1.566	−0.054	0.809	−0.400	−0.065	0.710	−0.309
	Venue	−0.244	-0.038	−0.054	0.011	0.139	0.020	0.035	0.245	0.020
	Content	−0.224	0.962	−0.808	−0.178	0.873	−0.113	−0.170	0.709	−0.217
	Net_deg	0.453	0.012	0.872	0.762	0.032	0.831	0.663	0.016	0.639
	Net_indeg	0.595	−0.116	**0.948**	0.846	−0.064	**0.881**	0.717	−0.025	**0.745**
	Net_eigen	0.027	0.039	0.019	0.250	0.070	0.039	0.193	0.052	0.034
	Net_page	0.386	0.013	0.894	0.616	0.021	0.687	0.658	0.017	0.567
Full	All	0.481	**0.963**	0.943	0.521	0.876	**0.881**	0.428	0.729	**0.730**

Figure 2 shows a separate visualization of the R^2 value of the CART model. Drop shows normal performance near 0.7 and 0.8 without significant difference in performance. On the contrary, the add group which selects only one feature group shows that degree, indegree, and PageRank centrality perform better. In other words, the network centrality is a relatively good feature when compared with existing features.

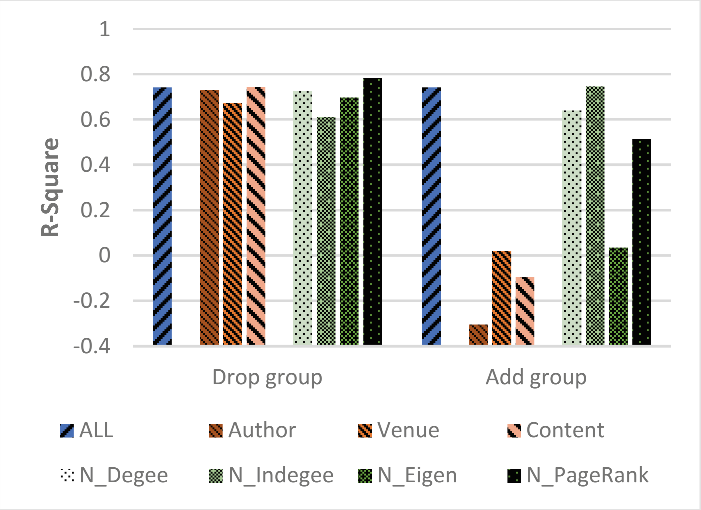


Fig. 2. Performance comparison when the group of features is created or removed by each feature group. ALL, the leftmost one, means **R**^{2} value when all features are used.

5.2 Importance Change Through Time

In analyzing the related research works, we felt that the data have a time series characteristic, whereas the experiment is fixed at a certain point due to the characteristics of the regression model. In the case of the Content feature, there is no problem, but in the case of Author, Venue, and Network, the features change over time, so we need to make sure that the predicted model with fixed viewpoint can be used in the future. Figure 3 shows the relative R^2 values of features in each of these years. According to the figure, the importance of the remaining models is not significantly changed by feature.

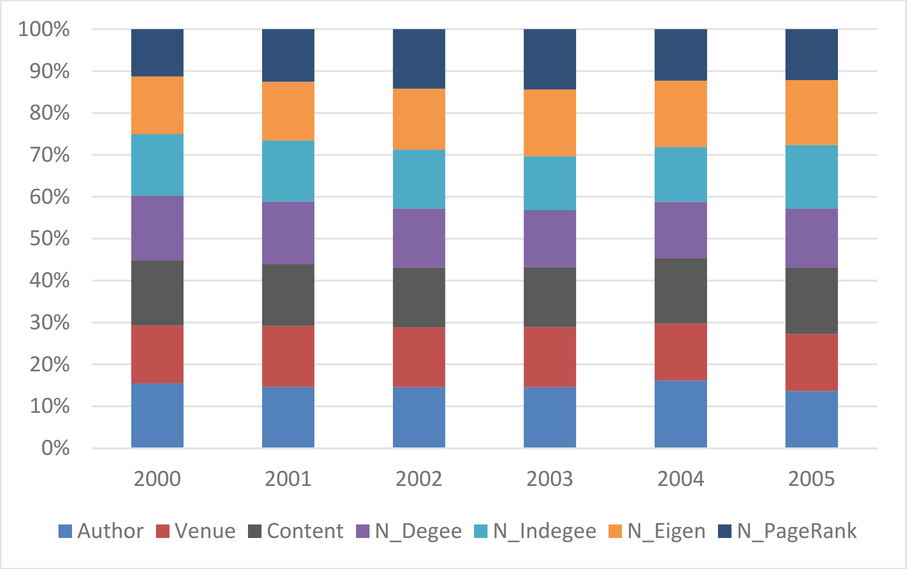


Fig. 3. Comparison of feature importance change from 200 to 2005 (**Δt** = 10), We used Drop $\mathbf{R}^2$ values as relative performance measure.

6 Conclusion

In this paper, we have added new features to the existing cited count prediction problem and analyzed its usefulness. The new features were obtained by centrality values of (in) degree, eigenvector and Page Rank algorithms from the citation network and the time transition of its value such as the slope, area, and top. The obtained features have higher. values compared to the existing Author, Venue, and Content class features, which means new features will help predict the number of citations in future papers. In addition, we found that the time-dependent changes did not significantly affect the importance of each feature.

Finally, we will further refine the features extracted by future research to improve the prediction system and make it an application that can be practically used and confirm how the actual results are applied. We also plan to provide a better way to solve the cited count prediction problem by finding a model that can more accurately reflect the time series characteristics of academic information data.

Acknowledgments. This research was supported by the MISP (Ministry of Science, ICT & Future Planning), Korea, under the National Program for Excellence in SW) supervised by the IITP (Institute for Information & communications Technology Promotion) (R2215161002000l002)

References

1. Bethard S, Jurafsky D (2010) Who should I cite: learning literature search models from citation behavior. In: Proceedings of the 19th ACM international conference on Information and knowledge management, pp 609–618
2. Yan R, Tang J, Liu X, Shan D, Li X (2011) Citation count prediction: learning to estimate future citations for literature. In: Proceedings of the 20th ACM international conference on Information and knowledge management, pp 1247–1252
3. Yan R, Huang C, Tang J, Zhang Y, Li X (2012) To better stand on the shoulder of giants. In: Proceedings of the 12th ACM/IEEE-CS joint conference on digital libraries, pp 51–60
4. Dong Y, Johnson RA, Chawla NV (2015) Will this paper increase your h-index? Scientific impact prediction. In: Proceedings of the eighth ACM international conference on web search and data mining, pp 149–158
5. Livne A, Adar E, Teevan J, Dumais S (2013) Predicting citation counts using text and graph mining. In: Proceedings of the iConference 2013 workshop on computational scientometrics: theory and applications
6. Pobiedina N, Ichise R (2014) Predicting citation counts for academic literature using graph pattern mining. In: International conference on industrial, engineering and other applications of applied intelligent systems, pp 109–119
7. Shibata N, Kajikawa Y, Takeda Y, Sakata I, Matsushima K (2011) Detecting emerging research fronts in regenerative medicine by the citation network analysis of scientific publications. Technol Forecast Soc Change 78:274–282
8. Kajikawa Y, Ohno J, Takeda Y, Matsushima K, Komiyama H (2007) Creating an academic landscape of sustainability science: an analysis of the citation network. Sustain Sci 2:221
9. Kajikawa Y, Takeda Y (2009) Citation network analysis of organic LEDs. Technol Forecast Soc Change 76:1115–1123
10. Cho T-S, Shih H-Y (2011) Patent citation network analysis of core and emerging technologies in Taiwan: 1997–2008. Scientometrics 89:795–811
11. Iwami S, Mori J, Sakata I, Kajikawa Y (2014) Detection method of emerging leading papers using time transition. Scientometrics 101:1515–1533
12. Tang J (2016) AMiner: mining deep knowledge from big scholar data. In: Proceedings of the 25th international conference companion on world wide web, p 373
13. Hirsch JE (2005) An index to quantify an individual's scientific research output. In: Proceedings of the national academy of sciences of the United States of America, pp 16569–16572
14. Park H-M, Sohn K-A (2016) Predicting emerging papers using time series citation network analysis. In: Proceedings of symposium of the Korean Institute of communications and Information Sciences, pp 66–67. Korea Institute of communication sciences
15. Hagberg AA, Schult DA, Swart PJ (2008) Exploring network structure, dynamics, and function using NetworkX. In: Proceedings of the 7th python in science conference (SciPy2008), pp 11–15, Pasadena, CA, USA
16. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830

A Secure Workflow-Net Model for Service-Specific Overlay Networks

Ismaeel Al Ridhawi(✉) and Yehia Kotb

College of Engineering and Technology, American University of the Middle East (AUM), Eqaila, Kuwait
{Ismaeel.Al-Ridhawi,Yehia.Kotb}@aum.edu.kw

Abstract. In this paper, a Workflow-net based mathematical framework called Secure Workflow-net is proposed to enhance security attributes of service-specific overlays (SSO). The framework checks for resource accessibility privileges on an overlay node and grants access whenever credible. A formal method is provided to determine the accessibility and availability of resources intended for service subscribers. Additionally, a closed form theorem for framework soundness and lemmas to study the characteristics of the framework are introduced. Simulation results demonstrate how task coverage can still be achieved in an adequate timely manner when considering security issues to construct service composition workflows.

Keywords: Petri-net · Workflow-net · Security · Overlay network · Service-specific overlay · Fog-to-cloud

1 Introduction

Overlay networks are created as an abstraction layer to the underlying physical network using software to run multiple virtualized network layers to provide application, networking, or security benefits [1]. With the emergence of the fog-to-cloud (F2C) computing paradigm [2], edge nodes such as mobile devices are used to provide computing, storage and networking services to achieve load balance among clouds and fogs, reduced network bandwidth usage, and energy efficiency for data centers [3, 4]. The composition of service-specific overlays (SSO) still plays an important role to achieve the requirements of F2C computing systems. Services are composed using edge nodes to provide composite and enhanced services needed for cloud subscribers.

Petri-net provides a solution towards service composition in which the available capabilities of edge nodes are merged together to achieve the requested task [5]. Workflow-net provides an extension to Petri-net and has been adopted lately for service composition to produce a more robust and sound solution [6]. Information system security has been a hot topic for many decades, e.g. [7]. An information system is considered to be secure if it has well-defined security measures and characteristics such as authenticity, confidentiality and integrity [8]. Security in service composition has been considered in the literature [9, 10], but has been overlooked when Workflow-nets are used to compose services. In this paper, a closed form Workflow-net based

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_42

mathematical model is proposed that ensures such security characteristics are satisfied both structurally and behaviorally. The solution is an extension to Workflow-nets in which we call it Secure Workflow-net.

This paper is organized as follows: Sect. 2 outlines some of the previous work in information security for service composition. Section 3 provides an introduction and overview to what Petri-nets and Workflow-nets are and their characteristics. Section 4 models the service overlay composition problem using Workflow-net. Section 5 illustrates the proposed Secure Workflow-net framework and presents a theorem and lemmas along with their proofs. Simulations are conducted in Sect. 6. Finally, Sect. 7 concludes the paper.

2 Related Work

Secure service composition solutions have been discussed earlier. The authors in [11] present a privacy preserving access control model and framework for secure service provisioning and composition. To create secure service compositions, the solution ranks possible chains of composite services according to the users' preferences and sensitivity level of their data. An access request for a service is permitted if the requester's attribute certificates, contextual conditions, and privacy preferences are in compliance with the access control policies specified by the service provider. In [12] the authors propose an information declassification mechanism used for secure service compositions. The declassification mechanism is based on cryptographic operations and information flow security requirements. Mobile service nodes then cooperate with each other to complete the secure composition procedure.

Varadharajan [13] proposed an extended Petri-net which can be used to model information flow security requirements. In his work, the author proposed the concept of *tokdata* which are tokens that flow according to the carried data based on a defined function. Castano et al. [14] proposed a tool for verifying, during the system security design, security properties of data used by complex applications to improve the overall security policies. Their research is based on colored Petri-nets [15]. Knorr [16] proposed an approach for access control using what he called an access control matrix using Workflow-nets. Juszczyszyn [17] addresses important problems met when implementing mandatory access control policies in complex distributed systems. The author also presents formal security models using colored Petri-nets. Mikolajczak also proposed a colored petri net approach for modeling security in information systems [8].

Despite the abundant existing studies on security in service composition for overlay networks, to our knowledge, the presented work is the first to employ a secure Workflow-net solution that can be used to compose SSOs which provides aid for the F2C computing paradigm.

3 Petri-Nets and Workflow-Nets

A Petri-net is a directed bipartite graph with two types of nodes, namely **places** (circles) and **transitions** (solid rectangles). Transitions model actions that may occur. Places are pre- or post-conditions for the transitions in which they are connected to. Places and transitions are connected via directed weighted **arcs**. If these arcs are not weighted then the **weight** is assumed to be one. These integer weights determine the number of activities that flow from places along the arcs per transition. Activities are called **tokens** (small solid circles that reside in places). The distribution of tokens over places is called a **marking**. Figure 1 depicts an example of a Petri-net structure.

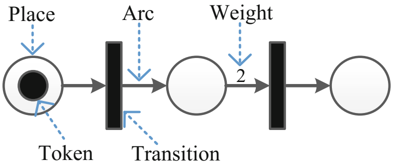


Fig. 1. The structure of a Petri-net.

When arcs run from places to a transition, these places are considered input places to the transition. On the contrary, when arcs run from a transition to places these places are considered output places to the transition. A transition in a Petri-net is enabled if and only if there are tokens in all the input places to the concerned transitions and each input place contains a number of tokens that is greater than or equal to the weight of its connecting arc. After a transition is enabled, it will eventually **fire** by consuming tokens from its input places and producing tokens in its output places.

Workflows have been used in the management of distributed information systems [18, 19]. Workflow-nets (WF_{net}) are used to model the structural and dynamic behaviors of workflows. The structural behavior of a workflow defines task dependencies and their structure which guarantees the desired output. The dynamic behavior of a workflow is how the structure of the workflow reacts online with activities that are handled by the workflow. A Workflow-net is a special type of Petri-net that has two special places, i and o, where i is the only place that does not have any input transitions and o is the only place that does not have any output transitions. Nodes i and o are called the **source** and **sink** nodes. Workflow-nets are preferred over normal Petri-nets in distributed systems due to the fact that they guarantee the success of a process. This supports the property of soundness which is described next. The following are definitions for Workflow-nets and their characteristics.

Definition 1 (Workflow-net). A Petri-net $\aleph$ is considered a WF_{net} if and only if:

1. $\aleph$ has an input place i, where $i = \emptyset$.
2. $\aleph$ has an output place o, where $o = \emptyset$.
3. If a transition $t*$ is added to $\aleph$ such that $t^* = o$ and $t^{\;*} = i$, the Petri-net $\aleph*$ becomes strongly connected, where $t*$ is a transition that connects the input to the output of WF_{net}.

In the above definition, i is the set of all input transitions to place i and o is the set of output transitions from place o. When a Petri-net is strongly connected, there is a sound path between any two transitions in the net. One of the Petri-net properties that is considered essential to Workflow-nets is the property of soundness. **Soundness** guarantees that a Petri-net will eventually terminate and, at that moment, there will be tokens in the output place and all other places will be empty. This means that all activities will reach the output place.

Definition 2 (Soundness of a Workflow-net). A WF_{net} is sound if and only if:

1. $\forall M \in |M_i\rangle, M_o \in |M\rangle$,
2. $\forall M_k \in |M_i\rangle$, if $M_k \in |M_o\rangle$ then $M = M_o$ and
3. $\forall t \in T, if\ \exists M \in |M_i\rangle$, then $t \in |M\rangle$.

where M is a marking of the WF_{net}, M_i is the input marking, M_o is the output marking, M_k is the marking at time k and t is a transition.

4 Modeling SSOs as Workflow-Nets

The service composition problem is modeled as a set of actions that must be performed in sequence to achieve the requested task. We assume that each action is performed by an overlay node. For instance, assume that a cloud subscriber is requesting a particular service that may not exist at a single edge node nor the cloud (e.g. a media content with certain modifications and enhancements). To provide the subscriber with the requested service, an overlay must be constructed using the edge nodes (Fig. 2). Each edge node performs an action, such that a sequence of actions performed by multiple edge nodes will lead to the composite requested service.

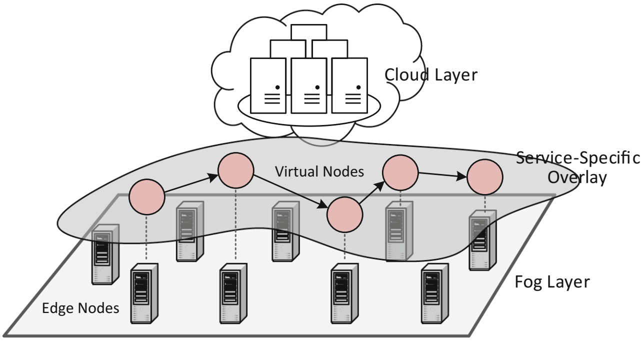


Fig. 2. Service composition through an overlay.

The composed SSO is modeled using a Workflow-net, in which workflow transitions perform actions (i.e. actions performed by an overlay node). As defined in Sect. 3 actions are represented as tokens residing in places (i.e. events preceding the appliance of actions). A transition executes (i.e. performs an action) after being enabled. The result of the execution is the removal of tokens from each of the transition's input places and the creation of tokens in each of its output places. Figure 3 depicts a Workflow-net model for the service composition process example outlined in Fig. 2. The figure outlines a media service composition problem that requires the addition of media enhancements to the original media content to produce a composite media service. Each transition characterizes an action that must be carried out by an overlay node. Each place depicts an overlay node with an action (service) awaiting to be applied (added) to the media content. As the media content (represented as a token residing in a place) is modified by the nodes, the token will eventually end up residing in the last place (i.e. last overlay node to add a media enhancement).

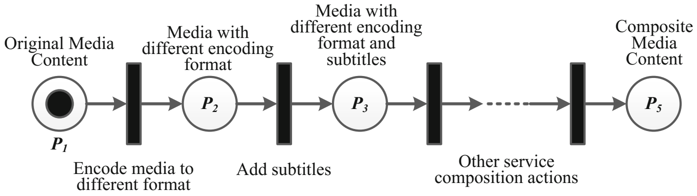

Original Media Content
Media with different encoding format
Media with different encoding format and subtitles
Composite Media Content
P_1
P_2
P_3
P_5
Encode media to different format
Add subtitles
Other service composition actions

Fig. 3. Workflow-net model for a service composition process.

5 Constructing a Secure Workflow-Net Framework

5.1 Security Constraints

According to [8], an information system is considered secure if it satisfies certain properties. The following outlines those properties:

Confidentiality and Data Secrecy (χ): is the availability of network resources (node services) and data for only those who are entitled to access such concerned resources and data. We mathematically define confidentiality as follows:

$$\chi(a,r) = \begin{bmatrix} x_1, x_2, \dots \dots \\ \dots \dots \dots \dots \\ \dots \dots \dots, x_k \end{bmatrix} \tag{1}$$

where $X = x_1, x_2, \ldots, x_n$ is a vector that represents different access levels. If $\chi(3,4) = x_6$, this means that node a_3 has access level x_6 on resource r_4.

Service Integrity (ψ): is the availability, reliability and completeness of the network. The availability of resources is demonstrated though a vector as follows:

$$V = [v_1, v_2, \dots \dots \dots, v_k] \tag{2}$$

where v_i is the availability of a resource identified by index i, and k is the maximum number of resources.

To assign a resource to a node, two conditions must be satisfied:

1. The resource has to be available.
2. The node must have an accessibility privilege to that resource.

which is mathematically represented by the following equation:

$$\forall S(a_i, r_k), \exists \chi(a_i, r_k) \text{ and } v_k \neq 0, \tag{3}$$

where S is the assignment probability, a_i is node i in the network, r_k is resource k, and v_k is the availability of that resource. The assignment probability matrix is the product of the availability vector and the transpose of the accessibility matrix:

$$S = \left[V \times \chi^T\right] \tag{4}$$

Data Integrity (δ): is the availability and reliability of the data. As mentioned earlier, in this work, we do not distinguish between data, software or hardware as they are all resources (services) and therefore the model proposed in Sect. 4 applies.

Authentication (α): is the process of checking assertions. In this context, we claim that α is the process of deciding whether or not to place a process token in the workflow input place as will be seen in the proposed solution.

Non-repudiation (ϖ): is a constraint that prevents the node from resource access denial in cases in which it has already been granted access. This is ensured by the structure of the framework as will be seen later in the next section.

5.2 Proposed Secure Framework

The proposed framework is an extension to Workflow-net that contributes to achieving security in SSOs. We call the new extension Secure Workflow-net WF_s, and is mathematically defined as follows:

$$WF_s =< WF_{net}, A, \chi, R, \Pi, \xi > \tag{5}$$

where WF_s is the structural workflow that defines the process, A is the set of nodes available for service composition in the network, χ is the accessibility of every node to resources, R is the set of available resources, Π is the routing mechanism that routes resources into their sub-workflows, and ξ is a sub-Workflow-net that binds the input to the output for access rejection or error handling cases.

For a WF_s to be structurally valid, the following constraints must be satisfied:

1. WF_{net} is a sound workflow net,
2. $\forall r \in R$ and $a \in A, \exists(a, r) \in \chi$,
3. $\Pi \cap WF_{net} \neq 0$,
4. ξ is a sound WF_{net} that binds input with output.

The main Workflow-net modeling the composition has to be a sound Workflow-net. The nodes have some defined access level on resources. The routing mechanism is tightly bound to the Workflow-net to guarantee the flow of tokens to the right sub-workflow. There is also a sub-workflow that drives the access rejection tokens to the output.

The following demonstrates the soundness of the secure Workflow-net though a proposed theorem of soundness and a set of lemmas.

Theorem 1. A secure Workflow net $\aleph$ is sound if and only if:

1. $\aleph$ is a structurally valid WF_s,
2. $\forall M_i, M_i \in |M_\Pi\rangle$ and $P_o \in |M_i\rangle$,
3. $\forall a_i \in A$ and $r_j \in R, \exists \chi(a_i, r_j)$ such that $M_\Pi(0) > 0$ or $M_\xi(0) > 0$.

We prove this theorem by showing that if $\aleph$ is sound, then the four structural validity conditions of WF_s are satisfied and vice versa.

Proof: First we proof that if $\aleph$ is sound then the four conditions are satisfied:

1. Since $\aleph$ is a sound Workflow-net,
2. then $\forall M(0), M(P_0) \in |M(0)\rangle$, or $\xi \in M(0)$ and $M(P_0) \in \xi$,
3. then WF_{Net} is a sound Workflow-net.
4. Since $\forall M(0), M(P_0) \in |M(0)\rangle$,
5. then there will always be $r \in R$ and $a \in A, \exists (a, r) \in \chi$.
6. Since the reachability is satisfied,
7. then WF_i in $|\Pi_o$,
8. then if $\aleph$ is sound then the four conditions are satisfied.

Now we proof that if the three conditions demonstrated in the theorem are satisfied, then $\aleph$ is a sound secure Workflow-net WF_s.

1. Since $\aleph$ is a structurally valid WF_s,
2. then $\forall M(P_0), M(P_0)$ in $|M(P_i)\rangle$,
3. Since $M(P_0)$ in $|M(\Pi)\rangle$,
4. then Π and WF_s are two connected workflows and $M(WF_i) \in M(\Pi_o)\rangle$,
5. Since $\forall a_i \in A$ and $r_j \in R, \exists \chi(a_i, r_j)$ such that $M_\Pi(0) > 0$ or $M_\xi(0) > 0$,
6. then there will always be marking M_o in the WF_s that corresponds to input Π_i,
7. then WF_s is a sound secure Workflow-net.

Lemma 1. If $\aleph$ is a sound secure Workflow-net then $\aleph$ is structurally valid.

Proof:

1. Since $\aleph$ is a sound Workflow-net,
2. then $\forall M \in \Pi$, M will eventually reach the output place P_o,
3. then WF_{Net} is a sound Workflow-net,
4. Since a marking M corresponds to the association of a resource to a node,
5. then $\forall r \in R$ and $a \in A, \exists (a, r) \in \chi$,
6. Since the structure is a sound workflow,

7. then $\Pi \cap WF_{Net} \neq 0$ for the token to reach the output,
8. Since there is a possibility of rejection and yet the token will end up being the output place in any case,
9. then ξ is a sound Workflow-net that binds the input with the output,
10. then if $\aleph$ is a sound secure Workflow-net then $\aleph$ is structurally valid.

We define the node demand coverage to be the property for the process to assign a set of resources to nodes that have access to them when required. If the Workflow-net is sound, then resource assignment will occur if resources are accessible. On the contrary, a rejection token is sent to the output of the Workflow-net if no resource assignment is possible.

Lemma 2. If $\aleph$ is a sound secure Workflow-net then $\aleph$ satisfies the property of node demand coverage.

Proof:

1. Since the WF_S is sound,
2. then for every resource access request there will eventually be an output,
3. then for nodes that have access to resources, a resource access grant will eventually occur,
4. Therefore the Workflow-net satisfies the property of node demand coverage.

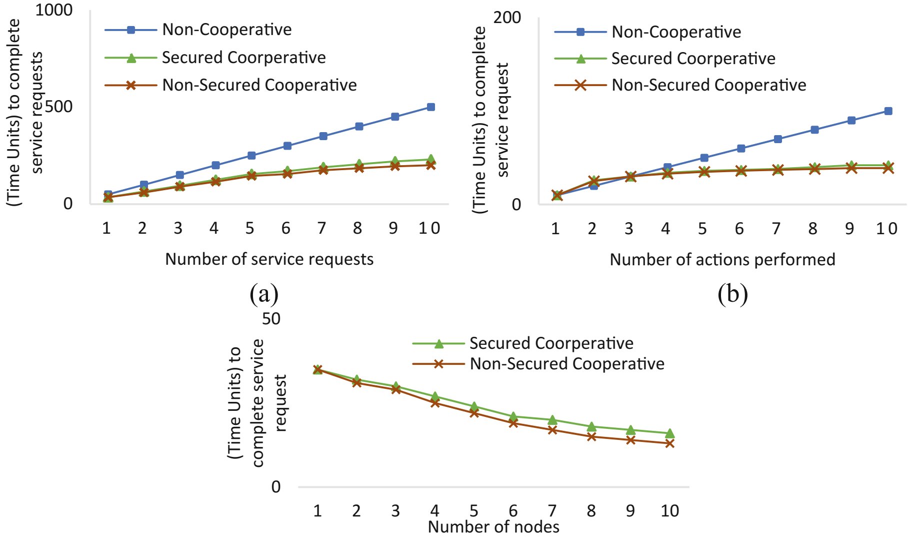
6 Simulation Results

We developed a simulator to test the proposed framework. The problem was generalized to test the system's capability regardless of the type of service requested. Three different solutions were considered: a cooperative non-secure Workflow-net, a cooperative secure Workflow-net, and a non-cooperative solution. The first considers a solution which uses a cooperative service composition technique developed in [18]. The second solution considers a similar cooperative model as in [18] but adds a layer of security as described in the previous section. The third solution disregards node cooperation and hence services are composed using a single overlay node. The goal of these simulation tests is to empirically demonstrate that our definition of secure Workflow-net is correct and that compositions can be adequately established.

The input to the simulator consists of a process in the form of a linear logic expression with operators described in [18]. Other input parameters consist of a set of nodes, each with a set of resources, expressed as Workflow-nets. Each resource corresponds to one action defined in the process, along with the cost associated with performing that action. Actions that are not part of a node's set of resources have their cost set to infinity. A uniformly distributed random variable is used to first determine the initial set of resources. When a node is granted access to a resource, the cost for executing the action is randomly determined with a normally distributed variable.

When the generated nodes' resources are insufficient to provide a complete composite service, the simulator terminates with infinity as a cost for execution. Otherwise, the cooperative process is constructed and executed. For simplicity reasons, the execution time is computed as the total execution cost in the Workflow-net.

We first examined the delay incurred to complete a certain number of service requests. The results depicted in Fig. 4(a) show that the non-cooperative solution incurs the most delay when compared with the other two cooperative solutions. Although both cooperative solutions show that the time needed to complete the service requests stabilizes as the number of service requests increase, the secure composition method incurs a small delay penalty compared to the non-secure composition approach (230 time units for the secured cooperative approach vs 200 time units for the non-secure cooperative method when 10 service compositions are requested).



(a)

(b)

(c)

Fig. 4. Service composition delay, (a) Time needed to complete requested services, (b) Time needed to complete service actions, (c) Time needed to complete service request using multiple nodes.

The second evaluation considers a service request which requires a set of actions to be performed to achieve the task. Results depicted in Fig. 4(b) show again that the cooperative approach outperforms the non-cooperative method. The secure cooperative approach incurs a small delay and shows similar time delays for the composition (42 time units for the secure cooperative approach vs 39 time units for the non-secure cooperative method when 10 actions are required to fulfill the composition request).

Finally, we evaluated the performance of the secure cooperative method as the number of nodes used for cooperation increases. Results shown in Fig. 4(c) prove that the secure approach does not incur an increased time burden compared to the non-secure approach. With the availability of 10 nodes to be used for cooperation, the delay is 16 time units for the secure approach vs 13 time units for the non-secure approach.

Overall, the secure service composition technique provides robust results with minimal delay overhead. Although the non-secure approach outperforms the secure method, the burden of using a non-secure approach outweighs the benefits.

7 Conclusion

This paper proposes a secure Workflow-net framework used to compose service specific overlays for cloud networks. The mathematical Workflow-net based model ensures that security characteristics are satisfied both structurally and behaviorally. The framework checks for resource accessibility privileges on an overlay node and grants access whenever credible. Simulation results show that the proposed technique incurs minimal delay overhead when composing service overlays.

References

1. Lua E, Crowcroft J, Pias M et al (2005) A survey and comparison of peer-to-peer overlay network schemes. IEEE Commun Surv Tutor 7:72–93
2. Masip-Bruin X, Marín-Tordera E, Tashakor G et al (2016) Foggy clouds and cloudy fogs: a real need for coordinated management of fog-to-cloud computing systems. IEEE Wirel Commun 23:120–128
3. Zhu J, Chan D, Prabhu M, Natarajan P, Hu H, Bonomi F (2013) Improving web sites performance using edge servers in fog computing architecture. In: Proceedings IEEE 7th international symposium on service-oriented system engineering, Redwood City
4. Jalali F, Hinton K, Ayre R et al (2016) Fog computing may help to save energy in cloud computing. IEEE J Sel Areas Commun 34:1728–1739
5. Xia Y, Luo X, Li J, Zhu Q (2013) A Petri-net-based approach to reliability determination of ontology-based service compositions. IEEE Trans Syst Man Cybern Syst 43:1240–1247
6. Kotb Y, Badreddin E (2005) Synchronization among activities in a workflow using extended workflow Petri nets. In: Proceedings 7th IEEE international conference on e-commerce technology
7. Salah K, Calero JMA, Zeadally S et al (2012) Using cloud computing to implement a security overlay network. IEEE Secur Priv Mag 11:45–53
8. Mikolajczak B, Joshi S (2004) Modeling of information systems security features with colored Petri nets. In: Proceedings of 2004 IEEE international conference on systems, man and cybernetics, The Hague, vol 5, pp 4879–4884
9. Kassmi I, Jarir Z (2016) Security requirements in web service composition: formalization, integration, and verification. In: Proceedings 25th IEEE international conference on enabling technologies: infrastructure for collaborative enterprises, Paris
10. Kassmi I, Jarir Z (2015) Towards security and privacy in dynamic web service composition. In: Proceedings third world conference on complex systems, Marrakech
11. Amini M, Osanloo F (2016) Purpose-based privacy preserving access control for secure service provision and composition. IEEE Trans Serv Comput 55:1
12. Xi N, Sun C, Ma J, Shen Y, Lu D (2016) Distributed secure service composition with declassification in mobile network. In: Proceedings international conference on networking and network applications, Hakodate
13. Varadharajan V (1990) Petri net based modelling of information flow security requirements. In: The computer security foundations workshop, Franconia, NH

14. Castano S, Samarati P, Villa C (1993) Verifying system security using Petri nets. In: Proceedings IEEE international carnahan conference on security technology, Ottawa, ON
15. Sakib K, Tari Z, Bertok P (2014) Petri nets. In: Verification of communication protocols in web services: model-checking service compositions. Wiley-IEEE Press, p 272
16. Knorr K (2000) Dynamic access control through Petri net workflows. In: Proceedings 16th annual conference on computer security applications, New Orleans, LA
17. Juszczyszyn K (2003) Verifying enterprise's mandatory access control policies with coloured Petri nets. In: Proceedings 12th IEEE international workshops on enabling technologies: infrastructure for collaborative enterprises
18. Kotb Y, Beauchemin S, Barron J (2012) Workflow nets for multiagent cooperation. IEEE Trans Autom Sci Eng 9:198–203
19. Tantitharanukul N, Natwichai J, Boonma P (2013) Workflow-based composite job scheduling for decentralized distributed systems. In: Proceedings 16th international conference on network-based information systems, Gwangju

Comparison of UML Sequence Diagrams to Trace Technical Specification Change

Supatra Insri and Yachai Limpiyakorn(✉)

Department of Computer Engineering, Chulalongkorn University, Bangkok 10330, Thailand
Supatra.I@student.chula.ac.th, Yachai.L@chula.ac.th

Abstract. Sequence diagrams are widely used to model the interactions between objects in an information system. This paper presents a method and develops a tool for comparing UML sequence diagrams to facilitate tracing the technical specification change. The change log is generated to report all the affected elements. The traceability graph associated with the new version of sequence diagram is also constructed to illustrate the updated dependency among components. The proposed automation approach would benefit software process improvement for change impact analysis. In addition, the output generated from the implemented system could fasten the tedious documentation chore.

Keywords: Requirements management · Traceability · Sequence diagram · UML · Software process improvement

1 Introduction

During the analysis and design phases, or the development of the evolving system, sequence diagrams can be used for behavioral modeling to provide understanding of the control flow of a use case scenario by time. Due to requirement changes, several versions of sequence diagrams will be created. For projects with the immature requirements management process, this would result in waste of resources, much of rework, and may end up with poor quality software product.

In some organizations, the new version of sequence diagram describing the changed technical specification is used as referred document for change impact analysis. This causes difficulty for manually tracing the changes. The rework process is also defect-prone due to human-error. This research thus presents a method and develops a tool for comparing UML sequence diagrams to facilitate tracing the technical specification change.

In literature, Grischick [1] presents an algorithm to compare two class diagrams and visualizes the differences. Filho and Lencastre [2] present a literature review about traceability visualisation techniques. A rule-based approach is described how to automate the generation of traceability relations and creation of HTML reports of traceability relations.

2 Background

2.1 UML Sequence Diagram

A sequence diagram is a dynamic model that shows the explicit sequence of messages passed between objects in a defined interaction [3]. An interaction basically involves

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_43

objects, operations, and messages. Objects have behaviors that are described by operations or implemented as methods. Each object also can send and receive messages which are function or procedure calls to tell another object to execute one of its behaviors. Since sequence diagrams emphasize the time-based ordering of activities taking place among a set of objects, they are generally used throughout software analysis and design phases for understanding real-time specifications and ensuring complete construction of all classes.

2.2 Traceability Graph

In a traceability graph, artifacts are represented as nodes. Nodes are connected by edges, if a trace link between the artifacts exists. Traceability graphs can be used to help manage requirements change. By navigating through the graph it is easy to identify the artifacts affected from changes or the missing links as a hint to create required artifacts. The probability to overlook related artifacts is then reduced. Referring to [4], traceability graphs allow getting an overview on the links exploratively and are characterized by a high information comprehension ratio.

3 Research Methodology

The proposed methodology to compare sequence diagrams for tracing the technical specification change is depicted in Fig. 1. The implemented system requires the input of UML sequence diagram is created by the free online tool– WebSequenceDiagrams. And the output of traceability graph is generated by Graphviz [5], which is a package of open-source tools that supports rendering graphs specified in DOT language scripts.

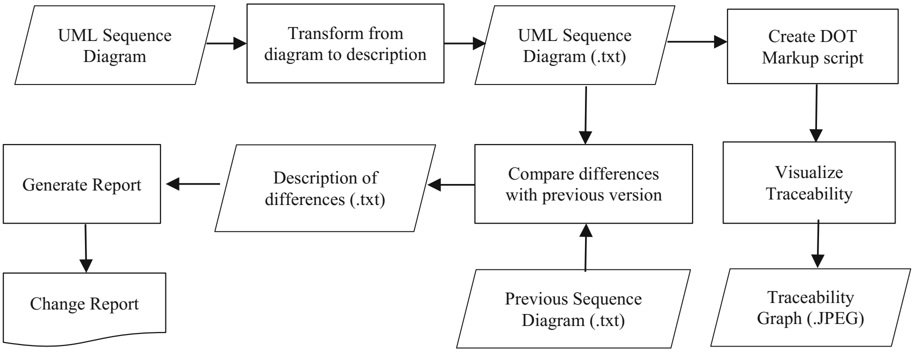
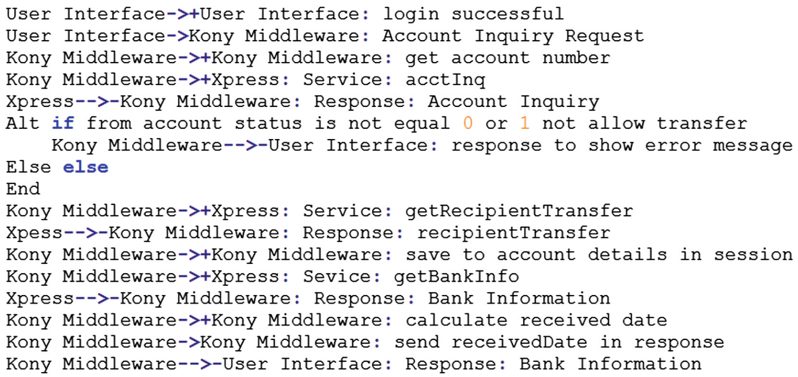

UML Sequence Diagram
Transform from diagram to description
UML Sequence Diagram (.txt)
Create DOT Markup script
Generate Report
Description of differences (.txt)
Compare differences with previous version
Visualize Traceability
Change Report
Previous Sequence Diagram (.txt)
Traceability Graph (.JPEG)

Fig. 1. Comparison of sequence diagrams for technical specification change.

3.1 Transform from Diagram to Description

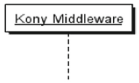
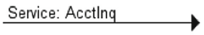
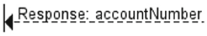
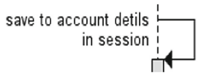
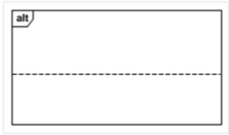
Initially, the input of UML Sequence Diagram is converted to description in text file (.txt). Example of a sequence diagram in .txt is shown in Fig. 2. Table 1 describes the notation and the syntax of some elements contained in the sequence diagrams created with WebSequenceDiagrams.



```
User Interface->+User Interface: login successful
User Interface->Kony Middleware: Account Inquiry Request
Kony Middleware->+Kony Middleware: get account number
Kony Middleware->+Xpress: Service: acctInq
Xpress-->-Kony Middleware: Response: Account Inquiry
Alt if from account status is not equal 0 or 1 not allow transfer
    Kony Middleware-->-User Interface: response to show error message
Else else
End
Kony Middleware->+Xpress: Service: getRecipientTransfer
Xpess-->-Kony Middleware: Response: recipientTransfer
Kony Middleware->+Kony Middleware: save to account details in session
Kony Middleware->+Xpress: Sevice: getBankInfo
Xpress-->-Kony Middleware: Response: Bank Information
Kony Middleware->+Kony Middleware: calculate received date
Kony Middleware->Kony Middleware: send receivedDate in response
Kony Middleware-->-User Interface: Response: Bank Information
```

Fig. 2. Example of sequence diagram in .txt file.

Table 1. Notation and syntax of sequence diagram elements by WebSequenceDiagrams.

Notation	Syntax of Sequence Diagram Elements
Life line 	***<web service> :*** Example User Interface :
Message Requset 	*<web service>* **->** *<web service> : message* Example Kony Middleware -> Xpress : Service: AcctInq
Message Response 	*<web service>* **-->-** *<web service> : message* Example Xpress -->- Kony Middleware : Response: Account Inquiry
Activation 	*<web service>* **->+***<web service>: message* Example Kony Middleware ->+ Kony Middleware : send receivedDate in response
Combined Fragement 	***alt*** *message* *<web service>* **-->-** *<web service>: message* ***else*** *message* *<web service>* **-->-** *<web service>: message* ***End*** Example alt If from account status is not equal 0 or 1 not allow transfer Kony Middleware -> User Interface : response to show error messge else *message* end

3.2 Compare Differences with Previous Version

The text comparison of sequence diagrams in .txt files is carried out line by line using Java function. The result is .txt file containing the description of differences, which in turn is the input for generating change report.

Figure 3 illustrates the example sequence diagram as the input for comparing with the previous version of sequence diagram as depicted in Fig. 4. The result of change report is generated as shown in Fig. 5. The change items reported in Fig. 5 are tantamount to those light grey frames denoting technical specification changes shown in Fig. 3.

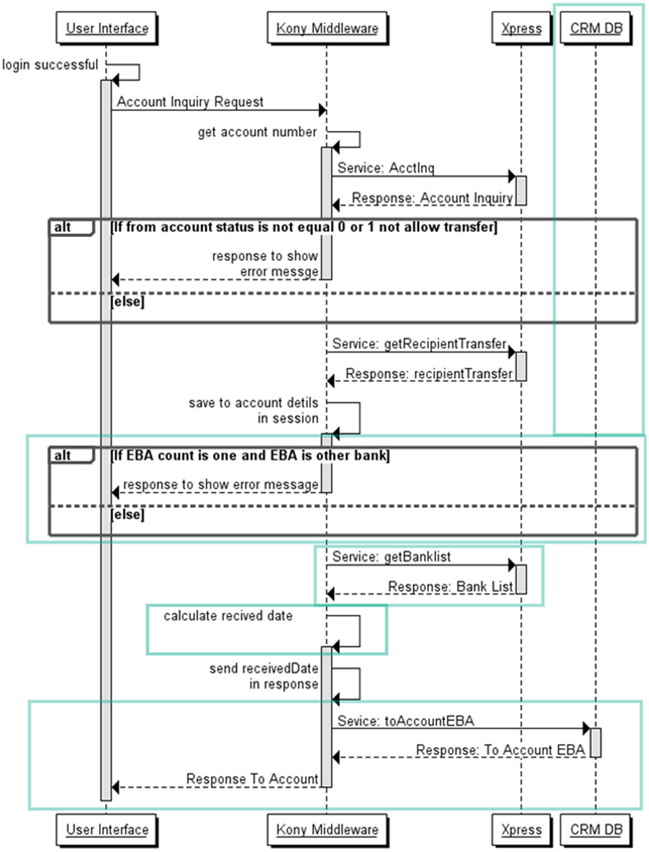

User Interface
Kony Middleware
Xpress
CRM DB
login successful
Account Inquiry Request
get account number
Service: AcctInq
Response: Account Inquiry
alt
[If from account status is not equal 0 or 1 not allow transfer]
response to show error messge
[else]
Service: getRecipientTransfer
Response: recipientTransfer
save to account detils in session
alt
[If EBA count is one and EBA is other bank]
response to show error message
[else]
Service: getBanklist
Response: Bank List
calculate recived date
send receivedDate in response
Sevice: toAccountEBA
Response: To Account EBA
Response To Account
User Interface
Kony Middleware
Xpress
CRM DB

Fig. 3. Example sequence diagram for comparison against the previous.

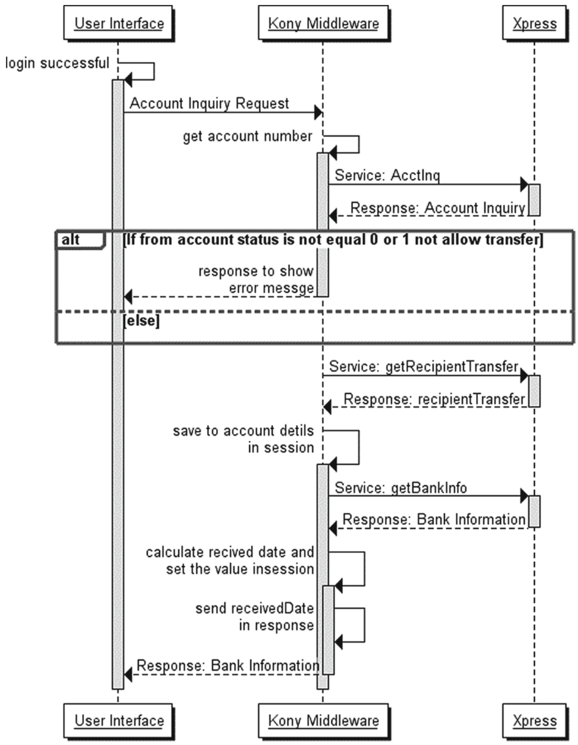


Fig. 4. Previous version of sequence diagram used for comparison.

Change Type	Previous version	New version
Add new service provider	Kony Middleware Xpress	Kony Middleware Xpress *CRMDB*
Rename web service	Service: getBankInfo	Service: *getBankList*
Rename Message Response	Response: Bank Information	Response: *Bank List*
Rename Activation	Kony Middleware: calculate received date add set the value in session	Kony Middleware: *calculate received date*

Add new alternative		*alt* If EBA count is one and EBA is other bank Kony Middleware -->- User Interface : response to show error message
		else
		end

Fig. 5. Result of generated change report.

3.3 Create DOT Markup Script

The graph description language, DOT [6], is a plain text for defining a graph but it does not provide facilities for rendering the graph. The mapping from sequence diagram components in .txt file to DOT syntax as shown in Table 2 is carried out using the algorithm described in Fig. 6. The result of DOT markup script associated with the input of sequence diagram (Fig. 3) is shown in Fig. 7.

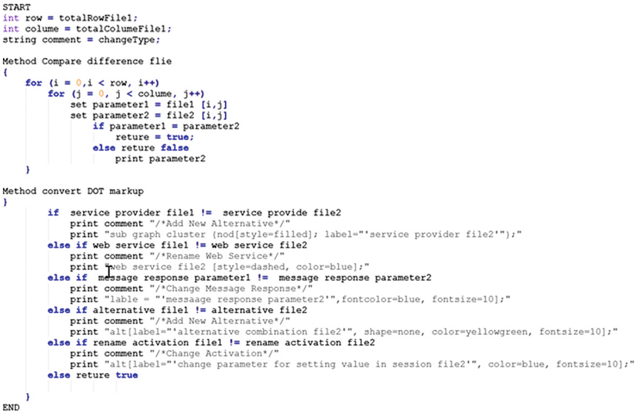
Table 2. Mapping from sequence diagram components (.txt) to DOT syntax.

Change Type	Sequence Diagram Component (.txt)	DOT Syntax

Add new service provider	Kony Middleware -> *CRMDB*	*Subgraph cluster {*
		Node [style=filled];
		Color = Green;
		Label = "CRMDB"
		……….
		}

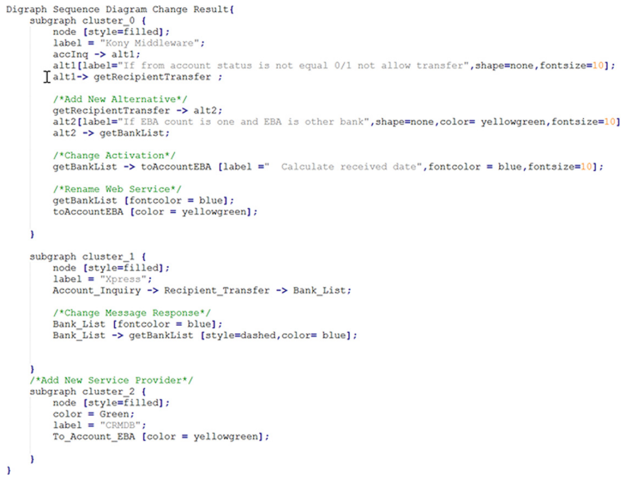
Rename web service	Kony Middleware -> Xpress: Service: *getBanklist*	*getBanklist [fontcolor = blue];*
Change Message Response	*Xpress -->- Kony Middleware : Response: Bank List*	*Bank_List -> getBanklist [style=dashed, color=blue];*
Change Activation	Kony Middleware ->+ Kony Middleware : *calculate received date*	*Label =”calulate received date”,fontcolor=blue, fontsize=10];*

Add new alternative	*alt* If EBA count is one and EBA is other bank
	Kony Middleware -->-User Interface: response to show error message
	else
	end
		Alt [label=”If EBA count is one add EBA is other bank, shape=none, color =yellowgreen, fontsize=10];



```
START
int row = totalRowFile1;
int colume = totalColumeFile1;
string comment = changeType;

Method Compare difference flie
{
    for (i = 0,i < row, i++)
        for (j = 0, j < colume, j++)
            set parameter1 = file1 [i,j]
            set parameter2 = file2 [i,j]
                if parameter1 = parameter2
                    reture = true;
                else reture false
                    print parameter2
    }

Method convert DOT markup
}
        if  service provider file1 !=  service provide file2
            print comment "/*Add New Alternative*/"
            print "sub graph cluster {nod[style=filled]; label="'service provider file2'"};"
        else if web service file1 != web service file2
            print comment "/*Rename Web Service*/"
            print "web service file2 [style=dashed, color=blue];"
        else if  message response parameter1 !=  message response parameter2
            print comment "/*Change Message Response*/"
            print "lable = "'messaage response parameter2'",fontcolor=blue, fontsize=10];"
        else if alternative file1 != alternative file2
            print comment "/*Add New Alternative*/"
            print "alt[label="'alternative combination file2'", shape=none, color=yellowgreen, fontsize=10];"
        else if rename activation file1 != rename activation file2
            print comment "/*Change Activation*/"
            print "alt[label="'change parameter for setting value in session file2'", color=blue, fontsize=10];"
        else reture true

    }
END
```

Fig. 6. Algorithm for comparing differences and creating DOT Markup script.



```
Digraph Sequence Diagram Change Result{
    subgraph cluster_0 {
        node [style=filled];
        label = "Kony Middleware";
        accInq -> alt1;
        alt1[label="If from account status is not equal 0/1 not allow transfer",shape=none,fontsize=10];
        alt1-> getRecipientTransfer ;

        /*Add New Alternative*/
        getRecipientTransfer -> alt2;
        alt2[label="If EBA count is one and EBA is other bank",shape=none,color= yellowgreen,fontsize=10]
        alt2 -> getBankList;

        /*Change Activation*/
        getBankList -> toAccountEBA [label ="  Calculate received date",fontcolor = blue,fontsize=10];

        /*Rename Web Service*/
        getBankList [fontcolor = blue];
        toAccountEBA [color = yellowgreen];

    }

    subgraph cluster_1 {
        node [style=filled];
        label = "Xpress";
        Account_Inquiry -> Recipient_Transfer -> Bank_List;

        /*Change Message Response*/
        Bank_List [fontcolor = blue];
        Bank_List -> getBankList [style=dashed,color= blue];


    }
    /*Add New Service Provider*/
    subgraph cluster_2 {
        node [style=filled];
        color = Green;
        label = "CRMDB";
        To_Account_EBA [color = yellowgreen];

    }
}
```

Fig. 7. DOT markup script associated with new version of sequence diagram.

3.4 Visualize Traceability

In this work, Graphviz is used as the tool for rendering the traceability graph from the DOT file obtained from the previous step. Figure 8 illustrates the traceability graph associated with the sequence diagram depicted in Fig. 3. The traceability graph of the previous version of sequence diagram (Fig. 4) is illustrated in Fig. 9.

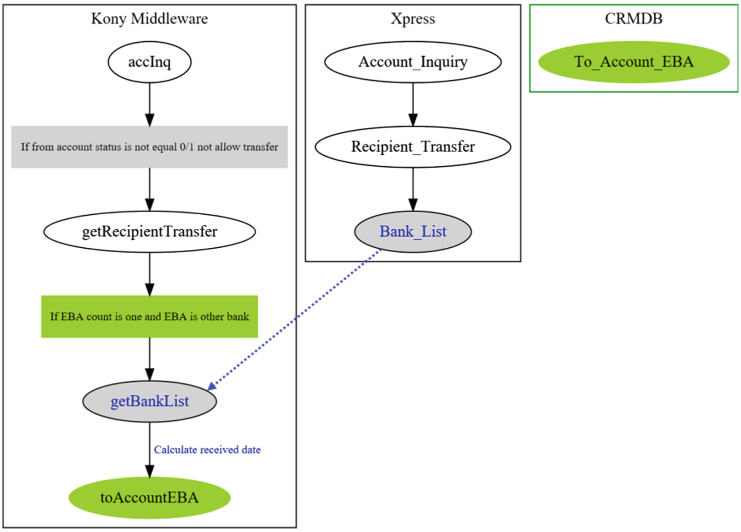


Fig. 8. Traceability graph associated with new UML sequence diagram.

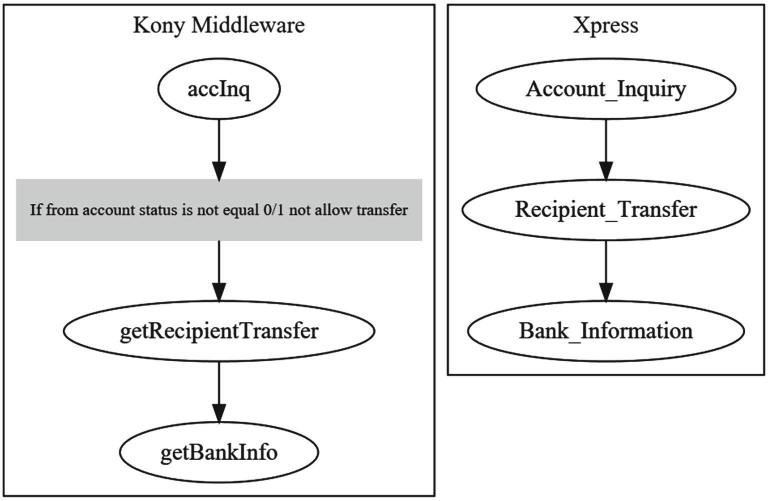


Fig. 9. Traceability graph of previous version of UML sequence diagram.

4 Conclusion and Future Work

Manually compare sequence diagrams to detect changes in technical specification is error-prone. The automation approach presented in this work would promote software process improvement in an organization, that is, reduce resource consumption, less rework and defects. Currently, the implemented system simply supports the comparison of sequence diagrams created with WebSequenceDiagrams. Further development would be the enhancement of the system to be able to compare UML sequence diagrams with the underlying XMI format.

References

1. Girschick M, Darmstadt T (2009) Difference detection and visualization in UML class diagrams. Technical report TUD-CS-2006-5, TU Darmstadt, Germany
2. GibertoFilho AA, Lencastre M (2012) Towards a traceability visualisation tool. International conference on the quality of information and communications technology. IEEE Computer Society, Washington, DC, pp 221–223
3. Dennis A, Wixom BH, Tegarden D (2012) System analysis and design with UML version 2.0: an object-oriented approach. 4th Edn. Wiley, New York
4. Li Y, Maalej W (2012) Which traceability visualization is suitable in this context? A comparative study. In: Proceedings of the 18th international conference on requirements engineering: foundation for software quality (REFSQ 2012). Springer, Heidelberg, pp 194–210
5. Gansner E, Koutsofios E, North S (2009) Drawing graphs with dot. http://www.graphviz.org/
6. Tantau T (2013) Graph drawing in tikz. Universität zu Lübeck, Germany

The Isolation Algorithm of Problem Location with Multi-agent Approach for End-to-End Network Performance Management

Buseung Cho[1], Kuinam J. Kim[2], and Hyuncheol Kim[3](✉)

[1] KREONET Operation and Service Division of Supercomputing,
Korea Institute of Science and Technology Information, Daejon, Korea
bscho@kisti.re.kr

[2] Department of Convergence Security, Kyonggi University, Suwon, Korea
kuinamj@gmail.com

[3] Department of Computer Science, Namseoul University, Cheonan, Korea
hckim@nsu.ac.kr

Abstract. With the dramatic increase in demands for high-performance network service, practically in advanced scientific research, as well as commercial internet service, the performance of data transfer for the huge amount of scientific experimental data and medical data, genome data is the key element that enables advanced global collaborative research. These activities require a guaranteed end-to-end (ETE) network performance and sophisticated network management framework for the ETE network. This paper suggests an isolation algorithm of problem location with multi-agent approach isolation algorithm of problem location with multi-agent approach ETE network performance management framework with Case-Base Reasoning (CBR) approach and multi-agent approach. It will enable a preliminary and proactive performance management for ETE network.

Keywords: Case-Based Reasoning · Network performance · Multi-agent approach · Flow monitoring

1 Introduction

With the dramatic increase in demands for high performance network service, practically in advanced scientific research, as well as commercial internet service, the performance of data transfer for the huge amount of scientific experimental data and medical data, genome data, and observation data is key element that enables advanced global collaborative research like high energy physics (HEP), astronomy, climate change, medical science, etc. If the performance of data transfer were not guaranteed, their researches couldn't achieve successful results and even more, the researches couldn't start to do. These applications require a guaranteed ETE (end-to-end) network performance and sophisticated network management framework for the ETE network. To cope with all these requirements, there is an increasing need for skilled experts to be able to rapidly

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_44

respond to issues and to use their experience, expertise, and knowledge to help solve users' problems quickly and efficiently.

The key issue to manage the network performance is how to simultaneously manage both network component and end-system component. It means, in order to build high-performance data transfer environment, the knowledge of network expert as well as the knowledge of end-system expert should be needed together. It is not easy for network operators alone to diagnose the fault of ETE network performance in totality. Also, the operator of the end-system couldn't comprehensively understand and troubleshoot the fault about the degradation of data transfer performance by himself. Furthermore, the administrators of network and system who is represented by experts in each field as well as application researchers who want to transfer huge amounts of data are mostly not familiar with cooperation and sharing the knowledge for figuring out the performance issue. For this reason, the concrete network performance management and a support system are definitely required [1].

This paper suggests a sophisticated ETE network performance management framework with Case-Base Reasoning (CBR) approach and multi-agent approach. It will enable a preliminary and proactive performance management for ETE network.

The organization of the paper is as follows. We first highlight some network performance enhancement and ETE flow measurement and monitoring technologies in Sect. 2. Detection the problem location with distributed multi-agent, isolation algorithm is presented in Sect. 3. Section 4 demonstrates the retrieval effectiveness of the proposed CBR system. Finally, the paper concludes in Sect. 5.

2 ETE Flow Measurement and Monitoring

2.1 Netflow and Argus Framework

In general, Netflow is used to collect flow data from Cisco. Though it is the most popular employed solution for flow measurement, there is a reliable issue about the gathered sampled flow records. Sampling normally decreases the amount of processed data and reduces the consumption of storage, but the reliability of the flow data depends on the sampling ratio. Tiago Fioreze, et al. investigated the trustfulness of measurement performed using the popular NetFlow monitoring solution with different sampling ratio when elephant flows are especially observed for the hybrid network [2, 3]. They showed that NetFlow provides reliable information regarding octets and packets. However, the flow duration reported when sampling is employed tends to be shorter than the actual duration.

As shown in Fig. 1, ARGUS is the network Audit Record Generation and Utilization System based on the flexible open source software developed by Carter Bullard [4]. ARGUS is next-generation network flow technology using advanced network flow data for network forensics and could be used for network operation, performance and security management. ARGUS is composed of the comprehensive network flow data generator, the ARGUS sensor that generates network bi-directional flow records at line rate.

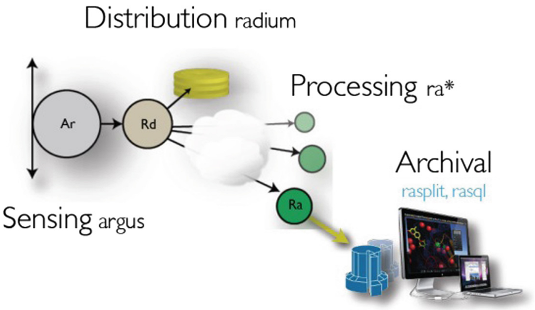


Fig. 1. ARGUS system design [4]

ARGUS is composed of sensing and sending network flow (argus), distribution and collecting argus flow data (radium), processing the argus flow data (ra*), archiving the flow data into the database (rasplit/rasql). ARGUS framework provides the direct use of argus data and databases are radium(), which collects the flow data and transmits it to a collecting node, and achieving the flow data into a database using API that is supported by ARGUS framework; the two MySQL programs, rasqlinsert() and rasql().

2.2 Flow-Based Large-Scale Network Monitoring

Richard A. Becker et al. attempted to visualize network data like network topology (link/node), network flow and geography in a single interface [5]. Practically in the visual cluster, displaying a great number of connection or flow with a single lie is challenging. SeeNet lets users adjust the visualization parameters of the map to manually reduce cluster and provides alternative designs to link maps. They also tried to display statistics from the internet that shows country-to-country traffic across the NSFNET/ANSnet2 backbone.

The Monitoring Agents in A Large Integrated (MonALISA) is global scale near real-time monitoring framework developed by Caltech and UPB to manage main data flows, traffic volume and quality of connectivity and used in USLHCNet production network including CMS, ALICE, ATLAS, and UltraLight [6]. It provides the presentation of each network topology like physical network layer, Layer 2 circuit network topology and Layer 3 routed network topology. But it's not easy to monitor the elephant flow data among all flow data on the specialized network with MonALISA framework.

GLORIAD Insight is developed by the GLORIAD (Global Ring Network for Advanced Application Development) Team to address improved operations, performance and security of research and education network as core cyber-infrastructure for deep monitoring of network traffic based on Argus network monitoring and measurement software and other open source software. As shown in Fig. 2, GLORIAD Insight supports rich array statistics about GLORIAD network about traffic load, packet loss, application protocols, and so on. It is offered as a community-built and maintained system-constructed solely on open-source tools and available.

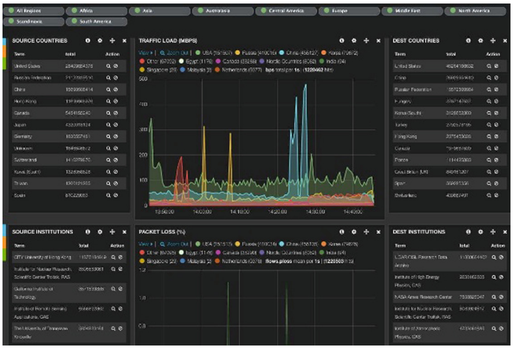


Fig. 2. GLORIAD insight system [7]

3 A Multi-agent Approach to ETE Network Performance Management

3.1 Distributed Multi-agent Architecture

In traditional media-shared LAN environment, there was significant performance degradation in simultaneously transferring data on several nodes. Today's the intelligent switch could separate collision domain with mesh topology among nodes and provides much high-performance transmission capability on more than 1Gbps network environment. The border network that interfaces between the campus network and external network like Internet service provider's carrier network has many issues about degrading network performance. For example, the performance of the border router including the queue size of the interface and security issues including the performance of the firewall and location of the firewall have become the potential cause of the performance degradation. In addition, the path MTU configuration could result in the performance degradation if it was not

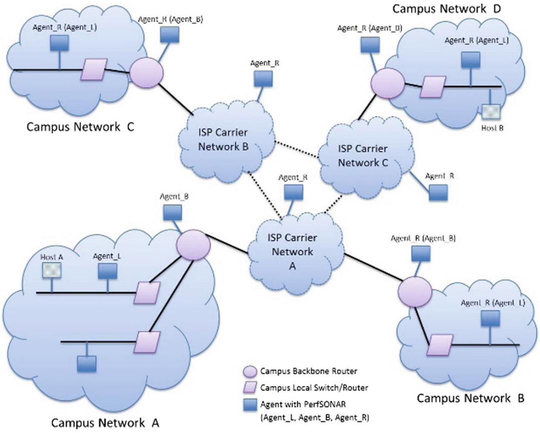


Fig. 3. General distributed multi-agent architecture

well configured in ETE whole path. The proposed multi-agent architecture, as shown in Fig. 3, detects the problems through testing the connectivity and the performance of the network and communicating with the agent in other network areas.

3.2 Isolation Algorithm for Problem Location with Multi-agent

This paper proposes an isolation algorithm for problem location with multi-agent for host x and host y can also be called, divide-and-conquer mechanism, in order to identify the problem location by interacting with agents. In order to identify the location of a performance problem, the result of throughput test from host x to host y and the result of throughput test to each agent is compared and then the problem location is determined. Practically by comparing throughput from host x to Agent_R and throughput from x to host y, problem domain is narrowed to a specific domain.

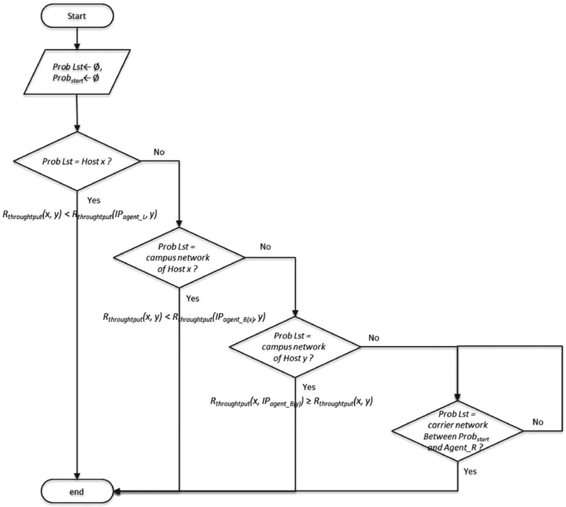

Start
Prob Lst← ∅,
$Prob_{start}$← ∅
Prob Lst = Host x ?
No
Yes
$R_{throughtput}(x, y) < R_{throughtput}(IP_{agent\_L}, y)$
Prob Lst = campus network of Host x ?
No
Yes
$R_{throughtput}(x, y) < R_{throughtput}(IP_{agent\_B(x)}, y)$
Prob Lst = campus network of Host y ?
No
Yes
$R_{throughtput}(x, IP_{agent\_B(y)}) \geq R_{throughtput}(x, y)$
Prob Lst = carrier network Between $Prob_{start}$ and Agent_R ?
No
Yes
end

Fig. 4. Flowchart of isolation algorithm for problem location

First of all, if the host itself has a problem issue or not is determined by testing throughput from an Agent_L to host y. Thus, the host that transfer the data is checked for the performance issue of the host like transfer tool, status of NIC, congestion control algorithm, tcp_timestamps, tcp_mem in the kernel and so on. Next through the throughput test from Agent_B that connect to backbone router of host x's campus network, if the campus network has a problem or not could be identified. Then through testing throughput to an Agent R that connect to a border router in host y's campus network, if there is a problem issue in campus network of host y or not is checked. Finally, the ISP's carrier network between campus network of host x and campus network of host y is checked and limit the range of a problem network domain. Figures 4 and 5 shows the flow chart and pseudocode for the isolation algorithm for problem location with multi-agent.

Agent_L Lst: a list of agent_L of host x, element (IP_{agent_L})
Agent_B(x) Lst: a list of agent_B of host x, element (IP_{agent_B})
Agent_B(y) Lst: a list of agent_B of host y, element (IP_{agent_B})
Agent_R Lst: a list of agent_R, element (IP_{agent_R})
$R_{throughtput}$(a, z): the result of throughput test from host (or agent) a to host (or agent) z
Prob Lst: a list of problem location between agent α and agent β , element (α, β) or (host)
$Prob_{start}$: an element (IP_{agent_R}) that start point of problem domain in ISP’s carrier network

```
Prob Lst ←∅
Probstart ←∅
for each Agent_L Lst do
  if Rthroughtput(x, y) < Rthroughtput(IPagent_L, y)
    Prob Lst ←host x
  end if
for each Agent_B(x) Lst do
  if Rthroughtput(x, y) < Rthroughtput(IPagent_B(x), y)
    Prob Lst ←Prob Lst ∪ (host x, Agent_B(x))
  end if
 for each Agent_B(y) Lst do
  if Rthroughtput(x, IPagent_B(y)) ≥ Rthroughtput(x, y)
    Prob Lst ← Prob Lst ∪ (Agent_B(y), host y)
  end if
 for each Agent_R Lst do
    if Rthroughtput(x, y) < Rthroughtput(x, IPagent_R)
      Probstart ← IPagent_R
    if Rthroughtput(x, y) ≥ Rthroughtput(x, IPagent_R)
      Prob Lst ← Prob Lst ∪ (Probstart, Agent_R)
    end if
```

Fig. 5. Pseudo code of isolation algorithm for problem location

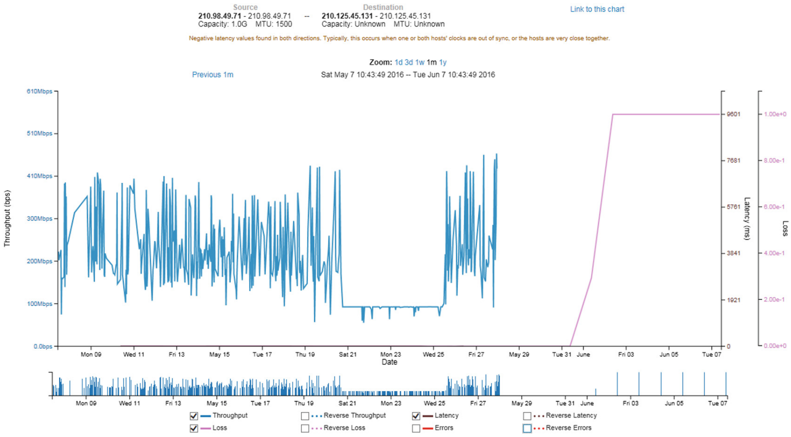
4 Experimental Results

This paper proposed a multi-agent architecture based on perfSONAR nodes and it is implemented by integration with perfSONAR in KREONET and GLORIAD network. PerfSONAR is a service-oriented architecture and all those services of perfSONAR communicate with each other using well-defined protocols called as perfSONAR. There are 16 perfSONARs that cover all KREONET regional networks in Korea and 4 perfSONARs in GLORIAD network is composed of 4 international PoP: Settle, Chicago, HongKong and Amsterdam. All perfSONAR are directly connected with backbone router or switch with 10GbE or 1GbE. In order to test throughput between two perfSONAR nodes, the Bandwidth Test Controller (BWCTL), a command line client application and scheduling demon, invokes the network measurement tool of perfSONAR including IPERF, PING, and TRACEROUTE.

4.1 Setting up Lightpath and Science DMZ

This case is related to performance enhancement activity using Lightpath and Science DMZ. Lightpath can provide guaranteed bandwidth with ETE circuit provisioning and Science DMZ can hinder the performance degradation owing to security system such as firewall, IPS, D-DOS system etc.

Figure 6 shows throughput and loss from a perfSONAR in APEC Climate Center (APCC) in Busan to a perfSONAR in Korea Meteorological Administration Supercomputing Center (KMASC) in Ochang. The perfSOANR in APCC connected to a backbone router of APCC. A firewall is located between the APCC backbone router and KREOENT (WAN). The perfSONAR in KMASC is located at DMZ zone of KMASC and firewall, IPS, D-DOS system is between the perfSOANR and KREOENT. The throughput was affected by several security systems in APCC and KMASC campus network.



Source Destination
210.98.49.71 - 210.98.49.71 -- 210.125.45.131 - 210.125.45.131
Capacity: 1.0G MTU: 1500 Capacity: Unknown MTU: Unknown

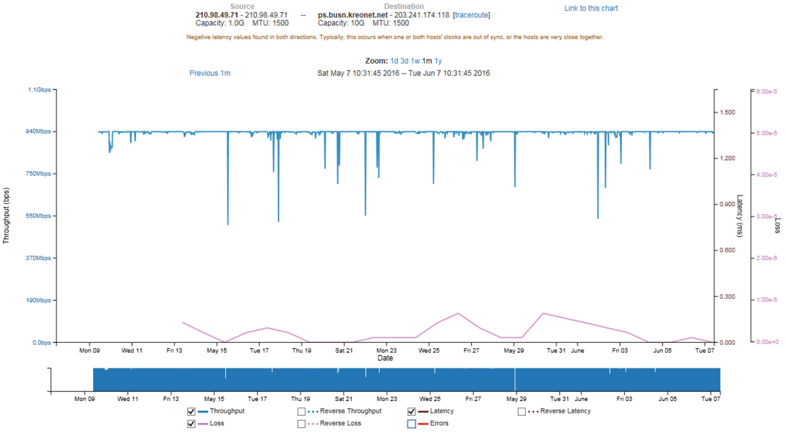
Negative latency values found in both directions. Typically, this occurs when one or both hosts' clocks are out of sync, or the hosts are very close together.

Sat May 7 10:43:49 2016 -- Tue Jun 7 10:43:49 2016

Fig. 6. Throughput and loss from APCC in Busan to KMASC in Ochang

Figure 7 shows that the throughput from APCC and KMASC is about 400Mbps. Throughput is tested and measured by the iperf tool in the perfSONAR. Figure 7 is a test result for throughput between the perfSOANR in APCC and a perfSOANR in KREONET Busan PoP just after the 1Gbps lightpath between APCC and KREONET

Busan PoP was provisioned. In this environment, the throughput is about 940Mbps. Then setting a lightpath between the perfSOANR in APCC and the perfSOANR in KMASC and Science DMZ for each campus network increased the performance of file transfer with Gridftp from about 300Mbps to 700Mbps. Figure 8 shows the performance for file transfer with Gridftp just before and after the configuration of lightpath and science DMZ.



Source — Destination
210.98.49.71 - 210.98.49.71 -- ps.busn.kreonet.net - 203.241.174.118 [traceroute]
Capacity: 1.0G MTU: 1500 — Capacity: 10G MTU: 1500

Link to this chart

Negative latency values found in both directions. Typically, this occurs when one or both hosts' clocks are out of sync, or the hosts are very close together.

Zoom: 1d 3d 1w 1m 1y

Previous 1m — Sat May 7 10:31:45 2016 -- Tue Jun 7 10:31:45 2016

☑ Throughput □ Reverse Throughput ☑ Latency □ Reverse Latency
☑ Loss □ Reverse Loss □ Errors

Fig. 7. Throughput and loss from APCC in Busan to KREONET Busan PoP in Busan

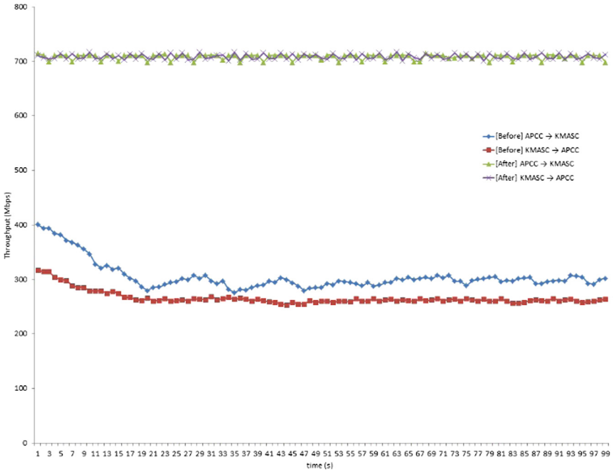


Fig. 8. Throughput from APCC in Busan to KMASC in Ochang, just after setting up Lightpath and Science DMZ

5 Conclusion

In the traditional network management framework, the performance was one of the managed objects among fault, configuration, accounting, performance, security (FCAPS) and it was handled by monitoring the network equipment itself. However, the performance has become a most important object that should be strictly managed today. Also, various other technologies such as artificial intelligent technology can be integrated to handle the performance issue.

In order to manage ETE network performance, a lot of activities in an automatic and intelligent method such as advanced artificial intelligent technology have been adapted. It's because of significant network complexity and heterogeneity. Especially demands high performance network service in advanced data-intensive scientific research are dramatically increased. Thus a sophisticated network management framework for the ETE network is required to manage the network performance. However, most support systems to monitor and troubleshoot the issues of network performance have not sufficient features to fulfill the requirement of the network operator and user.

This paper suggests a sophisticated ETE network performance management framework with CBR approach and multi-agent approach. It will enable a preliminary and proactive performance management for ETE network. The National Research Network of Korea, KREONET, as experimental environment and KREONET perfSONAR system as multi-agent are described. A case is introduced as a representative case in casebase.

References

1. Gingrich BL, Minden GJ (1990) MANDOLIN—a communications management expert system using a reduced form of the Dempster-Shafer uncertainty theory. In: The 3rd international conference on industrial and engineering applications of artificial intelligence and expert systems, vol 1, pp 76–85, January 1990
2. Fioreze T, Granville LZ, Pras A, Sperotto A, Sadre R (2009) Self-management of hybrid networks: can we 114 trust netflow data? IM 2009, IFIP/IEEE international symposium
3. Fioreze T, Wolbers MO, van de Meent R, Pras A (2007) Finding elephant flows for optical networks. In: 10th IFIP/IEEE international symposium on integrated network management, May 2007
4. Bullard C (2014) Building large network monitoring systems with argus. In: FloCon 2014, January 2014
5. Becker RA, Eick SG, Wilks AR (1995) Visualizing network data. IEEE Trans Vis Comput Graph 1(1):16–21
6. Dobre C, Voicu R, Legrand I (2011) Monitoring large scale network topologies. In: The 6th IEEE international conference on intelligent data acquisition and advanced computing system: technology and applications, September 2011
7. GLORIAD Insight System. https://insight.gloriad.org

Revised Virtual Resources Allocation Scheme in Network Function Virtualization (NFV) Enabled Networks

Hyuncheol Kim(✉)

Department of Computer Science, Namseoul University, Cheonan, Korea
hckim@nsu.ac.kr

Abstract. Network function virtualization (NFV) is a technology that abstracts all types of resources needed for networking and enables automatic management and control of network services by software. With the introduction of NFV technology, telecom operators want to gain the benefits and efficiency of increasingly complex network manageability, reduced administrative costs, and network agility. One of the most important considerations in NFV deployment is how to allocate the virtual resources that are needed to provide flexible virtual network services in an NFV-based network infrastructure. Thus, the most important prerequisite for NFV deployment is achieved fast, scalable and dynamic composition and allocation of networks functions (NFs) to implement network services (NSs). In this paper, we proposed a revised on-line RA algorithm that integrates embedding and scheduling of virtual network functions (VNFs) simultaneously.

Keywords: Combined virtual resource allocation · NFV-enabled network · Chain composition · Resource embedding

1 Introduction

As telecommunication service develops, support for new services reflecting the requirements of users in telecommunication network operators is fundamentally enormous (CAPital EXpenditure)/OPEX (OPerating EXpense) due to the introduction of new network equipment and change of physical infrastructure every time. Network service provider also could not guarantee the agility of providing new services because it required expenditures and once the installed equipment could not be moved or changed easily.

With the evolution of virtualization technologies such as hypervisors, network and storage virtualization, increased open-hardware utilization such as x86, and the proliferation of open software such as open stacks and open vSwitches, Network Function Virtualization (NFV), a new network infrastructure that runs network functions (NFs) such as transcoder, routing, and firewall on general high-performance servers such as blade servers, has emerged.

Funding for this paper was provided by Namseoul university.

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_45

Network function virtualization (NFV) is a technology that abstracts all types of resources needed for networking and enables automatic management and control of network services by software. With the introduction of NFV technology, telecom operators want to gain the benefits and efficiency of increasingly complex network manageability, reduced administrative costs, and network agility. In the NFV network, existing specific network functions are implemented as one or more software module(s), which is called VNF(s) (Virtual Network Functions). In other words, VNF is a virtualized instance of the existing NF, and VNF enables individual management with modularization and individualization of network functions. In addition, VNFs can be deployed anywhere on the network because VNFs can be installed and deployed on a general-purpose server and dynamically migrate between servers [1, 2].

One of the most important considerations in NFV deployment is how to allocate the virtual resources that are needed to provide flexible virtual network services in an NFV-based network infrastructure. Thus, the most important prerequisite for NFV deployment is achieved fast, scalable and dynamic composition and allocation of networks functions (NFs) to implement network services (NSs) [2]. In this paper, we proposed a revised on-line RA algorithm that integrates embedding and scheduling of virtual network functions (VNFs) simultaneously.

The organization of the paper is as follows. In Sect. 2, we first review previous research on VNF embedding and scheduling briefly. In Sect. 3, we present the improved online combined VNF embedding and scheduling algorithm proposed in this paper. Finally, Sect. 4 describes the conclusion and further study of this paper.

2 Research on VNF Embedding and Scheduling

As mentioned earlier, one of the most important considerations in NFV deployment is how to allocate the virtual resources that are needed to provide flexible virtual network services in an NFV-based network infrastructure. This virtual resource allocation related issues are called an NFV resource allocation (RA) problem. Thus, the most important prerequisite for NFV deployment is achieved fast, scalable and dynamic composition and allocation of networks functions (NFs) to implement network services (NSs). However, since NS requires multiple sets of VNFs, it is necessary to solve the following two problems in order to effectively perform network service coordination and management in NFV infrastructure. (1) How do you compose VNFs for a given NS? (2) How to effectively allocate and schedule VNFs that make NS on the substrate network (SN)?

The RA process generally consists of three steps as follows: (1) VNFs Chain composition (VNFs-CC), (2) As shown in Fig. 1, VNF Forwarding Graph Embedding (VNF-FGE) and (3) VNFs Scheduling (VNFs-SCH) [2–4].

In ETSI, NS is defined as an entity composed of an ordered number of VNFs. In other words, to provide NS, the packet must traverse the set of VNFs in a certain order. Because VNF is software, one of the major problems is how to effectively concatenate the different VNFs to form NS. Therefore, the first process is the chaining process, which is called CC [5, 6].

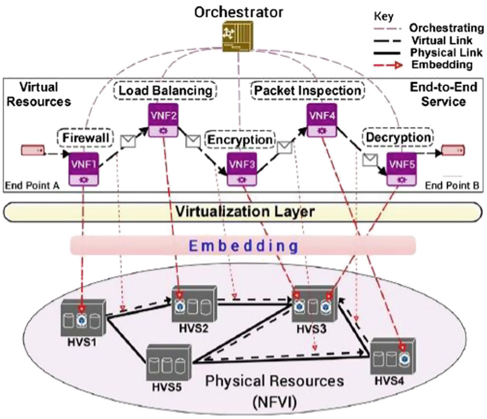


Fig. 1. An example of VNF forwarding graph embedding [2]

The VNF-FGE process is the second stage of NFV-RA. The VNF-FGE is looking for where to allocate VNFs in the network infrastructure in an appropriate way, taking into account several NS requests (requirements). In VNF-FGE, resource optimization must also be executed to perform various purposes of NS, such as maximizing remaining network resources, minimizing power consumption of SN, optimizing for specific QoS metrics, and so on [6, 7].

The last stage of the NFV-RA is the scheduling process and is also referred to as VNF scheduling. The VNF scheduling stage is to find a solution that minimizes the execution time without degrading the service performance and without violating the dependencies and precedence between the VNFs constituting the NS. Because the NFV infrastructure consists of several different HVSs, proper scheduling of VNFs execution can reduce the overall execution time and thus improve performance.

3 Revised Resource Allocation Algorithm

The definitions of the variables and functions required to describe the revised NFV-RA operation are as follows. Figure 2 describes the details of the proposed algorithm using these variables and functions.

- N: A set of all virtual nodes, $N = \{1, 2, 3, 4, \ldots, n\}$
- S: NS, consists of m sequential VNF, $F = \{1, 2, 3, 4, \ldots, m\}$
- $F = \{1, \ldots., \mathrm{m}\}$: function (VNF), function $1 <= i <= m$
- $\rho_{i,j}$: Processing time of VNF i at node j
- δ_i: The buffer used by the node to which function i is mapped
- B_j: At some point, the available buffer size at node j
- $\beta_{i,j}$: 1 if node j can handle function i, 0 otherwise
- t_l: Deadline time for which service to process
- t_i: Completion time of VNF i

```
function scale_ra ($S, N, T$)
{
    // $T = \{Network\ distance, Security\}$
    matrix[][] = init_matrix($N$);
    dist[][] = init_netdist($N$);   //$\omega_j$
    *graph[] = matrix;
    for (fucntion $i \in S$)
    {
        $N' = N'' = N''' = \{\square\}$
        if ($i = 1$)   $t_{i-1} = t_a$

        for (Node $j \in N$) {
            $t_e = \rho_{i,j} + \max(\pi_j, t_{i-1}$) + $\omega_j$// Ending condition
            if ($((\beta_{i,j} == 1) \wedge (B_j \geq \delta_i) \wedge (t_e \leq t_l)$) then
                $N' = N' \cup n$
        }
        if ($N' = 0) \{Reset\ Substrate\ network\ status; return\ 1;\}$
        for(Node $j \in N'$)   if ($((\mu_j - t_c) \geq \rho_{i,j}$) then   $N'' = N'' \cup n$
        if($|N''| == 1)$ then node_mapping($i, N''$)
        else if ($|N''| > 1)$ then
        {
                sort($N'', \omega_j$)
                if($|select\_tops(N'')| == 1$) then node_mapping($i, select\_tops(N'')$)
                else node_mapping($i, select\_random\_tops(N'')$)
        }
        else if ($(\mu_j - t_c) \geq 1$) then  //some TU, but not satisfiable
        {
                for(Node $j \in N''$)   $calc\_interference(i)$   //$\phi_i$
                sort($N'', \phi_i$)
                if($|select\_tops(N'')| == 1$) then node_mapping($i, select\_tops(N'')$)
                else node_mapping($i, select\_random\_tops(N'')$)
        }
        else
        {
                for(Node $j \in N'$)   if ($((\beta_{i,j} == 1)$) then $N''' = N''' \cup n$
                sort($N''', \pi_j$)   //expected completion time
                if($|select\_tops(N''')| == 1$) then node_mapping($i, select\_tops(N''')$)
                else node_mapping($i, select\_random\_tops(N''')$)
        }

        update $\pi_j$
        $t_i = max(\pi_j, t_{i-1}$)
        update $B_j$,  $t_{i-1}$
    }
}
```

Fig. 2. Revised VNF allocation algorithm

- t_c: Current time of VNF i
- t_a: The time at which mapping and scheduling requests for the service arrived on the physical network
- π_j: Expected completion time of the last function waiting for processing at the corresponding virtual node j
- μ_i: Start time of the first function waiting for processing at the corresponding virtual node j
- ω_j: The propagation delay time reflecting the network distance (or diameter, hop count) from node s to node j
- $\phi_{i,j}$: Degree of Interference, the number of VNFs to be relocated due to lack of time unit (TU) when assigning VNF i to virtual node j

A network example to illustrate the operation of the NFV-RA algorithm is shown in Fig. 3. As shown in Fig. 3, the network consists of 7 HVS nodes from *n1* to *n7* and supports 8 VNFs from VNF *f1* to *f8*.

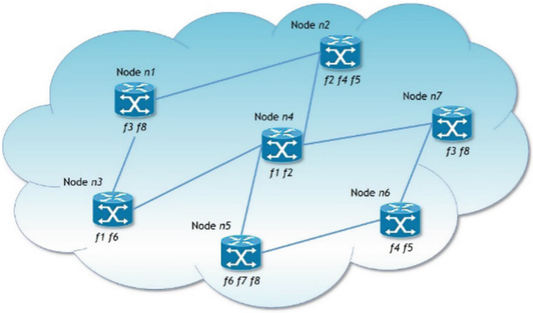


Fig. 3. Node capabilities and architecture of example node

To see the operation of the proposed algorithm, we assume that NS *S1* = {*f8* - *f2* - *f3* - *f6* - *f5*} requests arrive at T1 time. It is also assumed that 3, 2, 4, 2, and 3 TUs are required for VNF *f8, f2, f3, f6,* and *f5*, respectively.

1. If the scheduling TU space for VNF allocation in one or more nodes is empty

If only one node has space for VNF *f8*, the VNF is assigned to that node. As shown in Fig. 4 (a), at the time of T1, only *n1* that can support VNF *f8*. If the space is empty in one or more nodes as shown in Fig. 4 (b), the node with the lowest ω_j value representing the network distance from the previous VNF service node to the empty node (e.g., *n1, n5, n7*) is selected to assign VNF *f8*.

2. The space of nodes for VNF allocation is not as empty as the desired TU, but if the TU value is greater than 1 (TU > 1).

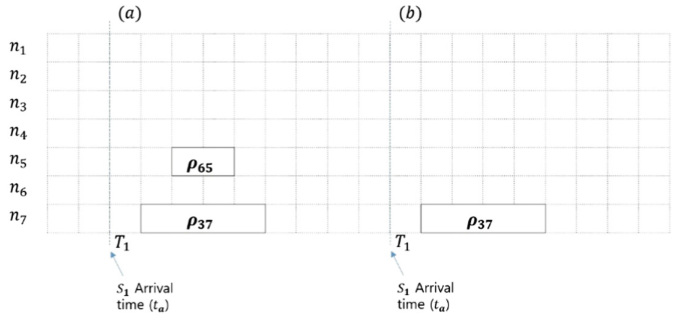


Fig. 4. If at least one space of nodes for VNF allocation is empty

Figure 5 shows that NS *S1* = {*f3* - *f2* - *f1* - *f5* - *f8*} and *S2* = {*f3* - *f4* - *f5*} are already scheduled when the NS *S3* request arrives at T1. In this situation, NS *S3* = {*f3* … .} arrives, only two TUs remain in the spare space because of the already scheduled *n1* ρ_{31} and *n7* ρ_{37}, the direct allocation is difficult due to the nature of VNF *f3* requiring 3 TUs and the relocation process must be performed.

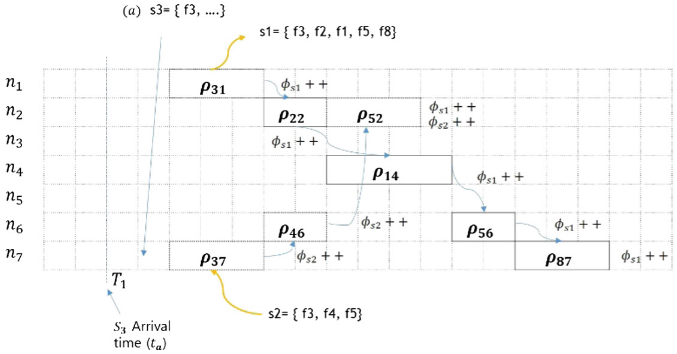


Fig. 5. If the space of nodes for VNF allocation is not empty

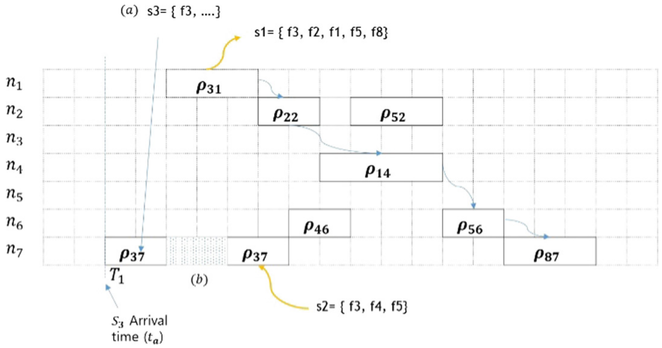


Fig. 6. The scheduling table after the VNF *f3* of NS *S3* is rescheduled

In order to reschedule *S1*, $\rho_{31}, \rho_{22}, \rho_{52}, \rho_{14}, \rho_{56}, and\, \rho_{87}$ must be rescheduled for 1 TU, and to reschedule *S2*, $\rho_{37}, \rho_{46}, and\, \rho_{52}$ must be rescheduled for 1 TU. Therefore, since $\phi_{3,1}$ is 6, and $\phi_{3,7}$ is 3, *S3* is scheduled to *n7* as shown in Fig. 6, and $\rho_{37}, \rho_{46}, and\, \rho_{52}$ belonging to *S2* are rescheduled for 1 TU. Figure 6 (b) shows the time in which the VNF supporting the *f3* in *S2* and the *f3* in *S3* are executed in parallel.

4 Conclusion

One of the most important considerations in NFV deployment is how to allocate the virtual resources that are needed to provide flexible virtual network services in an NFV-based network infrastructure. Thus, the most important prerequisite for NFV deployment is achieved fast, scalable and dynamic composition and allocation of networks functions (NFs) to implement network services.

In this paper, we introduces the concept of degree of interference $\phi_{i,j}$, the number of VNFs to be relocated due to lack of TU when assigning VNF *i* to virtual node *j*.

Simulation results show that the VNF completion time is similar to that of the revised RA algorithm and the Mijumbi algorithm when the NS arrival rate is low, but the VNF completion time is reduced by 14 ~ 16% when the NS arrival rate is concentrated.

References

1. Kim H (2017) Virtual Resources Allocation Scheme in ICT Converged Networks. Lecture Notes in Electrical Engineering, vol 424, pp 757–762
2. Herrera JG, Botero JF (2016) Resource allocation in NFV: a comprehensive survey. IEEE Trans Netw Serv Manage 13(3):518–532
3. Mijumbi R, Serrat J, Gorricho J, Bouten N, De Turck F, Boutaba R (2015) Network function virtualization: state-of-the-art and research challenges. IEEE Commun Surv Tutor PP(99):1–1
4. Riera JF, Hesselbach X, Escalona E, García-Espin JA, Grasa E (2014) On the complex scheduling formulation of virtual network functions over optical networks. In: International conference on transparent optical networks (ICTON), pp 1–5
5. Beck MT, Botero JF (2015) Coordinated allocation of service function chains. In: IEEE global communications conference, pp 1–6
6. Mijumbi R, Serrat J, Gorricho J-L, Bouten N, Turck F (2015) Design and evaluation of algorithms for mapping and scheduling of virtual network functions. In: IEEE conference on network softwarization (NetSoft), pp 1–9
7. Kao H-Y, Yang Y-M, Huang C-H (2015) Dynamic virtual machines placement in a cloud environment by multi-objective programming approaches. In: International conference on intelligent informatics and biomedical sciences (ICIIBMS), pp 364–365

Dynamic Information Extraction and Integrity Verification Scheme for Cloud Security

Hyunjoo Kim[1], Youngsoo Kim[1], Ikkyun Kim[1], and Hyuncheol Kim[2](✉)

[1] Information Security Research Division, Electronics and Telecommunications Research Institute, Daejeon, Korea
{hjookim,blitzkrieg,ikkim21}@etri.re.kr
[2] Department of Computer Science, Namseoul University, Cheonan, Korea
hckim@nsu.ac.kr

Abstract. To become a more popular cloud service, it is necessary to dynamically provision virtualized infrastructure resources and to automatically deploy and optimize workloads based on the state of the workload or the state of the entire infrastructure resource. However, in a cloud-based virtualization infrastructure, when multiple VMs work together to provide a customized virtualized network security service, existing debugging and profiling tools can no longer be used as performance measures or integrity verification tools. In order to solve these drawbacks, a tracing method is used. In the tracing, necessary information is simultaneously recorded with minimal overhead while executing the program. In this paper, we proposed a scheme to guarantee the integrity of the software that composes the VM in the cloud environment using Intel processor trace (PT).

Keywords: Processor trace · Cloud security · Information integrity

1 Introduction

To become a more popular cloud service, it is necessary to dynamically provision virtualized infrastructure resources across multiple workloads and user groups and to automatically deploy and optimize workloads based on the state of the workload or the state of the entire infrastructure resource. The life cycle of the virtual infrastructure should be systematically managed. In addition, development and technical advancement of cloud operating system and system management software that provide these functions must be supported.

Cloud computing has the advantage that users can use the desired service anytime if only the Internet environment is installed without having to install the program on a PC, and the data can be linked with various devices because it is located on-line. In addition, there is no need to install software for each device, which helps to save IT costs dramatically. On the other hand, even if the storage space of cloud computing is sufficient, users cannot provide all applications, so it can be difficult to install application or service, and if a server is attacked, personal information may be leaked. Therefore, security issues such as information leakage and personal information are becoming the most serious threats to cloud computing services as shown in Fig. 1 [1, 2, 4–6].

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_46

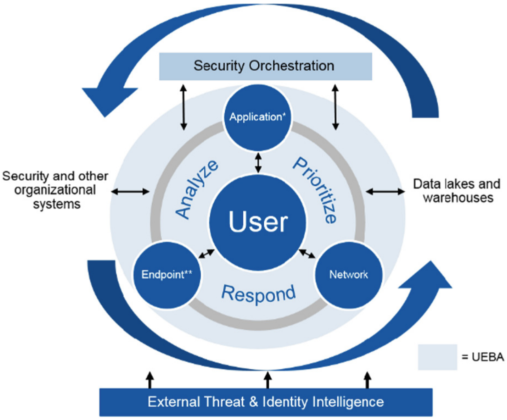

Security Orchestration
Application*
Analyze
Prioritize
User
Endpoint**
Network
Respond
Security and other organizational systems
Data lakes and warehouses
= UEBA
External Threat & Identity Intelligence

Fig. 1. User and entity behavior analytics for security intelligence (Gartner, 2015)

On the other hand, a virtualization technology that performs various programs on general-purpose hardware based on virtual machines (VMs) became the core technology of cloud infrastructure. However, in a cloud-based virtualization infrastructure, when multiple VMs work together to provide a customized virtualized network security service, existing debugging and profiling tools can no longer be used as performance measures or integrity verification tools [3, 7]. In order to solve these drawbacks, a tracing method is used. In the tracing, necessary information is simultaneously recorded with minimal overhead while executing the program. In tracing, tracepoint can be set statically/dynamically similar to breakpoints in order to gather the necessary information. The main uses of tracepoint are parts of complex and critical system software, such as OS kernel or security-related programs. Solaris, Linux, and Windows support tracing, and tracing can be run in both kernel mode and user space mode.

In this paper, we proposed a scheme to guarantee the integrity of the software that composes the VM in the cloud environment using Intel processor trace (PT). The formation of the paper is as follows. Section 2 describes the overall contents of the PT used to ensure the integrity of the VM software. Section 3 describes the flow trace analyzer structure that extracts VM dynamic information and reconstructs flow using PT. Finally, the paper concludes in Sect. 4.

2 Hardware-Based Tracing

Tracing can generally be defined as “a technique used to understand what is going on in a system in order to debug or monitor it”. In software engineering, tracing is a special form of logging that records information about the execution of a program. The information collected is typically used by the programmer for debugging purposes, and the system administrator or technical support personnel and software monitoring tools are used to diagnose common software problems according to the type and details of the information contained in the trace log.

2.1 Intel Processor Tracing

As described above, in a cloud-based virtualization infrastructure, when multiple VMs cooperate to provide one customized security service (e.g. NFV NS (Network Function Virtualization Network Service)), existing debugging or profiling tools no longer perform performance measurement or integrity verification cannot be used.

As shown in Fig. 2, Intel PT is a hardware-based function that stores all the information about software execution with minimal overhead to system operation. The PT software decoder can extract accurate software execution flow from the trace log with less than 5% overhead. PT can also store cycle count and timestamp information for synchronization with other traces. In addition, the PT does not require modification of source code and does not require OS sideband information such as context switches and address space modification and only the object code is needed for decode tracing. Figure 3 shows the various packet types used in the PT.

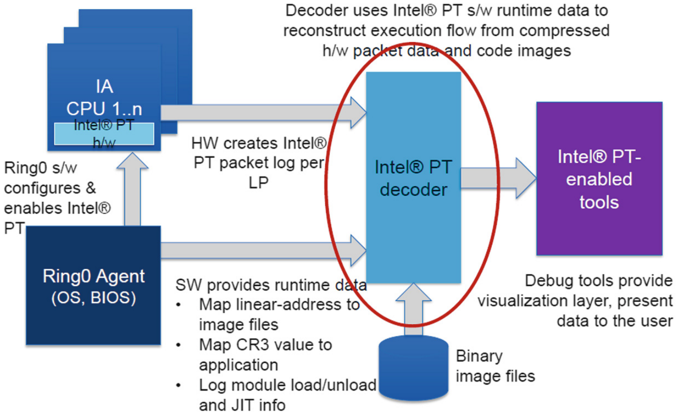

Decoder uses Intel® PT s/w runtime data to reconstruct execution flow from compressed h/w packet data and code images
IA CPU 1..n
Intel® PT h/w
HW creates Intel® PT packet log per LP
Intel® PT decoder
Intel® PT-enabled tools
Ring0 s/w configures & enables Intel® PT
Ring0 Agent (OS, BIOS)
SW provides runtime data
• Map linear-address to image files
• Map CR3 value to application
• Log module load/unload and JIT info
Binary image files
Debug tools provide visualization layer, present data to the user

Fig. 2. Intel PT architecture

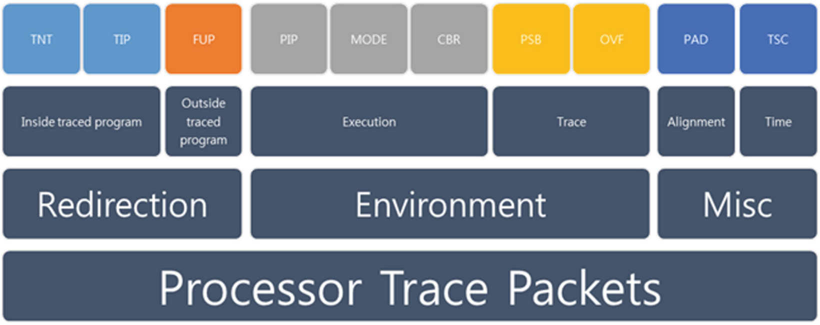


Fig. 3. Intel PT packet types

3 Proposed Flow Trace Analyzer Architecture

As shown in Fig. 4, the flow trace analyzer proposed in this paper is based on simple-pt and libipt, a PT decoder library [3, 7]. Simple-pt is a simple reference implementation on the Linux Kernel module and userspace by Andi Kleen to control the PT. Therefore, simple-pt uses a decoder based on libipt and uses a simple shell script to control the simple-pt module. Simple-pt consists of the following four elements: (1) kernel driver (2) sptcmd to collect data from the kernel driver (3) sptdecode for decoding PT information (4) Fastdecode for dumping Raw PT trace information [7].

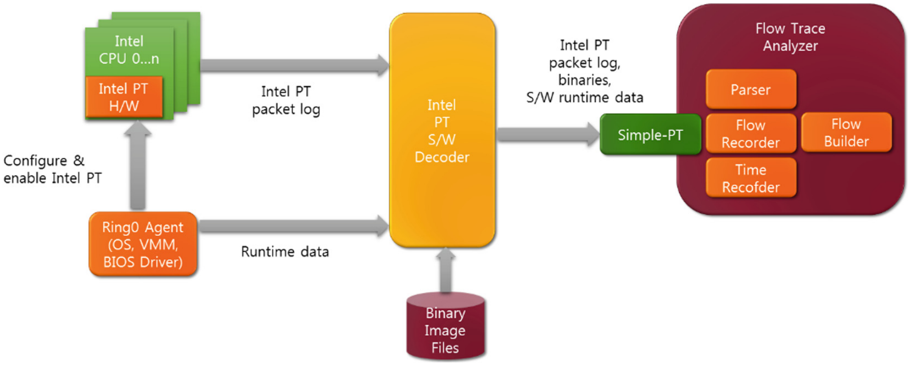


Fig. 4. Proposed Flow Trace Analyzer Architecture

As shown in Fig. 4, the proposed flow trace analyzer consists of a parser, flow recorder, time recorder, and flow builder. The parser module parses the results of simple-pt by function name, process name, and symbols along with time information. The flow recorder module records the flow by function name or process name by referring to the result of the parser module and the result of the time recorder module. Time recorder module records time information of functions, processes, and events, and provides recorded information to Parser module and Flow recorder module.

Finally, the flow builder module reconstructs the function flow or process flow. The flow builder module tracks the frequency at which a function or process is executed to check the integrity of a function or process.

The FTA software consists of several types of data structures such as struct FTA_master fx_master, struct flow_master, struct list, struct listnode, and struct fta, as shown in Fig. 5.

Struct FTA_master fx_master is a data structure that manages all FTA instances in the system, and pm is a pointer to fx_master in the system. Struct FTA_master fx_master * FTA is a pointer to a struct list that manages the FTA instances in the system. The struct list manages all FTA instances that exist in the system in the same form as struct listnode. In other words, this FTA implementation can support to run more than one FTA instance simultaneously.

In addition, struct FTA_master fx_master stores pointers to flow_master which manages all information such as functions, symbols, processes, and events used in each FTA instance.

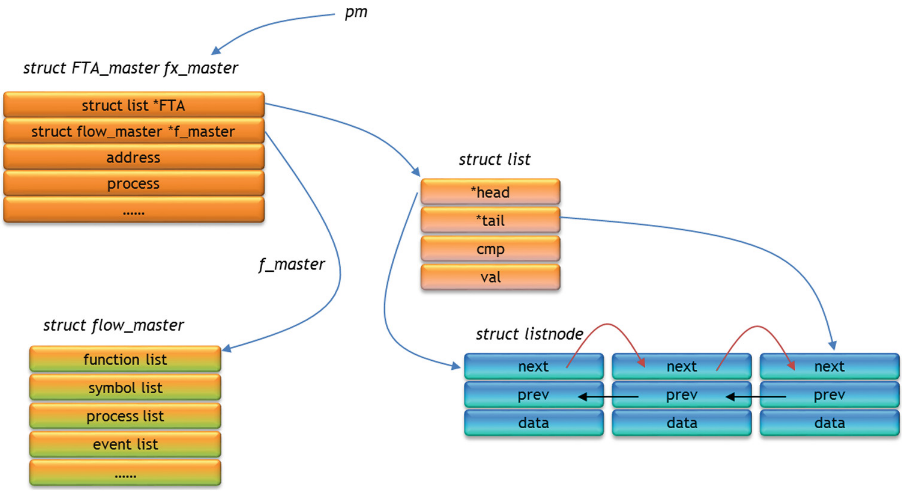


Fig. 5. Data Structure for Flow Trace Analyzer

4 Conclusion

Virtualization technology, a key technology in cloud computing, may be exploited as a means of passing data to non-moral people. To prevent this, the environment for implementation, unauthorized modification or activity of security policy should be monitored, which may be exemplary for installation or configuration. Efforts should also be made to encourage strong authentication and access control, to enforce service level agreements (SLAs) for patching and vulnerability improvement, and to conduct vulnerability checks and configuration audits.

In this paper, we proposed a scheme to guarantee the integrity of the software that composes the VM in the cloud environment using Intel processor trace (PT). Through the virtualization-based integrity analysis method proposed in this study, it is possible to check real-time integrity in a virtualized network based on SDN and NFV. It is possible to apply an efficient network policy such as only performing a verified process of integrity that is, performing only VNF so that the network can be configured more actively.

Acknowledgement. This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No. 2016-0-00078, Cloud-based Security Intelligence Technology Development for the Customized Security Service Provisioning).

References

1. Paxton NC (2016) Cloud security: a review of current issues and proposed solutions. In: International conference on collaboration and internet computing (CIC), pp 452–455
2. Mahboob T, Zahid M, Ahmad G (2016) Adopting information security techniques for cloud computing—a survey. In: International conference on information technology, information systems and electrical engineering (ICITISEE), pp 7–11
3. Thalheim J, Bhatotia P, Fetzer C (2016) INSPECTOR: data provenance using intel processor trace (PT). In: International conference on distributed computing systems (ICDCS), pp 25–34
4. Makkaoui KE, Ezzati A, Beni-Hssane A, Motamed C (2016) Cloud security and privacy model for providing secure cloud services. In: 2016 2nd international conference on cloud computing technologies and applications (CloudTech), pp 81–86
5. Duncan B, Bratterud A, Happe A (2016) Enhancing cloud security and privacy: time for a new approach? In: International conference on innovative computing technology (INTECH), pp 110–115
6. Lai S-F, Su H-K, Hsiao W-H, Chen K-J (2016) Design and implementation of cloud security defense system with software defined networking technologies. In: International conference on information and communication technology convergence (ICTC), pp 292–207
7. Kleen A, Simple Intel CPU processor tracing on Linux. https://github.com/andikleen/simple-pt

A Snapshot of 26 Years of Research on Creativity in Software Engineering - A Systematic Literature Review

Aamir Amin[1(✉)], Shuib Basri[2], Mohd Fadzil Hassan[2], and Mobashar Rehman[1]

[1] Faculty of Information and Communication Technology,
Universiti Tunku Abdul Rahman (UTAR), Kampar, Malaysia
{Aamir,Mobashar}@utar.edu.my

[2] Department of Computer and Information Sciences,
Universiti Teknologi Petronas (UTP), Teronoh, Malaysia
{Shuib_basri,mfadzil_hassan}@petrponas.com.my

Abstract. Creativity is important for software engineering. It is expected to gain more importance in coming decades. However the research work on creativity in software engineering is scattered and scarce. The current study aims to systematically review the existing literature on creativity in software engineering. As a result, the current study has highlighted 26 years of research work on creativity in software engineering. From the 49 selected studies, nearly half addressed creativity only in requirement engineering phase. Hence, it is safe to assume that there is a lack of research work on creativity in other phases of software development.

Keywords: Creativity · Software engineering · Systematic literature review

1 Introduction

Creativity is an essential element of almost all the human endeavors and has been progressively emphasized in the current knowledge centric workforce [28]. In one of our earlier publications [1], we defined creativity by combining the definition of creativity proposed by Boden [3] and that of Gaut [11] as *"an ability to come up with ideas or artifacts that are (a) new, (b) surprising and (c) valuable by flair"*. Before moving further, it is also important to understand that innovation is distinct from creativity. Innovation is defined as the *"successful implementation of creative ideas"* [73]. Hence, the focus of the present research is creativity and not innovation.

Winograd [30] compared development of software with the creation of art. Software engineering is an outcome of human knowledge and creativity [2, 26] and therefore successful and effective software engineering relies on knowledge collaboration and creativity of developers [29, 31]. Moreover, developers consider it interesting to work on those phases of development which they perceive as creative [14, 15, 21]. Furthermore, for software companies, the ability of its developers to generate creative solutions is critical [4, 10]. It is critical because software engineering entails complex problem solving and innovation, both of which indispensably require creativity [6, 12]. Likewise, the proponents of one of the well known and prevailing development approach, agile software development (i.e. eXtreme Programming (XP)), posit that the only solution of

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_47

complex software development problems, is creativity and not written rules [5, 8, 17, 18]. Hence, the role of creativity cannot be overlooked in software engineering [23].

Because of its importance, creativity in software engineering has been discussed by researchers since long [5]. About two decades ago, Winograd [30] stressed on paying more attention to human factors in software engineering especially creativity. More recently, Crawford et al. [6] posited that in coming decades the significance of creativity in software engineering will further boost. However, despite the importance of creativity in software engineering, it has been neglected [13]. Most of the research work on creativity in software engineering addresses the requirement engineering phase [14, 15, 68]. Therefore, it is important to systematically highlight and provide a clear picture of the existing work and research gaps on creativity in software engineering.

Hence, the present work will systematically review the available literature on creativity in software engineering. The paper follows the guidelines of Systematic Literature Review (SLR) as provided by Kitchenham [20]. The present research will answer following questions:

- What are the existing studies on creativity in software engineering?
- In which of the phases of software engineering, creativity has been addressed?
- Which phases of software engineering have been overlooked in reference to creativity?

The scope of the present research work is limited to the aforementioned purpose whereas defining creativity and the nature of it in software engineering is not addressed.

2 Methodology

2.1 Identification of the Need for SLR

In order to identify the need to conduct SLR on creativity in software engineering, a search was conducted on Google Scholar, Springer, ACM and IEEExplore to identify the existing SLR studies on creativity in software engineering between the period of 1990 and 2016. In order to avoid search bias, following search string was developed with the synonyms of systematic literature review (as in [9]).

((“Software Development” OR “Software Engineering”) AND (“creative” OR “creativity”) AND (“systematic review” OR “research review” OR “research synthesis” OR “research integration” OR “systematic overview” OR “systematic research synthesis” OR “integrative research review” OR “integrative review”))

Furthermore, so as to select the studies from the search results, the abstracts, keywords and titles were screened. As a result, five SLR studies were found on creativity in software engineering (see Table 1). Apart from the SLR studies mentioned in Table 1, another SLR study [19] was found. However, the study focuses on innovation and not creativity. Hence it was not included here.

Table 1. Available SLRs On Creativity In Software Engineering

SLR Focus	Reference
Creativity in requirement engineering phase	Saha et al. [27], Lemos et al. [22]
Motivational perspective of creativity in software engineering	Hedge and Walia [16]
Creativity in agile software development	Canboy et al. [5]

From Table 1, it is evident that the existing SLR studies have attempted to provide the existing literature in one aspect or phase of software engineering. Therefore, there is a need of a thorough systematic literature review on creativity in software engineering.

2.2 Search Protocol and Source Selection

In order to conduct the search for primary studies, at first, a search string was developed. For this purpose, combined with the word *'creativity'*, synonyms of software engineering as well as different phases of software engineering were used. These synonyms were adopted from Pirzadeh [24] (See Table 2).

Table 2. Search String And Keywords

Main body of search string	("Software Engineering" OR "Software Development") AND (Creativity OR Creative)
Lifecycle Phases	• "Requirement engineering" OR "Requirement elicitation" OR "Requirement analysis" OR "Requirement" • "Design", "Design phase", "architecture design" • "Implementation", "Implement", "Programming", "Programmer"

Furthermore, the search was conducted for the time duration between 1990 and 2016. At the same time, the search scope was limited to software development whereas the information system development was not included in the scope. Moreover, the search was conducted on major databases including Springer, ACM Digital Library, Science Direct and IEEExplore. In addition, the search was also conducted on Google Scholar.

2.3 Study Execution

The search on aforementioned databases resulted in 5,230 studies. At first, the screening of these studies was done based on titles. This resulted in 95 studies. In the second stage, further screening was done based on abstract reading. This resulted in the elimination of 30 studies. Hence, 65 studies were sent for validity.

2.4 Study Inclusion and Exclusion Criteria

The inclusion and exclusion of the retrieved studies was performed based on the criteria shown in Table 3.

Table 3. Study Inclusion /Exclusion Criteria

Inclusion Criteria	Exclusion Criteria
The studies which are related with creativity in any phase of software engineering	The studies in language other than English
The studies which report creativity tools and creativity procedures etc. in software engineering	The studies which were related with workshops, tutorials, mini reports etc
Studies which are related with creative style, contextual factors which effect creativity in software engineering	The studies which were not within the scope of this research
	The studies which were related with Management Information Systems (IS) and not software engineering. Both are distinct professions. MIS professionals work for the IS needs of a company whereas software engineers provide services to external customers [25]
	Studies which were conducted prior to 1990
	The studies which are related to innovation. As described in the introduction of this paper, creativity and innovation are distinct concepts

2.5 Validity

The selected studies were validated by three researchers in the field of software engineering. The researchers examined the list and provided their opinion on the relevance of the studies with the research questions and the inclusion/exclusion criteria. As a result, 49 papers, which were unanimously selected by the reviewers, were included in the final list.

3 Results and Discussion

This section will report the results of the SLR on creativity in software engineering. The section will be divided according to the research objectives.

1. What are the existing studies on creativity in software engineering?

As mentioned earlier, from the search result of 5,230 studies, 49 studies were selected for the current research. Table 4 shows the studies which are included in this research as well as the phase or topic which the selected studies address. The results are evident that the amount of work undertaken during the span of nearly 26 years is scarce compared to the importance of creativity in software engineering. This suggests that more research work is needed in the field of creativity in software engineering. Moreover, a lot of work on creativity in software engineering is undertaken by Maiden and team whereas in the domain of creativity in agile software development, Crawford and team has undertaken much of the research work. In addition, according to the results, most of the research work on creativity in software engineering is undertaken after the year 2000. The distribution of studies based on their year of publication is as follows: 1990 until 2000 (3); 2001 until 2005 (16); 2006 until 2010 (18) and from 2011 until 2016 (17). This shows that after 2000, although little, but the researchers' attention towards creativity in software engineering has increased. This increase in creativity research could be a result of demand in creativity of software products in recent years.

Table 4. SLR Studies

Topic	Reference
General	SLR [12, 14, 15, 32–39]
Requirement engineering /elicitation phase.	SLR [23–25, 40–61]
Software Design	SLR [62, 63]
Implementation Phase	SLR [1, 64–66]
Software Test and Reuse	SLR [67]
Agile methods, agile processes, agile teams and agile system development	SLR [5–7, 68–71]
Team creativity	SLR [29]
Software Process and Project	SLR [34, 72]

2. In which of the phases of software engineering, creativity has been addressed?

Lastly, the results highlight that the emphasis of researchers has been on creativity in requirement engineering phase of software development. 49 percent (24 out of 49) of the studies have addressed creativity in requirement engineering phase of software development. Moreover, after requirement engineering, 20 percent (10 out of 49) studies have addressed creativity in software engineering in general, 16 percent (8 out of 49) in agile software development, 6 percent (3 out of 49) in software implementation, 4 percent (2 out of 49) in software design, 4 percent (2 out of 49) in software process and projects, and 2 percent (1 out of 49) in software reuse and software teams each. Figure 1 highlights the number of studies for each phase or topic of software engineering.

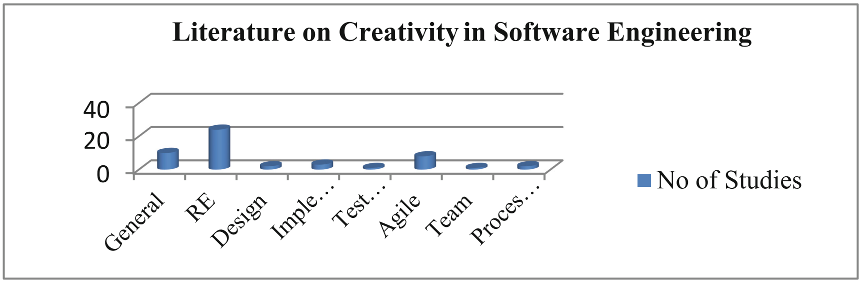


Fig. 1. No of studies for each phase

3. Which phases of software engineering have been overlooked in reference to creativity?

The above results (Fig. 1) clearly shows a lack of research work on creativity in the design, implementation, testing and reuse phases. Furthermore, there is a lack of research work on creativity in software processes and project.

In addition, there is no study which attempts to analyze the impact of contextual factors on software engineer's creativity. Moreover, there seems to be a lack of research work on creativity for specific roles in software engineering, such as creativity of programmers and designers. At the same time, there is no empirical study to examine creativity in software engineering phases such as design and implementation.

4 Conclusion and Future Work

The current study systematically reviewed the literature on creativity in software engineering. The study examined the literature between 1990 and 2016. The results of the systematic literature review revealed that majority of the current research work on creativity in software engineering has examined creativity in requirement engineering phase as well as agile software development. However, other phases of software engineering have been overlooked.

In future, there is a lot of room in the research on creativity in software engineering. For instance, researchers can address creativity in other phases such as design, implementation, testing and reuse. Moreover, creativity of individual roles (i.e. programmer, designer) should also be studied. Apart from this, factors which can inhibit or flourish creativity in software engineering are crucial to be understood in order to facilitate creativity in software development. In addition, there is also a need to generate a clear definition of creativity in software engineering as well as of different phases of software engineering. Lastly, it needs to be determined that which techniques and models of software engineering are conducive for creativity.

References

1. Amin A, Rehman M, Basri S, Hassan MF (2015) A proposed conceptual framework of programmer's creativity. In: ISTMET, pp 108–113
2. Bjørnson FO, Dingsøyr T (2008) Knowledge management in software engineering: a systematic review of studied concepts, findings and research methods used. Inf Softw Technol 50:1055–1068 Elsevier
3. Boden M (2004) The creative mind: myths and mechanisms, 2nd edn. Routledge, London
4. Ciborra C (1996) Improvisation and information technology in organizations. In: ICIS 1996, p 26
5. Conboy K, Wang X, Fitzgerald B (2009) Creativity in agile systems development: a literature review. In: IFIP- WG 8.2, Creative SME, pp 122–134
6. Crawford B, Barra CL, Soto R, Monfroy E (2012) Agile software engineering as creative work. In: Proceedings of CHASE. IEEE Press, pp 20–26
7. Crawford B, De La Barra CL (2007) Enhancing creativity in agile software teams. In: Proceedings of XP' 07. Springer, Heidelberg. pp 161–162

8. Crispin L, House T (2003) Testing extreme programming. Pearson, Boston
9. de Almeida Biolchini JC, Mian PG, Natali AC, Conte TU, Travassos GH (2007) Scientific research ontology to support systematic review in software engineering. Adv Eng Inf 21(2): 133–151
10. Dyba T (2000) Improvisation in small software organizations. IEEE Softw 17(5):82–87
11. Gaut B (2010) The philosophy of creativity. Philos Compass 5(12):1034–1046
12. Glass RL (1995) Software creativity. Prentice-Hall, Inc., Upper Saddle River
13. Graziotin D, Wang X, Abrahamsson P (2014) Software developer's moods, emotions, and performance. IEEE Softw 31(4):24–27. 10.1109/MS.2014.94
14. Graziotin D (2013) The dynamics of creativity in software development. In: PROFES 2013 - doctoral symposium proceedings, figshare. doi:10.6084/m9.figshare.703568
15. Gu M, Tong X (2004) Towards hypotheses on creativity in software development. In: Bomarius F, Iida H (eds) PROFES 2004. LNCS, vol 3009. Springer, Heidelberg, pp 47–61
16. Hegde R, Walia G (2014) How to enhance the creativity of software developers, a systematic literature review. In: SEKE 2014, Vancouver, Canada
17. Highsmith J (2004) Agile project management. Addison-Wesley, Boston
18. Highsmith J (2002) Agile software development ecosystems. Pearson, Boston
19. Juhola T, Hyrynsalmi S, Mäkilä T, Leppänen V (2013) Agile software development and innovation: a systematic literature review. In: 6th ISPIM innovation symposium, Melbourne, Australia
20. Kitchenham BA (2007) Guidelines for performing systematic literature reviews in software engineering. EBSE Technical report, Keele University and Durham University Joint Report
21. Knobelsdorf M, Romeike R (2008) Creativity as a pathway to computer science. ACM SIGCSE Bull 40(3):286
22. Lemos J, Alves C, Duboc L, Nunes G (2012) A systematic mapping study on creativity in requirements engineering. In: SAC' 12, pp 1083–1088
23. Mich L, Anesi C, Berry DM (2005) Applying a pragmatics based creativity fostering technique to requirements elicitation. Requir Eng 10(4):262–275
24. Pirzadeh L (2010) Human factors in software development: a systematic literature review
25. Rajeswari KS, Anantharaman RN (2003) Development of an instrument to measure stress among software professionals: Factor analytic study. In: Proceedings of the 2003 SIGMIS conference on computer personnel research: Freedom in Philadelphia–leveraging differences and diversity in the IT workforce, pp 34–43
26. Rizwan JQ, Sohayp AA, Fatima S (2014) Significance of the teamwork in agile software engineering. Sci Int (Lahore) 26(1):117–120
27. Saha KS, Selvi M, Buyiikcan G, Mohymen M (2012) A Systematic review on creativity techniques for requirements engineering. In: IEEE/OSA/IAPR, ICIEV'12 (2012)
28. Serrat O (2009) Harnessing creativity and innovation in the workplace. Knowl Solut 61:1–11 (Asian Development Bank, Manila)
29. Wang MH, Huang CF, Yang TY (2012) The effect of project environment on the relationship between knowledge sharing and team creativity in the software development con-text. Int J Bus Inf 7(1):59–80
30. Winograd T (1996) Bring design to software. Addison Wesley, Reading
31. Ye Y (2006) Supporting Software Development as Knowledge Intensive and collaborative activity. In: Foundations of software engineering conference, pp 15–22
32. Glass RL (2001) A story about the creativity involved in software work. IEEE Software
33. Connelly C (2001) Promoting creativity in software development. ASAC, London
34. Gallivan M (2003) The Influence of software developer's creative style on their attitudes to and assimilation of a software process innovation. Inf Manag 40(1):443–465

35. Maiden N (2010) Creativity in software engineering: a new research agenda? In: ICPC' 10, pp xiv
36. Kato N, Kunifuji S (1997) Consensus-making support system for creative problem solving. Knowl Syst 10(1):59–66
37. Guruge IT, Chinthaka AAJ (2015) The role of creative thinking in software development projects. In: RSEA 2015, SAITM, Malabe, Sri Lanka
38. Lee K, Scandura T, Kim Y, Joshi K, Lee J (2012) Examining leader-member exchange as a moderator of the relationship between emotional intelligence and creativity of software developers
39. Robertson J (2002) Eureka! why analysts should invent requirements. IEEE Softw 19(4):20–22
40. Maiden N, Robertson S, Gizikis A (2004) Provoking creativity: imagine What your requirements could be like. IEEE Softw 21(5):68–75
41. Maiden N, Robertson S (2005) Integrating creativity into requirements processes: experiences with an air traffic management system. In: RE 2005, France. IEEE Computer Society, Los Alamitos, pp 105–116
42. Robertson J (2005) Requirements analysts must also be inventors. Softw IEEE 22(1):48–50
43. Dallman S, Nguyen L, Lamp J, Cybulski J (2005) Contextual factors which influence creativity in requirements engineering. In: ECIS 2005 proceedings, 107
44. Grube PP, Schmid K (2008) Selecting creativity techniques for innovative requirements engineering. In: MERE' 08
45. Nguyen L, Shanks G (2009) A framework for understanding creativity in requirements engineering. Inf Softw Technol 51(3):655–662
46. Wen Y, Zhang H, Liu L, Yang H (2010) One bridge, two gaps - beyond an engineering approach: creativity in requirements elicitation. In: REV' 10, Sydney, Australia
47. Mahaux M, Mavin A, Heymans P (2012) Choose your creativity: why and how creativity in requirements engineering means different things to different people. In: Requirements engineering: foundation for software quality. Lecture notes in computer science vol 7195, pp 101–116
48. Elton RV, Alves C, Duboc L (2012) Creativity patterns guide: support for the application of creativity techniques in requirements engineering. In: Human-centered software engineering. Lecture notes in computer science, vol 7623, pp 283–290
49. Mahaux M, Gotel O, Mavin A, Nguyen L, Mich L, Schmid K (2013) Collaborative creativity in requirements engineering: analysis and practical advice. In: RCIS, 2013, pp 1–10
50. Sharma S, Walia G, Magel K (2014) Does domain knowledge increase creativity during requirements development: an empirical study. In: Proceedings of SERP: the steering committee of the world congress in computer science, computer engineering and applied computing (WorldComp)
51. Bhowmik T, Niu N, Mahmoud A, Savolainen J (2014) Automated support for combinational creativity in requirements engineering. In: RE' 14
52. Maiden N, Manning S, Robertson S, Greenwood J (2004) Integrating creativity workshops into structured requirements processes. In: Proceedings of the DIS' 04. ACM, pp 113–122
53. Karlsen IK, Maiden N, Kerne A (2009) Inventing requirements with creativity support tools. In: Requirements engineering: foundation for software quality. Springer, Heidelberg, pp. 162–174
54. Maiden N, Jones S, Karlsen K, Neill R, Zachos K, Milne A (2010) Requirements engineering as creative problem solving: a research agenda for idea finding. In: Proceedings of RE' 10, pp 57–66

55. Vieira ER, Alves C, Duboc L (2012) Creativity patterns guide: support for the application of creativity techniques in requirements engineering. In: Human-centered software engineering. Springer, Heidelberg, pp 283–290
56. Maiden N, Ncube C, Robertson S (2007) Can requirements be creative? Experiences with an enhanced air space management system. In: Software Engineering, ICSE 2007, pp 632–641
57. Maiden N, Robertson S, Robertson J (2006) Creative requirements: invention and its role in requirements engineering. In: Proceedings of the 28th international conference on Software engineering. ACM, pp 1073–1074
58. Svensson RB, Taghavianfar M (2015) Selecting creativity techniques for creative requirements: an evaluation of four techniques using creativity workshops. In: Proceedings of RE' 15. pp 66–75
59. Horkoff J, Maiden N, Lockerbie J (2015) Creativity and goal modeling for software requirements engineering. In: Proceedings of the 2015 ACM SIGCHI conference on creativity and cognition. ACM, pp 165–168
60. Nguyen L, Swatman PA (2006) Promoting and supporting requirements engineering creativity. In: Rationale management in software engineering. Springer, Heidelberg, pp 209–230
61. Mich L, Anesi C, Berry DM (2004) Requirements engineering and creativity: an innovative approach based on a model of the pragmatics of communication. In: Proceedings of the REFSQ, p 3-922602
62. Daughtry J, Burge J, Carroll MJ, Potts C (2009) Creativity and rationale in software design. ACM SIGSOFT Softw Eng Notes 34(1):27
63. Lirong Q, Hong L, Liping G (2004) A multi-agent system supporting creativity in conceptual design. In: Proceedings of CSCWD' 04, vol 1, pp 362–370
64. Mody RP (1992) Is programming an art? ACM SIGSOFT Softw Eng Notes 17(4):19–21
65. Greenfield GR (2006) Art by computer program = programmer creativity. Digit Creativity 17(01):25–35
66. Crawford B, Barra CL (2008) Does eXtreme programming support collaborative creativity? Comput J 1:19–21
67. Gomes P, Pereira FC, Bento C, Ferriera JL (2001) Using analogical reasoning to promote creativity in software reuse. In: Proceedings of the Workshop Programme of ICCBR, pp 152–158
68. Barra CL, Crawford B (2007) Fostering creativity thinking in agile software development. Springer, Heidelberg, pp 415–426
69. Hollis B, Maiden N (2013) Extending agile processes with creativity techniques. IEEE Softw 30:78–84
70. Barra CL, Crawford B, Soto R, Misra S, Monfroy E (2013) Agile software development: it is about knowledge management and creativity. In: ICCSA' 13. Springer, Heidelberg, pp 98–113 (2013)
71. Crawford B, Barra CL, Letelier P (2008) Communication and creative thinking in agile software development. In: Computer-Aided Innovation (CAI). Springer US, pp 205–216
72. Bobkowska A (2015) Balance between creativity and methodology in software projects. In: Proceedings of MIDI. ACM, p 3
73. Amabile TM, Mueller JS (2008) Handbook of organizational creativity: studying creativity, its processes and antecedents, an exploration of the componential theory of creativity. In: Zhou, Shelley CE (eds) Handbook of organizational creativity, pp 33–64. Lawrence Erlbaum, New York

Bilingual Word-Embedding for Korean and English Without Word Alignments

Min-su Kim, Moon-su Cha, and Kyung-Ah Sohn(✉)

Department of Computer Engineering, Ajou University, Suwon, South Korea
{Kms713,kashon}@ajou.ac.kr, ckanstnzja@gmail.com

Abstract. In spite of lots of cross-lingual word embedding models for various languages, approaches that support cross-lingual word embedding between languages that have different word order and different origin word are lacking. In this study, we address the problem of cross-lingual word embedding between Korean and English that have different word order and origin and perform experiments to examine its performance behavior. Cross-lingual models have different levels of supervision. For training between languages which have different word order, it is essential to reduce preprocessing time. Therefore, two sentence-level alignment cross-lingual models are chosen for our experiments. Our results show that cross-lingual embedding for Korean and English without word-alignment is possible. We also analyze which bilingual tasks are proper for each trained result by comparing characteristic of each model's trained result.

1 Introduction

Recently, Nature language processing (NLP) has been improved by word vector representations [7]. Word-embedding techniques using monolingual models are already popular and widely used for various tasks. However, by using monolingual models, our research is restricted to monolingual tasks because the semantic of similar embedding vector values from different languages do not match. Now, some cross-lingual models have been proposed, and we can find out relationship between different two paired languages.

In NLP, Korean has characteristics that word order of Korean is very different from other languages such as English or French. So, this raises the question of whether learning through cross-lingual models between Korean and another language is feasible or not. In this paper, we address this problem by training different cross-lingual models on Korean. Based on this, we can research about semantic relationship between Korean and the others and perform various tasks.

Each cross-lingual model requires different cross-lingual alignments: documents-level alignments, sentence-level, word-level, and both sentence and word alignments. In this paper, even though the result of using the most detailed cross-lingual supervision shows the best performance on semantic tasks, we will not use word-level alignment models for Korean considering that word-level alignment is very inefficient and time

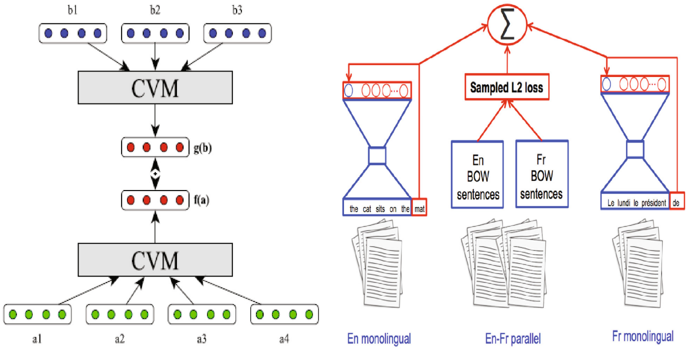
K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_48

consuming preprocessing work. Then we have to check that training result by cross-lingual model not requiring word-level alignment is useful.

Our experiment shows that cross-lingual training on Korean is meaningful by training Korean with another language which has different word order. We also show trained vector by the cross-lingual model not requiring word alignment is competitive by showing translation word-pairs are closely embedded in the vector space.

2 Bilingual Word-Embedding Without Word Alignments

In order not to suffer from time-wasting for word-alignment works, we choose two cross-lingual models which require sentence-level alignments. We briefly introduce algorithmic procedure for each bilingual word embedding model. The structure of each model is shown in Fig. 1.



(a)BiCVM [2] **(b)BilBOWA [4]**

Fig. 1. The structures of the bilingual models. They represent how each models train word vectors on their method. It is similar that they have joint training procedure after monolingual training. But it is different in that while BiCVM trains to minimize the distance of sentence, BilBOWA trains to learning features for word in sentence.

Bilingual Compositional Model (BICVM). Herman and Blunsom introduce a model that learns cross-lingual word vector using sentence-level alignment [1–3]. Because aligned sentence-pair have same meaning, BiCVM represent these sentence-pair as similar vector representation. For each language, in monolingual learning, BiCVM use compositional vector model (CVM) to semantically represent sentences [5, 6]. After monolingual training, BiCVM has monolingual representations of sentences for each language and it also knows information that which sentence-pair have same meaning. Then, BiCVM minimizes the distance of aligned sentence-pair works by its objective function.

Bilingual Bag-of-Words Without Alignments (BilBOWA). Stephan Gouws et al. presents BilBOWA, a simple and computationally-efficient bilingual word vector model [4]. The model does not require word-aligned but sentence-aligned parallel data like BiCVM. Moreover, this model can be used to unlimited amounts of monolingual raw text. Skip-gram word-embedding algorithm is used to monolingual word embeddings for each language. In cross-lingual stage, the goal of this model is learning features for each word. It is different from that BiCVM still focuses on sentence in cross-lingual stage.

3 Korean-English Word Embedding Using Bilingual Models

In this study, we train Korean-English paired data using two bilingual models. We use online bible corpus dataset. We use sentence-level aligned bilingual embedding model, and preprocessed our raw data to sentence-level aligned and parsed each sentence by semantic units. The used preprocessing tools are: KONLPY (available at http://konlpy.org/en/v0.4.4) for Korean, NLTK (available at http://www.nltk.org) for English. Then, we use thirty thousand sentence-aligned parallel en-kr corpuses.

We train en-kr corpus using two bilingual models. We use the tool (available at github.com/karlmoritz/bicvm) released by Hermann and Blunsom [2, 3] for BiCVM and use the tool (available at github.com/gouwsmeister/bilbowa) released by Stephan Gouws et al. [4]. After training, we compare the results of two models by comparing each result's Vector proximity. Also, for better quality's model, we check whether we can use our trained vector for some bilingual tasks without any additional process.

4 Experiment Results

Vector Proximity. In our experiment, the performance of bilingual models depends on proximity of translation pair words in vector space. So we measure the proximity of word vector by two means: Cosine-similarity, Euclidean distance. For measuring cosine-similarity and Euclidean distance, we use translation pair nouns from KorLex (Korean wordnet, available at http://korlex.pusan.ac.kr) which contains translation pair nouns between English. We measure average cosine-similarity and Euclidean distance between all translation pair nouns. The result is shown in Table 1.

We find that BilBOWA outperforms BiCVM absolutely. So we interpreted BilBOWA is a better model to train Korean between English. Since there is a great difference between their average performances, we do more evaluation test in a different way to validate our results.

Considering more evaluation in the perspective of vector proximity, to complement great difference of previous result, we use fake Wordnet-pair information as in a permutation test. With 5 randomly shuffled Wordnet-pair information, we measure the cosine similarity and Euclidean distance again for the same trained vectors. And we subtract score with fake information from real scores we got before. The results are shown in Table 2.

Table 1. Vector proximity test result from (a) BiCVM, and (b) BiBOWA. High Cosine similarity and low Euclidean distance represent pair-words' trained vectors are located closer. Best scores for each measure is shown in **bold**.

(a) BiCVM			
λ	dim	Cosine similarity	Euclidean distance
1	20	0.260	**0.922**
	40	**0.261**	1.313
	60	0.257	1.590
	80	0.260	1.831
	100	0.256	2.042
(b) BilBOWA			
λ	dim	Cosine similarity	Euclidean distance
1	20	**0.978**	**0.192**
	40	0.972	0.217
	60	0.969	0.226
	80	0.964	0.244
	100	0.944	0.306

Table 2. Vector proximity test result with fake translation-pair information $|RC - FC|$ represents the average absolute value of difference between cosine similarity score with real Wordnet information and cosine similarity with fake Wordnet information. $|RE - FE|$ represents the average absolute value of difference between Euclidean distance score with real Wordnet information and Euclidean distance with fake Wordnet information. High value represents each word is trained to locate close to its pair-word and locate far from unpaired-word. Best scores are shown in **bold**.

(a) BiCVM			
λ	dim	|RC – FC|	|RE – FE|
1	20	0.239	0.175
	40	0.225	0.243
	60	0.242	0.313
	80	**0.244**	0.366
	100	0.236	**0.391**
(b) BilBOWA			
λ	dim	|RC – FC|	|RE – FE|
1	20	0.009	0.046
	40	0.024	0.084
	60	0.029	0.097
	80	0.037	0.107
	100	**0.062**	**0.150**

In previous evaluation, the average scores of BilBOWA are absolutely high against BiCVM's. However, the performance of BiCVM is better in this aspect. With this result, although BilBOWA's scores with real information were absolutely high, it seems that all embedded vectors are located very close to each other. To use trained vectors from

BilBowa, it might be necessary to train more with different conditions to spread out evenly in vector space. With BiCVM's result, using the big difference of real and fake scores, BiCVM's trained vectors can be more useful to do some bilingual tasks.

Closest neighborhood newly considered perspective is the closest neighborhood test. We observed that BilBOWA gives us better vector proximity performance in terms of cosine similarity and Euclidean distance. To using this result, we have to check that when we choose closest English vector for a Korean word, its meaning must be same to chosen Korean word. Because we also obtained an expectation that every vectors trained by BilBOWA are located very close by previous result. So we use Cosine-similarity ranking measure. Cosine-similarity ranking represents that existence rate of the translation pair English word of a certain word in Korean in most 1,5,10 close word vectors. We measure on 10 different λ models (0.1–1). With this measure, higher rate represents better performance. Results are shown in Fig. 2.

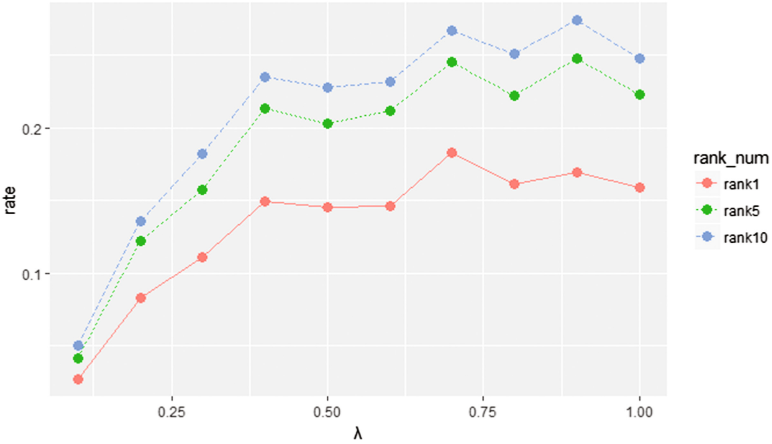


Fig. 2. Closest neighborhood test result. Y axis represents the existence rate of Wordnet-pair English words in most 1,5,10 closest words for every Korean word. We have 3 kinds of different closest vector numbers: 1 is solid, 5 is dotted, 10 is dashed respectively in figure.

Considering comparison number for each Korean word to all English word is about 10 thousand times, though the rate of result is not high (best score is about 30%), we can interpret performance is not promising. In some cases, we can use our trained vector information for several bilingual tasks without any additional process.

5 Conclusion

To summarize, we present training Korean corpus to vector with other languages which has different word order through cross-lingual embedding model is possible. Also, it gives useful information that results of using different two kinds of cross-lingual model provide different interpretations. Further experiments supplement how we can use those different results and what bilingual tasks are proper for each trained result.

In BilBOWA's case, more training with other condition (parameters, regularization term…) can supplement result of BilBOWA's weakness and we can try to apply trained vectors to bilingual semantic tasks.

Acknowledgement. This research was supported by the MISP (Ministry of Science, ICT & Future Planning), Korea, under the National Program for Excellence in SW) supervised by the IITP (Institute for Information & communications Technology Promotion) (R22151 610020001002).

References

1. Upadhyay S, Faruqui M, Dyer C, Roth D (2016) Cross-lingual models of word embeddings: an empirical comparison. In: Proceedings of ACL
2. Hermann KM, Blunsom P (2014) Multilingual models for compositional distributed semantics. In: Proceedings of ACL
3. Hermann KM, Blunsom P (2014) Multilingual distributed representations without word alignment. In: Proceedings of ICLR
4. Gouws S, Bengio Y, Corrado G (2015) BilBOWA: fast bilingual distributed representations without word alignments. In: Proceedings of ICML
5. Clark S, Pulman S (2007) Combining symbolic and distributional models of meaning. In: American Association for Artificial Intelligence (AAAI)
6. Hermann KM, Blunsom P (2013) The role of syntax in vector space models of compositional semantics. In: Proceedings of ACL
7. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Proceedings of NIPS

A Maturity Model for Implementation of Enterprise Business Intelligence Systems

Cheng Wai Khuen and Mobashar Rehman

Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman,
Kampar, Perak, Malaysia
{chengwk,mobashar}@utar.edu.my

Abstract. In the era of data explosion, organizations highly focus on the systematic management of business information. Various BI models exist for this purpose but the reliability of those models is questionable due to less focus on technical aspects. The purpose of this research is to build comprehensive and reliable Enterprise Business Intelligence Maturity (EBIM) model by using systematic and methodical design. This research investigated and explained problems of existing BI models and solved them by using systematic process through Delphi technique. Three dimensions namely process, technology and organization were considered to develop a mature model. This research concluded that comprehensiveness and reliability of EBIM model enable BI organizations to better plan, assess, and manage their BI initiatives.

Keywords: Business intelligence model · Process · Technology · Organization

1 Introduction

Organizations are becoming increasingly aware about the importance of information. This makes management of business information more proactive and systematic than before. Accurate and updated data are crucial for the survival of any organization in today's environment. This is becoming complex every day because of the huge volume of data produced by organizations on daily basis. To deal with this issue, organizations need some sort of solution, which can help them in handling large volume of up-to-date data with accuracy. BI can provide solution to organizations for this problem. BI covers technologies, applications and processes for gathering, storing, accessing and analyzing data to make decisions [1]. BI allows users to gather data from various sources across the organization for various purposes like making processes efficient and better decision making. However, implementation of BI across organization requires lots of resources including financial resources which make it difficult for some organizations to implement it. Even though, those organizations who have already implemented BI, many of them are unable to get full benefits [2, 3].

Quite a number of BI models exists but the reliability of those models is questionable [4, 5]. There are various reasons due to which the reliability of these models is questionable like less or no focus on model evaluation, technical guidelines. Another problem with the existing models is that they focus more on specific clients [4] and at the same

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_49

time focused on few concepts of BI. Based on these problems, it can be said that existing BI models are not mature and they lack comprehensiveness [6]. Therefore, this study will propose a more comprehensive BI model which will be suitable across various industries.

2 Literature Review

2.1 Benefits and Application of Business Intelligence

BI is combination of tools, technologies and architecture with the help of which individuals and organizations analyze the data in order to get better and in depth-analysis. This helps in better and quick decision making, improves business effectiveness and performance which may result in competitive advantage. Other benefits of BI include quick and accurate reporting, improved customer service, increase in revenue and saving in non-IT and IT related costs [7]. In [8], key benefits of using BI at enterprise level includes consistent information, provides actionable information and improvement in communication. Thus BI process can be described as the transformation of data to information, then to decisions, and finally to actions [9].

These benefits of BI can be achieved in almost every industry especially in retail and insurance industries. Besides banking, finance and securities, telecommunication and manufacturing industries are also benefiting from BI implementation.

2.2 Business Intelligence Maturity Models

Maturity refers to "*state of being complete, perfect or ready*" [10]. Various BI models claimed to be mature and after thorough research, seventeen of them are shortlisted for comparison in this study. These maturity models were analyzed based on 13 components namely data [11–14], architecture [15, 16], tools [17–20], impact [21], performance management [22–26], behavior [22], sponsorship and funding [27], change management [28], processes [21], staff [28], strategy [29], organizational structure [26] and efficiency [11]. These thirteen components can be categorized into IT and business domains. However, from these components few like data, architecture and tools which are related to IT or technology got more focus in comparison to business related factor especially processes and efficiency. Both of these components are important as well because if the processes are not well defined and BI usage is not efficient, organizations may face huge problems. Similarly, some important components which are not properly focused in the maturity models include strategy, staff and change management. Therefore, it will be unjustified to say that BI models are mature because most of them heavily focused on IT aspect and paying less or no attention to non-IT factors.

3 Research Methodology

Qualitative method was used for this study. Data was collected through Delphi technique. Fifteen BI experts were interviewed. Three rounds of the study were planned in order to get the consensus based on the feedback of the experts. Based on the results of interviews in round one, respondents were given with group median, interquartile range and his/her own response in round two and three and were asked to review their response if they wish to do so. Main purpose of conducting Delphi technique was to develop a consensus on the Key Process Areas (KPAs) for a Business Intelligence maturity model.

4 Data Analysis

4.1 Demographics

Respondents include managing director, deputy general manager, senior product manager, chief technology officer, application engineer and IT assistant manager. Majority of the respondents had more than 4 years of experience as shown in Table 1. The respondents were from different organizations including SMEs, MNCs and LPL and different sectors like financial sector, software, telecommunications and construction/architecture/engineering. Most of the organizations were from Malaysia (9 out of 15) but few were from Australia (2), United Kingdom (1), Indonesia (1), United States (1) and Singapore (1) as well.

Table 1. Demographic information

Participant	Position	Years of BI experience
1	Managing director	1 to 3 years
2	Academic (Lecturer)	1 to 3 years
3	Senior product manager	1 to 3 years
4	Application engineer	1 to 3 years
5	Manager	1 to 3 years
6	Deputy general manager	4 to 6 years
7	IT assistant manager	4 to 6 years
8	Assistant manager	4 to 6 years
9	Academic (Lecturer)	4 to 6 years
10	Team lead information management	4 to 6 years
11	Manager	4 to 6 years
12	Academic (Deputy director)	7 to 10 years
13	Academic (Head of school)	7 to 10 years
14	Chief technology officer	14 years and above
15	Partner	20 years and above

4.2 Results for First Round of Delphi Study

Forty-three KPAs were sent to the fifteen participants of the Delphi study. Participants were asked to categorize KPAs from level 1 to 5. From the results, it was analyzed that participants has consensus on five KPAs (shown in Table 2) which are Master Data Management (MDM) architecture, multiple BI applications developed, Little/No BI awareness, cross-departmental BI governance and BI as key revenue generator. These

KPAs are from organization, technology and process domains. Participants showed less consensus on eleven KPAs whereas twenty-seven KPA had moderate consensus.

Table 2. KPAs with high consensus.

KPA	Median	SD	Interquartile
MDM architecture	3	0.88	0.5
Multiple BI applications developed	3	0.83	0.5
Little/No BI awareness	1	0.99	0.5
Cross-departmental BI governance	3	0.74	0.5
BI as key revenue generator	4	0.88	0.0
KPAs with low consensus			
Spreadsheets	2	1.36	2.0
Chief level management actively sponsor BI	4	1.11	1.5
Define BI vision and roadmap	3	1.18	2.0
BI Service-Oriented Architecture (SOA)	5	1.44	1.5
Key Performance Indicators (KPI)	4	1.18	1.5
Advanced analytics	4	1.10	1.5
Programme management	4	1.18	1.5
Operational BI	3	1.03	1.5
Change management	3	1.01	1.5
Balanced scorecard	4	1.13	1.5

4.3 Results for Second Round of Delphi Study

Second round was also done through questionnaire just like round one and participants were allowed to review their opinion. During round two each participant was given the group median, interquartile range and his own response to the previous

Table 3. KPAs with high consensus in round two of Delphi study.

KPA	Median	SD	Interquartile
Little/No BI awareness	1	0.00	0.0
BI as key revenue generator	4	0.88	0.0
Spreadsheets	1	0.46	0.5
Centralized data warehouse	3	0.46	0.5
Chief level management actively sponsor BI	4	0.70	0.5
Multiple BI applications developed	3	0.70	0.5
Define BI vision and roadmap	3	0.93	0.5
BI Service-Oriented Architecture (SOA)	5	0.83	0.5
BI competency centre	4	0.88	0.5
Separated analytical process	2	0.83	0.5
Cross- departmental BI governance	3	0.88	0.5
Master Data Management (MDM) architecture	3	0.88	0.5

round for each KPA. In contrast to round one, more number of KPAs with high consensus were observed (shown in Table 3).

Furthermore, Standard Deviation of every high consensus KPA met threshold, hence they were shortlisted and included into the EBIM model. No KPA with low consensus was found in round two.

5 The Proposed Model - EBIM Model

The proposed EBIM model consists of three dimensions and five levels (Tables 4, 5 and 6). The three dimensions include process, organization and technology whereas five levels are Initial, Consolidate, Integrate, Optimize and Innovate respectively. An organization's enterprise-level BI maturity can be reasonably mapped in five evolutionary levels along these dimensions. Each maturity level is a prerequisite to the next higher one. Therefore each higher maturity level encompasses all previous lower levels. The definitions of five levels are:

Level 5 Innovate – Organization and BI is in oneness and effort of actively seeking innovation and breakthrough is carried out.

Level 4 Optimize – BI becomes a strategic enterprise resource and organization continues to optimize BI implementation to the fullness.

Level 3 Integrate – Connect and combine segmented BI effort across organization.

Level 2 Consolidate – Dynamically adopt BI and construct BI practices and systems.

Level 1 Initial – BI is strange to the organization or just start awareness of BI.

Table 4. Dimension process.

Level	KPA
5 (Innovate)	• Fully embedded advanced analytics and BI
4 (Optimize)	• Established a standard for best practices • Advanced Master Data Management (MDM)
3 (Integrate)	• Sophisticated data governance • Integrated analytics and BI within systems and workflows • Change management • Operational BI • Metadata Management
2 (Consolidate)	• Start Master Data Management (MDM) • Separated analytical process • Ad hoc BI
1 (Initial)	• Analytics is undefined • Inconsistent data

Table 5. Dimension organization.

Level	KPA
5 (Innovate)	• High quality information easily accessible
4 (Optimize)	• Align BI with business strategy • Maximum alliance between business and IT team • BI in R&D • Programme management • Chief level management actively sponsor BI • BI Competency Centre • KPI
3 (Integrate)	• Executive sponsors BI • Enterprise wide performance metric framework • Cross-departmental BI governance • Project-based roles and skills • Defined BI vision and roadmap
2 (Consolidate)	• Defined BI goals and scope • Begin to provide BI training • BI expertise remains in individuals • Low involvement from management level • Limited insight into customers, markets and competitors
1 (Initial)	• Little/No BI awareness

Table 6. Dimension technology.

Level	KPA
5 (Innovate)	• BI Service Oriented Architecture (SOA)
4 (Optimize)	• Enterprise data warehouse • Advanced analytics • Balanced scorecard
3 (Integrate)	• Centralized data warehouse • Master Data architecture • ETL & OLAP
2 (Consolidate)	• Data silos
1 (Initial)	• Spreadsheets

6 Verification of Results

Semi-structured interviews from the four selected case companies (mentioned in Table 7) were conducted. This sample size is consistent with the suggestions mentioned in [30–34]. At the beginning of the session, the interviewees were given a set of questionnaire in order to assess their current BI initiative by themselves, which served as foundation of subsequent questions. Predefined questions were used to conduct the interview which allowed in getting detailed information. The interview questions mainly revolved around themes including "accuracy of model", "satisfaction level of company's BI maturity", "motivating factors to move on or stay in current BI maturity level", "shift of focus as organization's BI maturity is increasing", "interviewee opinion toward comprehensiveness of EBIM model", "extra suggestion from interviewees" and other findings.

Table 7. Background of companies

Case	Years of company BI initiative	Company BI applications	Company type	BI investment from company annual revenues	Company location
C1	10 years	21 or more	SME	4% to 5%	England
C2	1 year	1 to 2	SME	1%	Malaysia
C3	10 years	21 or more	MNC	4% to 5%	United States
C4	2 years	1 to 2	MNC	11% or more	Singapore

6.1 Accuracy of Model

In order to gauge the accuracy of the EBIM model, results were verified from interviewees after completion of the assessment by using EBIM model. All of the four interviewees were satisfied with the assessment results. The results matched with the perception of their organization's BI maturity level. Another significant aspect to be considered was that the verification results confirmed the importance of non-IT aspects or dimensions besides the significance of technology component.

6.2 Satisfaction Level of Company's BI Maturity

Case C1, C3 and C4 were found to be satisfied with their current BI maturity in every dimension as well as the overall maturity in EBIM model. At the same time they showed desire to move to higher maturity level at individual and organizational level. One of the interviewee commented "being good is not enough, it is a must to do better".

In contrast case C2 was found to achieve Level 4 in "Technology" dimension; however it only achieved Level 1 in "Process" and "Organization" dimensions. It was observed that C2 paid attention and invested efforts and resources particularly to the dimensions with low maturity. It was interesting to know that, one of the interviewee expressed his urge to attain a higher level of BI maturity but had no clear idea how to align the organization with the desire. As reported by the interviewee, in the current scenario the propose EBIM model presents a clear pathway, regarding how to move forward to the next level thus proving the usability of EBIM model.

6.3 Shift of Focus as Organization's BI Maturity Is Increasing

For each level, the interviewees were asked to identify the most essential KPA. The answers were based on their experience rather than perception. It was observed that for "Level 1 – Initial" all four interviewees mutually reported and stressed on the issues created by "Inconsistent data". In "Level 2 – Consolidate", interviewees highlighted "Start Master Data Management (MDM)", "Start to define BI goals and scope", "Limited insight into customers, markets and competitors", "Low involvement from management level", "Start to define BI goals and scope" as utmost KPA.

From "Level 3 – Integrate", most of the KPAs were selected from dimension Organization. The selected KPA included "Executive Sponsors BI", "Project-based roles and skills", "Define BI vision and roadmap", "Change management", "ETL & OLAP" and "Centralized data warehouse". One of the interviewee chose "change management" as one of the most essential KPA in "Level 3 – Integrate" as it guarantees certain degree of transparency to people in the organization. This transparency is necessary in order to streamline the processes of BI implementation. In "Level 4 – Optimize", the most essential KPA included "Align BI with business strategy", "Advanced Master Data Management (MDM)" and "Advanced analytics". One of the interviewees remarked that advanced MDM is important as it does global consolidation and standardization of data within organization. In "Level 5 – Innovate" KPA such as "Fully embedded advanced analytics and BI in processes" and "BI-SOA" were chosen by interviewees. It was interesting to observe that interviewees considered "Level 4 – Optimize" and "Level 5 – Innovate" gelled to one another. It was further noticed that activities of "Level 4 – Optimize" affected "Level 5 – Innovate". Overall, most of the interviewees were found to be in favor of dimension Organization. This specified that organizational support is highly significant in the implementation of BI. Case C1, C3 and C4 were found to be aware of the dimension "Organization" as an essential driver of successful BI initiative. However in case of C2, lack of KPAs in dimensions "Process" and "Organization" was due to the management negligence in seeing data warehousing as IT solution to their problems. Irrespective of the organization's scale, "Technology" dimension was found as the most developed dimension. The three dimensions together with the KPAs in each cell were discussed sufficiently and interviewees reflected that the EBIM model is accurately telling the evolutionary path of BI maturity in a systematic manner.

7 Comprehensiveness of EBIM Model

All of the interviewees were found to be pleased with the EBIM model. One interviewee found KPA were well grouped in every level. Furthermore, KPA in every level matched with the descriptive name of the levels (Initial, Consolidate, Integrate, Optimize and Innovate). Another interviewee mentioned KPA of the EBIM provides a clearer and better view of different BI maturity levels. Previously, the interviewee only noticed technology as BI solution to business needs but now "Process" and "Organization" dimension were found to be significant too.

8 Conclusion and Future Work

This research developed a comprehensive BI maturity model for BI users in order to improve the BI maturity of organization. The proposed EBIM is not only comprehensive in comparison to existing BI maturity models but also presents a clearer pathway for organizations to attain higher BI maturity. It significantly supports enterprises in planning, assessing or undertaking large-scale BI initiatives by showing where currently they are and what should they do to improve BI implementation. The proposed EBIM model is also capable of allowing user to perform self-assessment.

Future work can be done to study people centric dimension to discover more KPAs about people. Further, a comprehensive software system may be developed for BI stakeholders to better measure and visualize the maturity levels of respective BI initiatives. A future model can be customized as per specific industry as opposed to the proposed EBIM model which is generalized for all industries.

References

1. Wixom BH, Watson HJ (2010) The BI-based organization. Int J Bus Intell Res 1(1):13–28
2. KPMG (2009) Does your business intelligence tell you the whole story, September 2009, UK
3. National Computing Centre: Data warehousing and BI systems let down 22% of users. http://www.ncc.co.uk/article/?articleid=16567. Accessed 7 Mar 2013
4. Lahrmann G, Marx F, Winter R, Wortmann F (2011) Business intelligence maturity: development and evaluation of a theoretical model. In: 44th Hawaii international conference on system sciences (HICSS), January. IEEE, pp 1–10
5. Shaaban E, Helmy Y, Khedr A, Nasr, M (2011) Business intelligence maturity models: toward new integrated model. In: The international Arab conference on information technology (ACIT 2011), Organized by Naif Arab University for Security Science (NAUSS), Riyadh, Saudi Arabia, 11–14 December 2011
6. Lahrmann G, Marx F, Winter R, Wortmann F (2010) Business intelligence maturity models: an overview. In: VII conference of the Italian chapter of AIS, Naples, Italy, p 6105
7. Thompson O (2004) Business intelligence success, lessons learned. Accessed 20 May 2015
8. Eckerson W, Howson C (2005) Enterprise business intelligence: strategies and technologies for deploying BI on an enterprise scale. The Date Warehousing Institute (TDWI), August. http://www.tdwi.org
9. Turban E, Sharda R, Aronson J, King D (2007) Business intelligence, 1st edn. Prentice Hall, New Jersey
10. Simpson JA, Weiner ESC (1989) The Oxford English dictionary. Oxford University Press, Oxford
11. Watson HJ, Ariyachandra T, Matyska RJ (2001) Data warehousing stages of growth. Inf Syst Manage 18(3):42–50
12. Cates JE, Gill SS, Zeituny N (2005) The Ladder of Business Intelligence (LOBI): a framework for enterprise IT planning and architecture. Int J Bus Inf Syst 1(1–2):220–238
13. Williams S, Williams N (2007) The profit impact of business intelligence. Morgan Kaufmann, New York
14. Eckerson WW (2004) Gauge your data warehousing maturity. DM Rev 14(11):34
15. SAS (2011) Information evaluation model. http://www.sas.com/software/iem/. Accessed Sept 2015
16. Eckerson W (2007) TDWI benchmark guide: interpreting benchmark scores using TDWI's maturity model. TDWI Research. http://onereports.inquisiteasp.com/Docs/TDWI_Benchmark_Final.pdf. Accessed 20 Apr 2009
17. Chamoni P, Gluchowski P (2004) Integration trends in business intelligence systems: an empirical study based on the business intelligence maturity model. Wirtschaftsinformatik 46(2):119–128
18. Schulze K-D, Besbak U, Dinter B, Overmeyer A, Schulz-Sacharow C, Stenzel E (2009) Business Intelligence-Studie. Steria Mummert Consulting AG, Hamburg

19. Kasnik A (2008) Model optimization infrastructure. Internal material of ZRSZ, Ljubljana
20. Töpfer J (2008) Active enterprise intelligence. In: Töpfer J, Winter R (eds) Active enterprise intelligence. Springer, Heidelberg, pp 1–28
21. Fisher T (2005) How mature is your data management environment? Bus Intell J 10(3): 20–26
22. Sen A, Sinha AP, Ramamurthy K (2006) Data warehousing process maturity: an exploratory study of factors influencing user perceptions. IEEE Trans Eng Manage 53(3):440–455
23. Hagerty J (2006) AMR research's business intelligence/performance management maturity model, Version 2
24. Deng R (2007) Business intelligence maturity hierarchy: a new perspective from knowledge management. Information management. http://www.information-management.com/infodirect/20070323/1079089-1.html. Accessed 24 Apr 2009
25. Sacu C, Spruit M (2010) BIDM: the business intelligence development model. In: Proceedings of the 12th international conference on enterprise information systems, Funchal, Madeira-Portugal
26. Gartner (2007) Business intelligence & information management, Sydney, Australia. http://www.gartner.com/ap/bi
27. Davenport TH, Harris JG (2007) Competing on analytics: the new science of winning. Harvard Business School Press, Boston
28. Hewlett Packard (2009) The HP business intelligence maturity model: describing the BI journey. Hewlett-Packard Development Company, L.P. http://h20195.www2.hp.com/V2/GetPDF.aspx/4AA1-5467ENW.pdf. Accessed 16 Dec 2009
29. Raber D, Winter R, Wortmann F (2012) Using quantitative analyses to construct a capability maturity model for business intelligence. In: Proceedings of the 45th Hawaii international conference on system science (HICSS), pp 4219–4228
30. Eisenhardt K (1989) Building theories from case study research. Acad Manage Rev 14(4): 523–550
31. Miles M, Huberman AM (1994) Qualitative data analysis: an expanded sourcebook. Sage, Thousand Oaks
32. Yin R (2003) Case study research: design and methods, 3rd edn. Sage, Thousand Oaks
33. Carson D, Gilmore A, Gronhaug K, Perry C (2001) Qualitative research in marketing. Sage, London
34. Perry C (2001) Case research in marketing. Mark. Rev 1:303–323

Autonomous Text Summarization Using Collective Intelligence Based on Nature-Inspired Algorithm

Kaleab Getaneh Tefrie and Kyung-Ah Sohn(✉)

Department of Computer Engineering, Ajou University,
San5, Woncheon-dong, Yeongtong-gu, Suwon 443-779, South Korea
{kaleab,kasohn}@ajou.ac.kr

Abstract. Thousands of years ago written language was introduced as a way of enhancing and facilitating communication. Fast forward to the twenty first century much has changed, especially the flow of data incrementing at fast rate and we should use the power of algorithms and hardware technology to understand text more clearly. With the Information age rising we are being cluttered with humongous data each day with no sign of it slowing. Humans have been trying to create ways on how to handle this continuous flow of text, image and video. And one of the categories of subjects regarding text is text summarization, given a document coming up with a reasonable summarized version of the original document. People have tried different aspects of summarizing to get a shorter yet an informative definition of document. This paper tries to utilize using nature inspired algorithms to implement an auto summarizer of text using pseudo-selected features. The main objective of this research is to use of cooperative nature-inspired algorithm specifically ant colony algorithm in text mining problems, in our case, text summarization. And throughout the paper we will try to show how this system can be achieved as well as show the performance and effectiveness of the measurement. We have used the standard data used to test summarization techniques, DUC data and at last comparing it to two algorithms for further analysis.

Keywords: Automated text summarization · Ant colony system · Natural language processing

1 Introduction

Summarization is a technique of getting valuable information while minimizing or condensing the data in size, meaning that getting the smaller version of the data should be at the cost of the size of the original data and not on the information being held on by the original data. The art of Summarizing is being used in our lives so much that we sometimes tend not to notice it. When we wake up from our sleep and browse our phones and look for news, when we meet new people we try to categorize and summarize their attitude and behavior based on their cloth or interaction with us, when trying to find a good restaurant to eat we check for good reviews on food quality or price, or you go to court where the judge sums up evidence found on a defendant or even scientists reading

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_50

the abstract of a paper to get a glimpse of what the paper is about. Although Summarization has good promise in making our lives easier there is another side of it. How do we know how good it is? There can never be a perfect summarizer because there is a problem on who summarizes it, and to get good summary the author is the one that knows the main points, although he or she may argue all of it is important and hence there is no need to summarize. So, the closest summarization lies in the hands or rather in the mind of the author after all he or she is writing it.

A summary can be seen in three ways based on target readers or function. The first is informative summary, which collects relevant or factual information in a concise manner to act as a substitute for the source. The other one is critical summaries (or reviews), where its main area will be to incorporate opinion statements from the content. The last being indicative summary, which provides content to alert users to relevant sources and hence help users make decisions to read the content in more depth [1].

2 Related Works

Looking back at history, at least written history, it was around 1958 that Luhn started working on summarizing scientific documents (research done at IBM) using word and phrase frequency to create a paradigm that started the work of auto summarization [2].

At the same year Baxendale and in 1969 [2] Edumndson published and tried different instances of summarization techniques. Most works published after that concentrated on different domains but mainly newswire data. In deeper level or sense they were not only battling the problem domain of text summarization but also the power of computers and limitation of data for testing.

The techniques used in the 1950's and 60's were simpler to process, largely due to the memory and power of computers and small corpora of text. But nowadays having both of the problems removed, at least comparing to the old times, different models and algorithms have been used to make a suitable solution for text summarization. For the sake of understanding we will try to see very well and general ways used by different researchers throughout time [3].

One of the most common and widely used techniques for problems in text mining is machine learning approach. A Machine Learning approach can be contemplated based on a set of documents and corresponding set of summaries to relate it to. And this works by building a trainable summarizer obtained from using a machine learning algorithm of both collection of documents and its respective summaries. The sentences within a document can be modeled as vectors of features extracted from the structure of the text. For instance, taking the summarization problem we can use a binary classification system, where the sentence can be categorized as "correct" if is in the extractive reference summary otherwise as "incorrect". The trainable summarizer "learns" the patterns used to classify each sentence into either sets of classes, and thus creates an extractive summary [4].

Other than machine learning systems, there are many semantic analysis techniques which are applied on text summarization to find the relation between different sentences. As examples of semantic and syntactic summary techniques we can name examples like graph representation, lexical chains, natural language processing. In the graph representation lexical graphs, graph matching, weighted graphs and unweighted graphs are used for summarization. In lexical chains word net, co-reference chains and lexical semantics etc. are used for summarization. Natural language processing used information extraction, part of speech tagger for the summarization [5] .

Outside of the above-mentioned approaches there are researches on using evolutionary algorithms to implement a summarization algorithm. One notable example will be a genetic algorithm based text summarization which will be later compared to our approach. Their algorithm tries to incorporate genetic algorithm by calculating a score function based on readability factor, cohesion factor and topic-relation factor [8].

3 The Proposed System Model

When dealing with text summarization and ant colony [9, 10] together, there are things that must gel and correspond in order to make sense and use ant colony to its fullest. During experiments, there were lots of behaviors that are seen and hence that helped us modify and build this model. And one of the things that was important was features extracted from the documents. In our case, we used different features for building this model. All the features have been selected as a pseudo candidate for the purpose of building this ant colony model text summarization on the basis of high recommendation on lots of papers together with different experiments to test their effectiveness. We regard them as pseudo candidate features because even though they were used for this model, we think extracting features should be different for different document categories. For our case, we used a different weighting system to make the features more adaptable to the documents.

3.1 Proposed Algorithm Methodology

The first thing that was done prior to using a customized ant colony system was given a document or a collection of documents we clean the data; this was done in regard of stemming and omitting some of the stop words. The next important step was extracting features from the document.

Each ant will save the path its going, the sentence (node) it has visited, each ant will also pick the next node based on pheromone factors, cost (score) functions, and other factors. The features were used because we needed to construct a scoring function for each sentence which will be interpreted as a node in a graph later, when integrated with the ant colony system. Below we will discuss some of the pseudo-features. Here we are using the context of pseudo to reflect that the idea of using only specific features and saying this will work for all does not make sense. In fact, using a static feature and integrating it to a dynamic system like ant colony might hinder in some way as will be seen in some part of our experimentation results later on.

(1) Word frequency: this is the most widely used feature where we record the number of times the word appears within the document. After collecting this we add the

word frequency score of each word found in a given sentence to give a score of a sentence based on the word frequency.

(2) TF-IDF: Using this feature we can calculate the relative frequency of words in a specific document compared to the inverse proportion of that word over the entire text document. This will help us in capturing the relevant words in specific document.

(3) TF-ISF: In summarization problems we can use the same idea where given a single document select a set of relevant sentences to be included in the extractive summary out of all sentences found within the document. The idea of collection of documents in information retrieval replaces the concept of a single document in text summarization. And the notion of document, as a member of a collection of documents relates to the idea of a sentence as an element of a document in summarization.

(4) Sentence Length: This basically is the length of the sentence. Its use for instance can be seen if we want to omit short sentence that have no significance in the construction of the summary. In some cases, we normalize the length of the sentence. To achieve that we take the ratio of the number of words occurring in the sentence over the number of words occurring in the longest sentence of the document.

(5) Sentence Position: This feature can involve several items, such as the position of a sentence in the document as a whole, it's the position in a section, in a paragraph, etc., different techniques were used here.one was to simply divide a paragraph into three groups, namely the top, middle and bottom of the paragraph. And more score was assigned to sentences that are found in the top and bottom of the paragraph. The other was using the percentile of the sentence position in the document, as proposed by Nevill-Manning [11] and the final value is normalized to take on values between 0 and 1.

(6) Similarity between sentences (intersection): For each sentence k we compute the similarity between k and other sentences in the same document, and then we add up those similarity scores, in the end obtaining the raw value of this feature for k. This process will be repeated for all sentences. At the end we will have a square matrix of values filled with similarity measure.

(7) The occurrence of nouns, pronouns and named entities

So based on that the score of a sentence i in a document will be

*S(i) = w1*word_frequency + w2 * TF-IDF + w3 * TF-ISF + w4 * Sentence Length + w5 * Sentence Position + w6 * Similarity between sentences(intersection) + w7 * nouns, pronouns and named entities*

As it can be seen each feature will have their own weight distribution for calculating the aggregate score a sentence. One reason for using weight distributions is to change the flexibility of the features to accommodate different documents.

The next step after this will be initializing the ant colony system parameters.at this stage when ants start traversing through nodes each sentence will act as an element of a node within a graph structure. Since we have simulated the sentence of a document as if they are nodes of a graph, we have to initialize pheromone score values from one node (sentence) to another node (sentence). The value of pheromone was mostly given a value of 1 in our experiments.

The first job within the ant colony module will be to march the ants. If it is the first time the ants are marching, then a random node (sentence) is chosen as a starting node position. This procedure will be done for every ant in the system.

In the picking next node (sentence) module, if it is the first iteration then the ant will choose a random sentence otherwise it will use the probability calculation module to help the ant guide it to the next node (sentence). Here the **Pr** controls the probability that an ant will simply wonder to any document.

In the calculation probability module, the ant will try to calculate the probability of going from one index (sentence) to another document.

$$p_{xy}^{k} = \frac{\left(T_{xy}^{a}\right)\left(N_{xy}^{b}\right)}{\sum_{y \neq x}\left(T_{xy}^{a}\right)\left(N_{xy}^{b}\right)} \tag{1}$$

This equation calculates the probability of moving from node x to node y at iteration k. Looking at the numerator, we have pheromone deposited between x and y given by T. We raise T to the power of alpha, as the training parameter alpha governs the effectiveness of pheromones. In the numerator, we also have N, which represents the value of moving from node x to node y. We raise N to the power of beta, as the training parameter beta governs the influence of cost on the ant movement. We divide the product of N and beta for the desired x and y by the summation of the T and N for all node transitions except the current node the ant resides in. The summation in the denominator of Equation above looks at the total value of the entire unvisited graph from the desirability of score and pheromone. We subsequently evaluate each potential move as the percentage of that total value to determine the probability of each potential move being chosen. Based on these probabilities, a random selection is made to decide the next cell for the given ant. Here we will also use the sentence score we have calculated within the eta variable.

$$N_{xy} = \frac{1}{Max\left(Score_{function(Node\,X, Node\,Y)}, R\right)} \tag{2}$$

The above score function will be the aggregate sentence score of x minus the aggregate sentence score of y.to avoid division by zero. The score function was made to be a maximum value between the aggregate sentence score of x minus the aggregate sentence score of y and value of R, so in case we have same score values between x and y (although it is very rare to happen), the value of the score function will be R. After picking next node the ant goes marches and adds the path from current sentence to the chosen sentence. The more a node (sentence) is chosen its rank within the document rises and hence will be chosen as one of the sentences to be in the summary. to control the convergence of this solution we used maximum iteration and another cost function that will handle if there is no instance new solution that are being added to the pool in order to see if it is stagnating. Following that the ant has to update pheromone scores. Once the ants have all marched through their complete paths, the pheromone trails must be updated. This two-part process considers both evaporation and pheromones deposited by the ants.

$$T_{xy} = RT_{xy} + \sum_{k} \Delta T_{xy}^{k} \tag{3}$$

The variable (T_{xy}) represents the pheromone strength between nodes x and y, in which we're calculating this value. The variable R specifies the evaporation rate training argument. represents the amount of pheromone left by an ant between x and y by ant k. Next we apply evaporation to all pheromone values. The evaporation rate will decrease amount of pheromone for every iteration, giving less value for the path from one document to another. For every sentence or node that is visited more by ants its rank gets higher a higher and hence will have more probability of being chosen as a summary based on when the iteration stops. To stop the iteration, we use two ways.one of them is setting the maximum iteration and the other being a stopping criteria based on how much of the sentence are being visited. For instance, if the highest ranked (more visited with ants) is increasing in a faster rate rather than other sentences then we can stagnate the iteration. For this purpose, we need to record the rate of growth for each sentence in each iteration to decide this.

4 Results and Discussions

4.1 Evaluation Metrics

Summary evaluation method attempts to determine how adequate (and reliable) or how useful a summary is relative to its source. We can evaluate summarization techniques following two ways. The first is using intrinsic method where humans evaluate the quality of summary. With intrinsic it tries to assess the coherence and contained information within the summaries. This works by which users judge the quality of summarization by directly analyzing the fluency and how well the summary covers the key ideas on how well it compares to some ideal summary written by the author or a "qualified person" [1, 12]. Whereas extrinsic methods measure the quality of a task based on performance measure and test the impact of summarization on tasks on reading assessment, reading comprehension [13]. And based on criteria for summary evaluation we have one based on precision and recall or compression Ratio and Retention Ratio. In our case we will be following the precision and recall measurement as our since the papers we are going to use to compare use this method [12].

4.2 DUC Data

When building this model we used DUC (Document Understanding Conferences) data [6, 13] of 2001 and 2002. one of the main reasons DUC is used is because it is the more official testing documents for text summarization and the other thing we choose DUC

2002 version specifically was that we wanted to test and compare to other well designed architectures.

Since the idea was to compare the evaluation of our approach with the genetic algorithm based approach and another algorithm noted on their paper, we selected the 10 documents each having average of 7 documents from the 'Document Understanding Conference (DUC), 2002 data. In each set we find two summaries, a 400 word and a 200-word summary. Both the 400 and 200-word summary are compared with two human made 400 and 200 word summaries respectively, as given by DUC.

One notable problem that arose when trying to evaluate and compare with other algorithms was each research paper uses different data. And the one who use data like DUC, they omit the documents they tested on it. After going through papers we found a paper entitled "Summarizing text with a genetic algorithm-based sentence extraction" which uses and tests DUC 2002 documents. In addition to that they specify the documents [7] they are using. In addition to that, during their experimentation, they compared to another algorithm (unnamed bur found in DUC 2002). With this in mind we made a request to organizers of the data and asked to get the data on the DUC 2002.

4.3 Experiment Setting and Results

We used precision and recall measurement in order to match the evaluation done within the other paper. The equations below show the way recall and precision have been calculated.

Overall both the genetic based and our approach score more or less the same but both are better than the third approach (previously used in data of DUC 2002). In the introduction we talked how hard it is to measure a summarizer. In a controlled experiment using DUC 2002 it showed a competitive result to that of the genetic algorithm version. But if we take them out in the real world where there is real fluctuation and no feedback system it will be hard to sustain good results. The main problem being outputting one summarizer for one document. One of the main advantages in using the proposed model is its flexibility in the output or final solution. Using the proposed approach, we can get one optimal summarizer chosen or elected (in aggregate manner) or use multiple individual ants results as an output or final solution. We can even merge some of the ant's solution and still go away with two or three different summaries. This effect can even be more relevant if the document is larger. Probabilistically living us with more chance of getting the main ideas while still minimizing the content. The below Figs. 1, 2, 3 and 4 show the results where the x axis is the document id and y axis being the average or maximum precision depending on the image. In the documents where the ant colony fails to have a better score. Firstly, this could have arisen from the difficulty of getting the correct value for the different parameters found in the system. And the other reason might be that the pseudo-features might not capturing the information as expected. And this is important because the way of using static features in text summarizer depends on what kind of document it is. In our case we used weighting techniques to lower that kind of mistakes. For both cases, integrating this approach with machine learning one where we can capture the patterns used might help us on deciding and getting a better subset of feature from that specific document or within the category of the genre of the document.

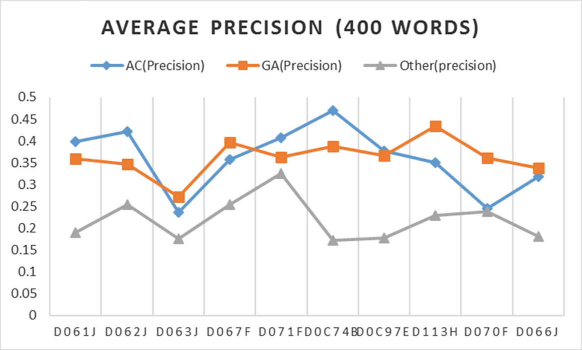


Fig. 1. Average precision of summaries of length 400

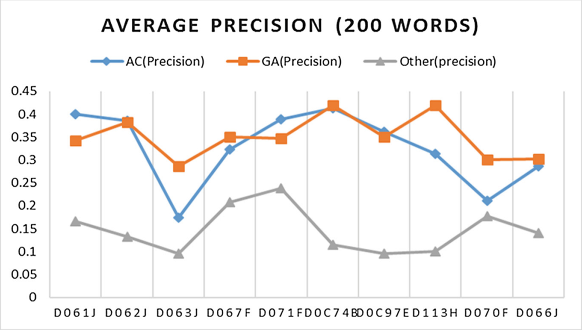


Fig. 2. Maximum precision value of summaries of length 400

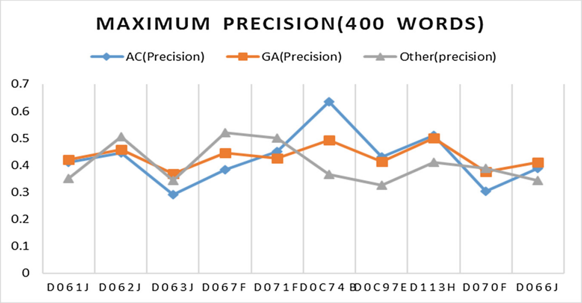


Fig. 3. Average precision value of summaries of length 200

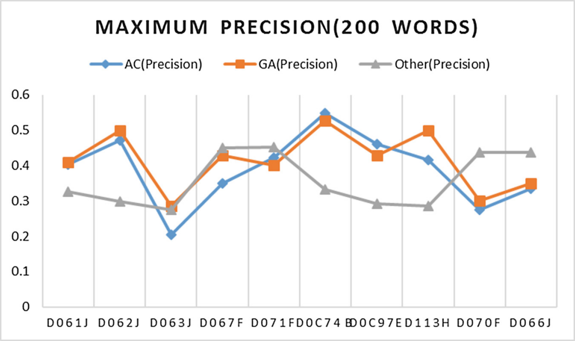


Fig. 4. Maximum precision value of summaries of length 200

5 Conclusion and Future Work

Throughout this paper, we have tried to show an ant colony system based auto text summarizer. And when comparing it to other algorithms it has a comparatively similar score overall to the genetic version and better score when compared to the other (past used in DUC 2002) algorithm. In respect to problems knowing the values of different parameters was a bit difficult. To enhance this system integrating it to other machine learning system will be the next step. The idea that we wanted to use a pseudo features gathered to best explain the majority documents also shows that given a set of features we need to create a flexible feature extraction in order to suit the behavior of a document.

One of the most important factors to choose the ant colony based approach is the fact that due its cooperative behavior we can control the number of summaries generated. And thus, on a real-world scenario instead of worrying if that single summary is a good one, we could assemble two or three summaries by using our approach. Since getting feedback for text summarization systems is difficult having multiple but strong summaries could be one way to minimize the mistake.

When people write a book or document they are guided with a set of rules, be it the grammar of the language or the style and behavior of the author reflecting it on his literature and hence using a set of static features is not something that will be suitable to understand the flexible behavior of authors. Our pseudo features might have covered some of the way of writing a summary with respect to written or unwritten law of English language but a problem arises when the pseudo features struggles to fined patterns in a document where it differs by some way to other documents. And these two problems will be the focus on the future work. Basing our idea on that we can trust to look at an author always following the grammar rule of English and knows how to write a good book will not be adequate. So accommodating some amount of error that an author might create will enhance the experiment result. How to track wrong way of writing and turn it to our advantage, this could one part where machine learning might come. And this will be one challenge we look for to and thereby enhance my idea even more.

Acknowledgments. This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education [NRF-2016R1D1A1B03933875].

References

1. The challenges of Summaries. http://www.cs.columbia.edu/~gmw/candidacy/HahnMani00.pdf
2. Das D, Martins AFT (2007) A Survey on Automatic Text Summarization, Language Technologies Institute Carnegie Mellon University
3. Hovy E, Lin C-Y (1999) Automated text summarization in SUMMARIST, In: Mani I, Maybury M (eds) Advances in automatic text summarization
4. Neto JL, Freitas AA, Kaestner CAA, Automatic text summarization using a machine learning approach
5. Sherry, Bhatia P (2015) A survey to automatic summarization technique. Int J Eng Res Gen Sci 3(5):1045–1053, September-October 2015. ISSN 2091-2730
6. Nenkova A, Columbia University Automatic text summarization of newswire: lessons learned from the document understanding conference
7. Qazvinian V, Hassanabadi LS, Halavati R (2008) Summarizing text with a genetic algorithm-based sentence extraction. Int J Knowl Manage Stud 2(4):426–444
8. Agarwal P, Mehta S (2014) Nature-inspired algorithms: state-of-art, problems and prospects. Int J Comput Appl (0975–8887) 100(14):14–21
9. http://www.cleveralgorithms.com/nature-inspired/introduction.html#problemdomains
10. http://www.heatonresearch.com/aifh/vol2/
11. Kundi1 FM, Asghar1 MZ, Zahra1 SR, Ahmad S, Khan A, A review of text summarization, MAGNT Research Report (ISSN. 1444-8939), Vol 2 (4), pp 309–317
12. Gupta V, Lehal GS (2010) A survey of text summarization extractive techniques. J Emerg Technol Web Intell 2(3):258–268
13. http://duc.nist.gov/data.html

Large Scale Text Classification with Efficient Word Embedding

Xiaohan Ma, Rize Jin(✉), Joon-Young Paik, and Tae-Sun Chung

Computer Engineering, Ajou University,
Worldcupro 206, Yeongtong-gu, Suwon 16499, South Korea
{maxiaohan,rizejin,lucadi,tschung}@ajou.ac.kr

Abstract. This article offers an empirical exploration on the efficient use of word-level convolutional neural networks (word-CNN) for large-scale text classification. Generally, the word-CNNs are difficult to train on large-scale datasets as the size of word embedding dramatically increases as the size of vocabulary increases. In order to handle this issue, this paper presents a de-noise approach to word embedding. We compare our model with several recently proposed CNN models on publicly available dataset. The experimental results show that proposed method improves the usefulness of word-CNN and increases the accuracy of text classification.

1 Introduction

In recent years, neural network models, *convolutional neural network* (CNN) [1] and *recurrent neural network* (RNN) [2], have achieved remarkable performance in many important applications such as text classification [4, 8], semantic parsing [3], sentiment analysis [5]. The CNN model has been shown to be effective for text classification in both word- and character-level. While, character-level CNN models work well with large datasets, as shown in [6], the word-level model is untested on the large-scale datasets.

Kim [4] proposed a shallow word-level CNN model with one layer of convolution on top of word vectors trained by Google News [7]. His model tested only on small datasets with largest vocabulary size being 21,323. Kalchbrenner et al. [8] introduced a convolutional network architecture named the *dynamic convolutional neural network* (DCNN) [7]. In his model, the largest dataset among the used datasets has 76,643 vocabularies. These previous word-level CNN models cannot afford to train on the large-scale datasets which has more than 260 k vocabularies with a 4 GB graphics card.

This paper presents a de-noise based word-CNN model (dn-CNN) which scales to large datasets. We tested the dn-CNN model on DBPedia dataset [5] with the vocabulary size being 610 k. Our experimental results showed better error rate over the existing word-CNN [3, 7] and char-CNN [5] models.

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_51

2 Motivation

The DBPedia corpus is a representative large dataset, which is extracted from the English edition of Wikipedia consists of over 400 million facts that describe 3.7 million things [6]. The DBPedia dataset used in our experiments is constructed from the DBPedia corpus, including 14 non-overlapping classes.

The DBPedia dataset contains 610 k distinct words, the size of the corresponding word embedding is 1.5 GB. Such large-size word embedding requires a large amount of memory, and it is difficult to train the large scale dataset on a single sever. That is why most of the previous word-CNN models were not tested for the DBPedia dataset.

On the other hand, the DBPedia dataset has many inadequate words for word-level training. We randomly selected 100 words from DBPedia. There are 48 misspelling words and uncommon used terms (such as person names, brands, etc.), as shown in Table 1. Therefore, we roughly estimated that 48% of words in DBPedia are insignificant words. These words will cause the waste of resources and have a negative effect to the performance of word-level CNN models.

Table 1. Randomly selected 100 words from DBPedia. (The shaded words are unusual or misspellings.)

scathing	socio	dimension	cavalry	crossthe
dillon	tuam	amongst	kenyan	180000
readership	copy	glatton	breastwork	zurich
automation	158th	forbes	lawson	rollins
virtuoso	fingerstyle	melodic	compositional	spans
classed	samba	bossanova	shred	constitutes
cottonwood	chair	occupational	licensure	labor
robur	pedunculate	anatolia	caucasus	shahkahan
shahkah	marz	dadkhoda	qaleh	ganj
nancy	dye	schrom	13th	oberlin
labeshka	labeshk	lishak	shk	dehshal
astaneh	ashrafiyeh	143	ndc	administrated
'colleges'	pipe	cylindrophis	opisthorhodus	cylindrophiidae
hammerbach	freital	yohanan	kuszyna	kie
czyg	paj	czno	68	symplocos
anomala	corolla	ripe	chiefly	900
6000	panay	varanus	mabitang	lizard
frugivore	destruction	deforestation	overhunting	icun
darker	coloration	bitatawa	agrij	egregy
patak	mellor	bc	overlooks	exists

Our analyses about DBPedia encouraged us to be interested in the reduction of vocabulary size by eliminating insignificant words.

3 De-noise Based Word2Vec

3.1 Basic Idea

To handle problems of large vocabulary size and misspellings of the DBPedia dataset, we propose a novel approach, called a de-noise based word2vec, for constructing word embedding. It aims at reducing the vocabulary size of large dataset by filtering out miss-spelled or uncommonly used words. For this, we can use statistics methods, such as chi-square and *term frequency-inverse document frequency* (TF-IDF), or reliable corpus, such as Google News [7] and PubMed biomedical dataset [10]. In this implementation, we limit the vocabulary size of DBPedia to 70 k by referring to the reliable corpus of Google News. In details, our word embedding is cross validated and extracted from word2vec [6] that were trained on the billion words Google News corpus. The proposed approach not only reduces the size of vocabulary, and also filters out the misspelled words from the resulting word embedding.

3.2 System Architecture

The model architecture is shown in Fig. 1, which consists of 5 layers. The first layer is the input (embedding) layer, with the embedding dimension being 300, the longest sentence length being 708, and batch size being 50. The second layer is a *convolution* layer with three different filters whose sizes are 3, 4, and 5. The number of filters is 100 for each size. That is, the convolutional layer generates three different shapes of feature maps: 706 × 100 by 3 × 300 sized filters, 705 × 100 by 4 × 300 sized filters and 704 × 100 by 5 × 300 sized filters. Then the *max-pooling* layer is applied to the feature maps and the corresponding results is reshaped to two dimensional shape (1 × 100) per feature size, which are then joined together to produce a single matrix (1 × 300). A fully connected layer with *dropout* probability of 0.5 generates the prediction and then a *SoftMax* layer normalizes the results (14-class probability).

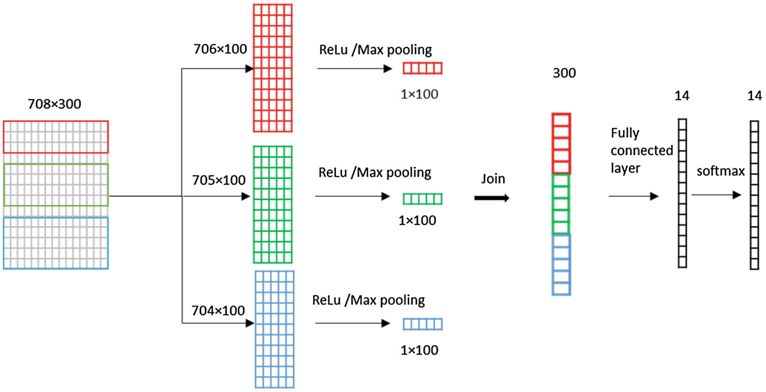


Fig. 1. Dn-CNN model architecture.

4 Experiment

Figures 2 and 3 show the comparison of test performance for our dn-CNN model, with CNN-static model of word level [4], and with character-level CNN model [6]. Due to the large word embedding of DBPedia, it is impossible to train the CNN- static model on DBPedia at once. Therefore, DBPedia is split into 4 smaller training sets (small batch), each with about 240 k distinct words. The performance of CNN-static model is shown in Fig. 2. In our dn-CNN model, we reduced the vocabulary size of DBPedia from 610 k to 70 k, therefore it requires less memory when training our model on DBPedia. We can train on DBPedia at once instead of separate trainings of the subsets. We cited the test performance of char-CNN model [6] in the Fig. 3. As shown in the figures, the best error rate for the CNN-static model is 1.49%, the best error rate for the char-CNN model is 1.95%, whereas, the best error rate for our dn-CNN model is 1.15%. We archived a better error rate over the state-of-the-art models.

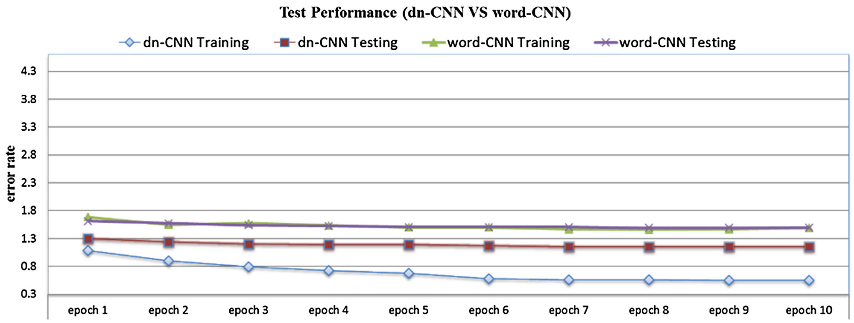


Fig. 2. Performance of dn-CNN model against batch based word-level CNN [4].

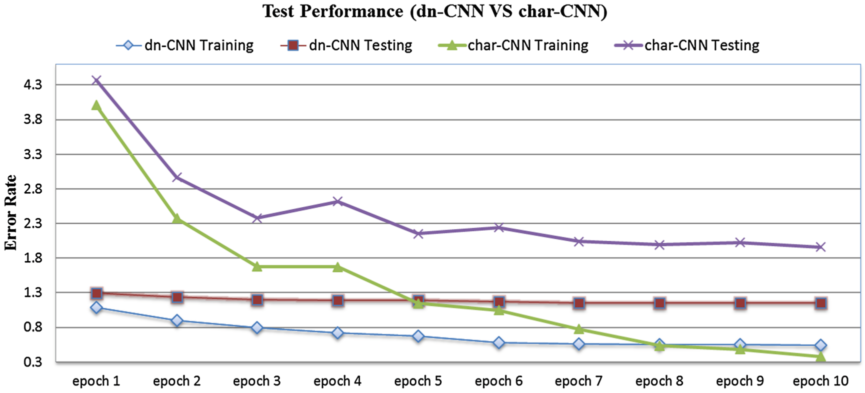


Fig. 3. Performance of dn-CNN model against character-level CNN [6].

5 Conclusion and Future Work

This paper has presented a de-noise based word embedding for word-level CNN text classification, which reduced the vocabulary size of large dataset in the training by selecting common used and correctly spelled words. Our model improved the accuracy of word-level CNN on DBPedia dataset.

For future work, we plan to explore other de-noising methods in order to gain better performance and faster training speed.

Acknowledgments. This research was supported by the MISP(Ministry of Science, ICT & Future Planning), Korea, under the National Program for Excellence in SW) supervised by the IITP(Institute for Information & communications Technology Promotion).

References

1. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324
2. Lai S, Xu L, Liu K, Zhao J (2015) Recurrent convolutional neural networks for text classification. In: Proceedings of AAAI, vol 333, pp 2267–2273
3. Yih W, He X, Meek C (2014) Semantic parsing for single-relation question answering. In: Proceedings of the 52th annual meeting of the association for computational linguistics, pp 643–648
4. Kim Y (2014) Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882
5. dos Santos CN, Gatti M (2014) Deep convolutional neural networks for sentiment analysis of short texts. In: Proceedings of 24th international conference on computational linguistics, pp 69–78
6. Zhang X, Zhao J, LeCun Y (2015) Character-level convolutional networks for text classification. In: Advances in neural information processing systems, pp 649–657
7. Mikolov T, et al. (2013) Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp 3111–3119
8. Kalchbrenner N, Grefenstette E, Blunsom P (2014) A convolutional neural network for modelling sentences. In: Proceedings of the 52nd annual meeting of the association for computational linguistics, pp 655–665
9. Lehmann J, Isele R, Jakob M, Jentzsch A, Kontokostas D, Mendes PN, Hellmann S et al (2015) DBPedia - a large-scale, multilingual knowledge base extracted from Wikipedia. Semant Web 6(2):167–195
10. PubMed. https://www.ncbi.nlm.nih.gov/pubmed

Study on the Strategies for Activating Silver Care in O2O Platform

SeungAe Kang(✉)

Department of Sport and Healthcare, Namseoul University, Cheonan, Korea
sahome@nsu.ac.kr

Abstract. The emergence of the O2O platform that advocates innovation and efficiency is emerging in the case of the elderly care service as well for the drastic increase in demands and service quality enhancement. The main format for the O2O service for the care of elderly is the two-sided business model method that mutually links care service user and provider by using mediation platform as the medium. Representative aged care helper mediation platforms in the US include 'ClearCare', 'HomeHero', 'Care.com', 'Honor', 'Kindlycare' and others. And a start-up company called, 'Please Take Care of My Mother' started service with the O2O mediation platform for the care of elderly in Korea. O2O service for the care of elderly can satisfy significant demands of the elderly who are limited in everyday life and the family members who need to take care of the elderly through the mediation medium called platform. Strategies that will help to improve stable profit structure, build trust that can satisfy both the consumers and provider and intuitive and instant convenience require multi-dimensional efforts for the vitalization.

Keywords: Startup · The elderly · O2O · Care service · Platform · Education

1 Introduction

Klaus Schwab, the founder and the Chairman of the World Economic Forum (WEF) that celebrated its 46th anniversary in 2016, adopted the 4th industrial revolution of the new era that is characterized by convergence between digital devices, human beings, and physical environment, as the key agenda. The concept of 4th industrial revolution's model is based on the concept of 'O2Oconvergence' which entails creating a better world by converging the virtual and reality. As innovation and efficiency are changing, the platform is evolving towards O2O market platform [1].

Domestic and overseas O2O markets are growing in size thanks to the advancement of the ICT technologies such as mobile, Internet of the Things, Big Data, Artificial Intelligence and others as the convenience of the online and field orientation of the offline are connected organically [2]. The Korean O2O market is expected to grow up

Funding for this paper was provided by Namseoul university

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_52

to 320 Trillion Won as the offline businesses of diverse domains get connected via online and mobile platforms [3, 4].

A number of smartphone users are increasingly drastically to the point that this age is called the smart age along with the ubiquitous age. A smartphone that offers superb portability, convenience and real-time that surpasses the computer is becoming an essential product in our everyday life that we cannot do without, and it is gaining attention as the medium that satisfies the educational need of the new age. In particular, a smartphone application which is a field of the ubiquitous era is used in very resourceful manner when it comes to obtaining various information and providing convenient services, and thus applications are being developed in various fields.

In particular, number of the elderly is increasing so fast to the point that the terms such as centenarian that refers to the elderly people who are 100 years old and super centenarian that refers to the elderly people who are 110 years old are emerging. Along with this, it is deemed that anyone can use the elderly care system of the application for the elderly such as the elderly who live alone and re-marrying elderly who are restricted in terms of behavior and social constraint, anytime, anywhere. This is important for the increase and maintenance of physical, social and emotional health of the elderly in Korea and to ensure successful in case of the Korea with the fasted aging speed in the world.

The O2O market for the elderly related industries is expected to grow as well. According to the 2015 National Statistical Office's statistics on the elderly in Korea, people who are at least 65 years old amounts to 6,624,000, which is 13.1% of the total population. One out of five households is headed by the elderly while one-person household consisting of the elderly comprises 7.4%. By 2020, the share of the elderly in Korea is expected to exceed 14% of the total population. As such, the Korean society will become an aging society in a full-fledged manner. In 2060, this share is expected to reach up to 40%. Drastic demand for the elderly's welfare and elderly care service is expected due to the drastic aging. The emergence of the O2O platform that advocates innovation and efficiency is emerging in the case of the elderly care service as well for the drastic increase in demands and service quality enhancement.

Accordingly, this paper seeks to study the representative cases of the elderly care service using O2O service technology characterized by repaid technology advancement and market expansion these days as a means of satisfying the needs of the elderly and their family members since the elderly people need others' help since they face restrictions in their everyday life due to the nature of aging, and to present the direction for the vitalization of the O2O service platform for the care of elderly that can connect care and protection service providers and users, conveniently and safely.

2 Current Status of the O2O Service in Korea and O2O Platform for the Elderly Care Service

O2O(Online To Offline) starts out with the concept in which customers are attracted via online and provided goods or services via offline, not merely shopping offline via online, and is now expanding into a broader concept that covers all types of forms that are connected to the offline from the online or from the offline to the online [5]. Since O2O

Table 1. Examples of O2O services in Korea [4]

Category	O2O Service	Description
Real estate	Dabang Zigbang Dukkubisesang	Real estate services which prevent falsely registered and listed items
	YourHome	A housing social commerce which provides curation services and customizing services
	BudongsanDiet	A real estate brokerage services mainly targeted for apartments and officetel deals
Food delivery	Baedalui-minjok	A food delivery intermediary platform for searching, ordering, delivering various food, and undeliverable restaurant food
	Yogiyo Baedaltonq	Food delivery intermediary platforms for searching, ordering, and delivering various food
Beauty	Mimibox	Sales of beauty products and own PB products
	Beauty-in-now	Real-time nail shop reservation information supply and demand and supply matching service
	Cut & cur	Find the best hair style and connect with a hairdresser
Transportation	Kakao-taxi Tmap-taxi	On-demand transportation services which provides taxis for hire with no commission
	Buttondaeri Kakao-driver	On-demand replacement driver services for drunken drivers
Accommodation	Yanolja Yeoki-eoddae Yeokiya	Intermediary platforms for a variety of accommodations such as hotels, motels, pensions, etc.
Home service	Doctorhouse	Residential space design service
	Daerijubu	Housework-related task service
	Cleanbaske	Laundry service
Car service	Cardoc	Quote and vendor selection service for automobile repairs
	HeyDealer BuyCar Chutcha	Transaction platforms for exchanging used cars
	Socar GreenCar	On-demand car sharing services
	Parkhere	On-demand parking space sharing services

service enables anyone with an idea to participate as a product or service supplier or platform business, start-up companies enter into the related markets and start the service as small and medium sized businesses based on the refreshing idea or innovative technology.

Recently, as the smartphone use is widespread and as the simple payment based on FinTech is getting started, size of the Korean O2O industry is as follow as of 2016; 12 Trillion Won for the food delivery application service, 10 Trillion Won for the quick courier service and cargo delivery services, 8.5 Trillion Won for the taxi service, and 4 Trillion Won for the Rent-a-Car service [6]. Service domains are expanding incrementally. Table 1 shows the types of the O2O services in Korea based on the platform and case studies.

The main format for the O2O service for the care of elderly is the two-sided business model method that mutually links care service user and provider by using mediation platform as the medium. Representative aged care helper mediation platforms in the US include ‘ClearCare’, ‘HomeHero’, ‘Care.com’, ‘Honor’, ‘Kindlycare’ and others. These O2O platforms are different by key customer groups (Fig. 1). Whereas the key customers for the ‘ClearCare’ and ‘HomeHero’ are specialized agencies and care hospitals, ‘Care.com’, ‘Honor’ and ‘Kindlycare’s key target customers are the service users and their family members. ‘Honor’ and ‘Kindlycare’ which are newcomer start-up companies, offer the following advantages; enhanced reliability by disclosing information such as profile and criminal record of the registered caretakers to differentiate from the existing mediation platforms, and the opportunity for the family members to check the contents and results of the service on a real-time basis.

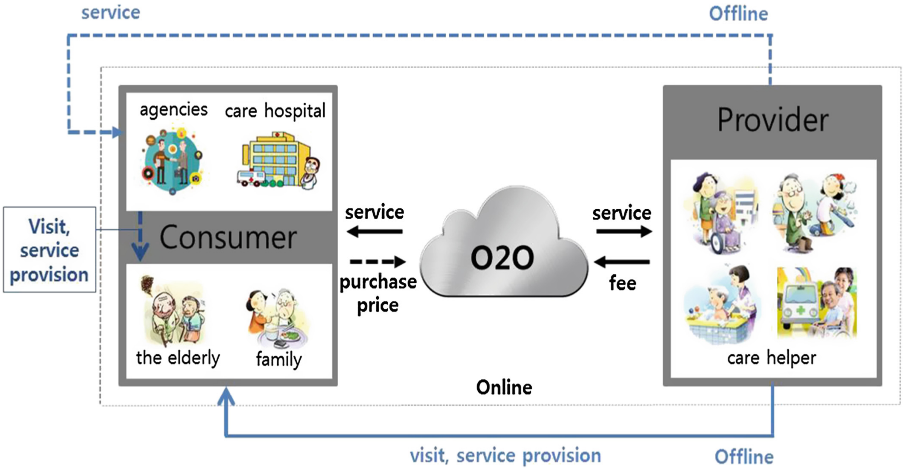


Fig. 1. O2O platform model

In Korea, a start-up company called, ‘Please Take Care of My Mother’ started service with the O2O mediation platform for the care of elderly. This service is comprised of seven services; accompanying to the hospital, taking care during the outing, taking care of the everyday chores, conversing while taking a walk, caretaking while bathing, nursing care, taking care during 24 h, and long-term care. Service provider pays a visit in person when user applies for the service needed, date/time and location after logging onto the platform [7].

A number of targets for the ‘long-term care system’ and ‘care service for the elderly’ that are supported by the government, is limited since targets need to meet certain criteria. Meanwhile, the elderly people often need help from others due to the limitations of the everyday life. Thus, this is the time when the expansion of the O2O service for the care of elderly is needed.

3 Strategy for the Vitalization of the O2O Platform for Elderly Care Service

- **Stable profit structure:** New services are being launched, but there is a problem; it is difficult to ensure a stable profit. In particular, older elderly people tend to need elderly care service more. Thus, they feel burdened when using online method. Instead, they prefer offline method. Accordingly, it is necessary to select the method of expanding main service users to the family members as well instead of merely the elderly. Moreover, presentations for attracting investments need to be held early on and the government's active support are needed to assist the start-up companies that have a promising business model for the care of elderly.
- **Reliability build-up:** Consumer trust is a critical element for the O2O service for the care of elderly. Consumption is bound to decrease if security is not guaranteed since the user needs to select a provider while relying on limited information available on the platform, which is an intangible space. O2O platforms that are in operation today provide information on the caretakers' profile, experience, the result of personality test, and criminal record. However, it is necessary to increase customer trust by providing increasingly detailed information that the customers want by gathering together consumer requirements. Moreover, it is essential to building trust with service providers as well by providing necessary training to enhance work capability to secure increasingly stable personnel pool for the service providers. Instead of merely providing a simple convenient function that connects the service providers and users, it is necessary to build trust by expanding the platform into a platform that encourages participation among many users.
- **Convenience:** Going forth, it is necessary to develop an intuitive and convenient platform by grafting together with diverse ICT technologies in order to provide the services that both the service providers and users can utilize easily. Moreover, it is necessary to select the on-demand economy method [8] which entails providing services that the consumers want immediately in order to grow into the business model that can react instantly in line with the personalized demand.

4 Conclusion

O2O technology that connects the virtual and reality is expected to create new opportunity when it comes to the elderly care service by linking the existing offline service providers and online users centered on the mobile platform. O2O service for the care of elderly can satisfy significant demands of the elderly who are limited in everyday life and the family members who need to take care of the elderly through the mediation medium called platform. Strategies that will help to improve stable profit structure, build trust that can satisfy both the consumers and provider and intuitive and instant convenience require multi-dimensional efforts for the vitalization.

References

1. Lee MH (2016) The 4th industrial revolution and O2O technology fusion revolution. Smart Device Trend Mag 22(4)
2. Kim J, Kwon H, Kim D (2016) Industry trends and challenges in on-demand services in South Korea. ICIC Express Lett Part B Appl 7(9):1933–1938
3. Kang D (2015) Domestic O2O platforms that grow rapidly. Industrial internet issue report
4. Kim DS, Kim KH, Choe DG, Jung JY (2016) Service issues and policy directions for promoting the O2O industry in Korea. J Soc e-Bus Stud 21(4):137–150
5. Rhee SK, Rhee KW (2016) Competition of online platforms in the O2O industry and its welfare effect. J Ind Organ 24(3):57–84
6. Seok HE, Kim SJ, Choi JH, Kim HM, Kang SM (2016) Consumer welfare from O2O-based designated driver service. Int Telecommun Policy Rev 23(4):1–28
7. http://www.caremother.co.kr
8. Ahn SH, Lee MH (2015) Fourth industrial revolution impact: How it changes jobs. Korean academic society of business administration article, pp 2344–2363

A Study on the Convergence Education of Science and Physical Activity Using IT Technology

SunYoung Kang[1] and SeungAe Kang[2](✉)

[1] Department of Physical Education, Korea University, Seoul, Korea
1010kang@hanmail.net
[2] Department of Sport and Healthcare, Namseoul University, Cheonan, Korea
sahome@nsu.ac.kr

Abstract. This study seeks to propose class model that is needed for executing education for the converged human resources and to increase the scientific knowledge of the students using IT technology at the field based on the theoretical framework of the education for converged human resources. During the introduction stage, students are presented with the study materials that can increase their interest. As such, contents of the previous stages – introduction to activity and study goal setting, organizing of the terms during the main activity and demonstration of the physical activity by applying IT technology, organizing and investigation – are re-verified, and students discuss about our body and movement in a diverse and creative manner while undergoing the problem solving process for the given themes for investigation. Through these series of processes, it is expected that the studies can come up with diverse ideas and grow their creative problem-solving ability while carrying out convergence education in the science theories and physical activities. Although this study suggested means to apply convergence education centered on the physical activities of specific grades in stages, it is hoped that a wider convergence education opportunity may be provided going forth by developing differentiated protocol in the domain that enables convergence education by each grade.

Keywords: Convergence · Physical activity · Experience game · Education

1 Introduction

Two papers, ‘New Vision for Education: Unlocking the Potential of Technology’ and ‘New Vision for Education: Fostering Social and Emotional Learning through Technology’ were presented with the theme of ‘New Vision for Education’ during the 2015 and 2016 WEF (World Economic Forum). Since technology change is bound to transform our society’s knowledge consumption pattern inevitably, WEF implies that there is a need to understand how this change in pattern gets connected with ‘education’ from the business perspective to lead to the creation of a new pattern [1].

As for the education paradigm of the 21st Century, educational orientation is changing such as strengthening of the multisensory participatory activity to increase convergence thinking ability in order to cultivate competent futuristic human resources

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_53

with multiple intellectual capability and sensitivity [2]. Information communication technology advancement enabled use of diverse teaching methods, going beyond existing education method in the education field. Due to the IT technology, students who are living in an age of knowledge and information flood need an educational environment that enables them to forecast and prepare for the future society that will be nearing going forth. Towards this end, distribution of the IT based convergence educational contents that enables students to experience and feel in person based on their dream and talent is called for at the field of public education [3]. The government is emphasizing the need for the development of the convergence educational program for the cultivation of converged science and art human resources from the aspect of cultivating creative competent human resources of the future [4].

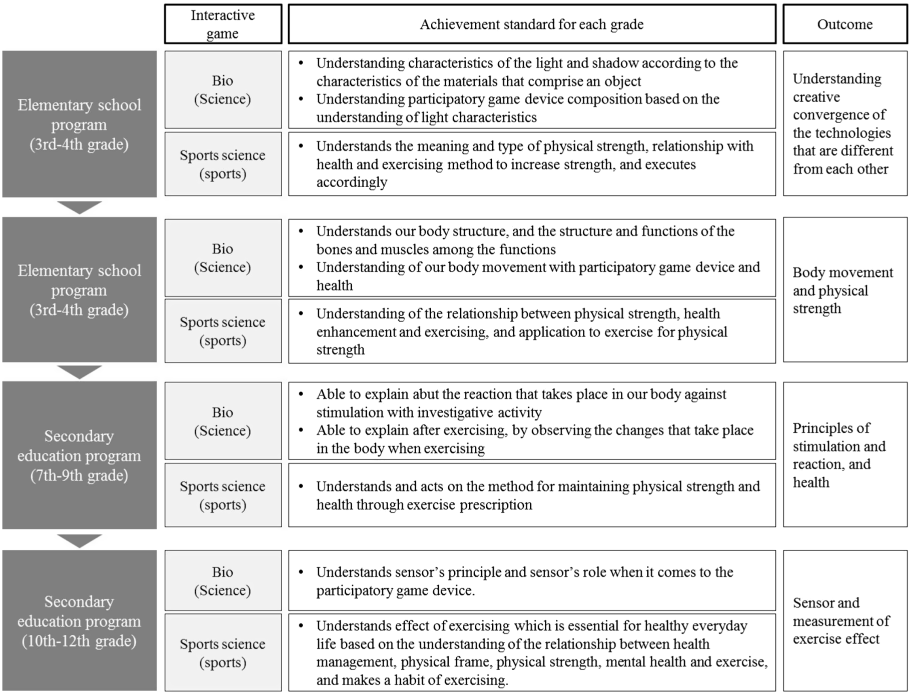
In particular, bio that uses existing applied science that is being carried out in the education field is limited in attracting students' interest and in increasing their understanding. In case of the sports science, the reality is such that the education opportunity is decreasing increasingly amidst the lack of the awareness following physical activity and education characteristics centered on university admission. Accordingly, approaching the bio domain that is suitable for the students' desire will enable approaching the participatory sensibility program as well using not only health and culture, but also game.

In the education field, the need for a new convergence education program is emerging that converges together the science and physical activity domains through applied science thinking by solving problems of science and physical activity education and by motivating students to study. This study seeks to propose class model that is needed for executing education for the converged human resources and to increase the scientific knowledge of the students using IT technology at the field based on the theoretical framework of the education for converged human resources.

2 Academic Achievement via Convergence Education

Convergence education refers to the applied education that converges at least two majors instead of referring to the knowledge of one major to ensure that the knowledge is closely related to the actual society of the current era [5, 6]. New idea is generated while knowledge of academic boundary get combined together and organized [4].

Convergence education needs to be improved so that it is an education centered on students using advanced technology instead of offering an educational environment centered on machine based merely on technology [2]. Through this, it is possible to expect benefits such as creativity development due to the increased awareness of science, cultivation of convergence thinking ability and problem solving ability, development and distribution of program to increase interest and to enable voluntary study, offering of the diverse convergence study materials, and meeting of the physical activity and science. Figure 1 showcases an example of the studying contents that enable achievement through convergence education during elementary, middle and high school curriculum. This paper presents the studying theme that needs to be achieved in the science and physical activity domains by each grade, contents of the



	Interactive game	Achievement standard for each grade	Outcome
Elementary school program (3rd-4th grade)	Bio (Science)	• Understanding characteristics of the light and shadow according to the characteristics of the materials that comprise an object • Understanding participatory game device composition based on the understanding of light characteristics	Understanding creative convergence of the technologies that are different from each other
	Sports science (sports)	• Understands the meaning and type of physical strength, relationship with health and exercising method to increase strength, and executes accordingly	
Elementary school program (3rd-4th grade)	Bio (Science)	• Understands our body structure, and the structure and functions of the bones and muscles among the functions • Understanding of our body movement with participatory game device and health	Body movement and physical strength
	Sports science (sports)	• Understanding of the relationship between physical strength, health enhancement and exercising, and application to exercise for physical strength	
Secondary education program (7th-9th grade)	Bio (Science)	• Able to explain abut the reaction that takes place in our body against stimulation with investigative activity • Able to explain after exercising, by observing the changes that take place in the body when exercising	Principles of stimulation and reaction, and health
	Sports science (sports)	• Understands and acts on the method for maintaining physical strength and health through exercise prescription	
Secondary education program (10th-12th grade)	Bio (Science)	• Understands sensor's principle and sensor's role when it comes to the participatory game device.	Sensor and measurement of exercise effect
	Sports science (sports)	• Understands effect of exercising which is essential for healthy everyday life based on the understanding of the relationship between health management, physical frame, physical strength, mental health and exercise, and makes a habit of exercising.	

Fig. 1. Example of the convergence education achievement standard by each grade

execution and the outcome that may be obtained through this process by using IT devices such as participatory game device as the medium.

3 Convergence Education Class Model of the Science and Physical Activity Domains

This study designed model by different stages that enables elementary and secondary school students to experience technology using IT device in case of the convergence education centered on the physical activity domain to learn science theory and physical activity. This study used the elementary school (5th–6th grade) education program as the standard, and it is shown on Table 1.

Science education program for the 5th and 6th elementary school students aims to build an understanding of basic concept of science and to cultivate creative and rational problem solving ability through the increased science investigation ability and scientific thinking mind set. Contents of the science education for the 5th and 6th elementary school students are comprised of the two areas; 'material and energy' and 'life and earth'. Among them, 'our body structure and function' among the 'life and earth' is the education program that enables use of participatory game device utilizing IT technology. Students can use participatory game device to learn about the structures and

Table 1. Education stages and themes by each class number

	Stages	Sub topic	Key class activities	Time (class number)
1	Setting up theme for investigation and investigation stage	Our body	• Introductory activity: Introduction to the program	80 min (1–2)
			• Activity 1: Understanding our body using model	
			• Activity 2: Understanding participatory game device's method for detecting our body movement	
			• Investigation stage: Guiding theme for investigation	
22	Stage for learning about the structure. function and principle (principle)	Our body's structure and function	• Activity 1: Investigation of our body structure and s function using model	80 min (3–4)
			• Activity 2: Investigation of bone and muscle structures and functions related to movement	
			• Investigation stage: Setting up theme for investigation and planning	
3	Investigation stage	Experiencing participatory game device	• Activity 1: Participatory game overview	160 min (5–8)
			• Activity 2: Investigation of physical activity and our body function using participatory game	
			• Investigation stage: Presentation on the theme for investigation and research plan by each group	
	Presentation and evaluation stage	How does our body move?	• Activity 1: Making evaluation standard	80 min (9–10)
			• Activity 2: Presentation to showcase our body structure and function, and movement using participatory game	

functions of the bones and muscles that control our body movement. A total of 10 class times were used after setting up our body structure and function as the theme for investigation. Students experience the investigation and learning stage that entails executing key class activities and execution stage that entails utilizing participatory game device, execution stage for experiencing physical activity and for investigating our body's function, presentation and evaluation stage.

Administer class for each class number with a theme for learning. Table 2 shows the example of the instruction for guiding class. Class is carried out as follows; introduction, main activity, organizing, and investigation. First and foremost, during the introduction stage, study materials that can increase students' sense of curiosity towards 'our body' which is a theme of the science domain is presented in order to increase their interest and attention towards the academic theme, and they are guided for carrying out activity and study goal is presented. Then, during the main activity, students learn about the terms and structure of our body's bones and muscles related to the movement. In conjunction, participatory game device is demonstrated to observe our body's actual movement. At this time, students are explained about the game device sensor, which is a method that is used by the participatory game device to detect our body movement, and the students undergo discussion process. During the organizing stage, students organize the key concepts learned, and themes for investigation are presented to encourage them to study in greater depth. During the investigation

Table 2. Education stages and themes by each class number

Stage	Stage for setting up theme for investigation and investigation	Class number	1–2
Study theme	• Our body structure and function (movement and bone muscle)		
Study goal	• Understand our body structure. • Able to understand our bone and muscle structure in relation to movement and to explain their functions. • Able to understand principles behind movement. • Able to understand principles behind, movement by experiencing participatory game device in person.		
Studying stages	Teaching-learning activity	Focus of guidance and precautions	
Introduction	• Adopt theme and increase students' Interest by using video material on our body • Guidance for activity and • check study goal Check study aid materials	• Increase students' interest in theme and incite their sense of curiosity. Encourage them to think flexibly and guide them so that they will know what they need to learn about our body structure and function	
Main activity	• Activity 1: Our bone and muscle structure and functions related to movement – Searching for the bone and muscle functions in our body – Bone related to movement	• Introduction to the terms of each structure	
	• Activity 2: Understanding of the method for detecting our body movement on the participatory game device	• Encourage students to become interested in our body movement and reaction of the game device sensor	
Summary	• Organizing the concept of body structure and function, and movement learned at this class • Completion of the theme for investigation by each group by the next time		
Investigation stage	• Setting up the theme for investigation by each group – Searching for the method for solving problems by using creative technique (brainstorming)	• Theme for investigation and research plan may be changed and form free ambience so that the students can come up with diverse ideas	

stage, brainstorming technique is used so that the students can generate diverse ideas through problem solving the themes for investigation.

4 Conclusion

This study made recommendation to enable convergence education that combines together the science and physical education domains by experiencing new technologies by utilizing IT devices. Towards this end, model for executing participatory game device with IT technology applied was presented centered on the curriculum for the 5th and 6th grades of elementary school by linking the science theories and body movements. During the introduction stage, students are presented with the study materials that can increase their interest. As such, contents of the previous stages – introduction to activity and study goal setting, organizing of the terms during the main activity and demonstration of the physical activity by applying IT technology, organizing and investigation – are re-verified, and students discuss about our body and movement in a diverse and creative manner while undergoing the problem solving process for the given themes for investigation. Through these series of processes, it is expected that the studies can come up with diverse ideas and grow their creative problem-solving ability while carrying out convergence education in the science theories and physical activities. Although this study suggested means to apply convergence education centered on the physical activities of specific grades in stages, it is hoped that a wider convergence education opportunity may be provided going forth by developing differentiated protocol in the domain that enables convergence education by each grade.

References

1. National Institute for Lifelong Education (2016) The future of education in the age of the fourth industrial revolution. Lifelong Educ Trend 2:1–15
2. Chong YS (2016) A plan for applying mobile augmented reality (AR) to art-based convergence education. J Image Cult Contents 10:113–126
3. Choi YM, Moon YS (2014) A study on the effective convergence education of physical education and science through the ‘KINECT’ based yoga contents - focused on the fostering core competence in arts & sports for elementary school students. Soc Des Convergence 13 (4):153–169
4. Kim JH (2015) Developing the governing principles of art-based STEAM curriculum. Art Educ Rev 54:101–135
5. Kim HY (2013) The proposition of the directions about convergence-based courses and basic-convergence subjects for systemed convergence education. Korean J Gen Educ 7(2): 11–38
6. You JA, Jin YK (2016) Exploring convergence instructional design for physical education activity in free semester system throughout Olympic education. Korean J Sport Pedagogy 23(1):23–39

A Proposal of Location Aware Shopping Assistance Using Memory-Based Resampling

Wan Mohd Yaakob Wan Bejuri[1,2](✉), Mohd Murtadha Mohamad[1],
Raja Zahilah Raja Mohd Radzi[1], Mazleena Salleh[1], and Ahmad Fadhil Yusof[1]

[1] Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bahru, Malaysia
mr.wanmohdyaakob.my@ieee.org,
{murtadha,zahilah,mazleena,ahmadfadhil}@utm.my

[2] Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka,
76100 Melaka, Malaysia

Abstract. The range of memory specifications of location aware shopping assistance poses difficulties for the developer (in terms of increased time and effort) when it comes to developing a resampling algorithm for mobile devices. Thus, a new resampling algorithm is required with a flexible capacity that would cater for a range of computing device memory devices specifications. This paper develops a memory based resampling in standard particle filter. The memory resampling is capable to read memory specifications of mobile devices before determines the most suitable resampling functions. The authors aim to extend this work in future by implementing their proposed method in a number of different emerging applications (in example, medical applications and real time locator systems).

Keywords: Particle filter · Resampling · Sequential implementation · Memory consumption

1 Introduction

The usage of Global Positioning Systems (GPS) inside building is making difficult for user to locate itself in shopping market. This because of unavailability of the signal in the building due to signal blockage [1–8]. In mobile devices, there many sensor that can be used to manipulated as location sensor such as; FM Radio, Bluetooth, WiFi and inertial sensor [9–12]. It is known the type of sensor is different based on mobile devices brand. Based on that, the FM radio and WiFi sensor is the most ubiquity in term of location determination compared to others [5, 10, 13–18]. Thus, previous research such mentioned in [3], they propose an integration WiFi and FM radio signal as signal locator inside building environment. The research is quite successful in term of ubiquity of location determination. However, it lacking in the implementation in different memory requirement mobile devices, since it require much effort and time to develop for such purpose. In this paper, we propose a memory resampling that can be used to implement in standard particle filter for location aware shopping assistance usage [19–23]. This paper is organised as follows: Sect. 2 outlines the concept of location aware shopping assistance in general. Section 3 examines the detailed design for the proposed method

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_54

which is knows as memory based resampling. Section 4 provides a conclusions and discussion about the implications of this study.

2 General Concept of Location Aware Shopping Assistance

Previous section discuss about introduction of the paper. This section is discusses about general concept location aware shopping assistance. Basically, it consists of three (3) sub system which are field subsystem, interface subsystem and database subsystem in mobile device (see Fig. 1) [24, 25]. This purpose of why it designed this is to ensure the ubiquity of the location aware shopping system [22, 24, 26]. The usage of wireless LAN and FM radio sensor inside field sub system is used to detect any signal surrounding and give it to interface subsystem to process it. At the interface subsystem, which is consist of CPU is used to process any code that written in software. In this time the interface subsystem will always lookup information in database subsystem. Finally, it will display user location shopping in mobile dives screen. Following section, we will discuss about our proposed method which is memory resampling.

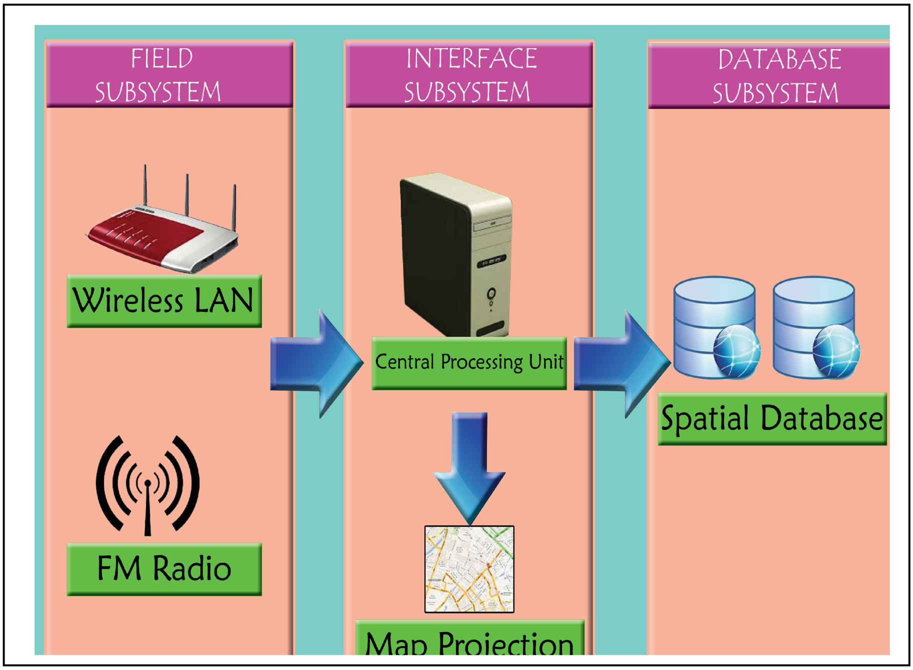


Fig. 1. Fundamental system architecture of location-aware shopping assistance

3 Memory-Based Resampling

Previous section discuss about basic concept of location aware shopping assistance. In this section, we will discuss about our proposed method which is called memory based resampling. By referring in Fig. 2, our proposed method is applied in standard particle filter which will embedded in location shopping aware assistance systems. The purpose of our method is to make easier the implementation of standard particle filter in different memory requirement of mobile devices, which is usually used as hardware for location shopping aware assistance systems. In our proposed method, basically it will observes required memory specifications of mobile devices. If the memory specifications is over than 1536 MB, then it will automatically choose rounding copy as resampling function. If not, it will choose systematic resampling as resampling function. The following section will present about conclusion and discussion of this research.

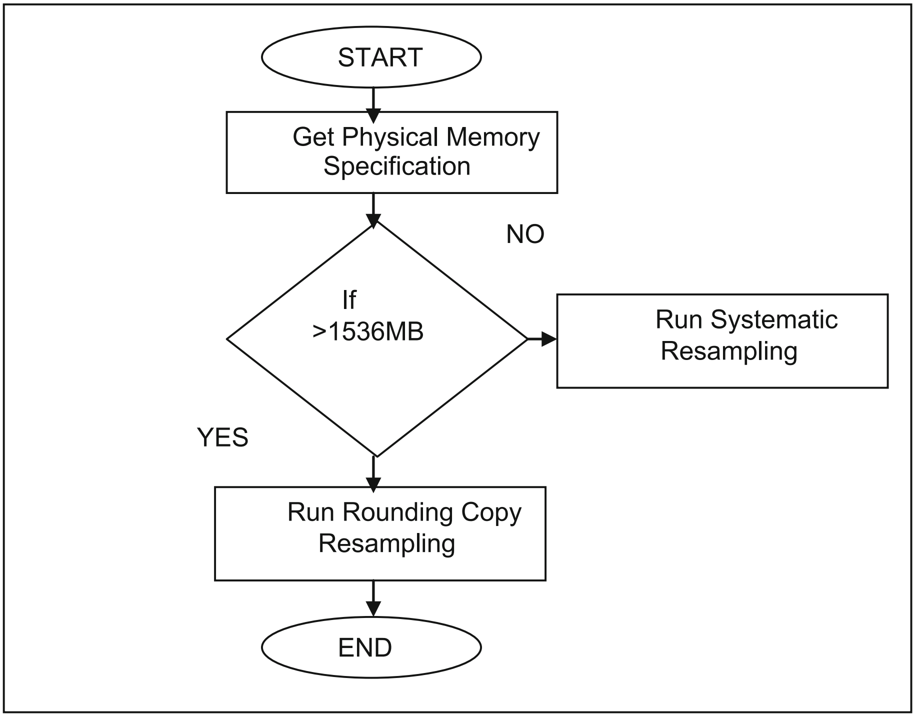

START
Get Physical Memory Specification
NO
If >1536MB
Run Systematic Resampling
YES
Run Rounding Copy Resampling
END

Fig. 2. Flow chart of memory-based resampling

4 Conclusions and Discussions

The varying memory specifications of mobile computing devices cause difficulties for developers due to the additional time and effort required to develop a resampling logarithm. This paper describes the development of a new single distribution resampling logarithm, entitled the AMSFC logarithm, that integrates the traditional resampling logarithm with the traditional variation resampling logarithm in the resampling architecture. The algorithm will switch the resampling algorithm based on memory specification. At the beginning of the operational process, the AMSFC selector is used to select a suitable resampling algorithm (for example, the systematic resampling or rounding copy resampling logarithm), according to the physical memory available in the computing device. Thus, the proposed algorithm (AMSFC) is capable of switching resampling algorithms to suit different physical memory requirements. Finally, this work can be extended by implementing this proposed method in a number of different emerging applications.

References

1. Bejuri WMYW, Mohamad MM, Sapri M (2011) Ubiquitous positioning: A taxonomy for location determination on mobile navigation system. Sign Image Process Int J SIPIJ 2(1):24–34
2. Bejuri WMYW, Mohamad MM, Radzi RZRM (2015) A proposal of emergency rescue location (ERL) using optimization of inertial measurement unit (IMU) based pedestrian simultaneously localization and mapping (SLAM). Int J Smart Home 9(12):9–22
3. Bejuri WMYW, Mohamad MM, Radzi RZRM (2015) Offline beacon selection-based RSSI fingerprinting for location-aware shopping assistance: A preliminary result. In: New trends in intelligent information and database systems, vol 598. Springer, pp 303–312
4. Bejuri WMYW, Mohamad MM, Sapri M, Rahim MSM, Chaudry JA (2014) Performance evaluation of spatial correlation-based feature detection and matching for automated wheelchair navigation system. Int J Intell Transp Syst Res 12(1):9–19
5. Falco G, Pini M, Marucco G (2017) Loose and tight GNSS/INS integrations: Comparison of performance assessed in real urban scenarios. Sensors 17(2):255
6. Ginsbourger D, Roustant O, Durrande N (2016) On degeneracy and invariances of random fields paths with applications in Gaussian process modelling. J Stat Plan Infer 170:117–128
7. Ingrassia S, Rocci R (2011) Degeneracy of the EM algorithm for the MLE of multivariate Gaussian mixtures and dynamic constraints. Comput Stat Data Anal 55(4):1715–1725
8. Li T, Bolic M, Djuric PM (2015) Resampling methods for particle filtering: Classification, implementation, and strategies. IEEE Signal Process Mag 32(3):70–86
9. Bejuri WMYW, Mohamad MM, Sapri M, Rosly MA (2012) Investigation of color constancy for ubiquitous wireless LAN/Camera positioning: An initial outcome. Int J Adv Comput Technol
10. Bejuri WMYW, Mohamad MM, Sapri M, Rosly MA (2012) Ubiquitous WLAN/Camera positioning using inverse intensity chromaticity space-based feature detection and matching: A preliminary result. ArXiv Preprint arXiv:1204.2294
11. Wang B, Yu L, Deng Z, Fu M (2016) A particle filter-based matching algorithm with gravity sample vector for underwater gravity aided navigation. IEEE ASME Trans Mechatron 21(3):1399–1408
12. Teixeira FC, Quintas J, Maurya P, Pascoal A (2016) Robust particle filter formulations with application to terrain-aided navigation. Int J Adapt Control Signal Process
13. Bejuri WMYW, Mohamad MM, Sapri M, Rosly MA (2012) Performance evaluation of mobile u-navigation based on GPS/WLAN hybridization. ArXiv Preprint arXiv:1210.3091

14. Bejuri WMYW, Mohamad MM, Zahilah R (2015) Optimization of rao-blackwellized particle filter in activity pedestrian simultaneously localization and mapping (SLAM): An initial proposal. Int J Secur Appl 9(11):377–390
15. Bejuri WMYW, Mohamad MM, Zahilah R, Radzi RM (2015) Emergency rescue localization (ERL) using GPS, wireless LAN and camera. Int J Softw Eng Appl 9(9):217–232
16. Albert MV, Shparii I, Zhao X (2017) The applicability of inertial motion sensors for locomotion and posture. In: Locomotion and Posture in Older Adults. Springer, pp 417–426
17. Belhajem I, Maissa YB, Tamtaoui A (2017) An improved robust low cost approach for real time vehicle positioning in a smart city. In: Industrial Networks and Intelligent Systems. Springer, pp 77–89
18. Jung S-H, Lee G, Han D (2017) Methods and tools to construct a global indoor positioning system. IEEE Trans Syst Man Cybern Syst 8(99):1–14
19. Bejuri WMYW, Mohamad MM (2014) Performance analysis of grey-world-based feature detection and matching for mobile positioning systems. Sens Imaging 15(1):1–24
20. Bejuri WMYW, Mohamad MM (2014) Wireless LAN/FM radio-based robust mobile indoor positioning: An initial outcome. Int J Softw Eng Appl 8(2):313–324
21. Bejuri WMYW, Saidin WMNWM, Mohamad MMB, Sapri M, Lim KS (2013) Ubiquitous positioning: integrated GPS/Wireless LAN positioning for wheelchair navigation system. In: Asian conference on intelligent information and database systems, pp 394–403
22. Retscher G, Roth F (2017) Wi-Fi fingerprinting with reduced signal strength observations from long-time measurements. In: Progress in location-based services 2016. Springer, pp 3–25
23. Xing H, Li J, Hou B, Zhang Y, Guo M (2017) Pedestrian stride length estimation from IMU measurements and ANN based algorithm. J Sens 2017
24. Li Y, Zhuang Y, Zhang P, Lan H, Niu X, El-Sheimy N (2017) An improved inertial/wifi/magnetic fusion structure for indoor navigation. Inf Fusion 34:101–119
25. Pak JM, Ahn CK, Shmaliy YS, Shi P, Lim MT (2016) Accurate and reliable human localization using composite particle/FIR filtering. IEEE Trans Hum-Mach Syst 47(3):332–342
26. Li X, Wei D, Lai Q, Xu Y, Yuan H (2017) Smartphone-based integrated PDR/GPS/Bluetooth pedestrian location. Adv Space Res 59(3):877–887

A Study on Video Stream Synchronization from Multi-Source to Multi-Screen

Hyojin Park[1], Kyuyeong Jeon[2], Jinhong Yang[2(✉)], and Kuinam J. Kim[3]

[1] Department of Information and Communications Engineering, KAIST,
Daejeon, Republic of Korea
gaiaphj@kaist.ac.kr

[2] R&D Team, HECAS Inc., Seongnam, Republic of Korea
{kyjeon,jinhong.yang}@hecas.co.kr

[3] Department of Convergence Security, Kyonggi University, Suwon, Republic of Korea
kuinamj@gmail.com

Abstract. With the advances of video services and user devices, now users can view multiple live video streams being taken at multiple angles on multiple devices at the same time. However, in order to actually perform such a service, a synchronization technique between plural video streams being played simultaneously on different screens is required. In this paper, we examine in detail the synchronization problems that may arise from the generation, transmission and consumption of multiple video streams. And we propose a structure and method to address the synchronization problem.

1 Introduction

The development of broadband networks and the spread of smart devices are making users to access video content anytime, anywhere using wired and wireless networks. The wired and wireless network, which has not been able to provide sufficient bandwidth, has been upgraded to NGN (Next Generation Network) and LTE (Long-Term Evolution) environment. Smart devices like smart phone or table PC, which are capable of video playback have become popular and diverse. Both evolution led the emergence of video services over IP-based network, e.g., IPTV and Mobile TV service that are mainly run by network operators [1–3].

Such changes affected existing broadcasting system as well to upgrade their television networks with IP-based network technology. Along with the traditional broadcasting station, the introduction of IP-based broadcasting service is underway [4–6]. These changes are not limited to the delivery of broadcasting content, but also in ways of generating broadcast content. In recent years, various devices like drone and smart phone have been able to capture live scenes and deliver them in an IP-based stream format. As a result, more efforts are required to combine the new types of video streams created from a variety of sources into one broadcast environment and to produce and serve vibrant content.

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_55

So far, conventional broadcasting systems have multiplexed several source streams into one video stream and each streams were mapped into one channel as an independent content. Now, owing to the spread of smart devices, it possible to watch a content on multiple screens rather than a single screen. That is, it is possible to receive multiple video streams rather than one video stream for one content and play them at the same time. When the different types of streams are delivered and played with current manner, asynchronization can be occurred due to the difference in quality and in delay between streams so that users can not be able to feel the stereoscopic effect of the content [7].

Figure 1 shows general situation of live broadcasting. The individually but concurrently shot sources are streamed over different network environments and delivered to the video headend. In this process, a time difference between the sources can be happened. Since it is a live broadcasting for multiple screens, the streams are not multiplexed into one stream but delivered as a group of streams through the video headend. In the process, a time difference in the individual section is additionally generated. Hence, a technology that can synchronously play the multiple streams need to be studied and developed.

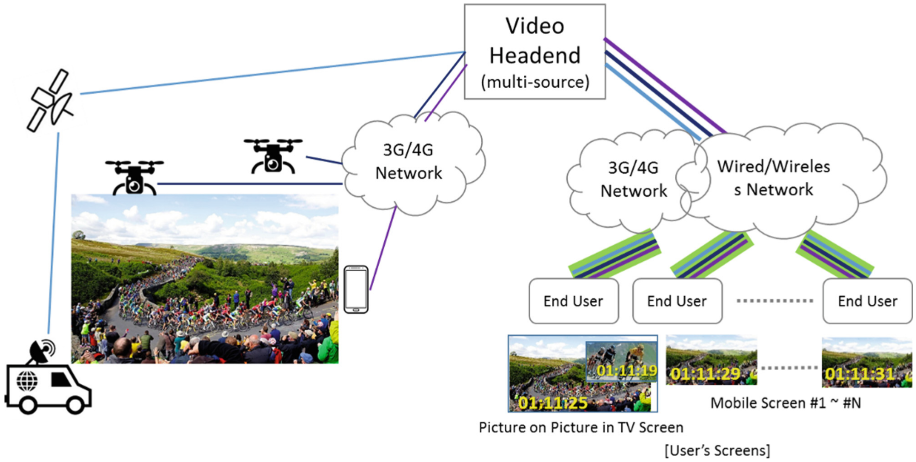


Fig. 1. Multiple video streaming environment

In this paper, we discuss the overall synchronization structure by examining where asynchrony can occur when IP is introduced in broadcasting environment. And then we propose a synchronized live multi-streamed live video transmission architecture. After representing the implemented and verified the synchronization function of multiple user terminals on multiple video streams from Video Stream Server to the final viewer, we conclude with future work direction.

2 Points to Synchronize

The reason why asynchrony occurs between videos shot at the same time and transmitted in real time varies and differs by the video streaming conditions. In this paper, we distinguish the four sections that causes asynchronization in the process of video generation, transmission, and watching as follows: (1) Inter source device synchronization - asynchrony can happen between individual source video generation devices, (2) Up stream synchronization – asynchrony can happen during up stream network transmission, (3) Down stream synchronization – asynchrony can happen during down stream network transmission, and (4) Client device synchronization – asynchrony can happen between individual client devices (Fig. 2).

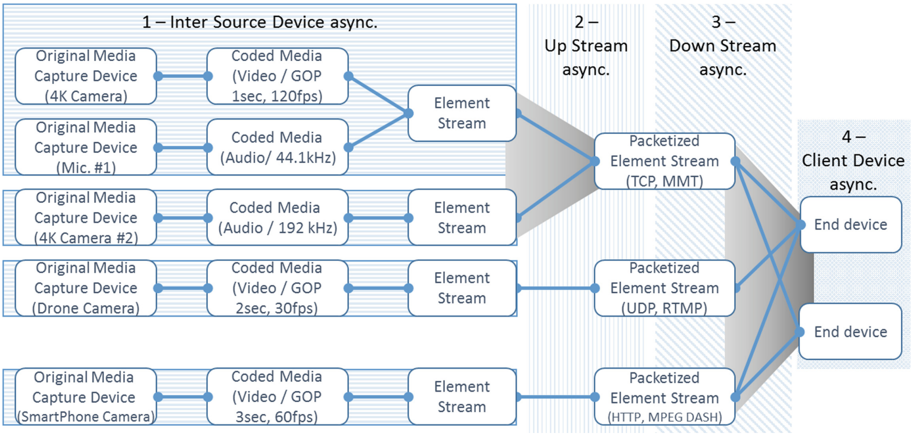

1 – Inter Source Device async.
2 – Up Stream async.
3 – Down Stream async.
4 – Client Device async.
Original Media Capture Device (4K Camera)
Coded Media (Video / GOP 1sec, 120fps)
Original Media Capture Device (Mic. #1)
Coded Media (Audio/ 44.1kHz)
Element Stream
Original Media Capture Device (4K Camera #2)
Coded Media (Audio / 192 kHz)
Element Stream
Original Media Capture Device (Drone Camera)
Coded Media (Video / GOP 2sec, 30fps)
Element Stream
Original Media Capture Device (SmartPhone Camera)
Coded Media (Video / GOP 3sec, 60fps)
Element Stream
Packetized Element Stream (TCP, MMT)
Packetized Element Stream (UDP, RTMP)
Packetized Element Stream (HTTP, MPEG DASH)
End device
End device

Fig. 2. Asynchronous elements in IP-based video streaming from multi-source to multi-screen

The environmental conditions, characteristics, and solutions to cause the delay for each asynchronous differs by each sections. First, in the inter source device asynchronous section, videos are taken and converted into coded media (usually encoding) to generate element streams. In this section, delay and asynchrony can occur between the element streams due to the physical performance of the video shooting devices, the interface of the encoding devices, and/or the difference in characteristics of the video format used in encoding. For example, when shooting and encoding is done by a high-performance video shooting device, it is possible to generate an element stream in the form of 120 fps, which is 4 K quality video. However, when the shooting is done by a general smartphone, a video stream of 30 fps level is generated in 2 K quality. In particular, the processing performance of codec’s depends on the hardware or software, and the delay is also caused by the length of the GoP (Group of Picture). Individual sampling rates of audio can also vary when sampling sound is done by separate audio channels.

Second, upstream synchronization addresses when the individual source streams are packetized according to a transmission medium such as satellite, wired/wireless Internet, LTE network, physical device, etc. and transmitted to a video server. It is the part about the synchronization of the converted video captured into the stream. In this section, the characteristics of the up stream module or software depend on what kind of network environment (wired or wireless) is used and which media transport streaming protocol is used. For example, it is possible to use UDP-based RTMP, TCP-based HTTP, or MMTP, and delay characteristics are generated depending on the characteristics of each protocol and the network conditions. Especially, in case of using multiple network environments, delay due to difference in network quality, and issues such as retransmission cause delays for each input stream on the video server, resulting in an asynchronous problem.

Third, down stream synchronization also corresponds to the network section like the up stream synchronization but addresses the transmissions from the video server to requested user devices. For video streaming with separate streaming channels from the video server without separate multiplexing, each media stream is generated for each source and each delay is sent to each source. Hence, when delay occurs while the transmission on any stream, asynchrony happens at the client devices.

Finally, it is about synchronization between client device. This is the process of decoding the individual video streams received from the network and buffering them on the user device and outputting them to the screen. The same delay issue can arise as the first encoding issue occurs. In case of high quality video stream, a lot of hardware and software resources are needed for processing, so that the postponement due to insufficient resource can occur in this process. This can cause asynchrony due to delays, even though users may be using the same channel video streaming on a separate owned device (Table 1).

Table 1. Asynchronous generation factors for each element

Features of video streaming	Point of delay	Modules of delay generation	Element of delay
(En)Coded media	synchronization between each source devices	Hardware, software	Codec (H.264, 265, etc.) Resolution (LD, SD, HD, FHD, 4 K, 8 K, etc.) Framerate (15 ~ 120 fps) GoP size (1 s, 2 s, automatic), etc.
Network	Synchronization between Up and down stream	Hardware, Software	Segmentation by delivery protocols (HTTP, UDP, and other protocols) Network Bandwidth and quality (bitrate, end to end delay and jitter, etc.)
(De)Coded media	Synchronization between each client devices	Hardware, Software	Codec, Resolution, Framerate and hardware performance (CPU, GUP, memory, etc.)

3 Proposed Video Stream Synchronization Architecture

In this section, we define a video service structure that users can access to video streams formed from multi-source based video streaming environment. Then we propose necessary functional modules and synchronization method for synchronized video play on multiple screens [8, 9].

IP-based video streaming environment generally used is composed of four modules [10, 11]. First, video capture device is a module that plays the role of generating element stream in video shooting mentioned in the previous chapter. Stream uploader is a module for transferring stream from video capture device to video server. The Media Stream Server is a module that delivers the received media to the user. The Client consists of a module that outputs the video stream received from the media stream server to the user's screen.

The following Fig. 3 shows the proposed structure for configuring video transmission environment in which multiple video channels are synchronized and played on a user device [12–14].

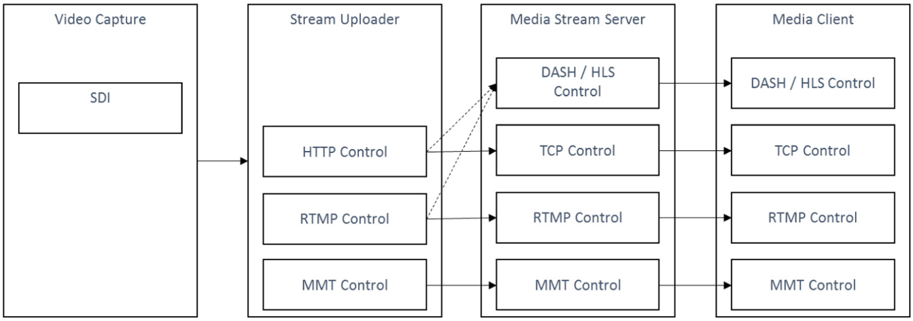


Fig. 3. General media stream system configuration

In order to provide synchronization in the existing video streaming environment, we propose the key functional module as shown in Fig. 4. First, for the video capture device, we define information based on the cross-definition of the device profile or manifest type. For the remaining stream uploaders, media stream servers, and media clients, a module to directly control the device's time and video playback related information is needed. This is because it is effective and scalable to use separate modules to be connected and to handle each of them in order to control the time and setting of the devices directly in the actual environment. The proposed profile and modules are described as follows.

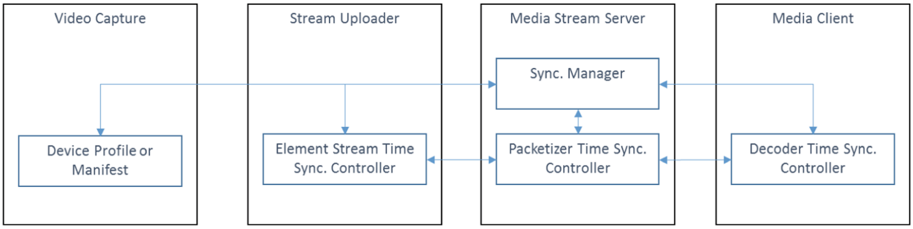

Video Capture
Stream Uploader
Media Stream Server
Media Client
Sync. Manager
Device Profile or Manifest
Element Stream Time Sync. Controller
Packetizer Time Sync. Controller
Decoder Time Sync. Controller

Fig. 4. Proposed profile and modules for synchronization

- Device Profile or Manifest

Device information on video capture may vary but it is not a dynamically changed form. So it can be defined as profile without additional module, and delay information of the module is recorded. The information to be recorded is information on types of codec, encoding options, resolution, framerate, and GoP size.

- Element Stream Time Sync. Controller

Element Stream Time Sync. Controller controls the stream uploader device, and controls the time stamp information and stream protocol settings of the element stream. Control information includes presentation time stamp (PTS) and decoded time stamp (DTS) information of the element stream. This module performs delay control through the buffer controller of the up streamer and transmission priority adjustment according to protocol characteristics.

- Sync. Manager

Sync. Manager, which is a key part of the Media Stream Server is the module that controls and supervises the total delay between multi-source and multi-client video streaming in an IP environment. Through this module, common time stamp (CTS), which is the common time information of the system, is controlled and time synchronization between each device is controlled.

- Packetizer Time Sync. Controller

Packetizer Time Sync. Controller, which is another key part of the Media Stream Server compensates individual time differences in transmission of input stream and output stream, and sets the maximum delay considering overall network conditions. By collecting information on the delay of the source or clients, this module sets the delay information so that can provide the synchronization service with minimum delay. In the case of a client belonging to the network environment exceeding this delay value, the synchronization function between clients can be turned off.

- Decoder Time Sync. Controller

Decoder Time Sync. Controller controls information related to the decoding performance at the client device. It is possible to select the appropriate resolution and frame rate for multi-source stream processing based on the maximum processing frame rate of the player. For this, time stamp information of the de-packetized video stream generated on the decoded buffer can be modified. Also, to control synchronization between individual clients, transmission delay and deviation of start time delay is controlled by managing playback speed at the player with common time.

4 Implementation and Test Result

In this chapter, we have verified the proposed function and synchronization architecture on multiple client devices through implementation of video server module and client device module [15]. To do this, the system consists of two video servers and test players on four smart phones, each of which is configured to be connected via WiFi or LTE as shown below. In addition, one of the client devices is composed of different devices in order to test the decoding performance difference of the client. The video streams used for synchronization test is the record of running stop watch with the frame rate of 30 fps. By doing so, the scale of synchronization can be verified in 30/1000 ms level when the actual video is judged by screen shot.

In the experimental test, two video servers simultaneously broadcast the same video content using console, and each client terminals are connected to the corresponding streaming servers. The following Fig. 5 shows the result of shooting a stop watch playing on the terminals connected to each video stream servers. It confirms that all client devices are playing the same video frame. As a result, with the proposed system, live video

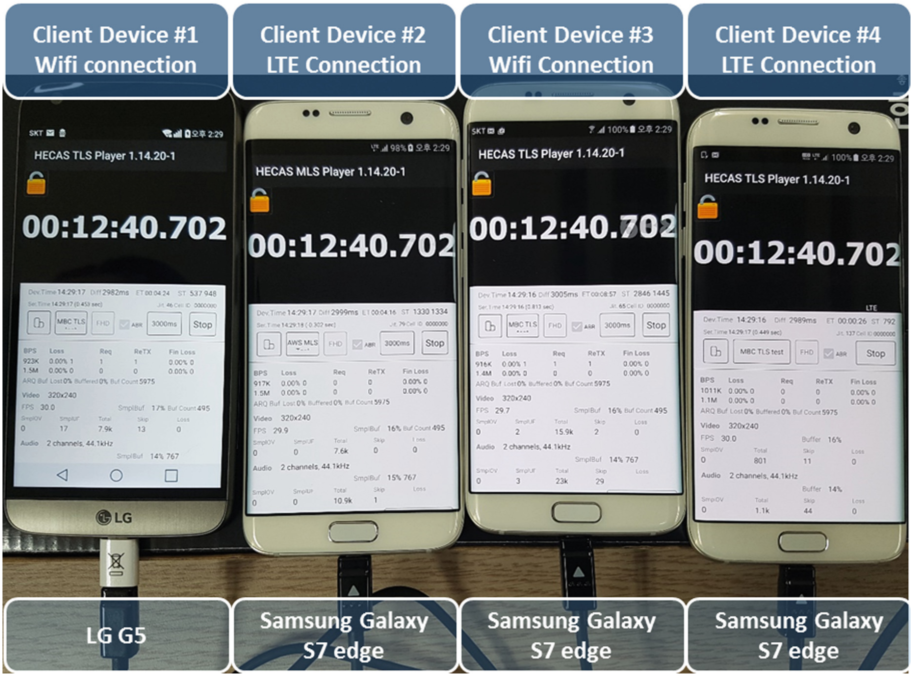


Fig. 5. Time sync. experimental result screen shot using stop watch video

synchronization between multi-video streams and multi-clients are achieved at about 30/1000 ms.

5 Conclusion

In the future, with the advancement of video services and user devices, users can view multiple live video streams from multiple angles simultaneously from multiple devices. To provide satisfying video consumption experience in these service environments, synchronization of the video stream must be provided. In this paper, we have analyzed the synchronization issues that can arise from the transmission and consumption of multiple video streams from a structural perspective. We have proposed a structural and functional methods for solving synchronization problems. However, synchronization of multiple streams proved in this study covers only from the video server to the user terminal. Additional research needs to be done in the future to provide end-to-end synchronization.

References

1. Global Internet Phenomena Report (2016), Sandvine Incorp., Waterloo, ON, Canada
2. Lee GM, Lee CS, Rhee WS, Choi JK (2009) Functional architecture for NGN-based personalized IPTV services. IEEE Trans Broadcast 55(2):329–342
3. Mikoczy E (2008) Next generation of multimedia services - NGN based IPTV architecture. In: 2008 15th international conference on systems, signals and image processing, Bratislava, pp 523–526
4. Lim Y, Aoki S, Bouazizi I, Song J (2014) New MPEG transport standard for next generation hybrid broadcasting system with IP. IEEE Trans Broadcast 60(2):160–169
5. Walker GK, Stockhammer T, Mandyam G, Wang YK, Lo C (2016) ROUTE/DASH IP streaming-based system for delivery of broadcast, broadband, and hybrid services. IEEE Trans Broadcast 62(1):328–337
6. Ohanian T (2016) Moving toward zero infrastructure broadcasting. SMPTE Motion Imaging J 125(8):49–59
7. Angrisani L, Capriglione D, Ferrigno L, Miele G (2013) An internet protocol packet delay variation estimator for reliable quality assessment of video-streaming services. IEEE Trans Instrum Meas 62(5):914–923
8. Ramos FMV (2013) Mitigating IPTV zapping delay. IEEE Commun Mag 51(8):128–133
9. Siebert P, Caenegem TNMV, Wagner M (2009) Analysis and improvements of zapping times in IPTV systems. IEEE Trans. Broadcasting 55(2):407–418
10. Wang J, Xu W, Wang J (2016) A study of live video streaming system for mobile devices. In: 2016 First IEEE international conference on computer communication and the internet (ICCCI), Wuhan, pp 157–160
11. Li X, Salehi MA, Bayoumi M (2016) VLSC: Video live streaming using cloud services. In: 2016 IEEE international conferences on big data and cloud computing (BDCloud), social computing and networking (SocialCom), sustainable computing and communications (SustainCom) (BDCloud-SocialCom-SustainCom), Atlanta, GA, pp 595–600

12. ISO/IEC JTC1/SC29/WG11 MPEG 23009-1 (2014) Information technology dynamic adaptive streaming over HTTP (DASH) Part 1: media presentation description and segment formats, May 2014
13. Parmar MTH (2012) Adobe's Real Time Messaging Protocol. Adobe Syst. Inc.
14. ISO/IEC JTC1/SC29/WG11 MPEG 23008-1 (2014) MPEG media transport, June 2014
15. HECAS Live Streamer and Player (2017). HECAS Inc. http://www.hecaslab.com

Electrical Engineering

A Nine-Switch Unified Power Quality Conditioner with Enhanced Repetitive Controller

Dang-Minh Phan and Hong-Hee Lee(✉)

School of Electrical Engineering, University of Ulsan, Ulsan, Korea
Pdm2708@gmail.com, Hhlee@mail.ulsan.ac.kr

Abstract. The 9-switch power converter can provide two sets of output terminals like the conventional back-to-back power converter. In this paper, the unified power quality conditioner (UPQC) is developed by using the 9-switch power converter with the enhanced repetitive controller (RC). The proposed 9-switch UPQC can mitigate concurrently harmonic current and voltage problems together with voltage sag/swell issues. And also, the switching loss is reduced significantly by using the discontinuous modulation strategy for the 9-switch UPQC. The characteristics of the proposed 9-switch UPQC are evaluated by simulation, and its superior performance is verified.

Keywords: 9-switch converter · Unified power quality conditioner · Repetitive controller · Power quality · Voltage sag · Harmonic current · Harmonic voltage

1 Introduction

Nowadays, the wide utilization of power electronics devices and nonlinear loads like adjustable speed motor drives, diode rectifiers, and power supply units causes the injection of a large amount of harmonic currents into the power distribution systems [1, 2]. This harmonic problem leads to many serious impacts on power systems such as voltage distortions, high power loss, malfunction of electronic equipment, and degradation of the networks power quality.

In order to solve these problems, many kinds of power quality compensators have been developed and applied in practical such as: passive power filters [3], active power filters (APFs) [4]. In recent decades, unified power quality conditioners (UPQCs) have been presented and adopted as a versatile and flexible solution for power quality mitigation [5] because it can independently and simultaneously operate as shunt and series APF to maintain both grid current and load voltage to be sinusoidal. However, the conventional UPQC has drawbacks due to its high cost, large volume, control complexity. In order to reduce the number of switching devices, 9-switch power converter has been considered as a promising solution to substitute the conventional UPQC topology [6] because it can provide two sets of output terminals. Authors in [6] demonstrate the effectiveness of 9-switch unified power quality conditioner (9S-UPQC). However, voltage distortion and voltage sag/swell are considered separately, and the APF includes many resonant controllers. So, it is very hard to implement the UPQC due to high computational burden as well as complexity to design the resonant controllers.

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_56

In order to solve this problem, repetitive controller (RC) is introduced as a bank of resonant controllers to track and regulate periodic frequencies [7]. In this paper, the RC based 9S-UPQC is developed to reduce both the number of power switches and computing burden. The RCs for both shunt and series APF controllers are designed and implemented to maintain the grid current/load voltage in *dq* frame. By reducing the RC's delay time by six times compared to that of the conventional one, the proposed 9S-UPQC can operate with high dynamic performance. And also, the computational burden is significantly reduced and the controller design process becomes simplified. Furthermore, the current harmonics, voltage distortion and voltage sag/swell issues are fully considered to improve the compensating performance. Simulation studies are carried out by PSIM software to verify the feasibility of the proposed RC-based 9S-UPQC.

2 9-Switch UPQC

Figure 1 shows the general circuit of the 9S-UPQC with the distorted grid voltage and nonlinear load in 3-phase system. The power converter is constructed by three semiconductor switches per-phase, and a micro-source is attached at the common DC-link to power for the converter.

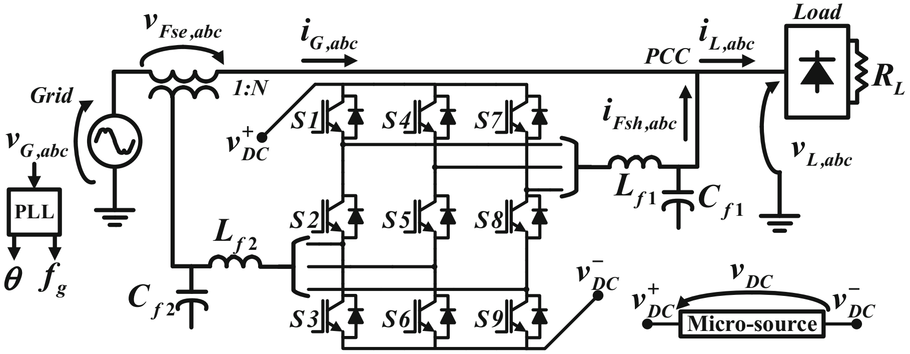


Fig. 1. General circuit of 9S-UPQC system.

Unlike the conventional back-to-back power converter where each three-phase full bridge side is independently controlled, the output terminals per phase of 9S-UPQC can be only connected to either v_{DC}^+ or v_{DC}^-: its upper terminal to the upper positive DC-link and lower terminal to the lower negative DC-link. In fact, it is impossible to connect the upper/lower terminals of 9S-UPQC to the negative/positive DC-link, respectively. Fortunately, we can step over this limit by modulating appropriately for both shunt and series APF. In this paper, we coordinate two modulating references per phase with only one carrier band. For simplicity, it is assumed that the reference for the upper terminal (shunt APF reference) is always placed above that of the reference for the lower terminal (series APF reference) and there is no intersection between these two reference signals to ensure the modulation restriction.

There are two kinds of reference frequency modes: different frequency (DF) mode and common frequency (CF) mode. DF mode means that the upper reference frequency is different from that of the lower one, while the reference frequencies are same in CF mode. The DF and CF modes are plotted in Fig. 2(a) and (b), respectively. In Fig. 2, h_1 and h_2 mean the modulation ratios of the shunt and series APF, respectively.

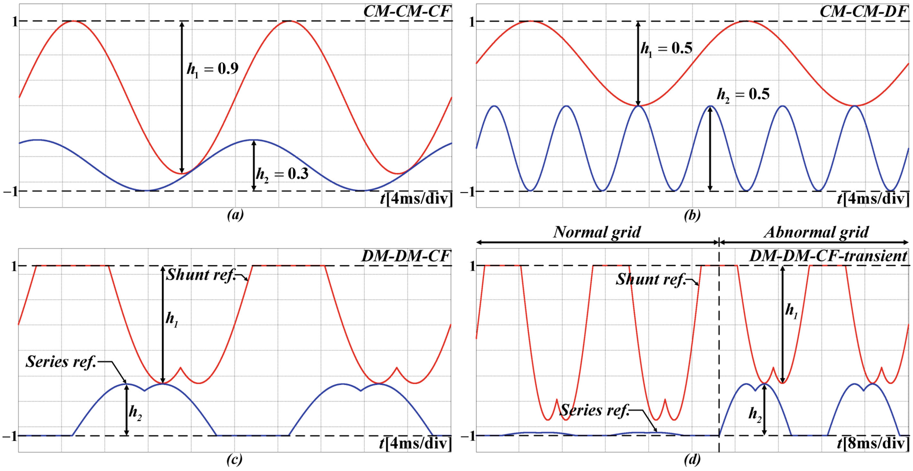


Fig. 2. Different modulation strategies: (a) CM-CM-CF, (b) CM-CM-DF, (c) DM-DM-CF in steady state and (d) DM-DM-CF in transient state.

Moreover, there are two types of modulation: continuous modulation (CM) and discontinuous modulation (DM). The main advantage of DM is the reduced commutation (up to 33%) as in [8]. As a result, the switching loss of DM is lessened dramatically compared with the CM. In this paper, we adopt the DM-DM for shunt and series APF references with CF mode. Figure 2(c) shows the modulation strategy in steady state. In 9S-UPQC, three switches S1, S2, S3 are turned off or on according to the following conditions:

$$S1 = \begin{cases} ON, \text{ if upper reference larger than carrier} \\ OFF, \text{ if upper reference lower than carrier} \end{cases} \tag{1}$$

$$S3 = \begin{cases} ON, \text{ if lower reference lower than carrier} \\ OFF, \text{ if lower reference larger than carrier} \end{cases} \tag{2}$$

$$S2 = XOR(S1, S3) \tag{3}$$

Due to the symmetry, the sets of S4, S5, S6 and S7, S8, S9 are switched with the same switching rule as switches S1, S2, and S3, respectively. In case of the ideal grid, the output voltage amplitude of the shunt APF is literally much larger than that of the series APF. In Fig. 2(d), when the series APF is in standby mode, there is no the compensating voltage across the coupling transformer in case of the normal grid.

Meanwhile, the shunt APF has to supply a large amount of harmonic current into system to compensate the current harmonics caused by nonlinear loads. During the normal gird period, the lowest three switches, i.e. S3-S6-S9, should be always kept ON. In case of the abnormal grid in Fig. 2(d), which means that sag/swell/distorted voltage happen, the large series voltage had to be injected, and the shunt modulating reference is reduced in order to increase the series APF reference as shown in Fig. 2(d) in abnormal grid period. In the 9S-UPQC, the shunt and series references are adjusted automatically since they share a common carrier band in order to mitigate simultaneously current/voltage harmonics as well as voltage sag/swell as a conventional UPQC.

3 The Proposed Control Scheme

3.1 Series APF Control Strategy

Theoretically, the load voltage becomes sinusoidal regardless the distorted/sag/swell grid voltage due to the series APF operated as voltage source. Figure 3(a) shows the control strategy of series APF in rotating *dq* frame. All the $(6n \pm 1)^{th}$ harmonics caused by the nonlinear load currents are converted to the $(6n)^{th}$ orders in *dq* frame. In the voltage control loop, proportional-integral (PI) controller and RC are adopted. Because of the limit bandwidth, PI controller only manages the DC component, which is the fundamental component in stationary frame, while the RC periodically tracks and forces the $(6n)^{th}$ harmonic voltages to be zero. The outputs of controllers are synthesized and transferred to the stationary frame, and then, they are forwarded to the DM block to generate the voltage reference. From this control diagram, the computational burden can be significantly reduced compared with the PR-based controller in stationary frame. Moreover, because it does not require any harmonic extractor or voltage sag/swell detector, the controller design process becomes much simpler with only one identified RC instead of a large number of PR controller gains.

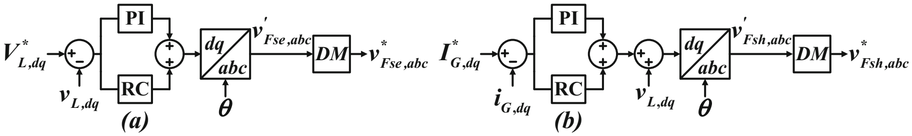


Fig. 3. The proposed control strategy for (a) series APF and (b) shunt APF.

Bode diagram of RC transfer function is plotted in Fig. 4 in order to show how to track the reference and to regulate harmonic components to be zero. The RC design procedure is described in detail in [7].

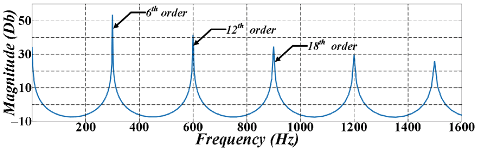


Fig. 4. Bode diagram of the utilized RC.

3.2 Shunt APF Control Strategy

Similar to the conventional shunt APF, the shunt APF in the proposed 9S-UPQC acts as a current source in order to actively injects a large amount of harmonic current into system via the point of common coupling (PCC). Consequently, the grid current is steadily maintained as sinusoidal waveform regardless of the nonlinear load condition. The power flow through the nonlinear load is expressed as following:

$$\begin{cases} p_L = P_L + \tilde{p}_L \\ q_L = Q_L + \tilde{q}_L \end{cases} \tag{4}$$

where $p_L/q_L, P_L/Q_L$ and $\tilde{p}_L/\tilde{q}_L$ are respectively active/reactive power, fundamental active/reactive power and oscillating active/reactive power flow through the load. Since the grid source only supplies the fundamental active power, the grid-side power flow is

$$\begin{cases} p_G = P_G \\ q_G = 0 \end{cases} \tag{5}$$

Therefore, the shunt APF power flow is obtained as follows:

$$\begin{cases} p_{Fsh} & = P_{Fsh} + \tilde{p}_{Fsh} \\ q_{Fsh} & = Q_{Fsh} + \tilde{q}_{Fsh} \\ P_G & = P_{Fsh} + P_L \\ \tilde{p}_{Fsh} & = -\tilde{p}_L \\ q_{Fsh} & = -q_L \end{cases} \tag{6}$$

From the Eqs. (4)–(6), it is clear that the shunt APF branch can totally handle all the harmonic load active power ($\tilde{p}_{Fsh} = -\tilde{p}_L$) and reactive load power ($q_{Fsh} = -q_L$). Additionally, the shunt APF has to consume a small amount of fundamental active power due to the power loss in devices ($P_G = P_{Fsh} + P_L$).

From the power analysis, even though the shunt APF rated power is quite high since it transfers such a large amount of power, the commutation loss of 9S-UPQC is significantly lessened, i.e. up to 33%. Hence, it is clear that the rated power of shunt APF in 9S-UPQC is much smaller than that in the conventional UPQC. Moreover, the number of switching devices is reduced together with the lowered DC-capacitor power rating. Therefore, cost and size of the proposed 9S-UPQC are significantly reduced compared with the conventional UPQC.

The detail control scheme for shunt APF is shown in Fig. 3(b), which is similar to the series APF control scheme. The desired quadrature component of grid-side current is set to be zero. Hence, the demanded reactive power is uniquely generated by shunt APF as explained before. In addition, the load voltage is feed-forwarded to the output of PI-RC to improve the dynamic response and the synchronization when the shunt APF is connected to the PCC.

4 Simulation Results

In order to investigate and verify the effectiveness of the proposed RC-based 9S-UPQC, the simulation is carried out for both steady state and transient-state with PSIM. The grid voltage is intentionally distorted together with sag/swell at specific moments. Meanwhile, the nonlinear loads are always connected to PCC to generate the large amount of harmonic currents into the network. The system parameters are listed in Table 1.

Table 1. System parameters.

Parameters	Values	Parameters	Values
Grid voltage v_G	$220V(rms)$	Sampling frequency f_s	10 kHz
$I^*_{G,d}$	$20A$	Switching frequency f_{sw}	10 kHz
$I^*_{G,q}$	$0A$	Output filter L_{f1}, L_{f2}	2.2 mH
$V^*_{L,d}$	310 V	Output filter C_{f1}, C_{f2}	40 µF
$V^*_{L,q}$	0 V	Transformer 1:N	50:100
Grid frequency f_n	50 Hz	Load R_L	15 Ω

4.1 Steady State Performance

Figure 5 provides the voltage and current waveforms in the proposed 9S-UPQC under highly distorted grid voltage and nonlinear load condition. From Fig. 5, the load voltage and grid current are close to the sinusoidal waveform, and the related total harmonic distortion (THD) values are given in Table 2. Even though the THD values of i_L and v_G are 22.28% and 14.43%, respectively, in the severe condition, the THDs of i_G and v_L are kept at 1.93% and 2.89% which are complied with the IEEE-512-1992 Std.

In addition, commutation of an upper switch in 9S-UPQC is compared with the one in conventional UPQC as shown in Fig. 6, where we can see that the number of the commutation with the proposed 9S-UPQC is significantly reduced. Therefore, the lower power loss and lower ratings of the devices are achieved finally by using the DM-DM-CF mode.

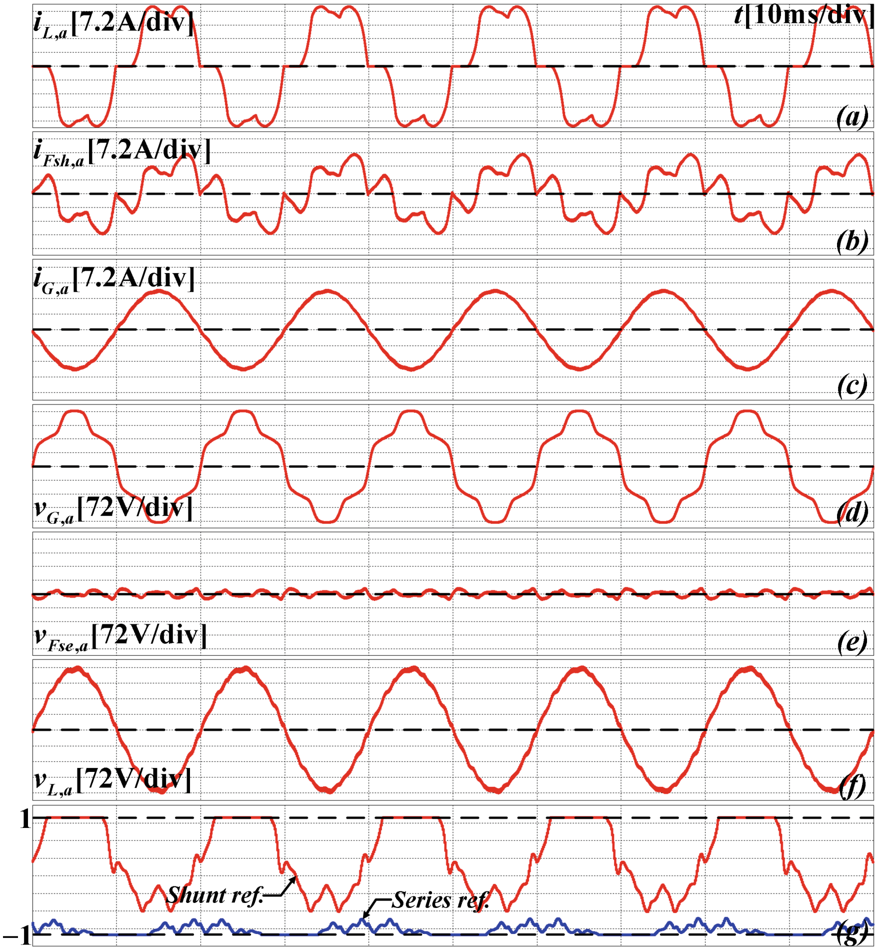


Fig. 5. Simulation results of proposed 9S-UPQC in steady state: (a) load current i_L, (b) shunt APF current i_{Fsh}, (c) grid current i_G, (d) grid voltage v_G, (e) series APF voltage v_{Fse}, (f) load voltage v_L and (g) references of shunt and series APF.

Table 2. THD of the voltage and current waveforms.

Parameters	i_L	i_{Fsh}	i_G	v_G	v_{Fse}	v_L
THD (%)	22.28	44.22	1.93	14.43	125.73	2.89

Throughout the steady state analyses, it is verified that the proposed RC-based 9S-UPQC can simultaneously and completely handle all current/voltage harmonics as well as voltage sag/swell issues in modern distribution system in spite of the reduced number and rated power of power electronic devices.

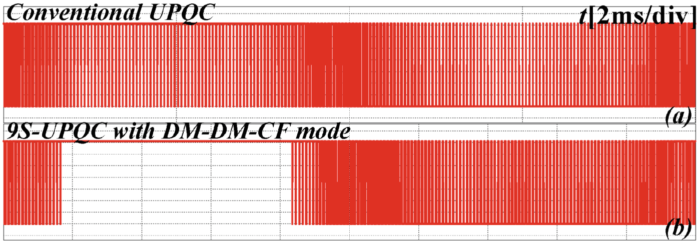


Fig. 6. Investigation of commutation in (a) the conventional UPQC and (b) the proposed 9S-UPQC with DM-DM-CF mode.

4.2 Dynamic Performance

Besides the steady state conditions, the proposed system is evaluated in the transient states when the load and grid are changed suddenly.

Firstly, Fig. 7 shows the dynamic performance when the power rating of nonlinear load is incidentally increased at a specific time. From the result, it takes less than one cycle before the grid current is maintained to be sinusoidal. In Fig. 7(d), the shunt and

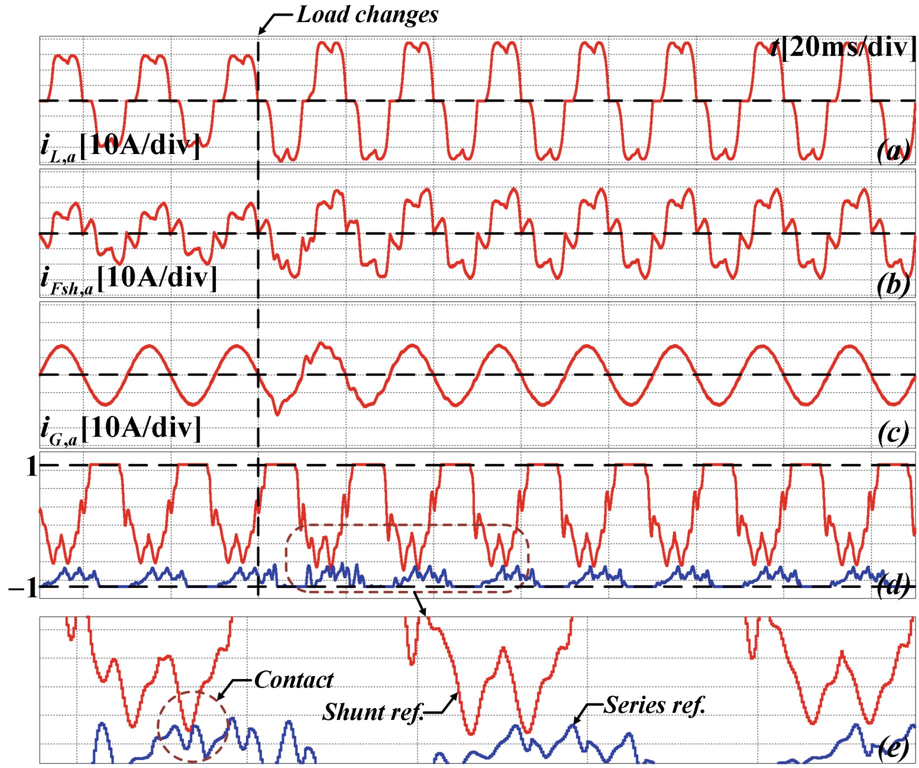


Fig. 7. Dynamic performance of shunt APF (a) load current i_L, (b) shunt APF current i_{Fsh}, (c) grid current i_G, (d) references of shunt and series APF, (e) magnification of (d).

series APF references are shown during the transient period, and two waveforms contact each other at just one point. Therefore, we can say that the modulation for 9S-UPQC is ensured properly with its stability margin.

Finally, the dynamic performance of the proposed 9S-UPQC is investigated by the deeply swelled grid voltage as shown in Fig. 8, where the waveforms of v^*_{Fsh} and v^*_{Fse} are also plotted in transient condition. From Fig. 8, the proposed system can deal with the severe swell together with highly distorted grid voltage; the response time to maintain v_L to be sinusoidal is less than a quarter of the fundamental period and only one contact between references is recorded. From the simulation results, we can conclude that the proposed system impressively offers fast dynamic response in spite of the load or voltage variations.

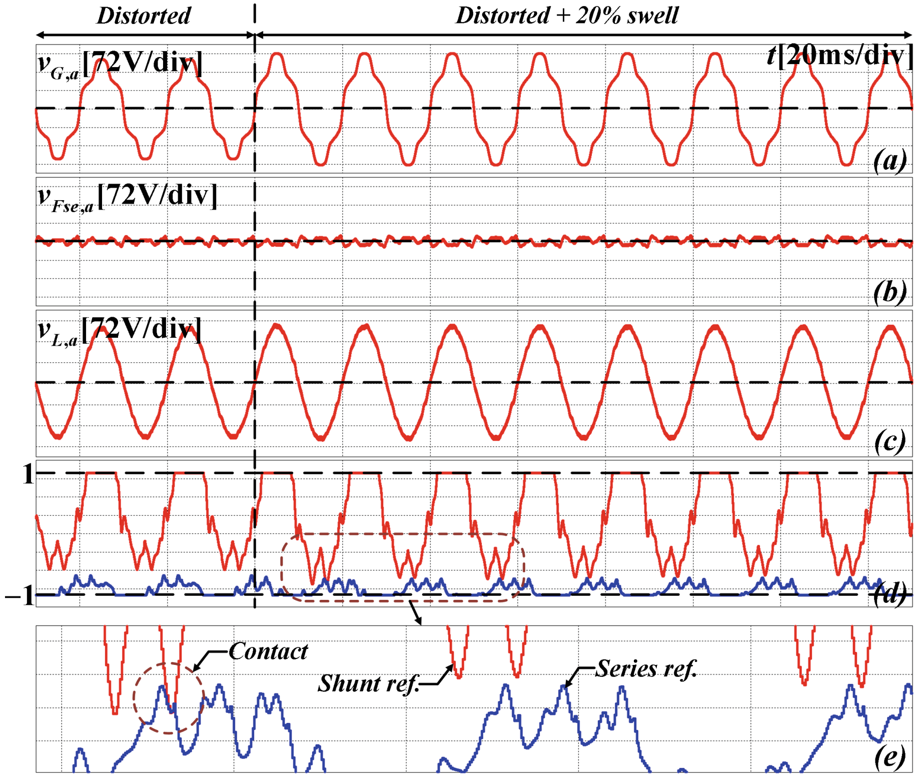


Fig. 8. Dynamic performance of series APF (a) grid voltage v_G, (b) series APF voltage v_{Fse}, (c) load voltage v_L, (d) references of shunt and series APF, (e) magnification of (d).

5 Conclusion

This paper proposed a low cost high performance 9S-UPQC to compensate harmonic current/voltage as well as voltage sag/swell in three-phase power system. The proposed system reduces not only the switching device number but also their power ratings compared to the conventional back-to-back UPQC. Moreover, by reducing the RC delay time, the dynamic performance is significantly improved and the grid current and load voltage are well regulated to be sinusoidal with low THDs regardless of the serious grid and/or load variations. The feasibility of the proposed 9S-UPQC is evaluated through the simulation, and it is verified that the proposed system is suitable to mitigate the power quality issues in distributed systems.

Acknowledgments. This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (2015R1D1A1A09058166).

References

1. Akagi H (1996) New trends in active filters for power conditioning. IEEE Trans Ind Appl 32(6): 1312–1322
2. Fujita H, Akagi H (1998) The unified power quality conditioner: the integration of series and shunt-active filters. IEEE Trans Power Electron 13(2):315–322
3. Yousif SNAL, Wanik MZC, Mohamed A (2004) Implementation of different passive filter designs for harmonic mitigation. In: PECon 2004 Proceedings National Power and Energy Conference, pp 229–234
4. Zou ZX, Zhou K, Wang Z, Cheng M (2015) Frequency-adaptive fractional-order repetitive control of shunt active power filters. IEEE Trans Ind Electron 62(3):1659–1668
5. Han B, Bae B, Kim H, Baek S (2006) Combined operation of unified power-quality conditioner with distributed generation. IEEE Trans Power Delivery 21(1):330–338
6. Zhang L, Loh PC, Gao F (2012) An integrated nine-switch power conditioner for power quality enhancement and voltage sag mitigation. IEEE Trans Power Electron 27(3):1177–1190
7. Trinh QN, Lee HH (2014) An enhanced grid current compensator for grid-connected distributed generation under nonlinear loads and grid voltage distortions. IEEE Trans Ind Electron 61(12):6528–6537
8. Gao F, Zhang L, Li D, Loh PC, Tang Y, Gao H (2010) Optimal pulsewidth modulation of nine-switch converter. IEEE Trans Power Electron 25(9):2331–2343

Rotor Flux Estimation for Low Speed Induction Motor Sensorless Drives with MRAS

Bigyan Basnet and Hong-Hee Lee(✉)

School of Electrical Engineering, University of Ulsan, Ulsan, Korea
bigyanjung123@gmail.com, hhlee@mail.ulsan.ac.kr

Abstract. In order to realize the sensorless induction motor drives with the vector control strategy, the rotor flux and the torque current are generally used to find the rotor speed. In this paper, an improved rotor flux estimation method is proposed to enhance the performance in low and zero speed condition. For accurate rotor flux and torque current estimation, the flux reference is obtained by using the modified low pass filter (LPF) and a high pass filter (HPF) is added in order to remove errors in magnitude and phase around the filter cut-off frequency. And also, the rotor speed is detected to achieve sensorless induction motor vector control with the aid of the model reference adaptive system (MRAS) scheme. Thanks to the accurate flux estimation, the steady state performance becomes better especially in low and zero speed conditions. Simulation and experimental results verify the effectiveness of the proposed system.

Keywords: Sensorless vector control · Flux estimation · Model reference adaptive system · Induction motor

1 Introduction

Induction motor drive systems are widely used in industrial, commercial and domestic applications due to their simple structure, low cost, rugged and easy implementation. In order to control induction motors, the scalar control scheme is simple with a good steady state response. But, its dynamic response is poor. Unlike the scalar control, the vector control provides excellent response in both steady state and transient state because it can drive induction motors like separately excited DC motors with sound performances. However, due to the speed sensor, it has some disadvantages such as extended shaft, reduced mechanical robustness and overall efficiency, and cost burden. To solve these problems, sensorless control is introduced by estimating the rotor speed without the speed sensor.

Plenty of research strategy has been proposed to estimate the rotor speed for sensorless induction motor drives. Among them, the model reference adaption system (MRAS) is the popular one due to relatively low computation effort and simplicity [1].

In the MRAS for the sensorless induction motor control, stator and rotor fluxes are generally used to estimate the rotor speed, and it is important to improve the flux estimation accuracy to guarantee the control performance. For this purpose, the integration algorithm is one of the widely-used methods because of its easy implementation [10].

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_57

But, the open loop integration causes dc offset and drifts, which results in erroneous speed estimation, especially in low speed range.

Various studies have been performed to overcome the limitations caused by the open loop integrator [2–9]. Authors in [11] proposed low pass filters with fixed cut off frequency to estimate the rotor speed. But, the low speed performance is very poor. Authors in [13], proposed a new rotor flux MRAS with the improved estimation of torque producing current estimation which improved the low speed estimation.

Usually, the model based induction motor drives have poor performance especially at low speed because of inaccurate flux estimation due to the DC offset and drift. This paper introduces an enhanced rotor flux estimation method to improve the performance of sensorless induction motor operation at low speed including the standstill. Basically, the rotor flux is estimated by using the flux and back EMF references. In this paper, the flux reference is obtained through the modified low pass filter (LPF) for accurate flux estimation, and a high pass filter (HPF) is added to remove the dc offset and drift caused by the LPFs. In order to implement the sensorless induction motor drive system, the rotor speed is detected with the aid of the MRAS scheme. Thanks to the MRAS with the enhanced rotor flux estimation, the performance of the low speed vector control is improved compared to the conventional system. The feasibility of the proposed sensorless scheme is verified through the simulation and experiment.

2 Conventional MRAS

The overall block diagram of the conventional MRAS is shown in Fig. 1, where the standard indirect vector control is carried out by using q and d current components assigned for torque and flux control, respectively. Rotor angle theta θ is the function of the estimated speed and slip speed, and the inverter utilizes the space vector pulse width modulation technique.

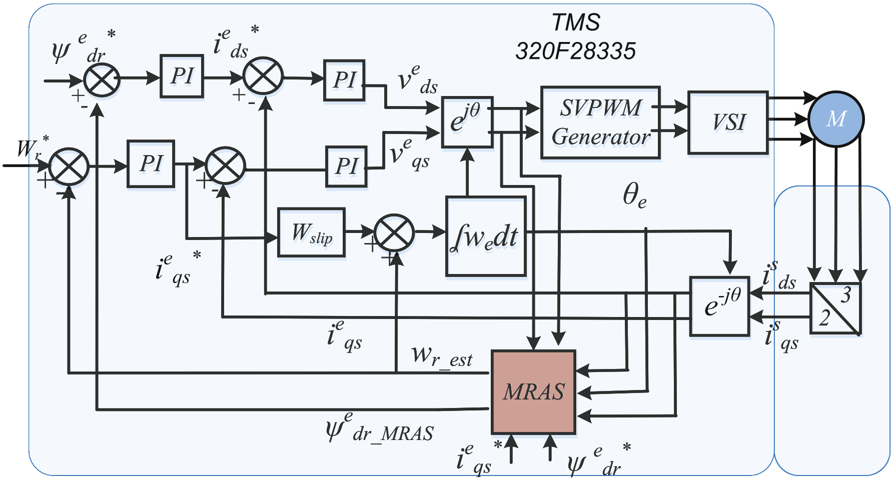


Fig. 1. Block diagram of the overall sensorless vector control

Figure 2 shows sensorless algorithm to estimate the rotor speed with MRAS scheme, and the rotor flux is estimated from (1) by integrating back EMF in stator frame [12].

$$\psi_r^s = \int \left(\frac{Lr}{Lm}(v_s - i_s R_s - (L_s - \frac{L_m^2}{L_r})\frac{di_s}{dt}) \right) \tag{1}$$

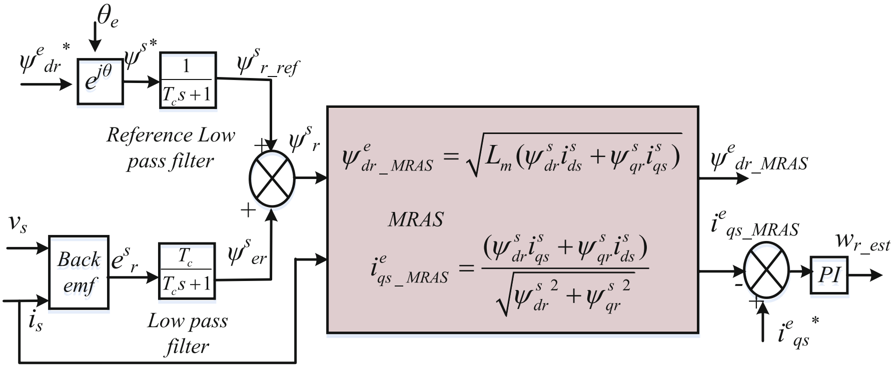


Fig. 2. Block diagram of the conventional MRAS scheme for sensorless vector control

To avoid the pure integration in (1), a LPF with fixed cut-off frequency is used as a feed forward reference term as shown in Fig. 2.

$$\psi_r^s = \frac{T_c}{T_c s + 1}\left(\frac{Lr}{Lm}(v_s - i_s R_s - (L_s - \frac{L_m^2}{L_r})\frac{di_s}{dt}) \right) + \frac{1}{T_c s + 1}\psi^{s*} \tag{2}$$

The rotor flux in (2) is estimated in the stator reference frame. In order to estimate the flux on in the rotating reference frame for the vector control, the rotor flux for MRAS, $\psi^e_{dr\_MRAS}$, is given as

$$\psi^e_{dr\_MRAS} = \sqrt{L_m(\psi^s_{dr} i^s_{ds} + \psi^s_{qr} i^s_{qs})} \tag{3}$$

After some basic trigonometric relationship for the rotor flux in (2), the q axis current in the rotating reference frame, $i^e_{qs\_MRAS}$, can be estimated as

$$i^e_{qs\_MRAS} = \frac{(\psi^s_{dr} i^s_{qs} + \psi^s_{qr} i^s_{ds})}{\sqrt{\psi^{s2}_{dr} + \psi^{s2}_{qr}}}. \tag{4}$$

Since the rotor flux in (3) is obtained from the stator voltage and stator currents, any error to measure these parameters affects the estimated value of the rotor flux which ultimately influences the speed estimation.

3 MRAS Scheme for Sensorless Vector Control

Figure 3 shows the proposed MRAS scheme to realize the sensorless vector control. As shown in Figs. 2 and 3, the rotor speed is estimated through the PI controller, which minimizes the error between the torque current components generated by the speed loop and estimated by MRAS scheme. Because the torque current is obtained from the rotor flux, it is important to estimate the rotor flux accurately for sensorless control.

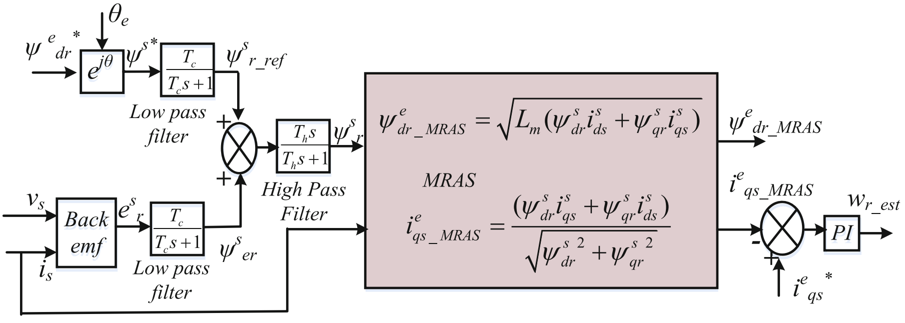


Fig. 3. Block diagram of the proposed MRAS scheme for sensorless vector control

From Fig. 3, the back EMF e^s_r is given as

$$e^s_r = \frac{Lr}{Lm}\left(v_s - i_s R_s - (L_s - \frac{L^2_m}{L_r})\frac{di_s}{dt}\right). \tag{5}$$

The flux in the rotating reference frame is converted to the stationary reference frame by the inverse park transformation:

$$\begin{aligned} \psi^s_q &= \psi^e_q cos\,\theta_e + \psi^e_d sin\,\theta_e \\ \psi^s_d &= \psi^e_d cos\,\theta_e - \psi^e_q sin\,\theta_e \end{aligned} \tag{6}$$

As shown in Fig. 2, because LPFs are used to obtain the reference flux, the DC offset and drift phenomena are inevitable. In order to avoid this problem, the HPF is used after the LPF applied to the flux reference is modified by multiplying the cutoff frequency as shown in Fig. 3. As a result, even though the estimated value from LPF induces errors in magnitude and phase especially in very low speed range, these errors can be removed through the HPF, and the estimated flux becomes much accurate.

From Fig. 3, the LPF for the reference flux is

$$\psi^s_{r\_ref} = \frac{T_c}{T_c s + 1}\psi^{s*}, \tag{7}$$

and the LPF for the back EMF is

$$\psi_{er}^{s} = \frac{T_c}{T_c s + 1} e_r^s. \tag{8}$$

By combining (7) with (8), the estimated rotor flux after the HPF becomes

$$\psi_r^s = \left(\psi_{r\_ref}^s + \psi_{er}^s\right)\left(\frac{T_h s}{T_h s + 1}\right) \tag{9}$$

where, T_c and T_h are the sampling times for LPF and HPF, respectively.

The estimated rotor flux in (9) is in the stationary reference frame, and it is expressed in the synchronously rotating reference frame as given in (3) and (4) to realize the vector control scheme of the induction motors.

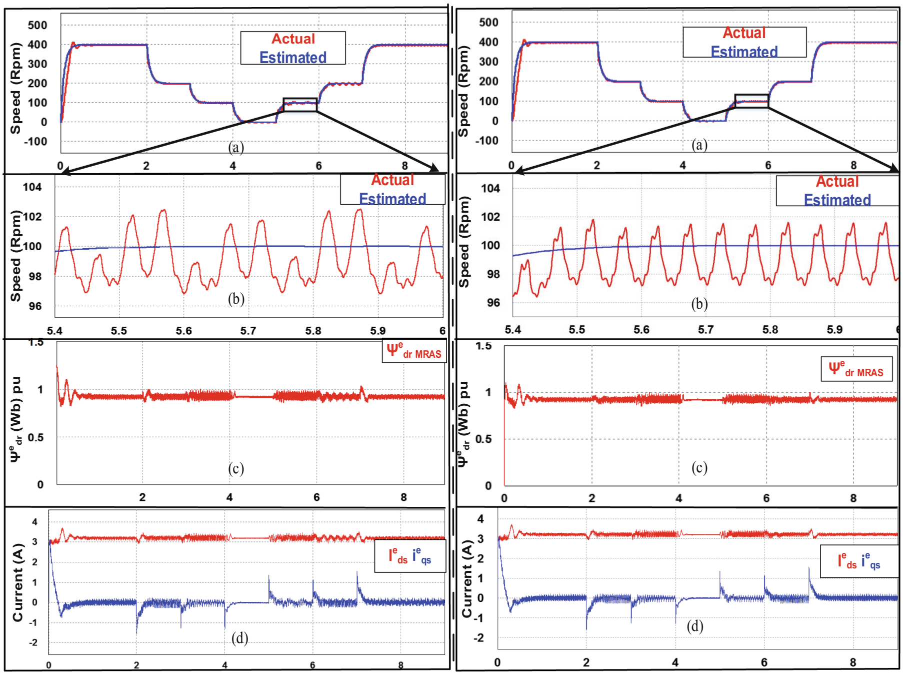
4 Simulation Results

The proposed MRAS is simulated to evaluate the performance based on 2.2 kW, 220 V, 50 Hz, delta connected induction motor together with the conventional one. The rotor resistance and stator resistance in all simulation are assumed to be nominal values regardless of speed change or load change. The induction motor is driven by the space vector modulated inverter with switching frequency of 10 kHz.

Figures 4.I and II show the simulated results for the conventional and proposed MRAS with no load for the reference rotor speed of 400-0-400 rpm, respectively. The rotor speed follows it reference very well with a good transient response. However, at instant when the speed drops to zero and increases, the speed response becomes worse with the conventional MRAS, which is zoomed in Fig. 4.I(b) from 5.4 secs to 6 secs. Due to the incorrect stator flux estimation, the rotor flux in Fig. 4.I(c) has slight oscillation during 5 secs to 7 secs. But, in case of the proposed MRAS, the speed response is kept with good performance at the zero speed thanks to the accurate flux estimation. Meanwhile, the synchronously rotating reference frame currents (Fig. 4(d)) are well settled except at transient state in both cases. So, we can say the proposed MRAS scheme has a superior performance compared to the conventional MRAS which uses the LFP to estimate the rotor flux as shown in Fig. 2.

In Fig. 5, 25% constant load torque is applied for similar reference speed as in Fig. 4. The simulation result shows that the system works well even when the load is held throughout the test. Less ripple in rotor speed is seen during loaded condition then with no load condition. However, the starting transient state is more sluggish than the identical no load condition.

Figure 6 shows the system response according to the variations of the load torque and the rotor speed. As shown in Fig. 6(a), the reference speed is increased from 100 rpm to 300 rpm and decreased to 100 rpm at 5 and 10 s, respectively. Meanwhile, the load torque is changed from zero to 0.25 Nm at 2.5 s, to 0.5 Nm at 7.5 s, and returned to zero at 12.5 s. In spite of the load or speed reference variation, the proposed system shows desired performance effectively.



(I)

(II)

Fig. 4. Simulated results with (I) conventional MRAS and (II) proposed MRAS with no load and speed reference 400-0-400 rpm: (a) rotor speed, (b) rotor speed at a particular time, (c) excitation reference frame flux estimate (pu), (d) reference currents in rotating frame.

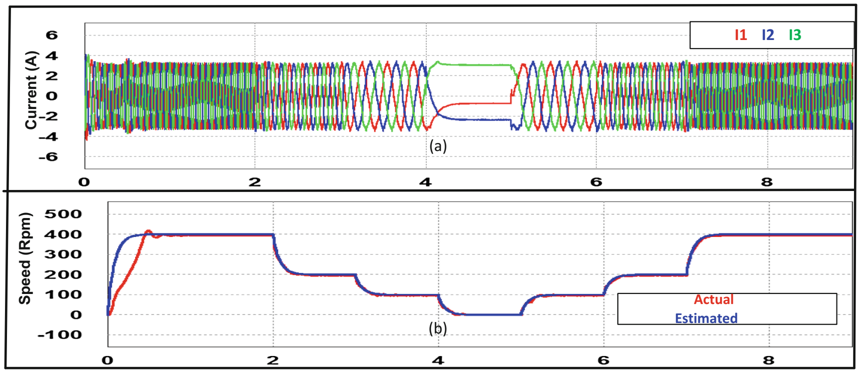


Fig. 5. Simulated results with the proposed MRAS under 25% load, 400-0-400 rpm. (a) 3 phase stator currents, (b) rotor speed responses.

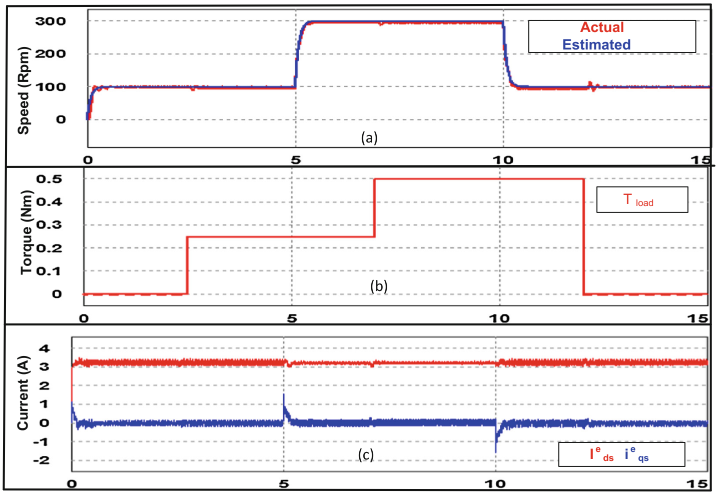


Fig. 6. Simulated results with the variation of speed and load commands. (a) rotor speed responses, (b) load torque, (c) dq currents

5 Experimental Results

In order to verify the implementation feasibility, the experiment with the proposed MRAS was carried out with the same parameters used in simulation. To verify the estimated speed response, the actual motor speed is measured by a 5000 pulses/revolution incremental optical encoder.

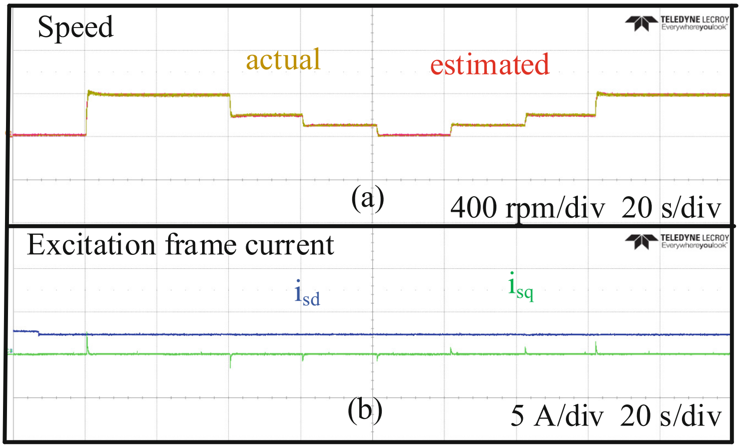


Fig. 7. Experimental result of proposed MRAS with no load, 400-0-400 rpm. (a) rotor speed, (b) excitation reference frame currents.

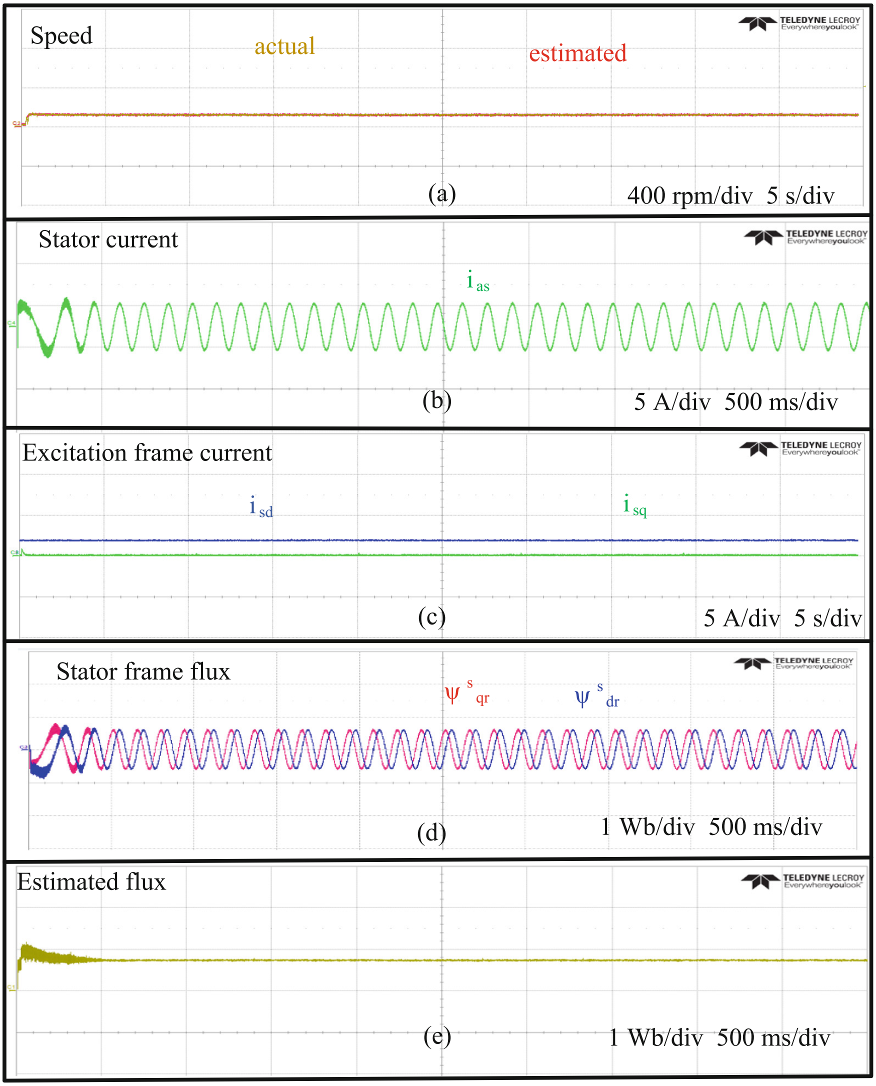


Fig. 8. Experimental results with the proposed MRAS under no load with 100 rpm speed reference. (a) rotor and estimated speed, (b) stator current, (c) dq currents, (d) estimated flux in stationary frame (pu), (e) estimated flux in rotating frame (pu).

Figure 7 shows the experimental result for the proposed MRAS with no load from 400 rpm to standstill. The result shows the stable performance in all speed range with slight rise and fall in q component current during respective rise and fall of rotor speed.

Figure 8 shows the experimental results with the proposed MRAS. During the start-up, roughness is seen in both flux and current response. However, it is maintained once the estimated speed reaches the reference speed. From Fig. 8, we can say that the experimental results coincide with those in simulated result and the proposed scheme can be applied to actual sensorless vector control of induction motor.

6 Conclusion

This paper has presented an enhanced rotor flux estimating method for the MRAS to realize sensorless induction motor drives. Because the HPF is designed to remove DC offset and drift caused by the conventional LPF, the proposed scheme is suitable for a low and zero speed operation with the improved performance compared to the conventional system. The simulation and experimental results have shown stable sensorless performance for different rotor speed conditions.

Acknowledgement. This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF- 2015R1D1A1A09058166).

References

1. Rashed M, Stronach AF (2004) A stable back-EMF MRAS-based sensorless low speed induction motor drive insensitive to stator resistance variation. Inst Electr Eng Proc Electr Power Appl 151(6):685–693
2. Schauder C (1992) Adaptive speed identification for vector control of induction-motors without rotational transducers. IEEE Trans Ind Appl 28(5):1054–1061
3. Orlowska-Kowalska T, Dybkowski M (2010) Stator-current-based MRAS estimator for a wide range speed-sensorless induction-motor drive. IEEE Trans Ind Electron 57:1296–1308
4. Maiti S, Chakraborty C, Hori Y, Ta MC (2008) Model reference adaptive controller-based rotor resistance and speed estimation techniques for vector controlled induction motor drive utilizing reactive power. IEEE Trans Ind Electron 55(2):594–601
5. Fengxiang W, Zhe C, Stolze P, Stumper J-F, Rodriguez J, Kennel R (2014) Encoderless finite-state predictive torque control for induction machine with a compensated MRAS. IEEE Trans Ind Informat 10(2):1097–1106
6. Lascu C, Andreescu GD (2006) Sliding-mode observer and improved integrator with DC-offset compensation for flux estimation in sensorless controlled induction motors. IEEE Trans Ind Electron 53(3):785–794
7. Stojic D, Milinkovic M, Veinovic S, Klasnic I (2015) Improved stator flux estimator for speed sensorless induction motor drives. IEEE Trans Power Electron 30(4):2363–2371
8. Alkorta P, Barambones O, Cortajarena JA, Zubizarrreta A (2014) Efficient multivariable generalized predictive control for sensorless induction motor drives. IEEE Trans Ind Electron 61(9):5126–5134
9. Wang K, Chen B, Shen G, Yao W, Lee K, Lu Z (2014) Online updating of rotor time constant based on combined voltage and current mode flux observer for speed-sensorless AC drives. IEEE Trans Ind Electron 61(9):4583–4593
10. Jun H, Bin W (1998) New integration algorithms for estimating motor flux over a wide speed range. IEEE Trans Power Electron 13(5):969–977

11. Ohtani T, Takada N, Tanaka K (1992) Vector control of induction-motor without shaft encoder. IEEE Trans Ind Appl 28(1):157–164
12. Ohyama K, Asher GM, Sumner M (2006) Comparative analysis of experimental performance and stability of sensorless induction motor drives. IEEE Trans Ind Electron 53:178–186
13. Smith A, Gadoue S, Finch J (2016) Improved rotor flux estimation at low speeds for torque MRAS-based sensorless induction motor drives. IEEE Trans Energy Conversion 31:270–282

An Enhanced Reactive Power Sharing and Secondary Voltage Restoration Control in Islanded Microgrid

Minh-Duc Pham and Hong-Hee Lee(✉)

School of Electrical Engineering, University of Ulsan, Ulsan, 680-749, South Korea
minhducpham2009@gmail.com, hhlee@mail.ulsan.ac.kr

Abstract. In the islanded micro-grid, parallel distributed generators (DGs) are normally controlled with the aid of the droop control scheme. However, the droop control method is still concerned to improve the accurate of reactive power sharing and variation of frequency and voltage at the point of common coupling (PCC). This paper proposes a control scheme to solve the limitation of microgrid in islanded operation such as reactive power sharing accuracy and PCC voltage and frequency drop by using a close-loop of the virtual impedance to compensate the feeder mismatch between DGs. Therefore, the power sharing performance is improved, and the circulating current between DGs is suppressed. In order to achieve the accurate reactive power sharing and PCC voltage restoration, a secondary control is implemented in both the microgrid central controller (MGCC) and local controller by using the low bandwidth communications. The effectiveness of the proposed control method is analyzed through the simulation.

Keywords: Islanded microgrid · Reactive power sharing · Secondary control

1 Introduction

Recently, small distributed generation (DG) systems units are growing popularly in the electrical power system together with the cut down of the traditional power plant [1, 2]. These DGs usually consist of the renewable energy resources such as fuel cell, photovoltaic cells, wind turbines, etc. Figure 1 shows the basic structure of the islanded microgrid. In islanded mode, the microgrid is isolated from the main grid when the static switch is open and each DGs has to work in autonomous mode. To achieve desired power sharing without the need of communication, the power sharing concept based on droop control is commonly used to share power between DGs [3]. The droop control scheme can be easily extended to a number of DGs in microgrid without changing the control strategy of DGs already setup in the islanded microgrid [4].

However, the droop control algorithm still has some disadvantages [3] such as circulating current and inaccurate reactive power sharing. Moreover, the output voltage amplitude and frequency of inverter are always smaller than nominal values in the P-ω and Q-E droop controls. In order to overcome these problems, the authors in [5] propose a transformation into the new virtual frame to decouple the active and reactive power. However, this method requires the exact value of the feeder impedance to calculate the ratio between the line resistance and line inductance, which is hard to be implemented

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_58

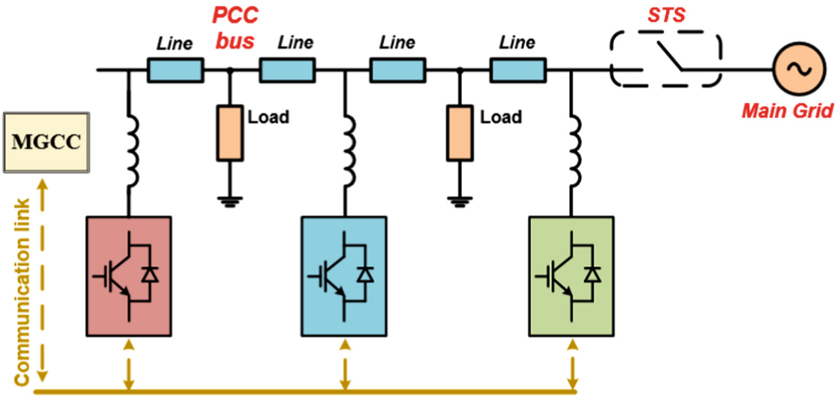


Fig. 1. Typical islanded microgrid configuration.

in practical applications. On the other hand, authors in [6, 7] improve the power sharing performance by adding the virtual impedance at the output of inverter. If the virtual impedance is properly selected, the output impedance of inverter is modified to ensure the active and reactive power decoupling. However, the exact value of virtual impedance of each DGs in islanded microgrid is difficult to detect since the feeder impedance is unknown and the mutual interaction between DGs in islanded microgrid.

To solve this problem, this paper proposes an enhanced power sharing control method that adaptively controls the virtual impedance to achieve the active and reactive power decoupling and reactive power sharing accuracy. Moreover, an external loop control is proposed to recover the voltage amplitude and frequency to those of nominal value. Furthermore, the low bandwidth communications are introduced to implement a secondary control in the microgrid central controller (MGCC). As a result, the quality and the stability of the islanded microgrid are improved significantly. The proposed control method is guaranteed and verified by investigating the dynamic response and stability of DGs in islanded microgrid with PLECS.

2 Islanded Microgrid System Analysis

2.1 Power Sharing with Droop Control

Figure 2(a) shows the equivalent circuit of parallel inverters working in islanded microgrid. Each DG works in voltage control mode and is connected to microgrid to maintain voltage and frequency at the PCC. To share the power between DGs autonomously, the active power versus frequency (P-ω) and the reactive power versus voltage (Q-E) droop control are used. As shown in [8], the active and reactive power can be expressed as following:

$$\begin{cases} P \approx \dfrac{EV}{X}\sin(\delta) \\ Q \approx \dfrac{EV\,\cos(\delta)\,-\,E^2}{X} \end{cases} \tag{1}$$

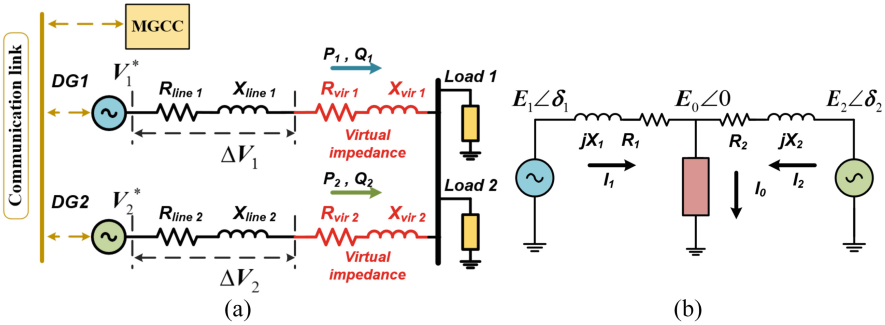


Fig. 2. (a) Simplified model of the islanded microgrid. (b) equivalent circuit model of DGs

From (1), it is obviously that the active power can be controlled by the phase angle while the reactive power can be controlled by the voltage. From the relationship between active and reactive powers with voltage amplitude and frequency, the droop control equations are defined as

$$\begin{cases} \omega = \omega^* - G_P P \\ E = E^* - G_Q Q \end{cases} \tag{2}$$

where ω^* and E^* are the nominal values of the output frequency and voltage, respectively, which are usually chosen to be same as those of the main grid. G_P and G_Q are the frequency and amplitude droop coefficients, respectively. The output voltage and frequency of inverter based on the droop concept are always smaller than nominal value. Since the frequency of the microgrid is not affected by the feeder impedance mismatches, the real power sharing capabilities can be achieved successfully. However, the reactive power sharing is not accurate because of unequal feeder impedance in microgrid configuration.

2.2 Circulating Current

From the equivalent circuit model of DGs in Fig. 2(b) the active and reactive circulating current of the single phase parallel inverter can be calculated as

$$\begin{cases} I_{cP} = \dfrac{R\,\Delta E + X\,E\Delta\delta}{2\,(R^2 + X^2)} \\ I_{cQ} = \dfrac{R\,E\Delta\delta + X\,\Delta E}{2\,(R^2 + X^2)} \end{cases} \tag{3}$$

where $\Delta E = E_1 - E_2$ and $\Delta\delta = \delta_1 - \delta_2$ [9]. The active and reactive circulating currents are mainly affected by the amplitude and phase difference of inverters. The deviation of voltage and frequency are caused by the feeder impedance difference between DGs in islanded microgrid. Hence, in order to remove the circulating current, the mismatched feeder impedance is needed to be compensated.

3 Principle of Enhanced Control in Islanded Microgrid

3.1 Enhanced Reactive Power Sharing with Virtual Impedance Control

In DG transmission system, the line impedance is mainly inductive, i.e., X≫R because of the leakage induction from transformer and output filter inductor, so the effect of R can be neglected [3]. An accurate approximation of the line voltage drop can be considered as follows:

$$\Delta V = V_1 - V_2 \cong \frac{XQ_1}{V_1} \tag{4}$$

To enhance reactive power sharing, the virtual impedance is introduced to compensate the mismatched voltage drop, and it is obtained by using the external loop control through PI controller:

$$L_{Vir} = K_{p\_vir}\left(Q_{average} - Q_i\right) + K_{i\_vir}\int\left(Q_{average} - Q_i\right)dt \tag{5}$$

where $Q_{average}$ is average value of the reactive power calculated from the total number of DGs, n, by the reactive power values from DGs through low bandwidth communication:

$$Q_{average} = \frac{\left(\sum_{i=1}^{n} Q_i\right)}{n}\left[\frac{1}{s}\right] \tag{6}$$

The virtual impedance is calculated based on the output current feedback:

$$X_{L\_Vir} = -I_{O\_\beta}L_{vir} \tag{7}$$

Figure 2(a) shows the virtual impedance at the output of inverter. By modifying output impedance, the voltage drop between DGs is compensated to maintain the output voltage at the PCC almost same, and the circulating current is significantly reduced.

3.2 PCC Voltage and Frequency Restoration

Because of the droop control characteristics and the virtual impedance, the voltage amplitude and frequency of DGs are always smaller than those of nominal value, which degrades the voltage quality at the PCC. To improve PCC voltage, the reference of the voltage and frequency droop control concept is modified by adding the recovering voltage amplitude and frequency terms:

$$\begin{cases} E_{ref} = E_{droop} + E_{res} \\ \omega_{ref} = \omega_{droop} + \omega_{res} \end{cases} \tag{8}$$

The MGCC measures the PCC voltage and frequency and implements a secondary control to obtain the additional E_{res} and ω_{res} value:

$$E_{res} = K_{p\_E}\left(E^*_{_{MG}} - E_{_{MG}}\right) + K_{i\_E}\int\left(E^*_{_{MG}} - E_{_{MG}}\right)dt \tag{9}$$

$$\omega_{res} = K_{p\_fre}\left(\omega^*_{_{MG}} - \omega_{_{MG}}\right) + K_{i\_fre}\int\left(\omega^*_{_{MG}} - \omega_{_{MG}}\right)dt \tag{10}$$

The control block diagram of the DG is shown in Fig. 3. Droop control reference is calculated based on the active and reactive power measured at the output of the inverter to obtain the voltage and frequency references.

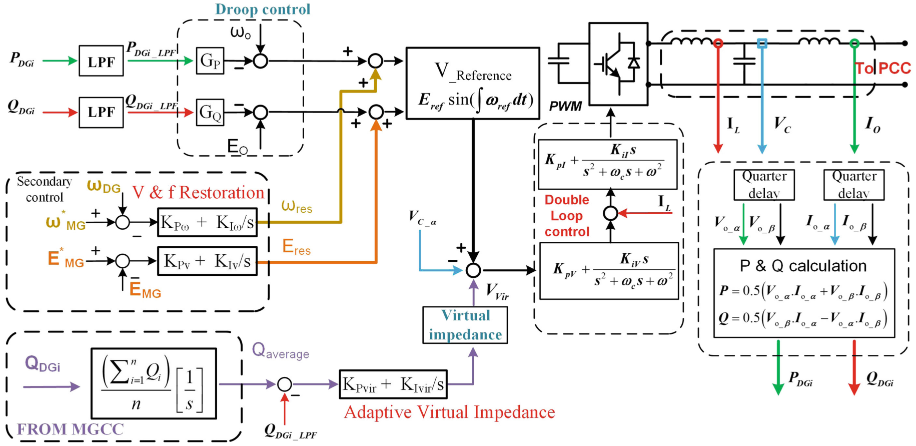


Fig. 3. The control block diagram of the system

4 Simulation Results

In order to verify the effectiveness of the proposed control method, the islanded microgrid with 3 DGs has been simulated using PLECS. The simulation parameters are given in Table 1.

Table 1. Simulation parameters

Parameter		Values
Microgrid configuration	Nominal voltage V_o^*	110 V
	Nominal frequency f_o	50 Hz
	Minimum microgrid voltage	104.5 V
	Minimum microgrid frequency	49.5 Hz
LCL filter	Inductor L_1	1.4 mH
	Inductor L_2	1.4 mH
	Capacitor C	20 uF
	Damping resister R_D	1.5 Ω
Droop coefficient	Droop coefficient G_P	0.01
	Droop coefficient G_Q	0.012
Line impedance	Line 1	0.2 Ω, 1.6 mH
	Line 2	0.1 Ω, 1 mH
	Line 3	0.15 Ω, 2 mH
Load	Load 1	15 Ω, 30 mH
	Load 2 (connected at 5 s)	15 Ω, 30 mH
	Load 3 (connected at 8.5 s)	30 Ω, 30 mH

Figure 4 shows the simulation result of DGs in islanded microgrid with the conventional droop control method and the proposed control method. Before the proposed control method is implemented, the active power is shared equally between DG1, DG2, and DG3 with the conventional droop controller. However, due to the mismatched impedance values between DGs, the reactive power sharing is not accurate. When the proposed control method is implemented at 2.5 s, the virtual impedance is adaptively tuned to compensate the unequal impedance between DGs and the accurate reactive power sharing is achieved after transient time about 2 s. Because of the tuning of the virtual impedance, the active power has a small oscillation but its variation is shortly decayed after 1.5 s, as shown in Fig. 4. Therefore, the power sharing performance is improved and the stability of the islanded microgrid is maintained with the proposed control method.

From Fig. 5(a), it can clearly be seen that DGs currents have the difference in both phase angle and magnitude before the proposed control method is applied. So, the circulating current rises between 3 DGs, which reduced the islanded microgrid quality and stability. With the proposed control method, the current sharing between DGs becomes balance, and the circulating current is suppressed, as shown in Fig. 5(b).

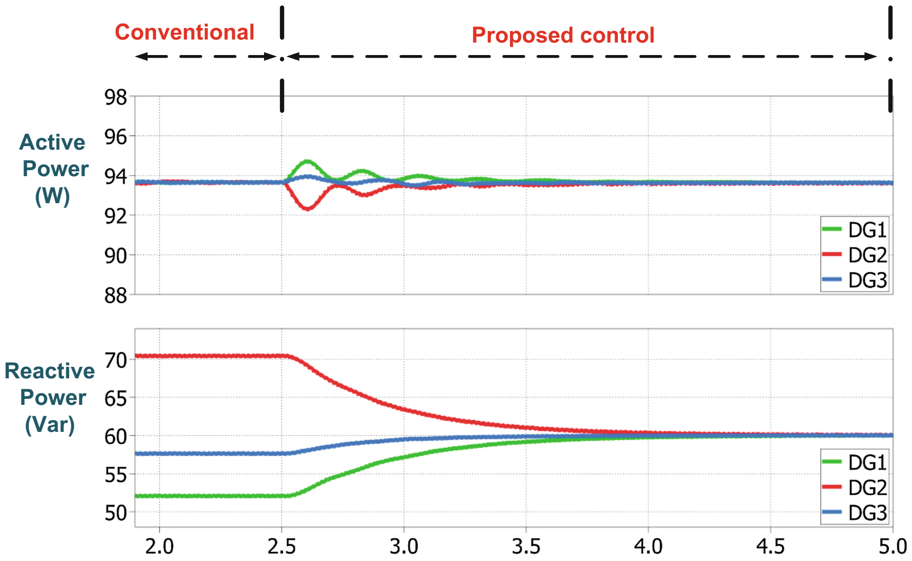


Fig. 4. The active and reactive power sharing between DG1, DG2, and DG3 with the conventional droop control method and the proposed control method.

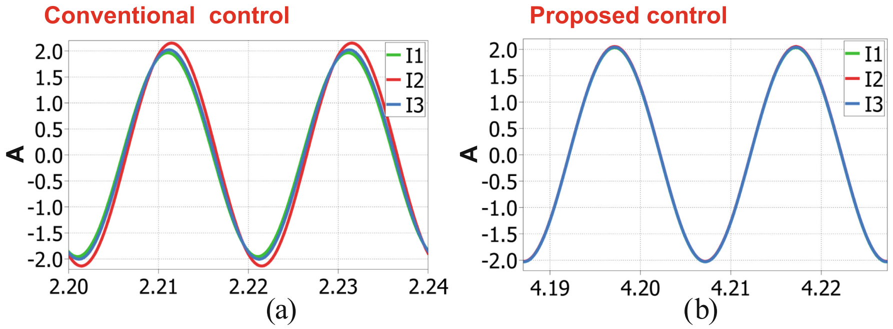


Fig. 5. The current sharing between DGs. (a) with the conventional droop control method. (b) with the proposed control method

The dynamic response of the proposed control method is examined under various conditions. In Fig. 6, the response of the proposed control method is good with no oscillation or overshoot even when the load 2 is connected at 5 s. Moreover, the reactive power sharing still has a smooth transient and achieves accuracy when a secondary control is enabled at 7 s. In addition, load 3 is connected at 8.5 s to examine more about the dynamic of the proposed control method when coordinated control with the secondary control. When load 3 is connected, the reactive and active power have the oscillation in the transient but still obtain the accurate power sharing in the steady state at some interval, made the whole islanded microgrid stable under various working conditions.

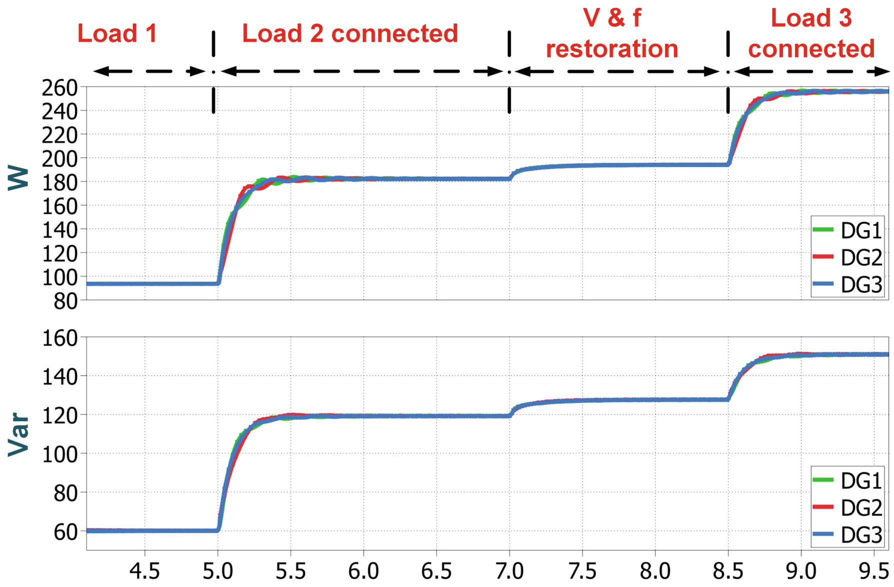


Fig. 6. The active and reactive power sharing with the proposed control method under load step condition.

Another problem in islanded microgrid with droop control concept is the inevitable output voltage and frequency drop. As can be seen in Fig. 7(a), the voltage magnitude and frequency at the PCC are smaller than those of nominal values because of the droop control characteristic. The voltage and frequency deviations are about $\Delta E = 3$ V and $\Delta f = 0.3$ Hz before the secondary control is implemented, as shown in Fig. 7(a) When the secondary control is implemented at 7.0 s, PCC voltage magnitude and frequency recover to those nominal values after a transient time around 0.8 s. In addition, the dynamic response of the secondary control is investigated when load 3 is connected at 8.5 s. The voltage and frequency at the PCC reduced when the load step but quickly recovered and achieved the steady state in less 1 s, as shown in Fig. 7(b). Hence, the

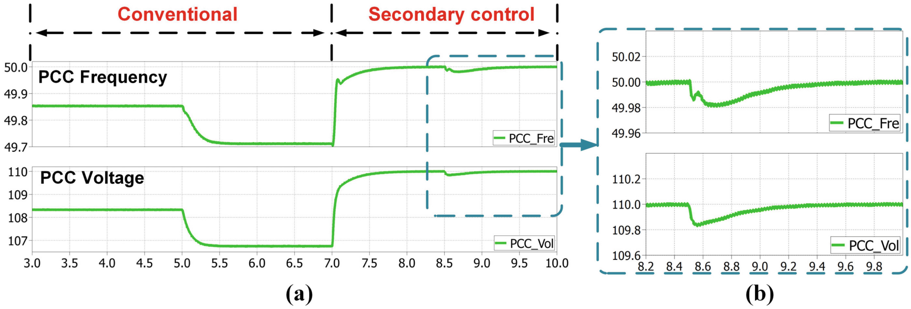


Fig. 7. (a) The voltage magnitude and frequency restoration at the PCC (b) the response of the secondary control when load 3 is connected at 8.5 s.

proposed control method achieved the objective of enhanced power sharing, voltage and frequency restoration, proving the efficiency of the proposed control method.

5 Conclusion

This paper has presented a control scheme for parallel inverters in islanded microgrid to achieve accurate reactive power sharing together with PCC voltage and frequency restoration. With the proposed control method, the virtual impedance is adaptively tuned to achieve accurate reactive power sharing together with minimizing the circulating current. In addition, a restoration of the microgrid voltage and frequency are also obtained using secondary control to recover PCC voltage and frequency. Furthermore, the proposed control method is stable and has a good dynamic response under the load variation. A series of simulation results show the effectiveness of the proposed control method.

Acknowledgment. This work was partly supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2015R1D1A1A09058166) and the Network-based Automation Research Center (NARC) funded by the Ministry of Trade, Industry & Energy.

References

1. Xue Y, Chang L, Kjaer SB, Bordonau J, Shimizu T (2004) Topologies of single-phase inverters for small distributed power generators: an overview. IEEE Trans Power Electron 19(5):1305–1314
2. Olivares DE, Mehrizi-Sani A, Etemadi AH, Cañizares CA, Iravani R, Kazerani M, Hajimiragha AH, Gomis-Bellmunt O, Saeedifard M, Palma-Behnke R, Jiménez-Estévez GA, Hatziargyriou ND (2014) Trends in microgrid control. IEEE Trans Smart Grid 5(4):1905–1919
3. Engler A, Soultanis N (2005) Droop control in LV-grids. In: 2005 International conference on future power systems, p 6
4. Mahmoud MS, Hussain SA, Abido MA (2014) Modeling and control of microgrid: an overview. J Franklin Inst 351(5):2822–2859
5. De Brabandere K, Bolsens B, Van den Keybus J, Woyte A, Driesen J, Belmans R (2007) A voltage and frequency droop control method for parallel inverters. IEEE Trans Power Electron 22(4):1107–1115
6. Guerrero JM, de Vicuna LG, Matas J, Castilla M, Miret J (2005) Output impedance design of parallel-connected UPS inverters with wireless load-sharing control. IEEE Trans Ind Electron 52(4):1126–1135
7. Wang X, Li YW, Blaabjerg F, Loh PC (2015) Virtual-impedance-based control for voltage-source and current-source converters. IEEE Trans Power Electron 30(12):7019–7037
8. Kawabata T, Higashino S (1988) Parallel operation of voltage source inverters. IEEE Trans Ind Appl 24(2):281–287
9. Xu S, Wang J, Xu J (2013) A current decoupling parallel control strategy of single-phase inverter with voltage and current dual closed-loop feedback. IEEE Trans Ind Electron 60(4): 1306–1313

A Predictive Current Control for Coordinate Control of Current and Power of Matrix Converter Under Unbalanced Input Voltages

Thanh-Luan Nguyen and Hong-Hee Lee(✉)

School of Electrical Engineering, University of Ulsan, Ulsan, South Korea
luanlik4@gmail.com, hhlee@mail.ulsan.ac.kr

Abstract. This paper presents a new predictive current control (PCC) method to achieve the coordinate control of current and power of the matrix converter under unbalanced input voltages. In order to avoid the complicated input voltage positive-negative sequence extraction, the flexible source current reference is constructed by filtering the square of input voltage vector with a notch filter. The optimal switching configuration to adjust source and load currents is selected by minimizating the cost function which is obtained from the sum of the absolute errors between the current references and their predictive values. Simulation results are given to validate the effectiveness of the proposed PCC method.

Keywords: Matrix converter (MC) · Predictive current control (PCC) · Unbalanced voltage · Power fluctuation

1 Introduction

Matrix converter (MC) is a single-stage direct AC-AC power converter, featuring no bulky capacitors on the dc bus. Compared with back-to-back voltage source converter, MC has many advantages such as sinusoidal input/output waveforms under normal conditions, controllable input power factor, regeneration capability and com-pact power circuit [1, 2].

However, due to lack of dc-link capacitors for energy storage, the MC is highly sensitive to the disturbance in the input voltage. In [3], when the input voltages are unbalanced, the low-order harmonics are induced in the output voltages and input currents. In order to reduce the effects of the unbalanced input voltage on the output performance, the most commonly used control strategy is the feed-forward compensation method based on the instantaneous value of input voltage proposed in [4]. This method is effective to provide the balanced output voltages and sinusoidal output currents. But, it causes severe harmonics in the input currents when the input voltages are unbalanced. In [5, 6], the improved methods, which dynamically modify the input power factor angle with the function of positive and negative sequence components of input voltages, are proposed to eliminate the undesirable harmonics in the input currents. In [7], the proportional integral resonant controller is designed in rotating *dq* reference frame to achieve a near unity input power factor and sinusoidal input current.

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_59

However, the reference input current calculation in [7] is highly dependent on the sequence component extraction of the input voltage.

Recent research has indicated the effective of predictive current control (PCC) method to control MC due to its advantages such as simplicity, fast dynamic response, and flexibility to control different variables [1]. In [8, 9], the PCC methods to achieve the sinusoidal output currents and unity input power factor are proposed by considering the output current regulation and the reactive power minimization on the source side. However, this method cannot ensure the source currents to be sinusoidal under unbalanced input voltage conditions. The PCC method which allows direct control of source and load currents has been presented in [10]. Even though the method in [10] can achieve the sinusoidal input current, it cannot control the input power fluctuation that is caused by the input voltage imbalance. This input power fluctuation could directly be transferred to the load, degrading the output performance of the MC.

In order to overcome these drawbacks, this paper presents an effective method to generate the source current reference by filtering the square of input voltage vector using a notch filter. The proposed method can simultaneously control the input and output currents along with input power fluctuation. Furthermore, the computation time and storage space are reduced by avoiding sequence component extraction. The proposed PCC method is verified by the simulation.

2 PCC of MC Under Unbalanced Input

Figure 1 shows the proposed PCC scheme to control the MC under unbalanced input voltages. The PCC method uses a discrete-time model to predict the source and load currents for the 27 switching configurations (SCs) of the MC, and then the SC which minimizes the cost function is applied. A discrete model of the input side is employed to predictive the next value of the source current. The input side is represented by a following state-space model:

$$\begin{bmatrix} \dot{\mathbf{v}}_{\mathrm{i}} \\ \dot{\mathbf{i}}_{\mathrm{s}} \end{bmatrix} = \boldsymbol{A} \begin{bmatrix} \mathbf{v}_{\mathrm{i}} \\ \mathbf{i}_{\mathrm{s}} \end{bmatrix} + \boldsymbol{B} \begin{bmatrix} \mathbf{v}_{\mathrm{s}} \\ \mathbf{i}_{\mathrm{i}} \end{bmatrix}, \tag{1}$$

where

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1/C_f \\ -1/L_f & -R_f/L_f \end{bmatrix}, \boldsymbol{B} = \begin{bmatrix} 0 & -1/C_f \\ 1/L_f & 0 \end{bmatrix}, \tag{2}$$

where L_f and C_f are the filter inductance and capacitance, respectively, and R_f is the leakage resistance.

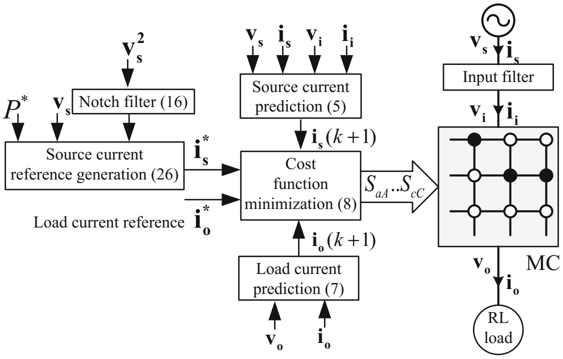

$\mathbf{v}_s^2$
Notch filter (16)
P^*
$\mathbf{v}_s$
Source current reference generation (26)
$\mathbf{i}_s^*$
Load current reference $\mathbf{i}_o^*$
$\mathbf{v}_s$ $\mathbf{i}_s$ $\mathbf{v}_i$ $\mathbf{i}_i$
Source current prediction (5)
$\mathbf{i}_s(k+1)$
Cost function minimization (8)
$S_{aA}..S_{cC}$
$\mathbf{i}_o(k+1)$
Load current prediction (7)
$\mathbf{v}_o$ $\mathbf{i}_o$
$\mathbf{v}_s$ $\mathbf{i}_s$
Input filter
$\mathbf{v}_i$ $\mathbf{i}_i$
MC
$\mathbf{v}_o$ $\mathbf{i}_o$
RL load

Fig. 1. Predictive current control scheme under unbalanced input voltages.

The discrete state-space model of the input side is determined as following:

$$\begin{bmatrix} \mathbf{v}_\mathbf{i}(k+1) \\ \mathbf{i}_\mathbf{s}(k+1) \end{bmatrix} = \mathbf{\Phi} \begin{bmatrix} \mathbf{v}_\mathbf{i}(k) \\ \mathbf{i}_\mathbf{s}(k) \end{bmatrix} + \mathbf{\Gamma} \begin{bmatrix} \mathbf{v}_\mathbf{s}(k) \\ \mathbf{i}_\mathbf{i}(k) \end{bmatrix}, \tag{3}$$

where

$$\mathbf{\Phi} = e^{AT_s},\ \mathbf{\Gamma} = A^{-1}(\mathbf{\Phi} - \mathrm{I}_{2\times 2})\boldsymbol{B}. \tag{4}$$

From (3), the source current prediction is obtained as

$$\mathbf{i}_\mathbf{s}(k+1) = \phi_{21}\mathbf{v}_\mathbf{i}(k) + \phi_{22}\mathbf{i}_\mathbf{s}(k) + \Gamma_{21}\mathbf{v}_\mathbf{s}(k) + \Gamma_{22}\mathbf{i}_\mathbf{i}(k) \tag{5}$$

On the load side, the dynamic model of the *RL* load is given as

$$\mathbf{v}_\mathbf{o} = R\mathbf{i}_\mathbf{o} + L\frac{d\mathbf{i}_\mathbf{o}}{dt}. \tag{6}$$

The output current prediction can be obtained by using forward Euler approximation for (6):

$$\mathbf{i}_o(k+1) = \frac{T_s}{RT_s + L}\left[\frac{L}{T_s}\mathbf{i}_\mathbf{o}(k) + \mathbf{v}_\mathbf{o}(k)\right]. \tag{7}$$

In order to regulate both the source and load currents, the cost function is given as following:

$$g = \left(\left|i^*_{o\alpha} - i^p_{o\alpha}\right| + \left|i^*_{o\beta} - i^p_{o\beta}\right|\right) + \lambda\left(\left|i^*_{s\alpha} - i^p_{s\alpha}\right| + \left|i^*_{s\beta} - i^p_{s\beta}\right|\right) \tag{8}$$

The superscript "*" denotes reference values, while the predicted values are denote by the superscript "*p*". The weighting factor λ handles the relative importance of the source current with the load current. The appropriate weighting factor is applied to minimize to total harmonic distortion (THD) of the source and load currents. The load current references $i^*_{o\alpha}, i^*_{o\beta}$ are imposed externally, while the source current references $i^*_{s\alpha}, i^*_{s\beta}$ are generated by the control strategy.

3 Proposed Source Current Reference

3.1 Input Current Harmonics and Power

A three-phase unbalanced input voltage can be expressed as the sum of positive and negative sequence components as follows:

$$\begin{bmatrix} v_{sa} \\ v_{sb} \\ v_{sc} \end{bmatrix} = V^+ \begin{bmatrix} \sin(\omega t + \theta_p) \\ \sin(\omega t - 120^0 + \theta_p) \\ \sin(\omega t + 120^0 + \theta_p) \end{bmatrix} + V^- \begin{bmatrix} \sin(\omega t + \theta_n) \\ \sin(\omega t + 120^0 + \theta_n) \\ \sin(\omega t - 120^0 + \theta_n) \end{bmatrix} \tag{9}$$

where V^+, V^-, θ_p, θ_n, and ω represent the positive and negative sequence voltage amplitude, phase angle, and angular frequency, respectively. The voltage vector (9) can be expressed on $\alpha\beta$ stationary reference frame by using the Clarke transformation as follows:

$$\begin{bmatrix} v_{s\alpha} \\ v_{s\beta} \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 2 & -1 & -1 \\ 0 & \sqrt{3} & -\sqrt{3} \end{bmatrix} \begin{bmatrix} v_{sa} \\ v_{sb} \\ v_{sc} \end{bmatrix} = \begin{bmatrix} v^+_{s\alpha} \\ v^+_{s\beta} \end{bmatrix} + \begin{bmatrix} v^-_{s\alpha} \\ v^-_{s\beta} \end{bmatrix} \tag{10}$$

with

$$\begin{bmatrix} v^+_{s\alpha} \\ v^+_{s\beta} \end{bmatrix} = V^+ \begin{bmatrix} \sin(\omega t + \theta_p) \\ -\cos(\omega t + \theta_p) \end{bmatrix} \text{and} \begin{bmatrix} v^-_{s\alpha} \\ v^-_{s\beta} \end{bmatrix} = V^- \begin{bmatrix} \sin(\omega t + \theta_n) \\ \cos(\omega t + \theta_n) \end{bmatrix}, \tag{11}$$

where $v^+_{s\alpha}, v^-_{s\alpha}, v^+_{s\beta}$, and $v^-_{s\beta}$ are the positive and negative sequence components of $v_{s\alpha}$ and $v_{s\beta}$, respectively.

According to the basic power theory, the instantaneous active and reactive powers of the converter are

$$\begin{bmatrix} p \\ q \end{bmatrix} = \frac{3}{2} \begin{bmatrix} v_{s\alpha} & v_{s\beta} \\ v_{s\beta} & -v_{s\alpha} \end{bmatrix} \begin{bmatrix} i_{s\alpha} \\ i_{s\beta} \end{bmatrix} \tag{12}$$

The source current references are derived from (12) as following:

$$\begin{bmatrix} i^*_{s\alpha} \\ i^*_{s\beta} \end{bmatrix} = \frac{2}{3} \frac{1}{v^2_{s\alpha} + v^2_{s\beta}} \begin{bmatrix} v_{s\alpha} & v_{s\beta} \\ v_{s\beta} & -v_{s\alpha} \end{bmatrix} \begin{bmatrix} P^* \\ Q^* \end{bmatrix}, \tag{13}$$

where P^* and Q^* are the input active and reactive power references, respectively. In order to achieve unity input power factor, Q^* is set to be zero. The source current references are rewritten based on the positive-negative sequence components of the input voltage:

$$\begin{bmatrix} i^*_{s\alpha} \\ i^*_{s\beta} \end{bmatrix} = \frac{2}{3}\frac{P^*}{(v^+_{s\alpha} + v^-_{s\alpha})^2 + (v^+_{s\beta} + v^-_{s\beta})^2}\begin{bmatrix} v^+_{s\alpha} + v^-_{s\alpha} \\ v^+_{s\beta} + v^-_{s\beta} \end{bmatrix}. \tag{14}$$

From (14), the source current references are not sinusoidal when the power reference P^* is constant. The inherent reason is that the denominators of (14) consists of a double input frequency oscillation term, as shown in (15):

$$\begin{aligned} &(v^+_{s\alpha} + v^-_{s\alpha})^2 + (v^+_{s\beta} + v^-_{s\beta})^2 \\ &\quad = (v^{+2}_{s\alpha} + v^{+2}_{s\beta}) + (v^{-2}_{s\alpha} + v^{-2}_{s\beta}) + 2v^+_{s\alpha}v^-_{s\alpha} + 2v^+_{s\beta}v^-_{s\beta} \\ &\quad = (V^+)^2 + (V^-)^2 + 2V^+V^- \cos(2\omega t + \theta_p + \theta_n). \end{aligned} \tag{15}$$

Therefore, the sinusoidal source currents can be obtained if the oscillation term is eliminated, which can be achieved with a notch filter *F(s)*:

$$F(s) = \frac{s^2 + \omega_n^2}{s^2 + \xi\omega_n s + \omega_n^2}, \tag{16}$$

where ω_n is set to 2ω and ξ is set to 1 for the good dynamic response as well as the filter performance. The sinusoidal source current reference can be obtained as following:

$$\begin{bmatrix} i^*_{s\alpha} \\ i^*_{s\beta} \end{bmatrix} = \frac{2}{3}\frac{P^*}{\left[(v^+_{s\alpha} + v^-_{s\alpha})^2 + (v^+_{s\beta} + v^-_{s\beta})^2\right]F(s)}\begin{bmatrix} v^+_{s\alpha} + v^-_{s\alpha} \\ v^+_{s\beta} + v^-_{s\beta} \end{bmatrix}. \tag{17}$$

From (17), the reference source currents can be rewritten in term of positive and negative sequence components, as follow:

$$i^+_{s\alpha} = \frac{2}{3}\frac{P^*}{(V^+)^2 + (V^-)^2}v^+_{s\alpha}, \tag{18}$$

$$i^+_{s\beta} = \frac{2}{3}\frac{P^*}{(V^+)^2 + (V^-)^2}v^+_{s\beta}, \tag{19}$$

$$i^-_{s\alpha} = \frac{2}{3}\frac{P^*}{(V^+)^2 + (V^-)^2}v^-_{s\alpha}, \tag{20}$$

$$i^-_{s\beta} = \frac{2}{3}\frac{P^*}{(V^+)^2 + (V^-)^2}v^-_{s\beta}. \tag{21}$$

The instantaneous input active power supplied to the converter is derived from (12), as follow:

$$p = \frac{3}{2}(v_{s\alpha} i_{s\alpha} + v_{s\beta} i_{s\beta}) = P^{+} + P^{-} + \tilde{p}, \tag{22}$$

where P^{+} and P^{-} are the positive and negative instantaneous active power, and $\tilde{p}$ is oscillating term at twice the source frequency. The formulation for the power component can be obtained by using (12), (18–21), as follows:

$$P^{+} = \frac{3}{2}(v_{s\alpha}^{+} i_{s\alpha}^{+} + v_{s\beta}^{+} i_{s\beta}^{+}) = \frac{P^{*}(V^{+})^{2}}{(V^{+})^{2} + (V^{-})^{2}}, \tag{23}$$

$$P^{-} = \frac{3}{2}(v_{s\alpha}^{-} i_{s\alpha}^{-} + v_{s\beta}^{-} i_{s\beta}^{-}) = \frac{P^{*}(V^{-})^{2}}{(V^{+})^{2} + (V^{-})^{2}}, \tag{24}$$

$$\begin{aligned} \tilde{p} &= \frac{3}{2}(v_{s\alpha}^{+} i_{s\alpha}^{-} + v_{s\beta}^{+} i_{s\beta}^{-} + v_{s\alpha}^{-} i_{s\alpha}^{+} + v_{s\beta}^{-} i_{s\beta}^{+}) \\ &= \frac{-2P^{*}V^{+}V^{-}}{(V^{+})^{2} + (V^{-})^{2}} \cos(2\omega t + \theta_{p} + \theta_{n}). \end{aligned} \tag{25}$$

By using (17), the reference source currents only consist the fundamental positive and negative components, excluding the harmonic components. However, the instantaneous input active power becomes fluctuation with the double frequency oscillation component, as shown in (22). Due to the absence of dc bus energy storage elements, this active power could directly be transferred to the output side, degrading the waveform quality of output voltages.

3.2 Flexible Source Current Reference

In order to achieve the coordinate control of the current and power, we introduce the flexible source current reference by combining both (14) and (17) with an adjustable coefficient $k\,(0 \le k \le 1)$. For simple description, the currents in (14) and (17) are assumed to be i_{s1}^{*} and i_{s2}^{*}, respectively. Then, the source current reference in the PCC method can be described as follows:

$$i_{s}^{*} = k i_{s1}^{*} + (1 - k) i_{s2}^{*}. \tag{26}$$

From (26), the proposed PCC method can flexibly control the input current and power by adjustment of coefficient k.

4 Simulation Results

In order to verify the effectiveness of the new PCC method, the numerical simulations were carried out using MATLAB-Simulink software. The system parameters are list in Table 1. A 10% decrease of phase voltage amplitude is applied to both phases B and C of the input voltages as shown in Fig. 2.

Table 1. System parameters

Variables	Description	Value
V_s	Input phase voltage (RMS)	220 V
f_s	Input frequency	50 Hz
L_f	Input filter inductance	0.4 mH
C_f	Input filter capacitance	22 μC
R_f	Input filter leakage resistance	0.5 Ω
R	Load resistance	10 Ω
L	Load inductance	30 mH
I_o	Output current amplitude reference	6 A
f_o	Output frequency	80 Hz
P	Input active power reference	540 W
T_s	Sampling time	10 μs

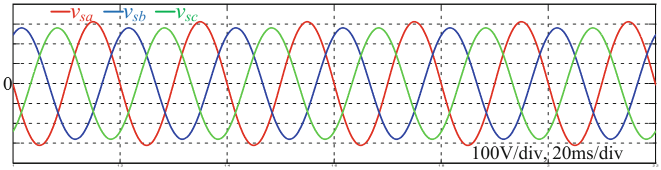


Fig. 2. Unbalanced input voltage.

Figure 3 shows the results of constant power control with $k = 1$. The weighting factor equal to λ = 13 which has been empirically adjusted as explained in [11]. Figures 3(a) show a correct tracking of the source current i_s to its reference i_s^* calculated from (14). The source currents are distorted with low-order harmonics as shown in Fig. 3(b). Figure 3(c) shows the good tracking of the load current i_o to its reference i_o^*. Figures 3(d) and (e) show that the input active power is almost constant and the source current i_{sa} is in phase with the input voltage v_{sa}, i.e., a unity input power factor is achieved.

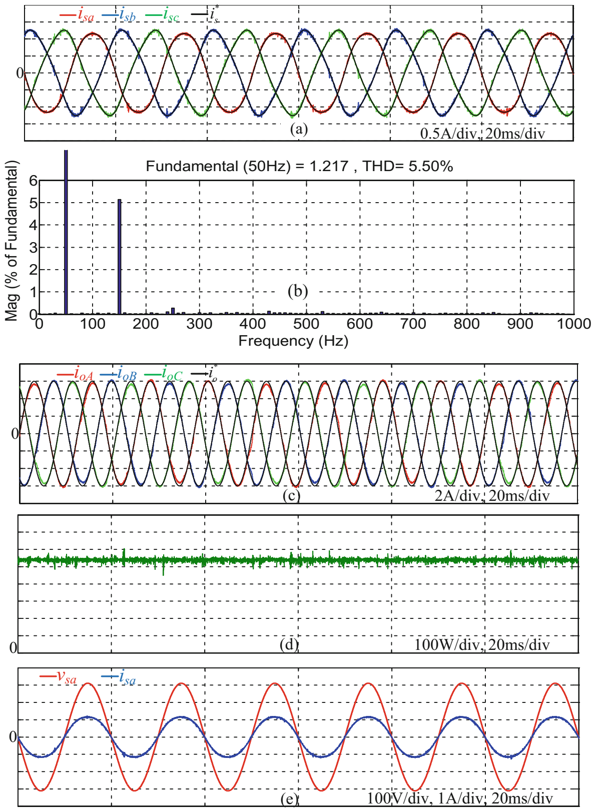


Fig. 3. Simulation results of constant power control (k = 1). (a) Source currents and their references, (b) source current harmonic spectrum, (c) load currents and their references, (d) input active power, and (e) source voltage and current phase a.

Figure 4 shows the results of sinusoidal source current control with $k = 0$. The good tracking of source current i_s to its reference i_s^* calculated from (17) and load current i_o to its reference i_o^* are shown in Fig. 4(a) and (c), respectively. Figure 4(b) shows that the source current harmonics are reduced, while the input active power is fluctuated as shown in Fig. 4(d). Figure 4(e) demonstrates that the unity input power factor is achieved.

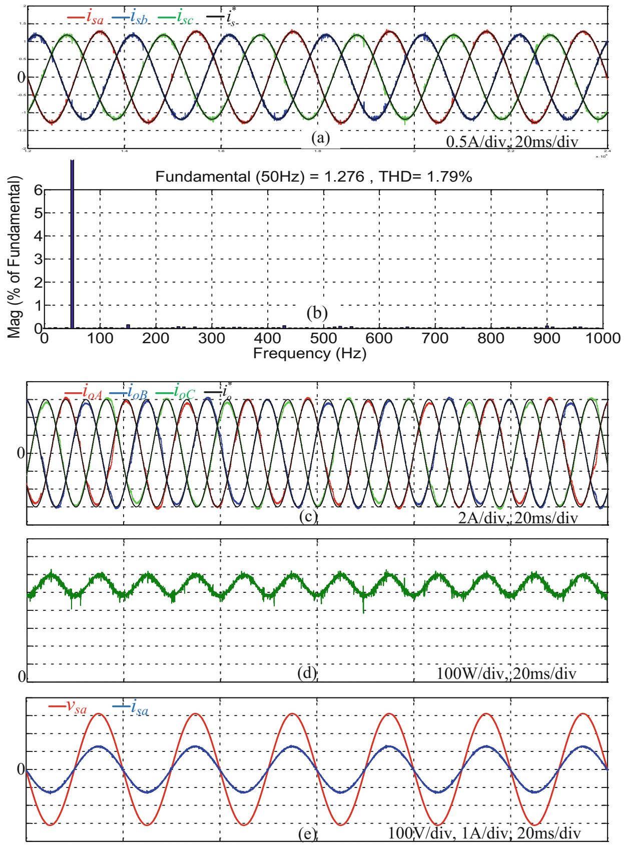


Fig. 4. Simulation results of sinusoidal source current control ($k = 0$). (a) Source currents and their references, (b) source current harmonic spectrum, (c) load currents and their references, (d) input active power, and (e) source voltage and current phase a.

In order to verify the flexibility of the control method, a step change of coefficient k is carried out with 20 ms interval by decreasing 0.2 step to reduce value of coefficient k from 1 to 0 as shown Fig. 5(a). It can be observed from Fig. 5(b) that the source current always in phase with its respective phase voltage, i.e., the unity input power factor is always achieved with this control method. Figures 5(c) and (d) show that the power fluctuation increases while the source current harmonics reduce as the value of k decreases, which is correspond with the theory analysis. In general, we cannot achieve good current waveform and constant power without fluctuation under the unbalanced condition. Therefore, a suitable coefficient k can be selected by considering the tradeoff between the current harmonics and power fluctuation.

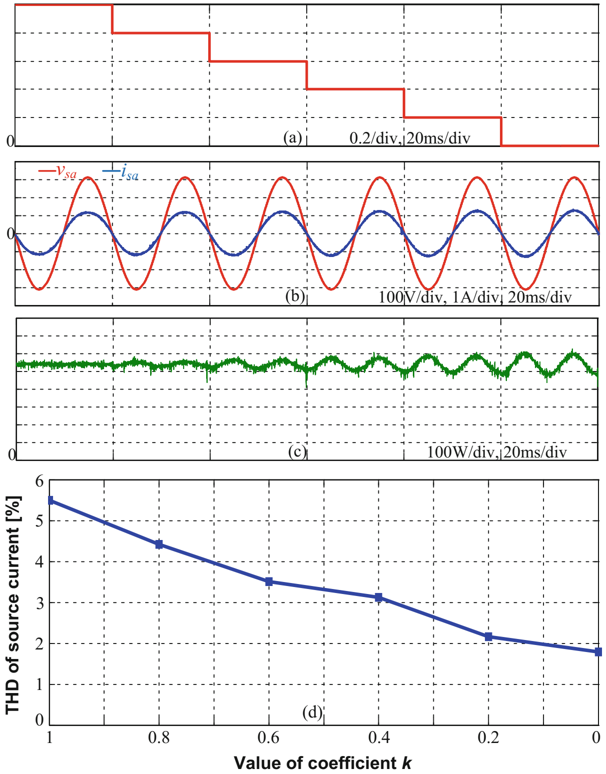


Fig. 5. Simulation results of flexible control method with step change of coefficient k ($0 \leq k \leq 1$). (a) Value of coefficient *k*, (b) source voltage and current phase a, (c) input active power, and (d) THD of source current.

5 Conclusion

This paper presents a new PCC method that can simultaneously control the current and power of MC under the unbalanced input voltage condition. The source current reference is flexibly generated by filtering the square of input voltage vector using a notch filter and an adjustable coefficient. Therefore, the computation time and storage space are reduced by avoiding sequence component extraction. The proposed PCC method achieves the good tracking performance of both source and load currents, and the unity power factor operation for the MC. The simulation results have demonstrated the effectiveness of the PCC method.

Acknowledgment. This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (2015R1D1A1A09058166).

References

1. Rodriguez J, Rivera M, Kolar JW, Wheeler PW (2012) A review of control and modulation methods for matrix converters. IEEE Trans Ind Electron 59(1):58–70
2. Nguyen HN, Lee HH (2016) A DSVM method for matrix converters to suppress common-mode voltage with reduced switching losses. IEEE Trans Power Electron 31(6):4020–4030
3. Casadei D, Serra G, Tani A (1998) A general approach for the analysis of the input power quality in matrix converters. IEEE Trans Power Electron 13(5):852–891
4. Nielsen P, Blaabjerg F, Pedersen JK (1996) Space vector modulated matrix converter with minimized number of switchings and a feedforward compensation of input voltage unbalance. In: Proceedings of the international conference on power electronic drives energy systems for industrial growth, January 8–11, 1996, vol 2, pp 833–839
5. Casadei D, Serra G, Tani A (1998) Reduction of the input current harmonic content in matrix converters under input/output unbalance. IEEE Trans Ind Electron 45(3):401–411
6. Lei J, Zhou B, Bian J, Qin X, Wei J (2016) A simple method for sinusoidal input currents of matrix converter under unbalanced input voltages. IEEE Trans Power Electron 31(1):21–25
7. Hamouda M, Blanchette HF, Al-Haddad K (2016) Unity power factor operation of indirect matrix converter tied to unbalanced grid. IEEE Trans Power Electron 31(2):1095–1107
8. Rivera M, Rodriguez J, Espinoza JR, Abu-Rub H (2012) Instantaneous reactive power minimization and current control for an indirect matrix converter under a distorted AC supply. IEEE Trans Ind Informat 8(3):482–490
9. Vargas R, Rodriguez J, Ammann U, Wheeler P (2008) Predictive current control of an induction machine fed by a matrix converter with reactive power control. IEEE Trans Ind Electron 55(12):4362–4371
10. Rivera M, Rodríguez J, Wheeler P, Rojas C, Wilson A, Espinoza J (2012) Control of a matrix converter with imposed sinusoidal source currents. IEEE Trans Ind Electron 59(4):1939–1949
11. Cortes P, Kouro S, La Rocca B, Vargas R, Rodriguez J, Leon J, Vazquez S, Franquelo L (2009) Guidelines for weighting factors design in model predictive control of power converters and drives. In: Proceedings of the IEEE ICIT, February 2009, pp 1–7

Comparison Between Three Different Level of Cascaded Multilevel Inverter with Separated DC Sources (CISS)

Suresh Thanakodi(✉), Yasotharan Visuvanathan,
Nazatul Shiema Moh Nazar, and Muhammad ‘Izzat Ahmad Sukeri

Department of Electrical and Electronic Engineering, Faculty of Engineering,
National Defence University of Malaysia, 57000 Kuala Lumpur, Malaysia
{suresh,nazatul.shima}@upnm.edu.my, yasotharan@live.com,
izzatsukeri@gmail.com

Abstract. The industrial revolution has been increased promptly at all over the world. With required suitable current, voltage and frequency for industrial purposes, multilevel inverter provides efficient power quality and continuous power supply for large industrial. Multilevel inverter is strongly known as an unconventional voltage medium, which converts DC power to AC power. However, the occurrence of harmonics would degrade the inferiority of the voltage produce by the inverter. This paper consists of three phase Cascaded multilevel Inverter with Separated dc Source (CISS) which eliminates the harmonics by three different levels. The switching technique used in this topology is sinusoidal pulse width modulation (SPWM). MATLAB/Simulink was used to generate the pulses to initiate harmonics. The percentage of harmonics produced will decrease in higher level. Three (3), five (5) and seven (7) level were selected to simulate the circuit topology by using MATLAB software. The results obtain in this paper will be compared and analysed with the previous research studies.

1 Introduction

Multilevel inverter has the ability to be used in the power industry and many interests have drawn into this concept because it is suitable to use in reactive power compensation. The quality of output waveform in a multilevel inverter depends on the value of output voltage produced by the inverter. The exclusive structure of different level in CISS allows them to achieve high voltage with minimum harmonics without using a transformer. The harmonic level decreases ominously, as the amount of voltage level escalates [1, 2].

Cascaded inverter also known as H - bridge inverter. CISS is the simplest inverter to produce the output waveform with least harmonics compared with conventional inverters. This inverter is made up of H-bridge topology that consists of four switches and each has their own DC source. Cascaded inverter able to produce complete sinusoidal output waveform. CISS able to construct up to a boundless number of levels, which eliminates more unwanted harmonics. Cascaded inverter became the significant inverter

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_60

to use in the industry because it has a great power performance with cheaper manufacturing cost [3]. The voltage level of CISS can be determined by the formula:

$$m = 2(N_s) + 1 \tag{1}$$

where = N_s are the numbers of DC source

The output voltage formula is:

$$V_{AN} = V_1 + V_2 + V_3 + \ldots + V_{(m-1)2} \tag{2}$$

The angle controlling at different level can define the quality of the output voltage. Figures 1, 2 and 3 are the topology that being used in this paper. There are few differences between these three levels of inverter in term of DC source, switch that being used, number of levels and the amount of H-bridge topology. These three different levels of inverters will be simulated and tested by using MATLAB/Simulink software.

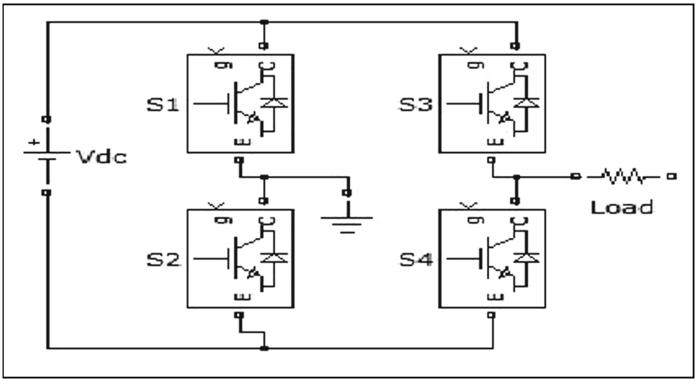


Fig. 1. 3-level of CISS

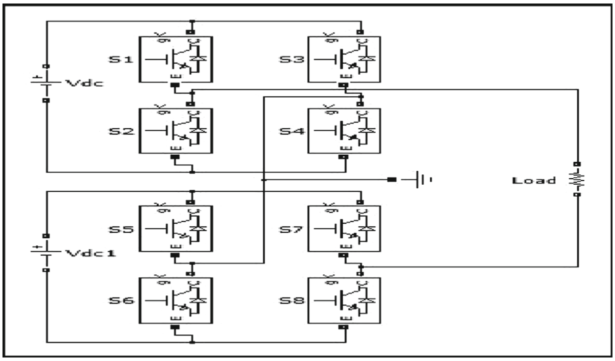


Fig. 2. 5-level of CISS

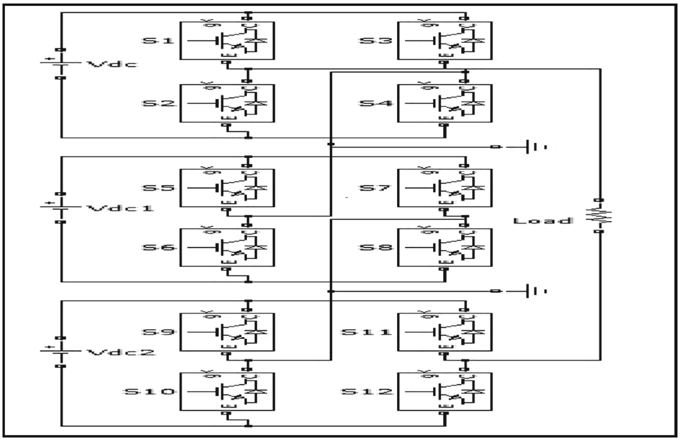


Fig. 3. 7-level of CISS

The circuit in Fig. 1 shows a single phase 3-level inverter that is connected in series as a single phase circuit topology. This circuit consists of four IGBT switches and single DC source. This level of the inverter can generate three voltages at the output, which are $+V_{dc}$, 0, and $-V_{dc}$. Four IGBT switches are connected to dc source at the output terminal. This occurs in each level of inverter and four IGBTs represent a single cell [4].

The circuit shown in Fig. 2 is a 5-level inverter that has two separate DC sources and eight IGBT. It generates an output voltage, which are $+2V_{dc}$, $+V_{dc}$, $0V_{dc}$, $-V_{dc}$, $-2V_{dc}$ [5]. A combination of two cells that has four IGBT diodes produce 5-level of output voltages.

The 7-level inverter as shown in Fig. 3 requires twelve IGBT switches and three DC sources. Each CISS has the same structure as a distinctive single phase inverter. This level of the inverter can generate an output voltage with seven levels of output with three cells IGBT combination [6].

1.1 Switching States

Three level inverter produces three different voltages at the output based on previous research [4]. Table 1 shows the output voltage produced which are $+V_{dc}$, 0, and $-V_{dc}$ when the switches are triggered.

Table 1. Switching Condition for 3-level of CISS

Voltage Output, Vo	Switching State			
	S1	S2	S3	S4
+V	1	0	0	1
0	1	1	0	0
−V	0	1	1	0

(1 = on, 0 = closed)

Since this is five level inverter, there are five output voltages were shown in Table 2 which are $+2V_{dc}$, $+V_{dc}$, 0, $-V_{dc}$, and $-2V_{dc}$ when the switches are triggered.

Table 2. Switching Condition for 5-level CISS

Voltage Output, Vo	Switching State							
	S1	S2	S3	S4	S5	S6	S7	S8
+2 V	1	0	0	1	1	0	0	1
+V	1	0	0	1	1	1	0	0
0	1	1	0	0	1	1	0	0
−V	0	1	1	0	0	0	1	1
−2 V	0	1	1	0	0	1	1	0

(1 = on, 0 = closed)

Table 3 shows the switching condition for 7-level CISS that produces seven output voltages. The output voltage produce consists of +3 V, +2 V, +V, 0, −V, −2 V, and −3 V.

Table 3. Switching Condition for 7-level CISS

Voltage Output, Vo	Switch States											
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
+3 V	1	0	0	1	1	0	0	1	1	0	0	1
+2 V	1	0	0	1	1	0	0	1	1	1	0	0
+V	1	0	0	1	1	1	0	0	1	1	0	0
0	1	1	0	0	1	1	0	0	1	1	0	0
−V	0	1	1	0	1	1	0	0	1	1	0	0
−2 V	0	1	1	0	0	1	1	0	1	1	0	0
−3 V	0	1	1	0	0	1	1	0	0	1	1	0

(1 = on, 0 = closed)

1.2 Advantages and Disadvantages of Cascaded Multilevel Inverter with Separated DC Sources (CISS)

The benefits of CISS is it needs the smallest number of components between all multilevel inverters to obtain the similar number of voltage level [7]. The circuit design can be modularized because each level consumes the same structure, and there are no additional clamping diodes or voltage balancing capacitors. CISS also can be used in this structure to evade lossy resistor-capacitor-diode snubbers [4, 5, 8]. The disadvantage of CISS is that it needs an isolated DC source for real power conversion and the applications are inadequate [8, 9].

1.3 Comparison on Total Harmonic Distortion (THD) Based on Levels of Inverter

Table 4 shows the value of THD value based on levels of inverter between single phase and three phases. As for single phase, the THD value decreases as the number of level increases, same as for three phase. It can be concluded that the number of levels and phase of inverter affects the THD value of the inverters. Based on previous studies, it shows that three phase inverter eliminates more harmonics compared to single phase inverter. Therefore, the simulation carried out in this paper were done in three phase circuit topology.

Table 4. Comparison on Total Harmonic Distortion (THD) Based On Levels of Inverter

Levels	3	5	7
THD (single phase)	48.34%	23.86%	16.16%
Reference	[10]	[11]	
THD (three phase)	35.33%	14.55%	12.22%
Reference	[12]	[11]	

2 Methodology

Figure 4 shows the flowchart for this project. In order to complete the workflow, the understanding of literature review and theory must be studied.

Figure 5 shows the overall circuit implementation for 3, 5 and 7 level inverter. The topologies of three different levels as shown in Figs. 1, 2, and 3 will be implemented inside the PWM cell. The simulation will be conducted separately for each level. One cell of PWM represents single phase inverter and total three cells of PWM represents three phase inverter. The result obtains for each level will be analysed and compared theoretically based on the study of previous research.

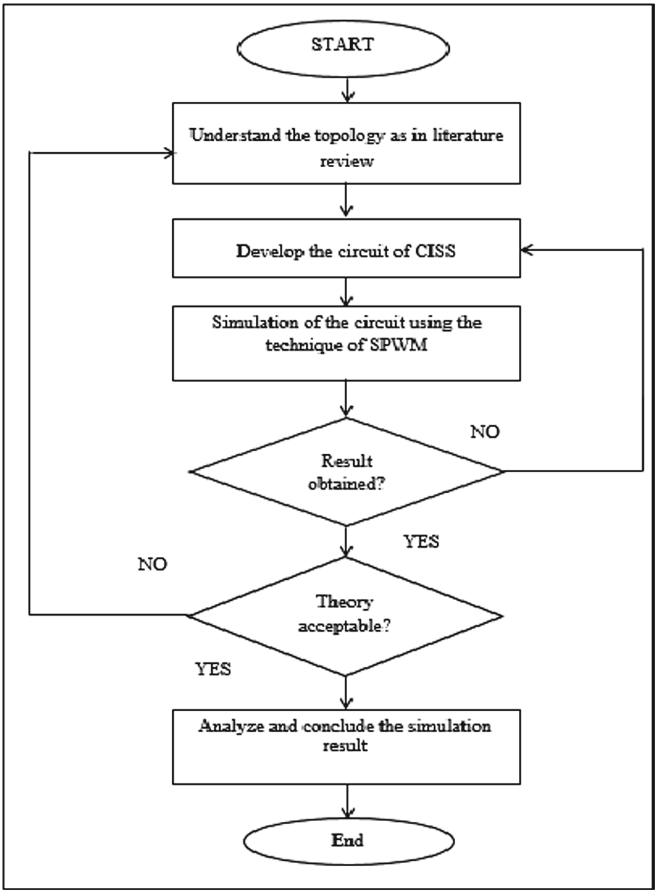

START
Understand the topology as in literature review
Develop the circuit of CISS
Simulation of the circuit using the technique of SPWM
Result obtained?
NO
YES
NO
Theory acceptable?
YES
Analyze and conclude the simulation result
End

Fig. 4. Flowcharts for project workflow

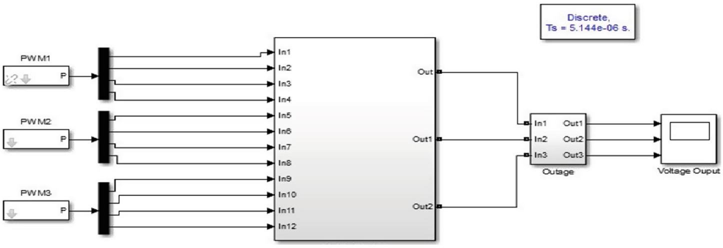


Fig. 5. Circuit Implementation for 3,5 and 7 level inverter

3 Results

Figure 6 displays the output waveform for 3-level CISS of three phases. The leg B and C were triggered at 120^0 and 240^0 phase delay correspondingly with respect to phase A. The voltage output is 200 V, 0 V, and −200 V. The total harmonic distortion value shown in Fig. 7 for three phases 3-level of CISS is 82.78%.

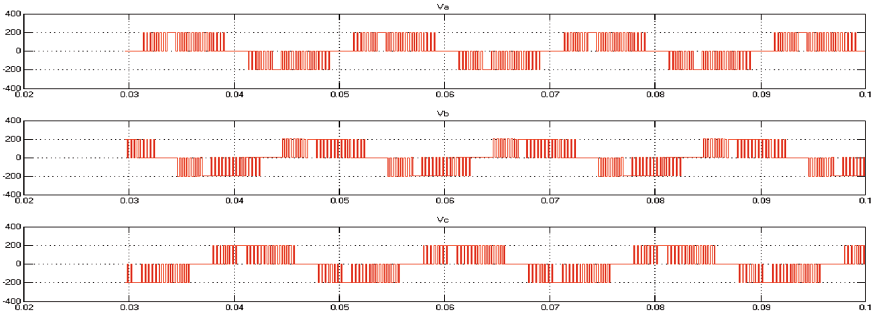


Fig. 6. Voltage output for 3-level CISS

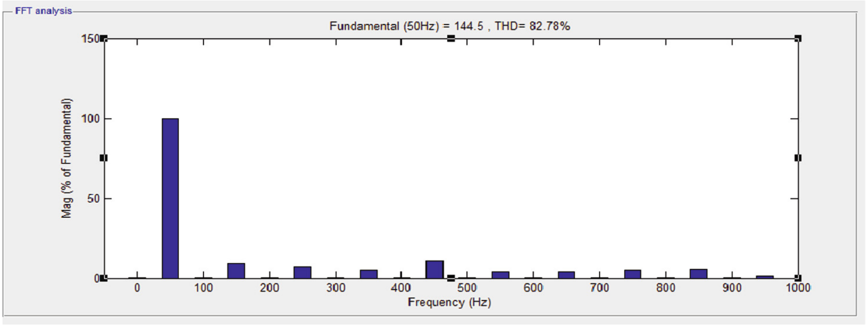


Fig. 7. THD value for 3-level

Figure 8 illustrates the output voltage waveform for 5-level CISS of three phases where each phase has a different of 120°. The voltage output is 400 V, 200 V, 0 V, −200 V and −400 V. The total harmonic distortion value shown in Fig. 9 for three phases 5-level of CISS is 37.16%.

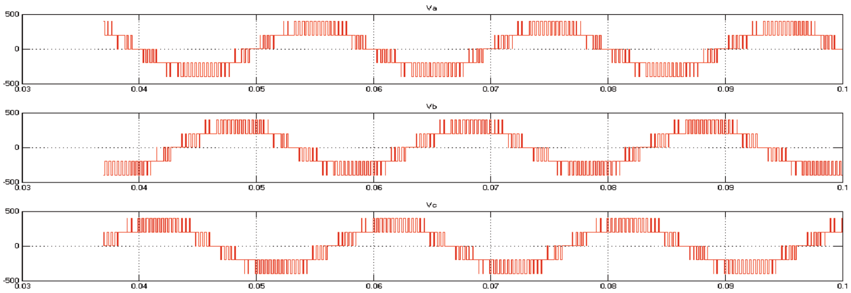


Fig. 8. Voltage output for 5-level CISS

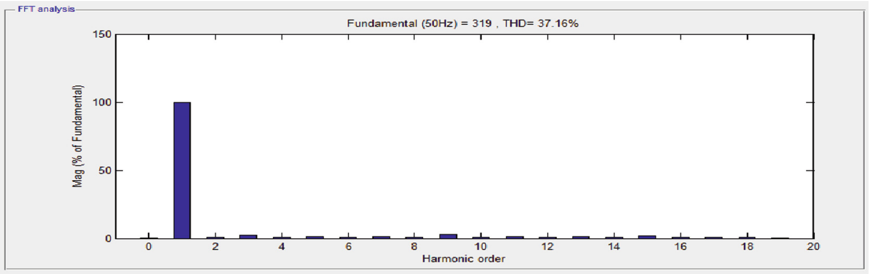


Fig. 9. THD value for 5-level

Figure 10 indicates the voltage waveform for 7-level CISS of three phases which each phase has a different of 120°. The voltage output is 600 V, 400 V, 200 V, 0, −200, −400 V and −600 V. The total harmonic distortion value shown in Fig. 11 for three phases 7-level of CISS is 23.45%.

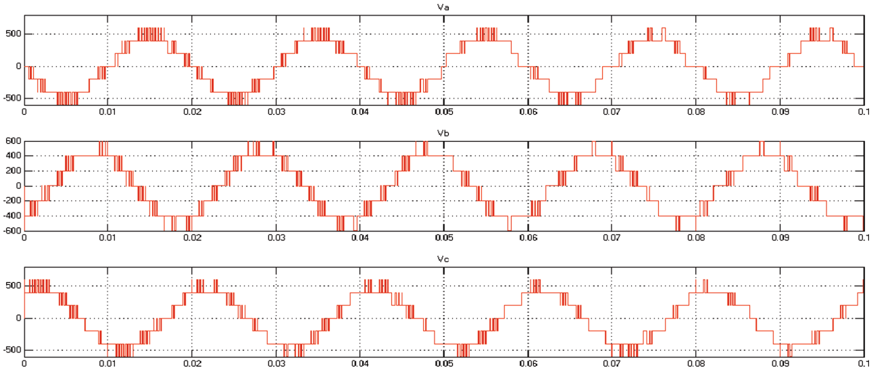


Fig. 10. Voltage output for 7-level CISS

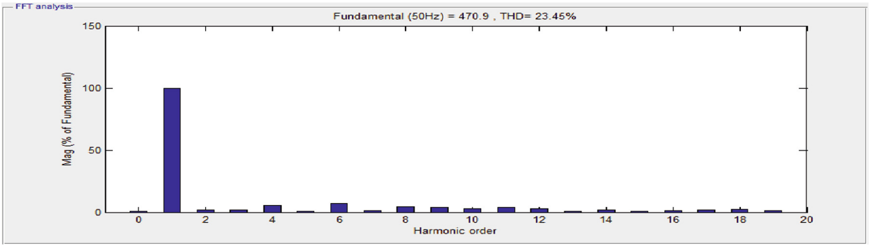


Fig. 11. THD value for 7-level

4 Discussion

In this section, the results for MATLAB simulation for 3, 5, and 7 level CISS are discussed and analysed. The results are shown for three phase simulation only because it produces a lesser percentage of Total Harmonic Distortion (THD) compared to single phase as identified in previous research [10–12]. The simulation was done to test the rationality of the results and compare it with the experimental result. The purpose of simulating 3, 5, and 7 level cascaded inverter is to prove that higher level of inverter eliminates more harmonics, thus producing lesser THD percentage. Figures 6 and 7 shows the 3 phase output voltage and Total Harmonic Distortion (THD) analysis for 3-level CISS with fixed frequency modulation (f_m) 60 Hz and modulation index (m_a) of 0.7. The purpose of using fixed modulation index and frequency modulation in 3, 5, and 7 level cascaded inverter is to make sure that the output result and THD does not affected by different values of modulation index and frequency modulation. From Fig. 6, the output voltages produced are three values that are 200 V, 0 and −200 V. The values produced are theoretically matches with previous study [4]. THD value for 3-level CISS is 82.78% can be seen in Fig. 7. The starting harmonics value for each level will be the highest as shown in Figs. 7, 9 and 11. This phenomenon occurs is because the starting harmonics represent the fundamental frequency (f_c) and it will overshoot at the starting point before the elimination of harmonics occur. Figure 8 shows the output voltage value for 5-level CISS. Five different values are produced which is 400 V, 200 V, 0, −200 V, and 400 V matches with previous analytical study [5]. THD value shown in Fig. 9 for three phase 5-level of CISS is 37.16%. Seven different output voltages produced in 7-level of CISS as shown in Fig. 10 and matches with the previous research analysis [6]. The THD value for 7-level CISS is 23.45% and it is the lowest value of THD compared to 3 and 5-level of the CISS. Table 5. summarises the comparison for 3, 5, and 7-level

Table 5. THD comparison

Number of levels	THD%
3 level	82.78
5 level	37.16
7 level	23.45

of THD. It is proven that high level of inverter eliminates more harmonics and producing lesser THD percentage in the system operation of each multilevel inverter. Hence, theoretical research study matches with the results obtained in this paper.

5 Conclusion

This paper presents the difference of THD between 3, 5, and 7-level three phases of a cascaded multilevel inverter with separated dc sources (CISS). Based on the papers of conventional multilevel inverter topologies given in the previous research, the THD value is decreased as the number of CISS level is increased. As the switches were triggered at steady intervals, plenty of computations are required to generate the pulses. This can be avoided by relating the PWM methods in multilevel inverters in the future. Percentage of harmonics can be reduced by applying the PWM techniques because it allows a decline in the switching frequency of each cell, thus reducing the switching losses. In this paper, the simulation was done to prove that higher level reduces the harmonics resulting in efficient power supply with less power loss. The simulation result theoretically matches with the results obtained in the previous study.

References

1. Doifode S, Dutt S, Shende R (2015) A study of multilevel inverter topologies. Int J Adv Res Electr Electron Instrum Eng 4(6), June 2015
2. Jamuna V, Gayathri Monicka J (2015) Multi carrier based multilevel inverter with minimal harmonic distortion. Int J Power Electron Drive Syst (IJPEDS) 6(2):356–361
3. Sultana WR, Sahoo SK, Singh HO, Dubey A (2014) Implementation of cascaded H-bridge multilevel inverter using MATLAB-DSP (ezDSP28335) interfacing. Res J Appl Sci Eng Technol 7(17):5
4. Mittal N, Singh B, Singh SP, Dixit R (2012) Multilevel inverters: a literature survey on topologies and control strategies. In: 2nd international conference on power, control and embedded systems
5. Subramanian D, Rasheed R (2013) Five level cascaded H-bridge multilevel inverter using multicarrier pulse width modulation technique. Int J Eng Innov Technol (IJEIT) 3(1), July 2013
6. Goel V, Kumar J, Gambhir J (2015) Different multilevel inverter topologies with reduced number of devices. In: Proceedings of 2015 RAECS UIET Panjab University Chandigarh, 21–22 December 2015
7. Daniel WH (2011) Power electronics. Tata McGraw-Hill Education, New Delhi
8. Roshankumar P, Rajeevan PP, Mathew K (2012) Five-level inverter topology with single-DC supply by cascading a flying capacitor inverter and an H-Bridge. IEEE Trans Power Electron 27(8), August 2012
9. Joca DR, Barreto LHSC, de Oliveira DS Jr (2013) Modulation technique based on CSV-PWM and HEPWM for THD reduction in flying capacitor multilevel inverters. Universidade Federal do Ceará Fortaleza, Brazil

10. Ahuja RK, Aggarwal L, Kumar P (2013) Simulation of single phase multilevel inverters with simple control strategy using MATLAB. Int J Adv Res Electr Electron Instrum Eng (An ISO 3297: 2007 Certified Organization) 2(10), October 2013. Department of Electrical Engineering, YMCA University of Science & Technology, Faridabad, Haryana, India
11. Javvaji HL, Varaja BB (2013) Simulation & analysis of different parameters of various levels of cascaded H bridge multilevel inverter. In: 2013 IEEE Asia pacific conference on postgraduate research in microelectronics and electronics (PrimeAsia), 19–21 December 2013, pp 62–67
12. Jain K, Chaturvedi P (2013) Matlab -based simulation & analysis of three - level SPWM inverter. Int J Soft Comput Eng (IJSCE) 2(1), March 2012. ISSN: 2231-2307

Survivor Tracking System Based on Heart Beats

Suresh Thanakodi[1(✉)], Nazatul Shiema Moh Nazar[1], Bryon Sim Phin Tzen[1], and Muhammad Muaz Mubasyir Roslan[2]

[1] Department of Electrical and Electronic Engineering, Faculty of Engineering,
National Defence University of Malaysia, 57000 Kuala Lumpur, Malaysia
{suresh,nazatul.shima}@upnm.edu.my, bryonsim@gmail.com

[2] Department of Mechanical Engineering, Faculty of Engineering,
National Defence University of Malaysia, 57000 Kuala Lumpur, Malaysia
commandermueh@gmail.com

Abstract. Survivor tracking system is a device that detects a human heartbeat by using pulse sensor and sends the information to a smartphone via Bluetooth. Apart from muscle sensor and accelerometer sensor that has difficulty in monitoring the condition of users, pulse sensor is one of the alternatives that can provide high portability and require a small surface area for analysis. In this paper, a wireless system was implemented. The design is achieved by utilizing a smartphone with Android Operating System, Peripheral Interface Controller (PIC) Microcontroller and Bluetooth. The objectives of the system are to monitor user heartbeat rate, durability and effectively SMS to obtain medical help for the user.

Keywords: Survivor tracking system · Bluetooth and PIC

1 Introduction

With the advancement of technology in this modern era, mobile technology such as smartphones have enabled a wide range of applications that can be used by people on the move. By integrating these devices with other inputs, we will be able to enhance the capability of the device. One of such integration is the use of sensors to monitor the condition of the user. This is important as it can be used to ensure the safety of the user by notifying the first responder, allowing faster and more efficient response taken.

In recent years, many researchers to monitor the condition of user have move towards the use of sensors such as an electromyogram (EMG) sensor, accelerometer sensor and pulse sensor. Sensor such as EMG is difficult to evaluate the practical contraction of muscle, and indirectly monitors the contraction of the muscle [1]. Comparing this to a pulse sensor, we will be able to obtain more valuable information to understand the health status of a human body for better analysis [2].

In this paper, we present a survival tracking system based on heartbeats rate, the system consists of a prototype device which integrate a pulse sensor with a PIC microcontroller and will transmit signal via Bluetooth to a connected smartphone with android operating system. The device will trigger when there is no detection of pulse and will transmit a distress signal via Short Message Service (SMS) or email to notify first

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_61

responder during an emergency. The signal sent will be able to detect the victims' location via Global Positioning System (GPS) for the rescue team to reach on time [3]. The smart phone must first be equipped with an application called "magnet code" for interface purposes.

The condition of the user is monitored based on the heart rate condition. The heart rate for a normal person during rest ranges from 60–100 bpm [4] while athletes will have a slightly lower range in 40–50 bpm. During exercises, a person heart rate limit can be monitored using the formula from Eqs. (1) and (2) [5].

$$(220 - \text{Age}) \times 0.60 = \text{Lower limit} \tag{1}$$

$$(220 - \text{Age}) \times 0.90 = \text{Upper limit} \tag{2}$$

The main objectives of this device are monitor heartbeat rate of users using a heart rate sensor, wearable devices which can withstand outdoor activities and effectively inform first respondent about the emergency situation of the victim.

The device is built for sports and outdoor activities such as hiking, mountain climbing, jungle tracking and etc. The device function in such a way that when the condition of the user heart rate changes to very high or very low for a set amount of time, it will trigger and send the preset message together with the user GPS location to a preset contact number. In an emergency, the user can also trigger the device manually.

The device can also be used for military survival training purposes such as Survival Army Basic Training whereby cadets are required to develop survival skill for approximately 72 h. During this period of time, cadets are only equipped with basic tools such as compass, ration and radio set. When an accident occurred, it is difficult for first responder to take action as they are unable to obtain information quickly. This device will be able to track and notify the location of the cadets when an accident occurred.

Besides that, the device can also be used in an emergency situation such as earthquakes or floods. The device is able to transmit the last track location of the user. This feature is important for search and rescue mission and also to count the number of losses to measure casualties of the disaster.

2 Technologies

2.1 Smartphone

A Smartphone is a portable device with its own operating system which combines some features that is available in computers with mobile phone. It usually consists of media player, GPS navigation unit, Bluetooth connectivity, digital camera, personal digital assistance (PDA) and etc. Most smartphones are equipped with internet access and can run third party apps.

There are currently three major operating systems used in a smartphone, which is Android, iOS and windows operating system. Android uses an open-source platform and is backed by Google. iOS is the operating system built for iPhone and is developed by Apple Inc. and windows phone is the OS built for windows mobile and is developed by Microsoft.

2.2 PIC Microcontroller

For this research, the controller used is the PIC16F767. PIC16F737/767 devices are available only in a 28 -pin configuration while PIC16F747/777 devices are for 40-pin and 44-pin configuration [6]. The block diagrams of the PIC16F737/767 and PIC16F747/777 devices are provided in Fig. 1:

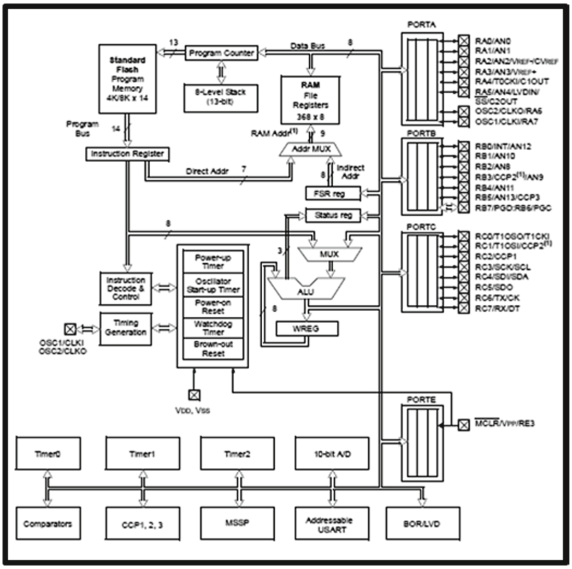


Fig. 1. PIC16F737 and PIC16F767 block diagram

2.3 Bluetooth

In this research, a Bluetooth connection is needed as a link between the two devices. The Bluetooth is a wireless technology that sends data within a short distance by using a short-wavelength UHF radio waves. The ISM band is within 2.4 GHz to 2.485 GHz. Bluetooth connection can connect to several devices which can overcome synchronization issues.

3 Methodology

3.1 Overall Process and Components

In this research, the system produced consists of a PIC-16f767 micro controller on a circuit board with App Link Bluetooth module, pulse sensor and an Android smartphone with Magnet code application [6]. These are the main components used to assemble and ensured that the objective of the research was achieved.

For the device to function optimally, it must first be connected to a smartphone via Bluetooth. The device is then attached to the user where the pulse sensor will monitor the pulse condition. The Android smartphone can orientate the data of the device by installing an application known as "m3". The magnet code and programming of the PIC-f767 to track either the heart is beating is done using two programs which are the PIC compiler and PIC kit 2 v2.55 [7]. Figure 2 shows the flowchart of system design.

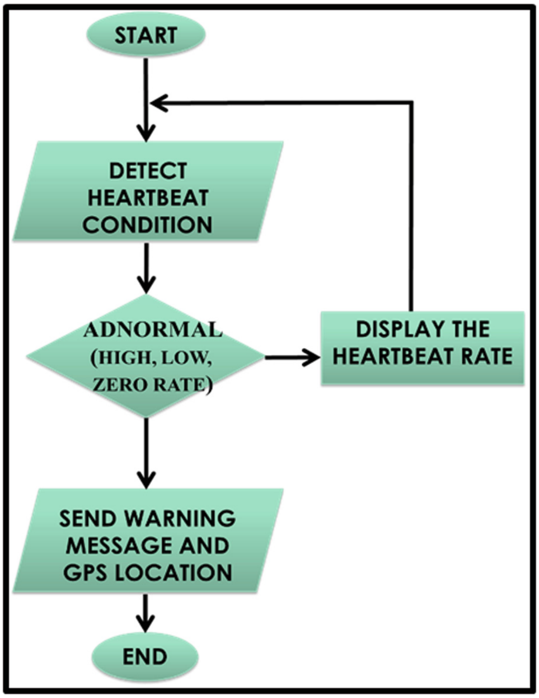
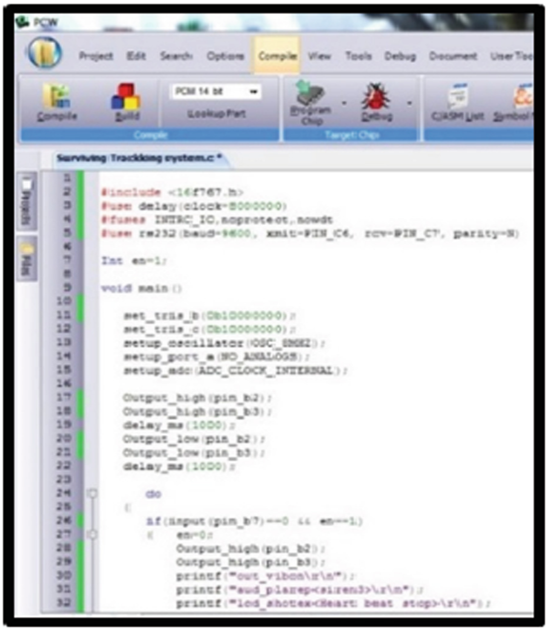

START
DETECT HEARTBEAT CONDITION
ADNORMAL (HIGH, LOW, ZERO RATE)
DISPLAY THE HEARTBEAT RATE
SEND WARNING MESSAGE AND GPS LOCATION
END

Fig. 2. Flow chart of system design

3.2 Assembling

Assembling the research are based two elements that are hardware & software. In the hardware assembling process the components of the device were collected & fitted onto the motherboard.

Meanwhile, the assembling process in software is planned, checked & installed into the PIC micro-controller using two Software Program which are 'PIC C Compiler' to construct the program logic & PICkit 2 v2.55 to install the program logic into the PIC [7, 8]. The logic of the program can be viewed in Fig. 3.

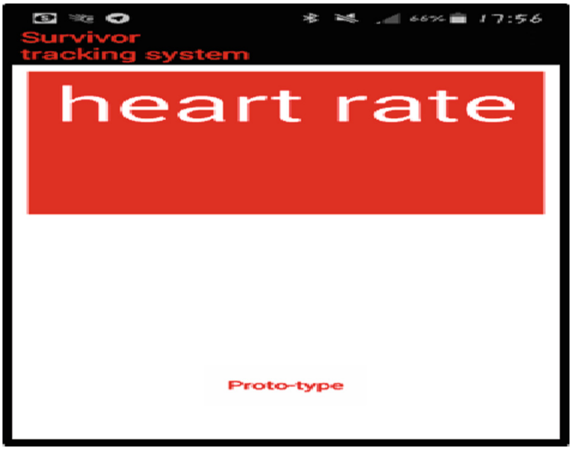


Surviving Trackking system.c *

```
#include <16f767.h>
#use delay(clock=8000000)
#fuses INTRC_IO,noprotect,nowdt
#use rs232(baud=9600, xmit=PIN_C6, rcv=PIN_C7, parity=N)

Int en=1;

void main()

   set_tris_b(0b10000000);
   set_tris_c(0b10000000);
   setup_oscillator(OSC_8MHZ);
   setup_port_a(NO_ANALOGS);
   setup_adc(ADC_CLOCK_INTERNAL);

   Output_high(pin_b2);
   Output_high(pin_b3);
   delay_ms(1000);
   Output_low(pin_b2);
   Output_low(pin_b3);
   delay_ms(1000);

      do
   {
      if(input(pin_b7)==0 && en==1)
      {    en=0;
           Output_high(pin_b2);
           Output_high(pin_b3);
           printf("out_vibon\r\n");
           printf("aud_playrep<siren3>\r\n");
           printf("lcd_showtex<Heart beat stop>\r\n");
```

Fig. 3. Program logic in PIC C compiler

Figure 4 shows the GUI application for this research. The program is then installed into the PIC microcontroller with PICkit 2 v2.55 using an adapter in Fig. 5.



Survivor tracking system

heart rate

Proto-type

Fig. 4. GUI of application

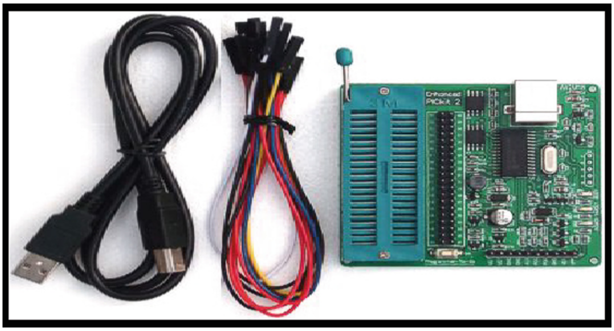


Fig. 5. PIC adapter

3.3 Testing Method

The test run was carried out on the assembled survivor tracking system on a user. Meanwhile the smart phone is held by another user to examine the reliability of the system based on connection, responds, and accuracy of the system reacting to the different situation made. Program code build is analyses and tested using PIC kit [9].

4 Discussion and Analysis

Although we have achieved the three main objectives of the research, there are still a few limitations in the project that could be improved in future. One of the limitation is in the coding of the device which is set to send signal continuously every 10 s. This setting will increase power consumption and result in less efficiency in power usage.

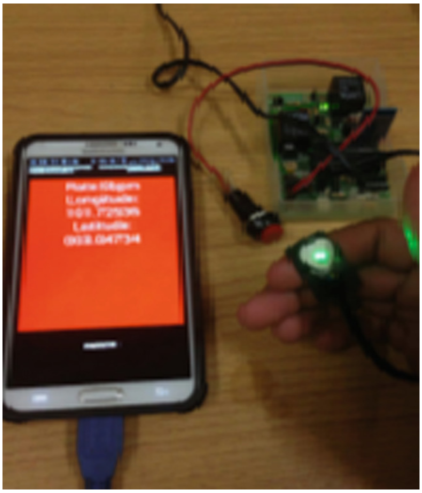


Fig. 6. The survival tracking system with turn on

This problem can be solved by modifying the coding of the PIC microcontroller. Figure 6 shows the completed survivor tracking system with the system in condition on, while Fig. 7 shows the completed survival tracking system when detects the heartbeat.

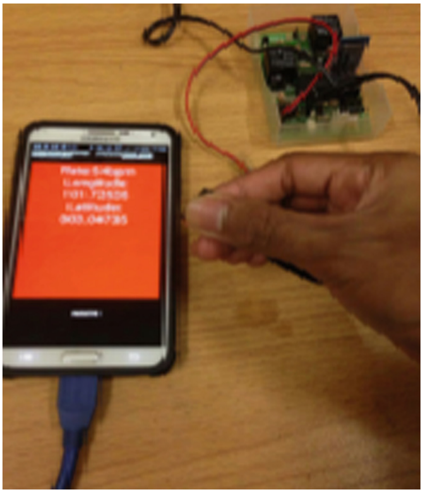


Fig. 7. The survival tracking system detect the heart rate

Table 1 shows the data that were extracted from the test. The data for the communication part presented firstly and followed by detection of the heart beat and the responsive of device to send SMS for help.

Table 1. Test Results

	Yes	No	Comment
Connection			
Communicate	✓		
Lagging	✓		0.2 s
Accuracy	✓		000 b/m

Next is durability of the device itself. Since the research that was held focused on obtaining the three main objectives of the research. The research has lacked qualities in this element. Durability element is important to ensure that the device is sustained and in good hands to function in its best condition.

Finally, the device tracking system has limited range of data connection through SMS, this is basically due to the budget provided. SMS ranges of connection are limited in certain areas. Improvements for further researches could apply the use of direct satellite connection where there is no limit in connection around the world. This will allow the device to send free and unlimited signal to track and monitor the condition of the user. Thus, a suggestion in the future is to allow free SMS services to provide limitless connectivity, especially in a region that is prompt to accident.

Last but not least, with further improvement of this system, it can be used to contribute in the medical field such as its facility to monitor the condition of patients and to prevent diseases before they occur [10]. This will lead to a safer and healthier life in the future.

5 Conclusion

The designed survivor tracking system was able to function accordingly, but with some minor glitches on the duration of taking heart rates. This research also manages to identify the limitations of the design and suggestions in enhancing it. Since this preliminary design exhibits possibility of the usage belief that this research could proceed to another level in nanotechnology. Overall the designed could be improved with enough research on the subject with high possibilities of success.

References

1. Tanaka M, Okuyama T (2011) Study on evaluation of muscle conditions using a mechanomyogram sensor. In: Systems, man, and cybernetics (SMC). IEEE
2. Wang L, Wen D (2009) Research on non-invasive arterial pulse sensor and detecting system. Biomedical Engineering and Informatics, IEEE (2009)
3. Yorozu Y, Hirano M, Oka K, Tagawa Y (1987) Electron spectroscopy studies on magneto-optical media and plastic substrate interface. IEEE Transl J Magn Japan 2:740–741
4. Target Heart Rates - AHA (2014) Target Heart Rates. American Heart Association, 4 April 2014. Accessed 21 May 2014
5. Izzo JL, Black HR (ed) Hypertension primer: the essentials of high blood pressure
6. Verle M (2009) PIC microcontrollers - programming in Basic, 1st edn. mikroElektronika, Virginia
7. Yousif I (2012) Design and implementation of electronic control trainer with PIC microcontroller. Intell Control Autom 3:222–228
8. Verle M (2010) PIC microcontrollers - programming in C, 1st edn. mikroElektronika, Virginia
9. CCS C Compiler Manual, October 2015
10. Nishiyama M, Watanabe K (2011) Unconstrained pulse pressure sensing for health management based on a hetero-core fiber optic sensor In: Sensors, IEEE

Boundary Effects to Accelerate Stochastic Moving Multi-Agents

Yuta Tsuruoka[1], Shihoko Tanabe[2], Marin Numazaki[2], and Isamu Shioya[1](✉)

[1] Graduate School of Science and Engineerings, Hosei University, Kajino-cho 3-7-2,
Koganei-shi, Tokyo 184-8584, Japan
shioyai@hosei.ac.jp

[2] Department of Science and Engineerings, Hosei University, Kajino-cho 3-7-2,
Koganei-shi, Tokyo 184-8584, Japan

Abstract. This paper considers the coordination of autonomous stochastic moving multi-agents consisting of finite cells arranged on a line. We assume there are interactions or coordination among agents such that (1) each agent can not move to a destination cell occupied by agents more than the agents of a current cell and (2) their agents have time-lag. Then it is ideal that every cells are always occupied by agents, because of the efficient use of resources, and it is desirable to be the fewest expected number of cells not occupied by agents. We show that the resource utilization becomes higher if every agents have appropriate average moving speed, also called fluctuations. In addition, the speed on boundaries, also called a boundary effect, further accelerates the cell resource utilization efficiency, where the boundary effects depend on the resource configurations. Then, they accelerate the resource utilization by which are giving fluctuation like to shake agents in a box. By giving some fluctuation, they can maximize the cell resource utilization.

Keywords: Autonomous moving stochastic multi-agents · Resource utilization · Boundary effect

1 Introduction

This paper discusses a resource utilization of autonomous stochastic mobile multi-agents with time-lag and average mobile speed. Every agents stochastically move on the cells along finite resources arranged over a line according to transition probabilities in synchronization. The moving of each agent at every time steps is affected by other agents such that it depends on the number of other agents on destination cells within specific ranged windows, and there is coordination among agents with time-lag. The interactions among agents are complicated to analyze the behavior of them, since each agent moves over the resources autonomously, and their combinations lay on a platform of huge spaces. The paper [7,8] showed that the stochastic moving multi-agent behavior becomes more stable if every agents take an appropriate average mobile speed. The paper

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_62

[7] discusses the resource utilization of agents, but does not consider the boundary effect on resources. Then, the average mobile speed of all the agents is the same, i.e. it is independent from resource locations. The other paper [6] presents more accelerated stable stochastic mobile multi-agent behavior by adjusting the average moving speed depending on the agent staying locations. Our analysis and experiments show that the variations with respect to the number of agents on cells become lower, that is more stable, when all the agents have an appropriate average mobile speed according to the agent staying locations, i.e. we can design more stable multi-agent configurations if we adjust their average mobile speed depending on the agent locations [6]. In this paper, we present the multi-agents behavior for higher resource utilization of cells considering boundary effect on resources.

We are in need of a simple model with no fat in mobile multi-agents for analyzing complex systems. Fortunately, Sen et al. [5] proposed a simple model for analyzing the behavior of mobile multi-agents, and Rustogi et al. [3] presented the fundamental results of the former model. Ishiduka et al. [2] also introduced a time lag and showed the relationship between time lag and stability in mobile multi-agents. The above models are intended to clarify how fast the mobile multi-agents fall into a complete stable state, i.e. a hole state in absorbing Markov chain [1], thus the goal is to design a coordinative system which falls into a stable hole in shorter passage time as soon as possible. The basis of this theory is Shilling model [4].

Our model, Multi-Agent behavior with Time lag and Moving Speed: MATMS, is based on Sen et al. [5] and the developed model with time-lag proposed by Rustogi and Singh [3]. We note that our purpose is different from the papers [2,3,5] which try to clarify the relationships between time-lag and stability in multi-agent systems. In other words, their papers try to find the multi-agent configurations satisfying autonomous uniform resource allocation in a shortest passage time. On the other hand, our multi-agents initially start from a most stable state, each agent on resource stochastically moves over cells, and it just likes atoms in a liquid. The agents are always moving on resources stochastically, and they never stop. In addition, we extend their models to have moving average speed. Our model satisfies Markov condition and irreducible so that the configurations do not depend on the initial configurations in the limit, and our problem is to find more stable multi-agent configurations accompanying agent movements. It just likes as a molecule has an energy so that it is always moving while the agents are alive, and it depends on the manner of substances. The paper [8] showed that a stochastic mobile multi-agent system, whose the agents move slowly as a whole on average, is more stable than other ones not having average moving speed as a whole theoretically and experimentally. This paper demonstrates higher resource configuration behavior when we consider the agent locations on resources, that is, boundary effects.

This paper is organized as the following. First, we define our model in Sect. 2. Section 3 shows that there exists a new distinct behavior based on theoretical analysis for small size 3×3, and Sect. 4 presents our experimental results to confirm the theoretical analysis. In the following section, we discuss the related works. Finally, we conclude this paper in Sect. 5.

2 Stochastic Moving Multi-agent Model with Time-Lag and Moving Speed

We shall define a multi-agent consisting of k agents, $k \geq 2$. All the agents are arranged over a finite resource consisting of cells $S(i)$, $i = 1, .., n$ on a straight line $[1, 2, .., n]$ instead of a circle [3,5], and move to synchronize over the resource according to the following transition probabilities $p_{i,j}$ in stochastic manner. In the following, sometimes we are simply expressed as i a resource $S(i)$.

First, we define a weight function $f_{i,j}$, $i, j = 1, .., n$ as

$$f_{i,j} = \begin{cases} 1, & i = j \\ 0, & i \neq j, r_i < r_j \\ 1 - \frac{1}{1+\gamma \exp(\frac{move(r_i - r_j, i, j) - \alpha}{\beta})}, & \text{otherwise}, \end{cases} \tag{1}$$

where r_i are the number of agents on i-th cell, and α, β and γ are constants. α is called an "inertia" which is the tendency of an agent to stay in its resource [3], and *move* is an accelerated function to give average moving speed on either left or right defined by it in later. Rustogi et al. model [3] does not satisfy the condition "irreducible" exactly, while our model satisfies Markov property under the condition not to restrict the moving directions of agents, and the model becomes irreducible (Fig. 1).

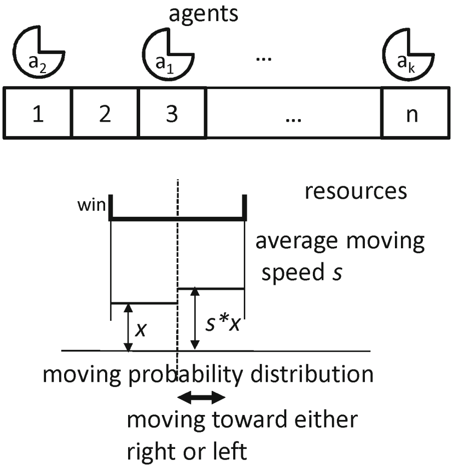

agents
a2
a1
...
ak
1
2
3
...
n
win
resources
average moving speed s
x
s*x
moving probability distribution
moving toward either right or left

Fig. 1. The model MATMS.

Our model also has an average moving *speed* such as every agents move with an average moving speed s_i ($s_i \geq 1$) either of left or right directions on the cells, where i are i-th locations of agents. Their agents move along the resources arranged over the straight line according to the probability $p_{i,j}$ in stochastic manner, where i and j indicate i-th and j-th cells, respectively. In the case of right average moving speed s_i, the function $move(x, i, j)$, which describes the ratio of imbalance from a cell i to a destination cell j with the difference $x(= r_i - r_j)$ in the numbers of agents on i and j, is defined by $move(x, i, j) = s_i \times x$ if $i < j$, otherwise x, wheres s_i is an average moving speed at i-th cell. On the other hand, for the left average moving, $move$ is defined by $move(x, i, j) = s_i \times x$ if $i > j$, otherwise x.

A moving transition probability $p_{i,j}$ from a cell $S(i)$ to a destination cell $S(j)$ is defined by the normalization of $f_{i,j}$ with probability 1 based on $f_{i,j}$. Rustogi et al. [3] introduced a window $win(i)$ with a fixed size for analyzing the behavior of multi-agent systems with time-lag. Then, a moving transition probability $p_{i,j}$ from a current cell $S(i)$ to a destination cell $S(j)$ is defined by

$$p_{i,j} = \begin{cases} \dfrac{f_{i,j}}{\sum_{k \in win(i)} f_{i,k}}, & i = 1, .., n, j \in win(i), \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where w is a window size, and $win(i)$ is the set $[i - w, i + w]$. A time delay which is local properties is proportional to the window size w (see [2,3]).

There are no constraints on the moving of agents such that each cell has a fixed upper limit capacity to occupy agents, while there is another constraint in the model, i.e. the moving transition probability $p_{i,j}$ is 0 if the number of agents on a cell $S(i)$ is less than the number of agents on a destination cell $S(j)$.

Our proposed model, Multi-Agent behavior with Time delay and Moving Speed: MATMS, is similar to the models [2,3]. The resources in MATMS are arranged over a straight line $[1, 2, .., n]$ as in [2], and the wind function $win(i)$ is the set $[i - w, i + w] \cap [1, n]$ if w is a window size. We note that there are two choices on the moving average directions which are either left or right. Suppose an agent moves towards left on average at the previous step. Which is the moving direction at the next step? If we exclude the cases that the agents stay on boundaries, there are two exclusive cases(or a model protocol) for each agent independently: (1) we inherit the directions at the previous steps, i.e. left on average in above, or (2) we randomly select it at each step according to even probability either left or right, i.e. half to half rule for the direction. The second case (2) is suite to Markov property. The first case (1) does not satisfy Markov property so that the systems depend on the initial configurations.

3 Theoretical Analysis of 3 × 3 Model

In this section, we present a concrete mobile multi-agent such that the multi-agent taking an appropriate average speed achieves higher resource utilization than a multi-agent not taking moving average speed, i.e. every cells are occupied by agents in many cases on average.

Suppose the multi-agent of which the number of cells and agents are 3 together. This is a minimal model to examine a coordination among agents. We first use the parameter values $\beta = 2$ and $\gamma = 1$, and fix the window size w to 1.

An example of the agent moving configuration is represented by $(a, r, 1)$ if the agent a on the cell 1 moves towards right with average moving speed s_b. The multi-agent moving configuration consists of three agent configurations in this minimal model. For an example, $[(a, r, 1), (a, r, 1), (a, r, 1)]$ is a multi-agent moving configuration.

In our minimal model, the directions of average agent moving are stochastically chosen at every steps so that the multi-agent becomes Markov chain. In this setting, there are 10 multi-agent states shown in Fig. 2, and we must consider 136 probabilistic transition rules shown in Appendix in [8]. That is, the number of the states (Fig. 2), the state transition rules (Appendix in [8]) and the multi-agent moving configurations (The top items of Appendix in [8]) are 10, 136 and 20, respectively. The illustration of the transition rule (d-1) is shown in Fig. 3. Also the details of the transition rules (g-2) in Appendix in [8] is shown in Fig. 4.

This simple model satisfies Markov condition and it is irreducible, so we easily compute the eigenvectors of the state transition matrix with the size 10×10 using Appendix in [8], and compute the transition probabilities among every states in

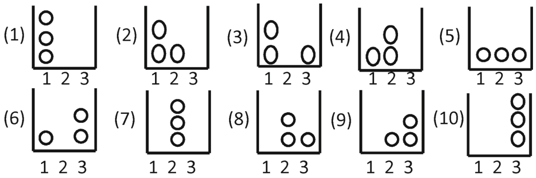


Fig. 2. The states of the multi-agent: *cells* = 3, *agents* = 3 and $w = 1$.

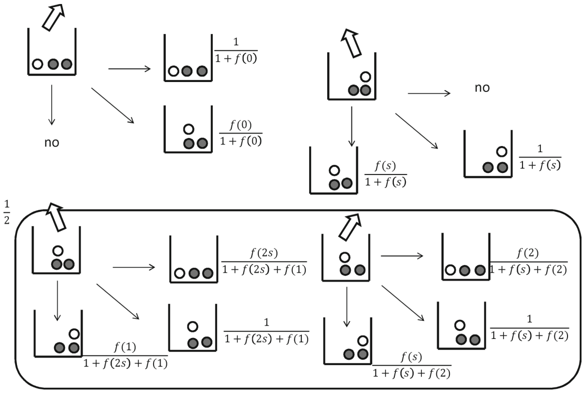


Fig. 3. The transition rules, Appendix in [8].

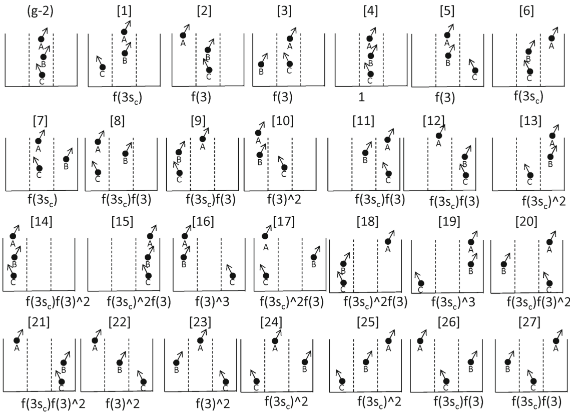


Fig. 4. The details of the transition rules (g-2), Appendix in [8]. The denominators $(1 + f(3s_c) + f(3))^3$ are abbreviated.

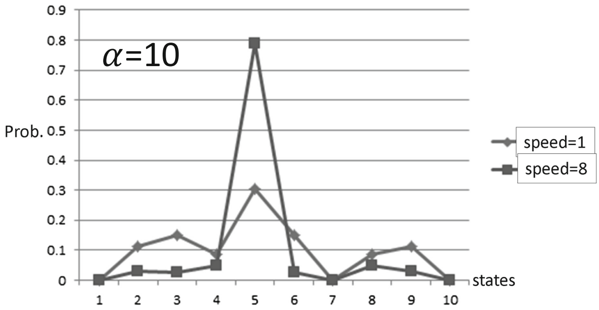


Fig. 5. The probabilities staying the states in the case $cells = 3$, $agents = 3$ and $w = 1$ based on theoretical computation.

the limit by changing the moving speed. The theoretical computational results of the existence probabilities for every states are shown in Fig. 5.

The expected average number m_1 of the cell 1 occupied by agents is given by the following:

$$m_1 = p_1 + p_2 + p_3 + p_4 + p_5 + p_6, \tag{3}$$

where p_i are the probabilities of the correspondence states i shown in Fig. 5. In other words, m_1 is the average resource utilization on the cell 1.

By similar way, we can compute the expected average numbers m_2 and m_3 of the cells 2 and 3, respectively, occupied by agents:

$$m_2 = p_2 + p_4 + p_5 + p_7 + p_8 + p_9, \tag{4}$$

$$m_3 = p_3 + p_5 + p_6 + p_8 + p_9 + p_{10}. \tag{5}$$

The whole expected average number v_m of cells occupied by agents, i.e. the resource utilization over the resource, is given as

$$\begin{aligned} v_m = {} & (p_1 + p_7 + p_{10}) \\ & + 2(p_2 + p_3 + p_4 + p_6 + p_8 + p_9) \\ & + 3p_5. \end{aligned} \tag{6}$$

4 Experiments

In this section, we present our experimental results and make the comparisons of the theoretical analysis and the experimental results.

We initially configure the multi-agent which is staying at most stable state, i.e. the state (5) in Fig. 2, and observed it during 50,000 steps. We computed the variance with respect to the number of agents between 10,001 and 50,000 steps. Then, we used Mersenne twister random number generator for a long period.

We show our experimental results in Figs. 9, 10 and 11 by changing α, s_b and s_c. Compare Figs. 6, 7, 8 and Figs. 9, 10, 11 to the order in which the have

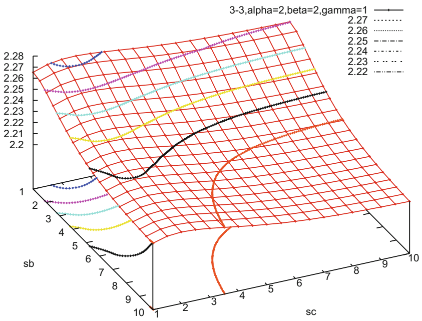


Fig. 6. The theoretical analysis of the resource utilization, $\alpha = 2$.

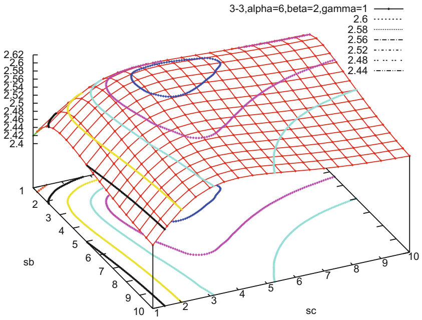


Fig. 7. The theoretical analysis of the resource utilization, $\alpha = 6$.

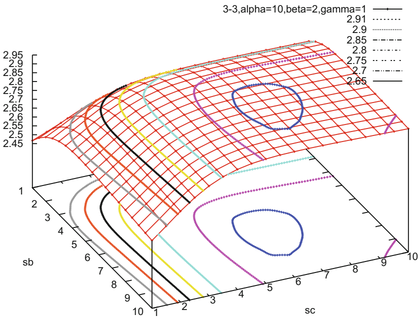


Fig. 8. The theoretical analysis of the resource utilization, $\alpha = 10$.

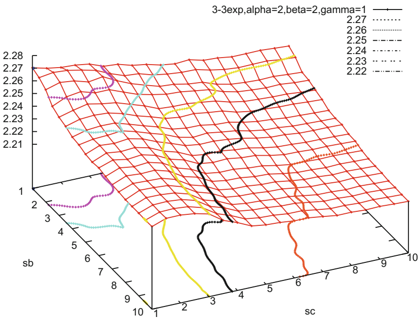


Fig. 9. The experimental analysis of the resource utilization, $\alpha = 2$.

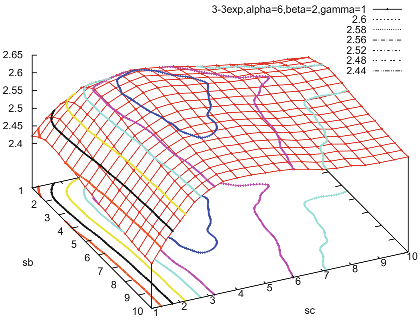


Fig. 10. The experimental analysis of the resource utilization, $\alpha = 6$.

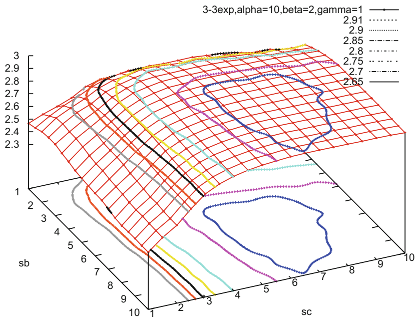


Fig. 11. The experimental analysis of the resource utilization, $\alpha = 10$.

Table 1. The experimental results of the optimal moving speed in resource utilization.

α	s_b	s_c	v_m
2	1	2	2.27
4	2	3	2.41
6	3	4	2.61
8	5	5.5	2.8
10	7	7.5	2.92

listed up. The experimental results and the theoretical analysis are almost the same. Table 1 shows the optimum moving speed, and Table 1 shows the optimum points of the v_b versus v_c. While the boundary effect is greater in small α, the effect is small in large α.

5 Conclusions

In this paper, we considered a stochastic mobile multi-agent model, and presented that the model, Multi-Agent behavior with Time delay and Moving Speed: MATMS, having appropriate average stochastic moving speed become higher resource utilization than ones not having average moving speed. Then, we considered the boundary effects on resources. This shows that each agent needs the moving acceleration to achieve hight resource utilization. The acceleration is a *speed* in this paper.

In our model, we showed that there is an appropriate speed to achieve the most stable or utilizable configuration for each inertia α. Since individual objects in nature are governed by the lower entropy, all the objects which move randomly with interactions over resources cause naturally flow. Then, each object may independently take a different direction for moving rather than coordination, i.e. no need to control agents. We may extract the energy from the multi-agents which move randomly in a closed region, then we may think them just like atoms.

This kind of work has done by Toyabe et al. [9] in single agent. The applications of this research are online algorithm and numerical analysis.

References

1. Grinstead CM, Snell JL (1997) Introduction to probability. American Mathematical Society
2. Ishiduka Y, Iwanuma K (2003) A relationship between time delay and the amount of knowlege in distributed cooperation of multi-agent systems. IEICE D-I, J86-I, pp 117–120. Japanese
3. Rustogi SK, Singh MP (1999) Be patient and tolerate imprecision: how autonomous agents can coordinate effectively. In: 6th IJCAI
4. Schelling Thomas C (1971) Dynamic models of segregation. J Math Sociol 1:143–186
5. Sen S, Roychowdhury S, Arona N (1996) Effects of local information on group behavior. In: Proceedings of the second international conference on multiagent systems, AAAI
6. Shioya I (2013) Accelerated stable stochastic mobile multi-agents using boundary effects. In: The third international conference on digital information and communication technology and its applications (DICTAP), vol 3, pp 54–67
7. Shioya I (2014) An accelerated resource utilization in autonomous stochastic mobile multi-agents. Int J Innovation Digital Econ 5:1–14
8. Shioya I, Miura T (2012) An autonomous accelerated coordination of stochastic moving multi-agents under variations and resource utilization. Int J Digital Inf Wireless Commun (IJDIWC) 2:943–957
9. Toyabe S, Sagawa T, Ueda M, Muneyuki E, Sano M (2010) Experimental demonstration of information-to-energy conversion and validation of the generalized jarzynski equality. Nat Phys 6:988–992

Technology and Applications of IoT with Wireless Advanced Networking

Discovering Similar Music for Alpha Wave Music

Yu-Lung Lo(✉), Chien-Yu Chiu, and Ta-Wei Chang

Department of Information Management, Chaoyang University of Technology,
168, Jifeng E. Road, Wufeng District, Taichung 41349, Taiwan, R.O.C.
yllo@cyut.edu.tw, ciouyu09@gmail.com, mylyfwy771@gmail.com

Abstract. When people close eyes to relax, an alpha wave in the frequency range of 8–13 Hz appears from brain signals. There were many medical reports proofed that some specific music can resonate with the alpha wave and strengthen the wave. Therefore, this alpha wave music can improve more relaxing for people and are very helpful when they need to take a rest. Due to the alpha wave music is classified manually by experts only, it is not popular in the market currently. In this paper, we will investigate the content-based features of the alpha wave music and use them to analyze the similarity between alpha wave music and existing music genres. The purpose of this research is to find the music which is similar to alpha wave music, such that we can recommend to users for relaxing before the automatic classification scheme for alpha wave music being developed.

1 Introduction

Listening music can stimulate the brain's functioning and music therapy uses music to help patients to improve or maintain their physical and spiritual health [7]. People usually listen to music to relieve stress when they feel under the pressure. However, which music can help people to relieve the pressure? The Dr. Hans Berger discovered four major types of brain waves exist, including β (Beta) wave, α (Alpha) wave, θ (theta) waves, and δ (delta) wave [3]. There are different frequencies of brain wave detected while humans are in different state of mind. Among them, the frequency of alpha wave between 8 Hz and 13 Hz, as shown in Fig. 1, is measured by EEG (Electroencephalography) when people close their eyes for a short rest. There were many medical reports proofed that some specific music can resonate with the alpha wave and strengthen it [1, 2, 7]. Therefore, the alpha wave music can improve more relaxing for people and is very helpful when they need to take a rest. That's why most people like to listen to music when relaxing. Although it rare, there still a few alpha wave music albums can be found on the market, such as:

- Masterworks
- The Journey Home
- Into The Deep
- Gaia
- Cloudscapes.

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_63

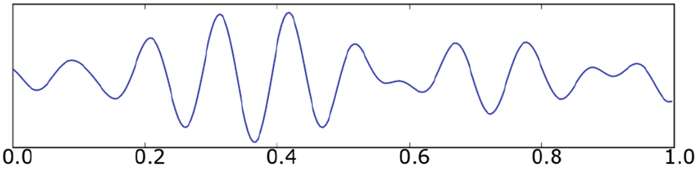


Fig. 1. Frequency of alpha brain waves detected by EEG [16]

The content of digital music provides many features which can be used for music analysis and retrieval. The music features, such as melody, rhythm, and chord, can represent the music styles and characteristics. Therefore, content-based music classification as well as music retrieval is an important research field for music databases. There were approaches for content-based music classification, such as music classification using significant repeating patterns by [8], hierarchical genre classification for large music collections by [4], automatic chord recognition for music classification and retrieval by Cheng et al. [5], content-based multi-feature music classification by Lo et al. [9], and so forth. However, these existing music classification approaches are almost all categorized by styles and genres, such as pop, classical, jazz, folk, etc. Lo et al. [10] also proposed content-based classification of alpha wave music, however, this study emphasized on analyzing the common features of already identified alpha wave music. It can not substantiate the accuracy for further application on classifying of alpha wave music. Accordingly, till now, the alpha wave music is classified manually by expertise only and is very rare on the market.

In this paper, we will investigate the content-based features of music and use them to analyze the similarity between alpha wave music and existing music genres. The purpose of this research is to find the music which is similar to alpha wave music, such that we can recommend user to listen such music for relaxing before the scheme of automatic classification of alpha wave music being developed. We hope our effort will help people not only to find more relaxed music but also to aid of music therapy.

2 Related Works

In recent years, automatic classification of music data can be discriminated into two categories. One is based on analysis of music content for classification, such as SRP-Based Classification by Lin et al. [8], Hierarchical Genre Classification by Brecheisen et al. [4], content-based multi-feature music classification by Lo et al. [9], and so on. The other category is the application of training by learning machines in which naive Bayesian, linear and neural network are employed to build classifiers for styles, such as Extreme Learning Machine by Loh et al. [11], Multiple-Instance Learning by Mandel [12], Automatic Chord Recognition by Cheng et al. [5], Multi-modal Music Genre Classification Zhen et al. [15], Optimized Feature Vector by Deepa et al. [6], and so forth. In addition, the Multi-Label Music Mood Classification proposed by Myint et al. [13] uses new mood taxonomy model to classify music.

These music classification approaches are just about all categorized by styles and genres, such as pop, classical, jazz, folk, etc. Lo et al. [10] studied the common features of the already identified alpha wave music. However, their work cannot be verified consistent with the consequence of classifying by expertise. Accordingly, till now, the alpha wave music is classified manually by expertise only. Therefore, to find the music which is similar to alpha wave music may also be a good way to recommend for people.

3 Research Method

3.1 Music Features

Our research method based on music features and comparison schemes. A musical composition consists of three basic elements - note, rhythm and harmony. Chords are a part of harmony as well. Moreover, the pitch change is also an important characteristic to compose music.

Notes – A melody was composed by notes. The fundamental frequency of musical note A above middle C is usually set at 440 Hz [14]. The pitch ratio between any two successive notes of the scale is exactly equal to $\sqrt[12]{2}$ (about 1.05946). The A an octave above that is 880 Hz because they are twelve notes apart. We would like to explore the alpha wave music to ascertain whether there are specific notes existing most likely to come about the harmonic resonance in the brain.

Rhythms – Rhythm is the pattern of musical movement through time. It is formed by a series of notes differing in duration and stress. For example: the 2/2 time signature means two half-note (crotchet) beats per bar and the beat pattern is strong-weak, the 3/4 time signature means three quarter-note beats per bar and the beat pattern is strong-weak-weak, and the 4/4 time signature means four quarter-note beats per bar and the beat pattern is strong-weak-strong-weak. Most of Waltzes music are the 3/4 time signature. Generally, quick tempo can boost the human's spirit and slow tempo can make people to feel relaxing. We would also like to analyze the connection between music rhythm and alpha wave music.

Chords – A chord in music is any harmonic set of two or more notes that is heard as if sounding simultaneously. The most frequently encountered chords are triads, so called because they consist of three distinct notes, further notes may be sevenths, ninths, and so forth. There are also four types of triads - major, minor, augmented, and diminished. People listening to various chords have distinct feelings such as sorrowful for major chords, suddenly enlightened for diminished seventh chords, and unexpectedly flying overhead for major second chords. The affection of chord may be an interesting direction for studying alpha wave music.

Pitch change – Pitch change is the variation of two adjacent notes. For example, a melody segment of the Little Bee is "So Mi Mi Fa Re Re Do Re Me Fa" such that

> the pitch changes will be "−2 0 +1 −2 0 −1 1 1 1". Since the pitch change is not effected by music key up and down, it is a favorable feature for query by example in music retrieval. Normally, the pitch change of hot music is more significant that may inspirit people. On the contrary, the pitch change of lyrical music is smoother that may allow people to relax. The alpha wave music seems to have the same effect as lyrical music does.

Among them, the chord is complicate in variety. Therefore, only notes, rhythms and pitch changes will be investigated in our studies.

3.2 Comparison Schemes

Our study used distance functions and machine learning to explore which music is more similar to alpha wave music.

3.2.1 Distance Function

In [10], we can first analyses the music content, such as notes and rhythms, as features for individual genre (ex: alpha, classical, and so forth). Let the frequencies of n highest occurrences are x_1, x_2..., x_n for a feature of a music genre then these values can be the coordinates of the centre as in an n dimensional space. Thus, alpha wave music can be examined by the distance from the centre of each music genre. Suppose a music has been analyzed and the n highest occurrences of a music feature are y_1, y_2,..., y_n with in decreasing order. The distance function $d(y_1, y_2,..., y_n)$ for the music to the centre can be derived as Eq. (1).

$$d(y_1, y_2, \ldots, y_n) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{1}$$

Thus, the most closest genre of an alpha wave music then can be decided when the distances to the centre of all music genres have been examined.

3.2.2 Distance Function with Weight

We derived another distance function for our experiment as shown in Eq. (2). It is similar to Eq. (1) except that a weight factor w_i is added. Where w_i denotes the weight for the ith music feature.

$$wd(y_1, y_2, \ldots, y_n) = \sqrt{\sum_{i=1}^{n} (w_i(x_i - y_i))^2} \tag{2}$$

3.2.3 Machine Learning

Machine learning is an algorithm that analyzes the rules from the sample data and uses the rules to automatically predict the unknown data. It is also often used in data classification such as [11, 12]. Therefore, we hope that through the machine learning technology to analyze the characteristic regularity of the music genre and to carry out the classification of alpha wave music analysis. Our experiment will use support vector machine through LIBSVM [17] and MATLAB [18] to achieve.

Support Vector Machine (SVM) – It is an approach for statistical classification and regression analysis. In which, the classified data is trained to find a hyperplan to establish a classification model. Then, such model is used to exam the data that has not yet been classified.

LIBSVM (A Library for Support Vector Machines) – Proposed in [17], it supports diverse classifications for easy used of SVMs (such as C-SVC, nu-SVC) and regression analyses (such as epsilon-SVR, nu-SVR).

We denote that we will exam music data in two ways for SVM experiment. The first one uses the highest frequency of the *n* characteristics of music in a genre for SVM training to establish a classification model. The second way uses Eq. (2) to evaluate the distance for music to the center of the belonged genre such that the distances can be used for SVM training to establish a classification model.

4 Experiment Analyses

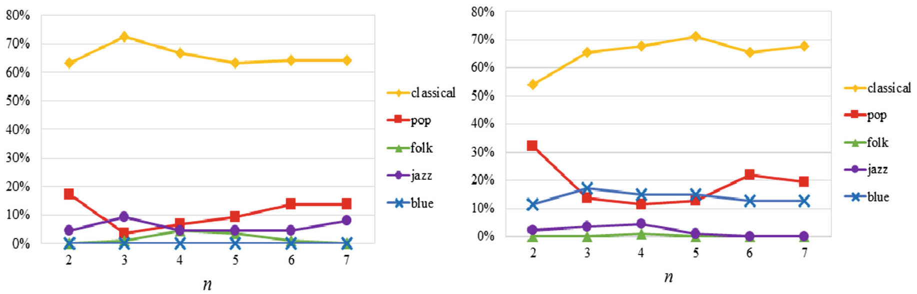
4.1 Experimental Setting

We collect classical, folk, pop, jazz, and blue five music genres in our database and each genre has 150 pieces of music. Since there have been not numerous alpha wave music albums classified by experts in the market, we have merely collected 87 scores of alpha wave music for our music database. Therefore, our experimental database contains total 837 of music. We also extracted notes, rhythms, and pitch changes of collected music as the features and analyzed their occurrence frequencies for experiment. Thus, the distance equations and LIBSVM can be applied in our experiment. The experimental results are shown in the following sessions.

4.2 Experimental Results

4.2.1 Analysis of Notes

To start experiment, we first analyze the occurrence frequencies of notes of music data in our database and then the centre coordinates of features for each music genre (except alpha wave music) can be established. Having these centre coordinates, the alpha wave music can be examined by Eqs. (1) and (2) one by one to find the nearest centre coordinate which may be the most likely similar music genre. The numbers of the highest occurrences (n) for the centre coordinate of a music genre examined are varied from 2 to 7. We used the center coordinate value corresponding to each music feature as weight (w_i) in Eq. (2), so that it with high frequency has a relatively high weight value to strengthen it in calculating the distance. The experimental results for analysis of notes are shown in Fig. 2. The results show that there are 63%–72% of alpha wave music similar to classical music.

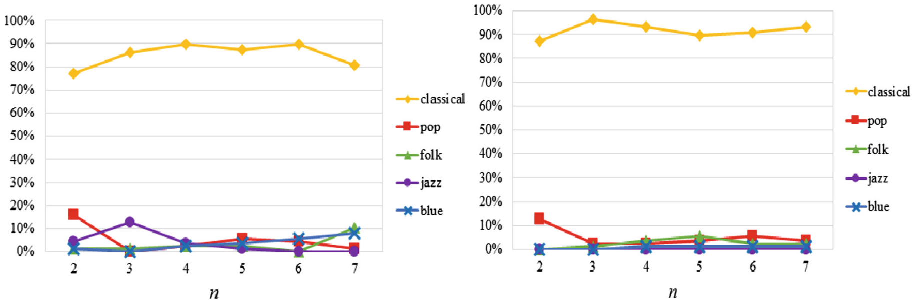


(a) by equation (1)

(b) by equation (2)

Fig. 2. Note analysis by distance equations

Furthermore, we used LIBSVM to analyze which genre the alpha wave music is most likely close to. The top 2 to 7 of highest occurrence notes are used for training to establish classification model for each music genre. Then, we can use these classification models to exam alpha wave music. The experimental results are shown in Fig. 3(a). This result demonstrates that there are 77%–90% of alpha wave music similar to classical music. In addition, the distances of each music to the center coordinate of belonged genre is computed by Eq. (2) also used to train for building classification models. The alpha wave music is also investigated in this classification model. This experimental result is shown in Fig. 3(b) and it demonstrates that there are 87%–97% of alpha wave music being similar to classical music.



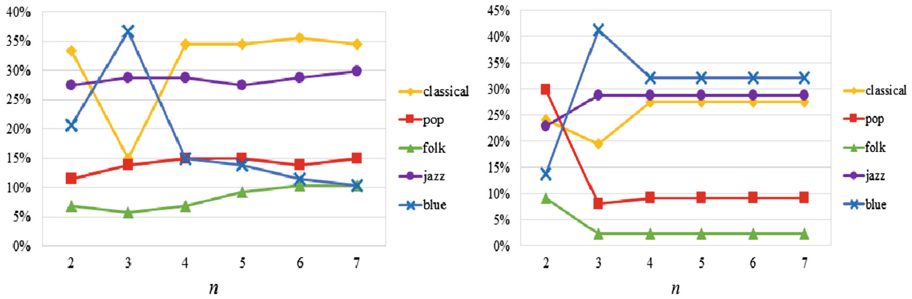
(a) note occurrence for training

(b) distance for training

Fig. 3. Note analysis by LIBSVM

4.2.2 Analysis of Rhythms

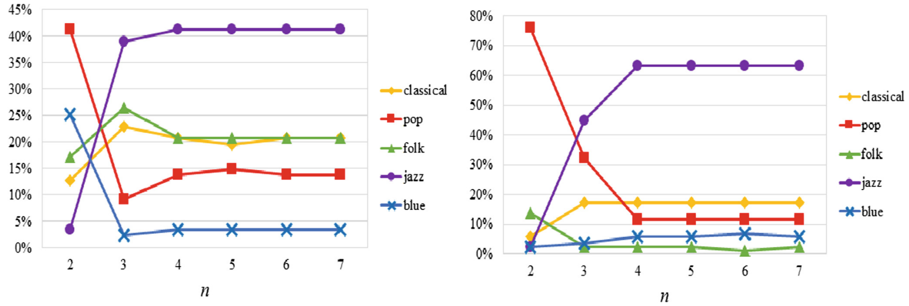
This experiment is the same as analysis of notes except that rhythms is instead of notes. The experimental results are shown in Figs. 4 and 5. The result is worse than analysis of notes. The alpha wave music is not quite close to a certain music genre.



(a) by equation (1)

(b) by equation (2)

Fig. 4. Rhythm analysis by distance equations



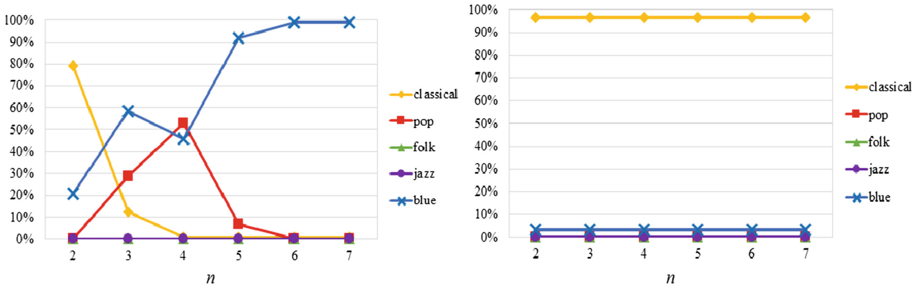
(a) rhythm occurrence for training

(b) distance for training

Fig. 5. Rhythm analysis by LIBSVM

4.2.3 Analysis of Pitch Changes

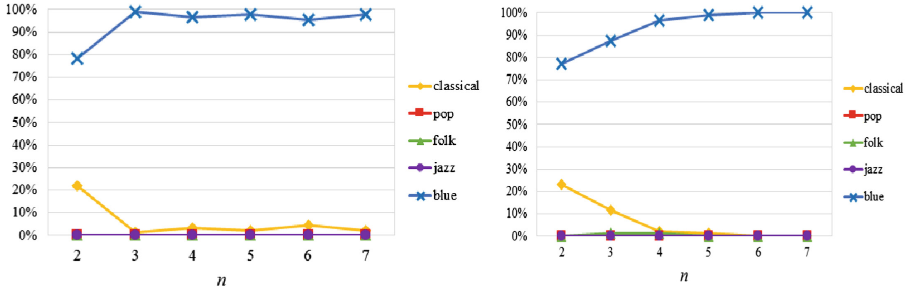
This experiment is the same as analysis of notes except that pitch changes is instead of notes. The experimental results are shown in Figs. 6 and 7. There are up to 96% of alpha wave music similar to classical in Fig. 6(b) and up to 100% of alpha wave music similar to blue in Fig. 7(a) and (b). The experimental results of Figs. 6 and 7 are inconsistent which needs more further studies.



(a) by equation (1)

(b) by equation (2)

Fig. 6. Pitch change analysis by distance equations



(a) pitch change occurrence for training

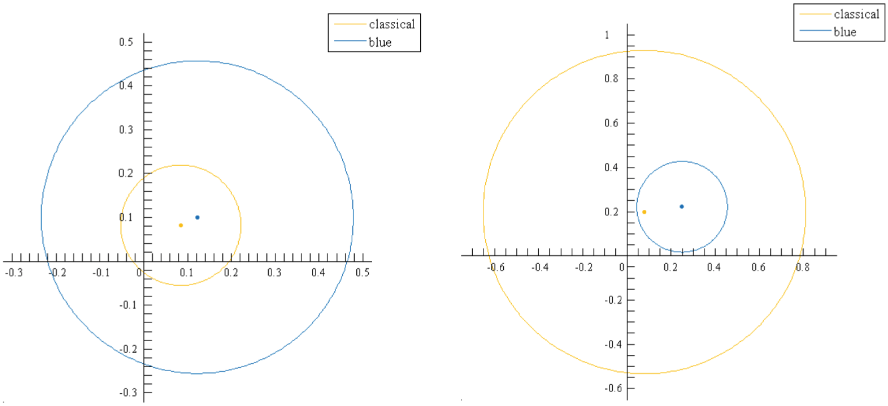
(b) distance for training

Fig. 7. Pitch change analysis by LIBSVM

4.3 Further Analysis for Classical and Blue

From the previous experimental results can be found that the rhythm of the alpha wave music is not biased towards a specific genre. However, the notes of alpha wave music is closer to classical music genre, as well as the pitch changes of alpha wave music is closer to the blue music genre. In this section we further analyze which music to recommend in classical and blue, and such music may be able to achieve the effect of alpha wave music.

Since the notes of alpha wave music are closer to classical music, we use the top two highest occurrence frequencies of classical notes for the two-dimensional center coordinates. In addition, we also use the occurrences of the same two notes in blue music as a two-dimensional center coordinates. Then, at each center coordinate, draw a circle for classical and blue music genres in which each covers 90% of belonged music in the database, as shown in Fig. 8(a). The main purpose of taking only cover 90% for drawing music circle is to exclude some music with special or exceptional features in their belonged genres. It can avoid the radius of circle being too large and becoming a sparse circle. We also used pitch change instead of notes for classical and blue to draw circles again, as shown in Fig. 8(b).



(a) by notes (b) by pitch changes

Fig. 8. Circles for classical and blue music

In Fig. 8(a), we find that the circle of classical music is included in the circle of blue music. That means the blue music which falls in the domain of classical music circle has common features in both music genres. Such blue music with common features may be closer to the alpha wave music and are worth to recommend for users. On the contrary, the circle of blue music is included in the circle of classical music in Fig. 8(b). There is also some blue music with common features similar to alpha wave music and are worth to recommend.

We only used two-dimensional space ($n = 2$) in this study. However, we proposed this approach which can be deduced to the higher dimension analyses ($n > 2$) in order to obtain the music closer alpha wave music to be recommended.

5 Conclusion

When people take a short rest with closed eyes, an alpha wave appears with brain signals. There were many medical reports proofed that some specific music can resonate with the alpha wave and strengthen the wave to improve more relaxing. Although there are many existing schemes for music classification, to categorize alpha wave music has not succeeded yet. Till now, the alpha music is classified manually by expertise only and rarely to be found in the market. In this research, we explored the contents of classical, pop, jazz, folk, blue, and alpha wave music by distance equations and learning machine approaches. We found that the notes of alpha wave music are closest to classical music as well as the pitch changes of alpha wave music are closest to blue music. Our further studies discovered that some music is similar to alpha wave music containing common features of classical and blue music. We would like to recommend such music to people for relaxing.

References

1. Basar E (1980) EEG brain dynamics. Elsevier Science, Amsterdam
2. Basar E (1988) Dynamics of sensory and cognitive processing by the brain. Springer, Berlin
3. Berger H (1969) On the electroencephalogram of man (Electroencephalography and clinical neurophysiology supplement No. 28). In: Gloor P (ed) Elsevier Science Ltd. ISBN-10: 0444407391
4. Brecheisen S, Kriegel H-P, Kunath P, Pryakhin A (2006) Hierarchical genre classification for large music collections. In: IEEE 7th international conference on multimedia and expo, pp 1385–1388
5. Cheng HT, Yang YH, Lin YC, Chen HH (2008) Automatic chord recognition for music classification and retrieval. In: IEEE international conference on multimedia and expo, pp 1505–1508
6. Deepa PL, Suresh K (2011) An optimized feature set for music genre classification based on support vector machine. In: Proceedings of IEEE conference on recent advances in intelligent computational systems (RAICS), pp 610–614
7. Goodman KD (2011) Music therapy education and training: from theory to practice. Charles C. Thomas, Springfield. ISBN 0-398-08609-5
8. Lin C-R, Liu N-H, Wu Y-H, Chen ALP (2004) Music classification using significant repeating patterns. Lecture notes in computer science, vol 2973. Springer, Heidelberg, pp 506–518
9. Lo YL, Lin YC (2012) Content-based multi-feature music classification. In: International conference on innovation and management, Republic, Palau
10. Lo YL, Lai Z-Y (2014) Content-based classification of alpha wave music. In: 2014 international conference on business and information (BAI 2014)
11. Loh QJB, Emmanuel S (2006) ELM the classification of music genres. In: 9th international conference on control, automation, robotics and vision, pp 1–6
12. Mandel M, Ellis DPW (2008) Multiple-instance learning for music information retrieval. In: 9th international conference on music information retrieval, pp 577–582
13. Myint EEP, Pwint M (2010) An approach for multi-label music mood classification. In: 2nd international conference on signal processing systems (ICSPS), pp 290–294
14. Rayleigh JWS, Lindsay RB (1945) The theory of sound. Courier Corporation, New York
15. Zhen C, Xu J (2010) Multi-modal music genre classification approach. In: 3rd IEEE international conference on computer science and information technology (ICCSIT), pp 398–402
16. Gamboa H (2005) A wave. Wikipedia. http://en.wikipedia.org/wiki/Alpha_wave
17. http://www.csie.ntu.edu.tw/~cjlin/LIBSVM/
18. http://www.mathworks.com/

A Novel Vibration-Based Real-Time Monitoring System for Illegal Logging

Jiun-Jian Liaw, Chun-Cheng Peng(✉), Yu-Yan Chen, and Wen-Chung Tsai

Department of Information and Communication Engineering,
Chaoyang University of Technology, Taichung, Taiwan
{jjliaw,goudapeng,s10330621,azongtsai}@cyut.edu.tw

Abstract. In this paper, aiming to tackle monitoring issues of illegal loggineevents, such as time latency, system maintain and evidence providing, a novel vibration based real-time system is proposed. With the developed mechanism of power-consumption reduction, the resulted wireless sensor network consists of gravity-, audio- and visual- sensing nodes in order to effectively and efficiently recognize and notify illegal logging events. Preliminary experimental results positively indicate that the proposed monitoring system is with strong potentials of tiny time latency of event notification, light effort of system maintain and great prevention of illegal logging events.

Keywords: Illegal logging · Monitor system · Power-consumption reduction · Raspberry Pi · Real-time data transmission · Wireless sensor networks · ZigBee

1 Introduction

As it might be costing nations tens of billions of dollars each year [1], illegal logging (IL) has become an extremely important issue for carbon dioxide emissions globally [2–5]. Since the noises of chain saws and caused vibrations can be easily sensed day and night while timber pirates are working, an audio and vibration based IL monitoring system is proposed. In addition, in order to notify the law enforcement agencies such IL events effectively and timely, accompanied with physical evidence, one of the feasible ways is to construct sensors within forests.

Benefitted from rapid development of integrated circuits, volumes of wireless communication modules, induction components and the corresponding data processing units are able to be reduced and combined into individual sensor nodes. As the result, wireless sensor networks (WSN) are easily implemented [6, 7]. With the capabilities of automatic recognition and information granting, the desired environmental monitoring is then realized. For solving the problem that the transmission mechanism is difficult to be revised, software-defined network (SDN) is applied, in order to attain the greater efficacious network with the adventures of better efficiency, flexibility and lower cost [8]. More precisely, the integration of WSN and SDN can effectively complete the data transmitting task for implementation of environmental monitoring systems [9].

The rest of this paper is organized as following. After the whole big picture of the proposed mechanism is reviewed in Sect. 2, the scheduling design of the conducted

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_64

sensors is proposed in Sect. 3. With the implementation of specific sensors is discussed in Sect. 4, the proposed scheme of energy saving is presented in Sect. 5, while conclusions and future works are drawn in Sect. 6.

2 Proposed Monitoring Mechanism for Illegal Logging

By means of audio and vibration recognitions, in-progress IL events can be discovered immediately, with the proposed integration of SDN and WSN. Considering physical difficulty of system maintain and battery changing, power consumption of the conducted sensor nodes, i.e., audio sensor nodes (ASNs) and gravity sensor nodes (GSNs), should be efficiently and effectively reduced. Therefore, the two proposed distinct work scheduling approaches for ASNs and GSNs are predetermined-time and pin-controlled schemes, respectively.

Figure 1 illustrates a layout example within a forest. In the proposed mechanism, the whole sensor network consists of GSNs, ASNs, video sensor nodes (VSNs) and Fog-computing node with Lightweight SDN controllers (FLCs), while each blue box indicates the monitoring area for a certain FLC. Via the proposed scheduling schemes, a GSN can be waken to analyze the specific vibration and return the conducted results to the FLC, while after triggered by specific noises, ASN shall send information back to the corresponding FLC as well.

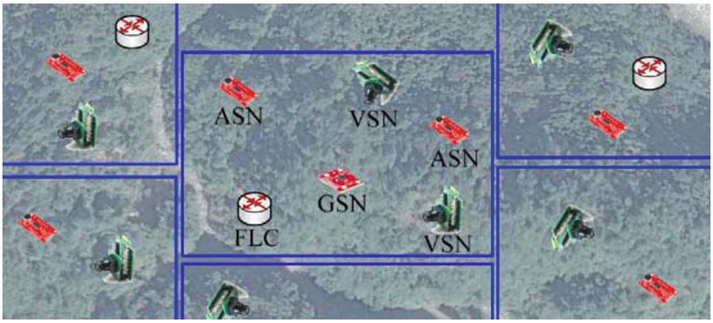


Fig. 1. A layout example of the proposed monitoring mechanism

More precisely speaking, since the working IL noises and vibrations are quite conspicuous, the proposed detection processes for GSNs and ASNs are summarized as in Figs. 2 and 3, respectively. Since FLC is responsible of information transmission and relative control, once the notified signals from GSN and/or ASN are confirmed, commands of photo shooting are sent to the specific VSN, as exhibited in Fig. 4, in order to complete the monitoring task of both the location detection and evidence preservation.

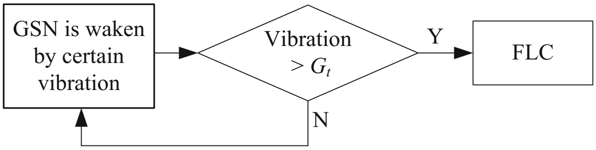


Fig. 2. Working process of GSNs

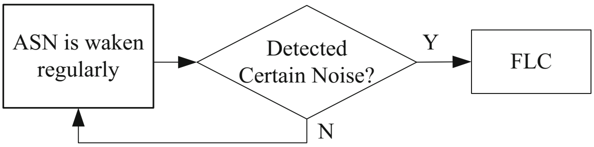


Fig. 3. Working process of ASNs

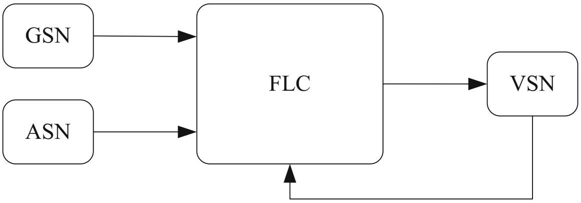


Fig. 4. Working process of FLCs

3 Proposed Dormancy Scheduling Approaches for Sensors

In order to effectively reduce power consumptions of sensors and increase system durability of the proposed monitoring mechanism, this paper presents two dormancy scheduling approaches for GSNs and ASNs, as illustrated in Figs. 5 and 6, where *t* stands for time instance.

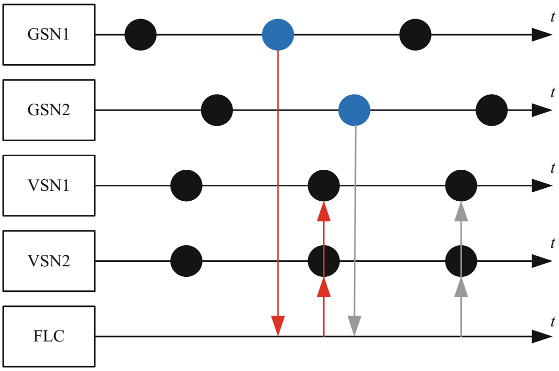


Fig. 5. A vibration-sensed example of signal transmission

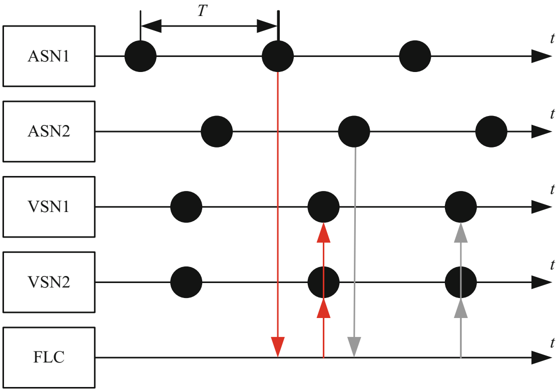


Fig. 6. Anaudio-sensedexample of signal transmission

When IL events are happening, the certain vibration signals can be sensed by MPU6050 and awake the whole GSN, under the working process described in Fig. 2. If the vibration amount is larger than the given detection threshold (i.e., blue circles in Fig. 5), an alarm signal will be sent to FLC (i.e., red and grey arrows from GSN1 and GSN2 to FLC), the specific periodical-activated VSNs will take photos while receives the granted command from the FLC (i.e., red and grey arrows from FLC to VSN1 and VSN2).

As exhibited in Fig. 6, each dots on *t* time axis for both ASN1 and ASN2 indicate that, for the sake of energy saving, the proposed ASNs are activated every *T* time instances. If timber sawing noises are detected and recognized, alarm signals are transmitted to FLC right away (i.e. red and grey arrows from ASN1 and ASN2 to FLC). The specific VSNs will take photos as illustrated in Fig. 5.

4 Implementation of GSNs and FLCs

This study applies the Atmega328P single chip to drive other components and programming, as presented in Fig. 7, the oscillator helps the chosen single chip working properly and controlling vibration sensing device. Through the two control pins, i.e., SDA and SCL, to communicate with the three-axis accelerometer, i.e., MPU6050, the sensed vibration is conveyed to Atmega328P. If the vibration amount is greater than the specific threshold, a notification signal will be sent to the ZigBee part in order to transmit the vibration to Raspberry Pi 3 for further analysis. The resulted PCB diagram is produced by Altium Designer software and demonstrated as in Fig. 8, where blocks U1, U2 and U3 are respectively MPU6050, ZigBee and Atmega328P.

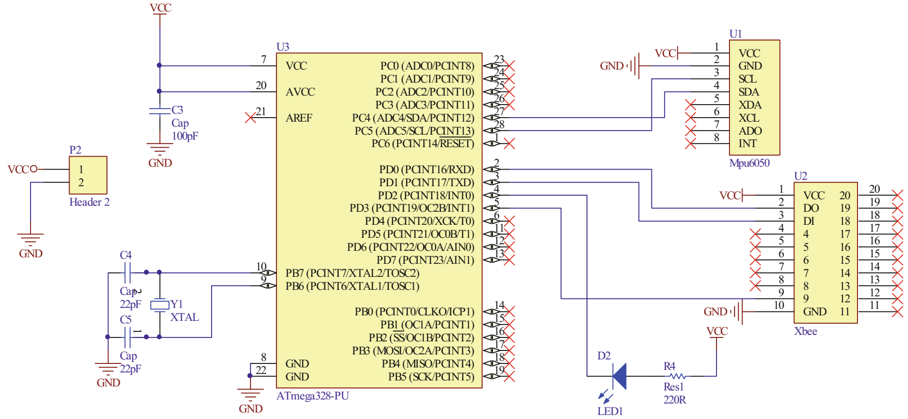


Fig. 7. Circuit Diagram of the proposed GSNs

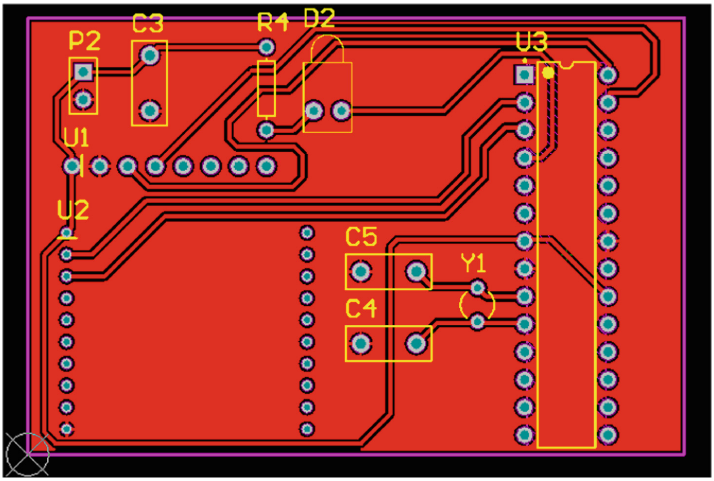


Fig. 8. PCB diagram of the proposed GSN

Being the core controller of the proposed sensor network and implemented on Raspberry Pi 3, the FLC is responsible of message transmission and the corresponding actions, between GSNs, ASNs and VSNs. Due to the reason that Raspberry Pi3 is with build-in UART interface and GPIO pin, communications with ZigBee can be easily achieved [10, 11], while the convenient environment of programming and fast processing speed, related signals are able to transform from time domain to frequency domain by fast Fourier transform, in order to analyze received GSNs' and ASNs' information.

5 Discussion of Power-Consumption Reductions

Current measurements are conducted serially, as summarized in Table 1, under the three working conditions, i.e., not transmitting data, transmitting and sleeping. According to Eq. (1), the power consumption of each device can be easily calculated,

$$Q = It, \tag{1}$$

where Q, I, t represent electric charge, current and electrified time.

Table 1. Working currents of related devices (in mA)

Device	Not transmitting	Transmitting	Sleeping
LED	3.5	3.5	–
MPU6050	9.4	9.4	–
Atmega328P	6.3	6.7	–
ZigBee	32.5	32.8	0.0233
GSN	52.9	53.4	20.4

According to the above measured results, it can be easily observed that the ZigBee device consumes the most electricity power. As the result, it is believed that applying sleep modes can effectively reduce the power consumption amounts. Table 2 reviews the three situations of sleep modes to control the ZigBee device [12], where SM = 2 and SM = 4 represent the GSN and ASN operational modes, respectively. It can be found that the difference between SM = 1 and SM = 2 is the wake-up time. After the sleep-mode settings mentioned in Table 2 are completed, the corresponding power consumption amounts under different operational voltages are granted, as presented in Table 3.

Table 2. Three sleep modes of the ZigBee device

Sleep mode (SM) setting	Characteristics	Typical power-down current (μA)	Wake-up time (ms)
Pin Hibernate (SM = 1)	Pin/Host-controlled	<10	13.2
Pin Doze (SM = 2)	Pin/Host-controlled	<50	2
Cyclic Sleep (SM = 4)	Pre-determined time (Cyclic: 1 - 0xFFFF)	<50	2

Table 3. Power consumptionof the ZigBee device(inμA)

Vcc(V)	Pin hibernate (SM = 1)	Pin doze (SM = 2)	Cyclic sleep (SM = 4)
3.0	3	35	34
3.1	8	37	36
3.2	32	48	49
3.3	101	83	100
3.4	255	170	240

In order to verify the above simulated results, the real time clock (RTC) [13] is applied to calculate the total time of data transmission, as described in Fig. 9. Experimental results indicates that, calculations of both the RTC's and Eq. (1)'s are identical.

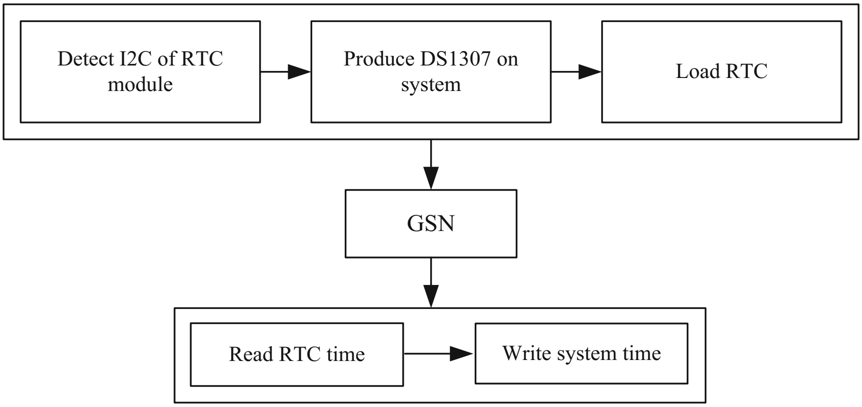


Fig. 9. Time duration calculation for applying RTC

In addition, simulations of using two different-type batteries (i.e. two C-size alkaline and one lithium) are conducted to analyze their time durations under the two working conditions (i.e. normal mode and ZigBee-sleep). The alkaline battery provides 3v and 3000mAH, while sustained working time with normal and sleeping modes are respectively 56.2 and 147.1 h. For the lithium battery with 3.7 V and 7200mAH, the certain working time durations are 134.2 and 347.4 h, respectively. In other words, the sleeping-mode design for ZigBee devices can raise their working durations about 3 times better than the ones in just normal mode. As the result, reductions of power consumption of the proposed mechanism are extremely reasonable.

6 Conclusions and Future Works

This paper proposes a real-time audio-and-vibration based IL monitoring system. With integration and construction of wireless communication module, sensors, data processing units, the resulted WSNs can be effectively employed within forests and mountain areas. In addition, the proposed designs of vibration- and periodical-trigger for GSNs and ASNs improve efficiently the battery duration. Simulated experiments via RTC shows that power consumptions of the proposed dormancy mechanism are three times smaller than the ones of normal modes. As the result, the proposed real-time IL monitoring system is proved to have its generalization and feasibility.

In the coming future, besides hardware implementation of ASNs and improving the recognition method for sounds of chain saws with consideration of relatively similar nature sounds within forests, integration of GSNs, ASNs, VSNs and FLCs and in-field experiments within forests are necessary to deploy and conduct, in order to further enhance system accuracy and verify robustness of the proposed IL monitoring mechanism.

Acknowledgment. This study was supported by the Ministry of Science and Technology, Taiwan (No. MOST 105-2221-E-324-008-MY2).

References

1. Lynch J, Maslin M, Balzter H, Sweeting M (2013) Choose satellites to monitor deforestation: illegal logging threatens tropical forests and carbon stocks. Governments must work together to build an early warning system. Nature 496(7445):293–295
2. Dudley RG (2004) A system dynamics examination of the willingness of villagers to engage in illegal logging. J Sustain Forest 19:31–53
3. Lewis JD (2007) Enabling forest people to map their resources & monitor illegal logging in Cameroon. Before Farming Archaeol Anthropol Hunter-Gatherers 2(3):1–7
4. Xu JH (2014) Review of China's forest CoC certification system against its illegal logging and trade. BSc Essay, Forest Resources Management, U. British Columbia
5. Wijaya A (2005) Application of multi-Stage classification to detect illegal logging with the use of multi-source data. Master Thesis. Int'l Inst. Geo-information Science and Earth Observation
6. Liaw J-J, Chou C-W, Dai C-Y (2013) The lifetime extension of wireless sensor networks using adaptive energy allocation by distance. Int J Distrib Sens Netw 9(8):982573
7. Kaur J, Grewal R, Saini KS (2015) A survey on recent congestion control schemes in wireless sensor network. In: Proceedings of IEEE international advance computing conference, pp 387–392
8. Chowdhury SR, Bari MF, Ahmed R, Boutaba R (2014) Payless: a low cost network monitoring framework for software defined networks. In: Proceedings of IEEE network operations and management symposium, pp 1–9, May 2014
9. Gante AD, Aslan M, Matrawy A (2014) Smart wireless sensor network management based on software-defined networking. In: Proceedings of 27th biennial symposium on communications, pp 71–75
10. Raspberry Pi Foundation. https://www.raspberrypi.org/products/raspberry-pi-3-model-b/. Accessed 13 Feb 2017
11. Raspberry Pi Foundation. https://www.raspberrypi.org/magpi/raspberry-pi-3-specs-benchmarks. Accessed 13 Feb 2017
12. Digi International Inc., (2009) Product Manual of XBee RF Modules. https://www.sparkfun.com/datasheets/Wireless/Zigbee/XBee-Datasheet.pdf
13. Maxim Integrated Inc., datasheet of DS130764*8, serial, I^2C real-time clock. https://datasheets.maximintegrated.com/en/ds/DS1307.pdf

The Development of Skin Image Analysis Device Based on Embedded System

Chuan-Pin Lu(✉), Yu-Wen Liu, Zi-Qing Fang, Zi-Yu Chen,
Tzu-Ching Wu, and Yun-Jie Zhang

Department of Information Technology, Meiho University,
23, Pingguang Rd., Neipu, Pingtung, Taiwan, R.O.C.
chuan.pin.lu@gmail.com, prinsen520@gmail.com,
qoo831221@gmail.com, dsiney105@gmail.com,
duck860122@gmail.com, a0976368301@gmail.com

Abstract. Skin appearance is a key determinant of perceived age. Medical studies have verified the importance of skin care, stating that functional anti-aging products can defer skin aging. However, the effects of these products are largely self-perceived by users. These self-perceptions are difficult to quantify. The most common equipment for measuring skin quality is the dielectric analyzer, while fine lines and luster are largely measured visually with a color chart. However, the data produced by dielectric analyzers are not directly associated with skin appearance, and visual measurements contain considerable bias. Therefore, in this paper, a digital image processing method was developed for skin analysis; wherein an embedded system was used to develop a device for measuring luster and fine lines. Users can use this device to extract, analyze and compute skin images, and compare the differences in skin quality before and after using skin care products. The skin area affected by high ankle strains was selected as the target of analysis. The research outcomes are presented in the "Experiment" chapter.

1 Introduction

With the perpetual advancement in aesthetic medicine technology, aesthetic medicine has become the third largest industry in the world, second only to the aviation and automobile industries. A report published by Medical Insight indicated that the global aesthetic medicine industry achieved an average annual compound growth rate of 10.9% in the recent decade and that the growth rate in Asia exceeded the global average growth and achieved between 13% and 15%. Due to cultural influences, the demand for anti-aging products is extremely high in Asia. Perceived age is highly valued in Asia. Despite the inability to stop aging, functional anti-aging skin care products can be used to defer skin aging and keep healthy. After all, the soft and smooth skin that has luster and elasticity is a sign of good health. Skin can be characterized into three layers [1]: epidermis, dermis, and hypodermis. The stratum corneum is the outermost layer of the epidermis and serves as the body's first line of defense. Corneocytes are non-living cells that stack on top of one another to form the stratum corneum. Aging causes corneocytes

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_65

to build up in the stratum corneum, increasing dryness, shedding, and fine lines and expediting wrinkles, dullness, coarseness, and stiffness. Medical studies [2–4] have verified that retinoic acids, alpha hydroxy acids, antioxidants (vitamin C, vitamin E, catechin, and curcumin) and stem cell products can protect the skin from aging.

Currently, many commercial anti-aging, anti-wrinkle, and moisturizing products are available on the market. However, users and aesthetic medical professionals have always been skeptical about whether these products are effective, leading to the rise in demand for skin analyzers. Currently, the market does not offer a consumer skin analyzer centered on measuring wrinkles, luster, whitening, and coarseness. Take the luster measurement for example, complexion is currently measured visually with the help of color charts. This method is extremely inaccurate, which is caused by several factors: inconsistent ambient lighting causes discoloration; color charts are only good rough estimates and outcomes cannot be quantified; previous observations cannot be clearly recorded; and people perceive color differently. In addition, conventional skin analyzers use dielectric constants as the measure for determining skin properties. However, water content in the stratum corneum influences conductivity and, by extension, dielectric constants. These analyzers are unable to show skin appearance results directly, and only present relative outcomes to reflect skin appearance.

Based on the above-mentioned reasons, this paper developed a method in which users of skin care products can effectively and scientifically track the improvement of their skin. To achieve this objective, a skin analyzer for measuring luster and fine lines was developed using a system-on-chip ARM architecture coupled with computer vision technology. The device offers image processing, feature capturing, imaging storage, and color quantization and comparison functionalities. In terms of methodology, a consistent optical imaging device was used to capture skin images. The images were then processed using an image processing algorithm run on the ARM processor, embedded Linux operating system, and Node.js [7] server. The skin images were processed using a self-developed algorithm coupled with a color quantization algorithm, geometric features, and fine lines measurements to obtain quantified skin image data. The data were then analyzed to determine the level of improvement. In the software and hardware architecture, the proposed device featured a web interface to enable users to conveniently operate it. Users need only access their web browser to connect to and operate the device, making it extremely convenient and cross-platform compatible while overcoming hardware limitations. The hardware configuration and image processing algorithm of the proposed device are discussed in the “Methodology” chapter. The performance of the skin image quantization algorithm is demonstrated in the “Experiment” chapter.

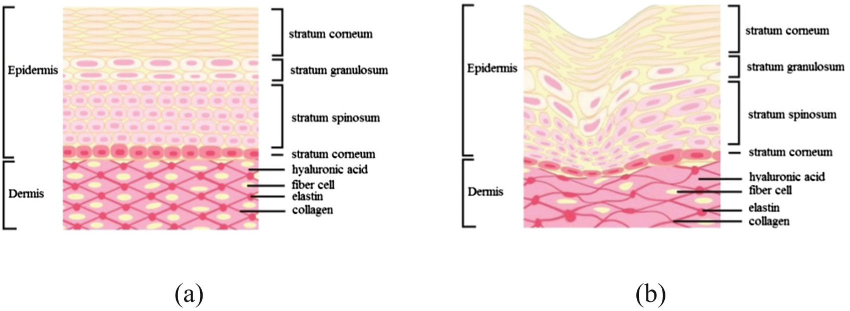
2 Skin Structure

The skin [1, 5] is the largest organ on the human body. It is connected to the digestive, respiratory, and urinary systems to provide protection, prevent dehydration, regulate temperature, and absorb nutrients. The skin consists of the epidermis, dermis, and hypodermis, and each skin layer contains hair follicles, arrector pili muscles, nails, sebaceous glands, and sweat glands (see Fig. 1(a)). The skin system comprises skin, sweat glands, sebaceous glands, hairs, and nails.

Skin ages [5] intrinsically and extrinsically. Intrinsic aging refers to aging over time. The proliferation of skin fibroblasts decelerates, leading to a decline in the collagen fibrin, elastic fibrin, and hyaluronic acid content in the dermis. Slowly, the skin loses elasticity and forms static fine lines (fine lines that are observable without facial movement). Intrinsic aging is inevitable. Extrinsic aging is associated with life habits, such as staying up late, smoking, excessive drinking, exposure to sunlight, and prolonged exposure to dry air, leading to the formation of wrinkles. Among these habits, exposure to sunlight is the root cause of skin aging, and they are unrelated to age. Wrinkles are also formed from the excessive use of facial muscles. Depending on wrinkle depth, wrinkles can be categorized into fine lines, coarse lines, and grooves. Fine lines mainly appear on the epidermis, while dynamic wrinkles and grooves extend to the dermis. Skin aging deforms the epidermis, changes the size and shape of cells, expedites cell degeneration, and reduces melanin and Langerhans cells (see Fig. 1(b)). In the dermis, skin aging causes the loss of stromal cells and atrophy; reduces fibroblasts, mast cells, and blood vessels; and induces nerve ending degradation. In the skin appendages, it lightens hair color, induces hair and glandular loss, and reduces sebum secretion leading to dry skin. The functional influences of skin aging include stunted immune functions, dry skin, poor temperature regulations, and paresthesia. Using topical skin care products such as aloe-vera, AHA, Q10, and vitamin-A-acid can improve aging skin [6]. Moisturizing skin care products activate the stratum corneum. A-acid can improve skin coarseness, reduce fine lines and pigmentation, and mitigate skin cancer.

3 Methodology

The hardware composition of the proposed skin analyzer primarily comprises an ARM chip that serves as the core processor and an embedded Linux operating system that runs the Node.js web server. The overall software architecture is written in HTML5 and JavaScript was adopted to run the image processing algorithm. The procedures of the algorithm included image preprocessing, color quantization, fine lines detection, and appearance improvement computations. The objectives of the image processing algorithm were to detect luster and fine lines. Previous skin observations revealed that after using moisturizer, the moisture content of the stratum corneum increased, wrinkle depth reduced, appearance became smoother, and luster increased. Therefore, increased luster and reduced fine lines were selected as the criteria for skin improvement. Image preprocessing was required to remove skin textures, fine hairs, and moles from the images, and to retain the remaining portion of images for skin appearance analysis. Texture was used as an alternative feature for evaluating skin quality. Color analysis was performed to identify the complexion of the epidermis. Pixels with different color values are dispersed throughout the skin images. Therefore, a color quantization process [8] is required to calculate the principle complexion value. Once the representative complexion value is obtained, the Euclidean distance between the before and after images to obtain suitable appearances. To identify the optimal color quantization method, three algorithms were compared in this paper, specifically the Median-cut [9], K-means [10], and Self-organizing Maps (SOM) [11]. Finally, the time performance of the complexion analysis was tested. The skin image processing procedures can be broadly characterized into two stages, namely, preprocessing and feature capturing. The preprocessing stage included image capturing, gray-scale conversion, Sobel edge detection, image enlargement, and texture removal. The feature capturing stage included color quantization, color feature calculation, and feature documentation. Finally, the level of improvement was calculated.

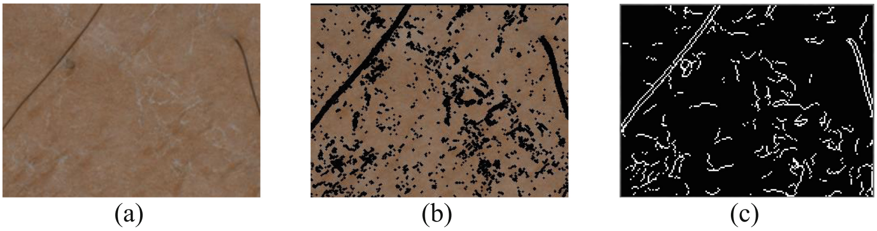


(a) (b)

Fig. 1. Skin structure: (a) normal skin;(b) aging skin.

3.1 Skin Image Preprocessing

To prevent the system from incorporating skin texture, fine hairs, and moles into color evaluation, image preprocessing was performed to remove unnecessary noise. In this paper, noise is defined as the texture of the skin. Skin is essentially an uneven surface with considerable texture. Such texture is a blend of moderate and thin geometric lines. Additionally, arm hair and melanin cells (commonly known as moles) are also beyond the scope of color quantization. The Sobel algorithm [8] was used for edge detection during preprocessing. The edges that were detected reflected the texture, fine hairs, and moles on the skin image. The Sobel algorithm can highlight the high-frequency data in images. Therefore, it is ideal for detecting texture pixels in the images. In this paper, to enhance edge detection, the range of edges was expanded by using morphological



(a) (b) (c)

Fig. 2. Skin lines removal: (a) original skin image; (b) skin image after removing skin lines; (c) skin fine lines image.

expansion [8], the obtained edges were then removed, and the remaining portion underwent color quantization (as shown in Fig. 2). Figure 2(a) is original skin image, and Fig. 2(b) is the skin image after removing skin lines.

3.2 Color Quantization Algorithms

Color is distributed throughout an image in the form of pixels. To identify the primary color of the skin, a color quantization algorithm is required to calculate principal color. This is achieved by determining the color with the most pixels (largest area) in the image. This color is selected as the representative color and color feature. Color quantization algorithms can be broadly characterized into two types, namely, clustering algorithms and splitting algorithms. Clustering algorithms can produce more useful and accurate quantization outcomes, but require more time. Moreover, researchers that apply clustering algorithms also take into account the problems of convergence and computation speed. Splitting algorithms are the opposite. They have faster computation but produce less favorable quantization outcomes than their clustering counterparts. To identify the ideal method, we compared the quantization performance of three algorithms: median-cut algorithm, K-means clustering algorithm, and SOM algorithm.

3.3 Median-Cut Algorithm

Median-cut algorithm was proposed by Heckbert in 1982 [9] and is a type of non-average color quantization method. This method uses analysis techniques to segment the color space repeatedly until a similar color is reached. The surface is segmented vertically with a single color axis, and the element with the largest pixel difference range out of the three color axis will be used as a standard for segmenting. Sorting is used and then the middle gray scale value is used as the segmenting point.

3.4 K-Means Clustering Algorithm

K-means is a clustering algorithm proposed by MacQueen in 1967. The amount of clusters must be defined before using the algorithm. Afterwards the center point of the cluster, the sum of the distance between the vector points and the extreme values need to be found to achieve the goal of optimal clustering. In 1995, Verevka applied this method to color quantization [10].

3.5 Self-organizing Maps Algorithm (SOM)

In 1994, Dekker proposed a color quantization method using SOM [11]. It has relatively quick calculation time, and is able to raise sample counts and significantly improve quantization quality. This method mainly employs a one dimensional self-organizing map where the network contains every cluster's neurons. Through a self-learning process, every neuron obtains a weight vector which has representative value. After self-learning, the pixels are also reflected in the nearest weight vector. The algorithm runs in two stages, specifically, the comparison stage and the learning stage. In the comparison stage, the distance between the selected weighted vectors and pixel groups can be determined.

3.6 Representative Color, and Root Mean Square Error

Due to the fact that after color quantization, the coloring is not only of one single color, in order to find the main color, this method takes the color with the most pixels to be the representative color for the examined section. This method complies with the method recognized by testing personnel. The root mean square error (RMSE) [8] is used to evaluate the color quantization outcomes. This is also to compare the outcomes of same units. The similarity between the quantized image and original image increases as the difference between the RMSE values of the two images decreases.

3.7 Skin Fine Lines Detection

A novel edge drawing algorithm (ED) is needed in fine line detection for skin image. Consequently, the edge drawing algorithm proposed by Topal et al. [12] was adopted. The edge detection method was applied to execute this procedure. A commonly adopted edge detection methods are the Sobel, Canny, Prewitt, Roberts, Laplacian etc. This method was adapted from a traditional edge detection algorithm, incorporating direction maps, gradient maps, and edge maps to identify anchor points. Finally, anchor points were connected to form a boundary line. Because the width of the ED boundary line was one pixel, which was insufficient as a cutting line, the corners of the boundary line were enhanced to obtain the skin fine lines (see Fig. 2(c)).

4 Experiments

This section entails the detection and analysis of the skin images. The skin images of ten participants aged between 16 and 53 (male and female) were selected as the experiment samples. These images were the skin area affected by high ankle strains, where there is less sebum secretion and the common over-drying leads to skin coarseness phenomenon, thus it is easy to observe the skin changes before and after using skin care products. All tests were performed using the same imaging equipment (white LED lighting, image spatial resolution is 0.01367 mm/pixel) and moisturizer (NIVEA extra white firming body lotion). The moist, oil, and soft outcomes of the high ankle strains area produced by the skin analyzer (MyB ST-100) were compared (Table 1). Three tests were designed. The first test was an RMSE evaluation using a complexion analysis with three color quantization algorithms. Six quantization colors were defined. The second test was a comparison and evaluation of luster improvement outcomes. The third test was a comparison and evaluation of fine lines improvement outcomes. Table 1 is shown the measurement results of moist, oil and skin soft by using ST-100 device. In addition, time was also evaluated in the second and third tests. All tests were conducted on a PC.

The PC is equipped with Intel(R) Core (TM)2 Quad CPU 2.40 GHz and Win7 OS. The front-end calculation refers to using the Chrome browser.

Table 1. Measurement results of moist by ST-100 device (moist/oil/soft).

No	Tester age	Sample A	Sample B	Moist improvement level
1	16 years male	Moist 33% (1/3/1)	Moist 34% (2/3/2)	1 level
2	21 years female	Moist Lo (1/1/1)	Moist 34% (2/3/2)	1 level
3	23 years male	No data	Moist 33% (1/2/1)	1 level
4	24 years male	No data	Moist 33% (1/2/1)	1 level
5	53 years male	No data	Moist 34% (2/3/2)	2 level
6	53 years female	No data	Moist 34% (2/1/2)	2 level
7	22 years female	Moist Lo (1/1/1)	Moist 40% (3/2/3)	2 level
8	22 years male	Moist Lo (1/1/1)	Moist 35% (3/2/3)	2 level
9	21 years female	Moist 33% (1/3/1)	Moist 35% (2/1/2)	1 level
10	22 years female	Moist 33% (1/2/1)	Moist 35% (3/1/3)	2 level

Three experiments are described as following:

- *Experiment 1: comparison of the RMSE results for three color quantization algorithms*

The skin on the dorsum of the hand of five participants was selected as the test samples in this test. The skin texture was removed before applying the three color quantization algorithms. Then, the RMSE between the quantization outcomes and the original images was calculated. The KM algorithm produced the most favorable results (achieving the smallest RMSE all ten times), followed by the Median-cut algorithm. The SOM algorithm produced the least favorable results. Therefore, the KM algorithm was selected as the color quantization algorithm in this paper. Table 2 is experimental results.

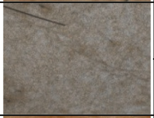
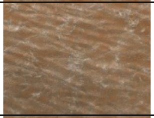
Table 2. Comparison of the RMSE results for color quantization of skin images.

Sample no	Color quantization algorithms		
	Median-cut	KM	SOM
1	RMSE: 10.56	RMSE: 9.81	RMSE: 10.86
2	RMSE: 7.64	RMSE: 6.80	RMSE: 9.06
3	RMSE: 13.01	RMSE: 12.17	RMSE: 13.15
4	RMSE: 11.00	RMSE: 10.45	RMSE: 14.52
5	RMSE: 11.74	RMSE: 10.43	RMSE: 12.66
6	RMSE: 11.59	RMSE: 11.17	RMSE: 16.92
7	RMSE: 5.90	RMSE: 5.48	RMSE: 7.49
8	RMSE: 9.40	RMSE: 9.11	RMSE: 12.5
9	RMSE: 8.58	RMSE: 7.96	RMSE: 9.72
10	RMSE: 12.56	RMSE: 11.63	RMSE: 20.34

- *Experiment 2: comparison experiment of skin gloss improvement*

According to the outcomes of the first test, the KM algorithm was applied to evaluate the level of luster improvement. The ten skin images in Table 1 were examined and compared

Table 3. Experimental results of improvement level (ms: millisecond)

NO	Sample A	Representative Color	Sample B	Representative Color	Level V (%)
		Computation Time		Computation Time	
1		RGB(142,104,74)		RGB(150,111,80)	4.8%
		7769 ms		7527ms	
2		RGB(128,106,88)		RGB(128,108,93)	1.8%
		6049 ms		6020 ms	
3		RGB(142,95,60)		RGB(149,100,62)	2.5%
		9055 ms		9937ms	
4		RGB(144,98,66)		RGB(135,109,91)	7.2%
		6498 ms		6218ms	
5		RGB(132,93,63)		RGB(147,99,62)	3.3%
		8755 ms		8231ms	
6		RGB(144,96,60)		RGB(142,97,62)	0.4%
		10489 ms		14473ms	
7		RGB(136,108,87)		RGB(141,99,67)	-6.6%
		5347 ms		5652ms	
8		RGB(132,98,72)		RGB(132,109,92)	7.0%
		5998 ms		5078ms	
9		RGB(133,100,74)		RGB(137,103,78)	2.3%
		6454 ms		5864ms	
10		RGB(155,100,58)		RGB(156,101,59)	0.4%
		7482 ms		6282ms	

to determine whether luster was recovered after using moisturizer. “Sample A” is the image of skin before using moisturizer, and “Sample B” is the image of skin after using moisturizer. The similarity between pure white RGB (255,255,255) and the complexion value before using moisturizer was adopted as the K factor for current state of skin quality (denominator), the similarity between pure white and the complexion value after using moisturizer was adopted as the D factor for current state of skin quality (numerator), and the skin improvement value (V) was adopted as the V = (1-(D/K))*100%. The overall test results are summarized in Table 3. Except Sample 7 which exhibited negative improvement, all the other samples showed positive improvement at different levels. Sample 4 exhibited the highest improvement of 7.2%. In comparison with outcomes in Table 1, before applying moisturizer, no data were measured for four samples tested (No. 3 ~ 6) with the ST-100 and three samples (No. 2, 7, 8) produced low values (no data); after applying skin care products, all samples could be measured. The levels of improvement are listed in the right column of the table. A comparison of the data in the two tables shows that improvement levels produced by the two methods were inconsistent. For the ST-100, effective data could not be retrieved from seven of the samples prior to applying the moisturizer. Moreover, improvement could only be quantized in levels rather than precise values. The outcomes produced by the proposed image detection method indicated that most of the samples exhibited improvement in luster after applying the moisturizer. Only Sample 7 showed negative improvement. This was because the imaging position was different, which caused discrepancies in the image analysis outcomes. Subsequently, the image processing time for each sample was within 15 s.

- *Experiment 3: skin fine lines improvement*

In this test, improvement in fine lines was evaluated. The ST-100 was unable to detect the improvement of fine lines. Therefore, only the proposed image detection method was used in this test. The ED algorithm was adopted for wrinkle detection. ED results were then converted to obtain the sum of the fine lines, which served as the fine lines data. The improvement level of skin fine lines is shown in Table 4, where $S_1(S_2)$ is pixel number of skin fine lines in ED line image. The computation time of skin fine lines detection is about 16 ms.

Table 4. Experimental results of skin fine lines improvement.

No	Sample A: Skin fine lines (S_1)	Sample B: Skin fine lines (S_2)	Level (%) $100*(S_1-S_2)/S_1$
1	3611	3389	6%
2	9693	5014	48%
3	8520	5966	29%
4	6890	2046	70%
5	4827	12323	−155%
6	12146	7855	35%
7	3119	8369	−168%
8	11528	2476	78%
9	2583	1058	59%
10	7946	3616	54%

5 Conclusions

This paper introduced a skin image analyzer equipped with an ARM processor to detect skin appearance. The device is coupled with a self-developed image processing algorithm to calculate luster and fine lines, enabling users to compare the difference before and after using skin care products and evaluate the effectiveness of these products. Compared to conventional dielectric skin analyzers, the proposed image detection method can better measure luster and fine lines. A number of image processing method were compared in this paper. Outcomes revealed that KM outperformed the other algorithms. It was combined with self-designed features, to remove skin texture, fine hairs, and moles on the skin images and obtain luster results accurately. Moreover, an edge detection method was employed to measure fine lines. This method produced excellent results. In this paper, we found that the accuracy of the imaging position significantly influences detection outcomes. In future, researchers can aim to improve skin imaging position and investigate the feasibility of using the proposed device in testing other skin areas.

Acknowledgment. This work was supported by the Ministry of Science and Technology, Taiwan, R.O.C., under grant MOST 105-2221-E-276-002.

References

1. Millington PF, Wilkinson R (2009) Skin, 1st edn. Cambridge University Press, New York
2. Stephens TJ, Sigler ML, Hino PD, Moigne AL, Dispensa L (2016) A randomized double-blind, placebo-controlled clinical trial evaluating an oral anti-aging skin care supplement for treating photodamaged skin. J Clin Aesthet Dermatol 9(4):25–32
3. Jiang J, Kong F, Li N, Zhang D, Yan C, Lv H (2016) Purification, structural characterization and in vitro antioxidant activity of a novel polysaccharide from Boshuzhi. Carbohydr Polym 147:365–371
4. Herndon JH Jr, Jiang LI, Kononov T, Fox T (2016) An open label clinical trial to evaluate the efficacy and tolerance of a retinol and Vitamin C facial regimen in women with mild-to-moderate hyperpigmentation and photodamaged facial skin. J Drugs Dermatol 15(4):476–482
5. Proksch E, Brandner JM, Jensen J-M (2008) The skin: an indispensable barrier. Exp Dermatol 17:1063–1072
6. Kuehne A, Hildebrand J, Soehle J, Wenck H, Terstegen L, Gallinat S, Knott A, Winnefeld M, Zamboni N (2017) An integrative metabolomics and transcriptomics study to identify metabolic alterations in aged skin of humans in vivo. BMC Genom 18(1):169
7. Farah S, Benachenhou A, Neveux G, Barataud D, Andrieu G, Fredon T (2015) Flexible and real-time remote laboratory architecture based on node.js server. In: 2015 3rd experiment international conference, pp 155–156
8. Gonzalez RC, Woods RE (2002) Digital Image Processing, 2nd edn. Prentice-Hall, Upper Saddle River
9. Heckbert PS (1982) Color image quantization for frame buffer display. Comput Graph 16:297–307

10. Verevka O (1995) Color image quantization in window system with local K-means algorithm. In: Proceedings of western computer graphics symposium, pp 74–79
11. Dekker AH (1994) Kohonen neural networks for optimal colour quantization. Netw Comput Neural Syst 5:351–367
12. Topal C, Akinlar C (2012) Edge drawing: a combined real-time edge and segment detector. J Visual Commun Image Represent 23:862–872

A QoS Channel Allocation Scheme for Enhancing the Reliability in 6TiSCH Networks

Tsung-Han Lee(✉), Lin-Huang Chang, and Yan-Wei Liu

Department of Computer Science, National Taichung University of Education,
No. 140, Minsheng Road, West District, Taichung 40306, Taiwan (R.O.C.)
{thlee,lchang}@mail.ntcu.edu.tw, BCS104103@gm.ntcu.edu.tw

Abstract. The recent IEEE 802.15.4-2015 standard proposes a number of Medium Access Control (MAC) layer protocols for low-power and lossy networks. The Timeslotted Channel Hopping (TSCH) is one of the proposed MAC operation modes, which take the advantage of channel hopping to make the transmissions more reliable and deterministic. In TSCH, slotframe is sliced into a timeslots and random channel hopping is used to mitigate the effects of co-channel interference and multipath fading. In this paper, we focus on the issue of random channel hopping in the error-prone wireless channel condition. In this research, a QoS Channel Allocation Scheme (QCAS) is proposed to enhance the reliability of emergency packet transmission in 6TiSCH networks. Based on our simulation results, we have observed that the proposed QCAS can achieve higher reliability of transmission over selected channel to provide higher throughput and higher delivery ratio than the original TSCH.

Keywords: IEEE 802.15.4-2015 · 6TiSCH · 6top · TSCH

1 Introduction

The wireless sensor technologies are widely used in industrial applications today, such as industrial automation, environmental monitoring, and condition monitoring. These applications require the larger number of low-power wireless sensor devices provides the capability to be monitored and controlled in a wide area. Many industrial wireless sensor networks [1] below to the IEEE 802.15.4-2015 standard [2] and most of them are required for low power consumption and the Internet networking capabilities. In the trend of IP-enabled industrial wireless sensor devices, a new IETF 6TiSCH [3] standard is being proposed at the IETF to enable IPv6 over the TSCH [4] mode of the IEEE802.15.4e standard [5]. 6TiSCH uses a pre-scheduled timeslotted channel hopping sequence, which repeats in time to reduce the influence of co-channel interference. Each channel of timeslot is selected from the sequence. The pair of sending and receiving devices can able to exchange information and acknowledgments during one timeslot. Therefore, if the current timeslot suffers from co-channel interference or multipath fading, the transmission failure probability will decrease in next timeslot since the channel hopping mechanism. Although, TSCH provides increased reliability using channel hopping mechanism. However, the channel suffers from co-channel interference

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_66

or multipath fading still exists in the fixed channel hopping sequence, therefore the link will susceptible to periodic interference. Therefore, we propose a QoS Channel Allocation Scheme (QCAS) when the communication channel of nodes in IWSNs subject to transmit emergency packets, node can change to a good quality channel in order to maintain the stability of communication for emergency packets. Simulation experiment proved that the QCAS had made the 6TiSCH can achieve higher reliability of transmission to provide QoS for emergency packets.

The rest of the paper is organized as follows. In Sect. 2, we review some of the relevant studies on 6TiSCH and 6top [6] sub-layer. In Sect. 3, we describe the proposed QCAS for 6TiSCH networks. In Sect. 4 aims at compared transmission performance between the QCAS and the original 6TiSCH. Finally, we draw conclusions and future works in Sect. 5.

2 Relative Works

This section will describe the 6TiSCH, TSCH and 6top sub-layer and related literature.

2.1 6TiSCH

IEEE 802.15.4 proposed a new Time Slotted Channel Hopping (TSCH) mechanism. In the same year, IETF also propose new 6TiSCH architecture. The physical layer of the IEEE 802.15.4e TSCH is still the underlying protocol that uses the original IEEE 802.15.4 and can be operated on any hardware architecture that is compatible with the IEEE 802.15.4 standard. IEEE802.15.4e adds TSCH MAC protocol between IPv6 and LLNs. However, the TSCH MAC Layer and 6LoWPAN Adaptation Layer have some issues between routing and cells scheduling. Therefore, in order to amend this issue, IETF has proposed the 6iTSCH Working Group. The main goal of IETF 6TiSCH is to integrate IEEE802.15.4e TSCH function and IETF6LoWPAN and ROLL [7] standards. The IETF 6TiSCH standard is based on the existing architecture, re-examining issues that may arise when integrating existing protocol, and providing LLC operations to abstract IPv6 links. In addition, the RFC 7554 documents for IETF 6TiSCH are still in the draft stage. However, the relevant technologies have been concerned by international industry, academic research and other institutions, such as ITU, Cisco, IBM, etc.

2.2 Time Slotted Channel Hopping

IEEE802.15.4e TSCH is a low power and reliable network solutions. The time is sliced into timeslots and the group of continuous timeslots called the slotframe in IEEE802.15.4e TSCH. TSCH defines a slot counter which called the Absolute Slot Number (ASN). The Eq. (1) presents the relationship between the selected channel and ASN.

$$\mathrm{frequency} = \left\{(\mathrm{ASN} + \mathrm{ChannelOffset})\%\, \mathrm{N}^{\mathrm{Freq}}\right\} \tag{1}$$

The TSCH will use Eq. (1) to calculate the transmission channel to reduce the impact of interference and multipath Fading.

2.3 6top Protocol

6top is one of sub-layers in 6TiSCH stack. The main propose of 6top is used to provide the upper layer function from IEEE 802.15.4e TSCH MAC layer to strengthen the original IEEE 802.15.4e TSCH MAC Layer for the construction of the network topology. 6top will maintain the operation of the link and provides each node synchronization mechanism. 6top can be adjusted according to TSCH MAC routing path selection, topology information, energy consumption and latency requirements and provide relevant information to the 6LoWPAN network layer. 6top Protocol [8] (6P) is defined in the 6TiSCH distributed scheduling network, so that neighboring nodes in the 6TiSCH network can add or delete cells. In this paper, the proposed QCAS is also based on the 6P command.

The OpenWSN [9] is the first wireless sensor network simulation has implemented the fully IEEE802.15.4e standard. In addition, openWSN also implement Internet of Things [10] standards, such as 6LoWPAN [11], RPL and CoAP to achieve low power and reliability of the mesh network.

In [12], authors introduce a bandwidth request module use on the 6TiSCH network. The main purpose of this research is to dynamically match the link layer according to the requirements of the network application. A threshold has been set to prevent the preventing ring effects in the parallel cell allocations. 6top is used to monitoring, and add or delete the cell through 6P when the threshold is not met. The OTF performance analysis through the simulation with 50 sensor nodes to show that the end-to-end delay time within one second and more than 99% of end-to-end reliability.

A pair-housekeeping mechanism has been proposed in [13]. The cell with poor packet delivery ratio will delete through the 6top cell reallocation procedure to improve the link reliability.

3 Design of Proposed QoS Channel Allocation Scheme

This section will discuss the proposed QoS Channel Allocation Scheme (QCAS) and how to implement into the On-the-Fly (OTF) [12] module in 6TiSCH networks.

3.1 Channel Quality Estimation

The characteristic of the QCAS is to monitor the quality of each channel by using the Packet Delivery Ratio (PDR) during timeslots. The first metric measure the packet error rate for each channel including packets transmitted and received in a timeslot. The value of the metric is the ratio between the number of packets being acknowledged divided by the number of packets being sent by a sender node for each channel.

3.2 The Best Channel Candidate

We use channel quality estimation to select the best channel when the sender node require the emergency packet transmission. To determine the best channel candidate, we first rank the measured channels i.e., the maximum PDR channel that above the QoS PDR threshold 95%, and next we choose among the top one PDR channel. Clearly the use of this second metric makes sense only if it is smaller than a given QoS PDR threshold, otherwise the best selection is simply the channel with the maximum PDR.

3.3 Channel Allocation Using 6top Protocol

Let's consider the schedule in 6TiSCH, at least one shared cell [14] and serialRx cell in the front of slotframe. All other cells in slotframe may be reserved by any pair of nodes randomly. Furthermore, the OTF module is used to schedule required bandwidth to network. OTF asks the 6top protocol to reserve cells for pair of nodes. In the distributed 6TiSCH mode, the node sends a request to its next-hop node. The request message includes the list of free cells and the number of required cells. The next-hop node will accept the request if it has available resource, and then reply a response to the sender node. The 6TiSCH scheduling will monitor cells PDR to decide when a cell has to be inserted or removed in the schedule. Thus, a strategy to allocate the best channel on-demand based on the OTF scheduling in 6TiSCH networks has been proposed in this paper.

The Fig. 1 shows the flowchart of QCAS. The sender node will generate a QoS request to its next-hop when emergency packets are inserted into the queue. The QoS request message includes the list of best channel candidate. The next-hop node will accept the request and select a proper channel from the list of best channel candidate, and then reply a response to the sender node. The next-hop node will change the receiving channel to the selected best channel from next timeslot. The steps describe above will complete the procedure of QCAS.

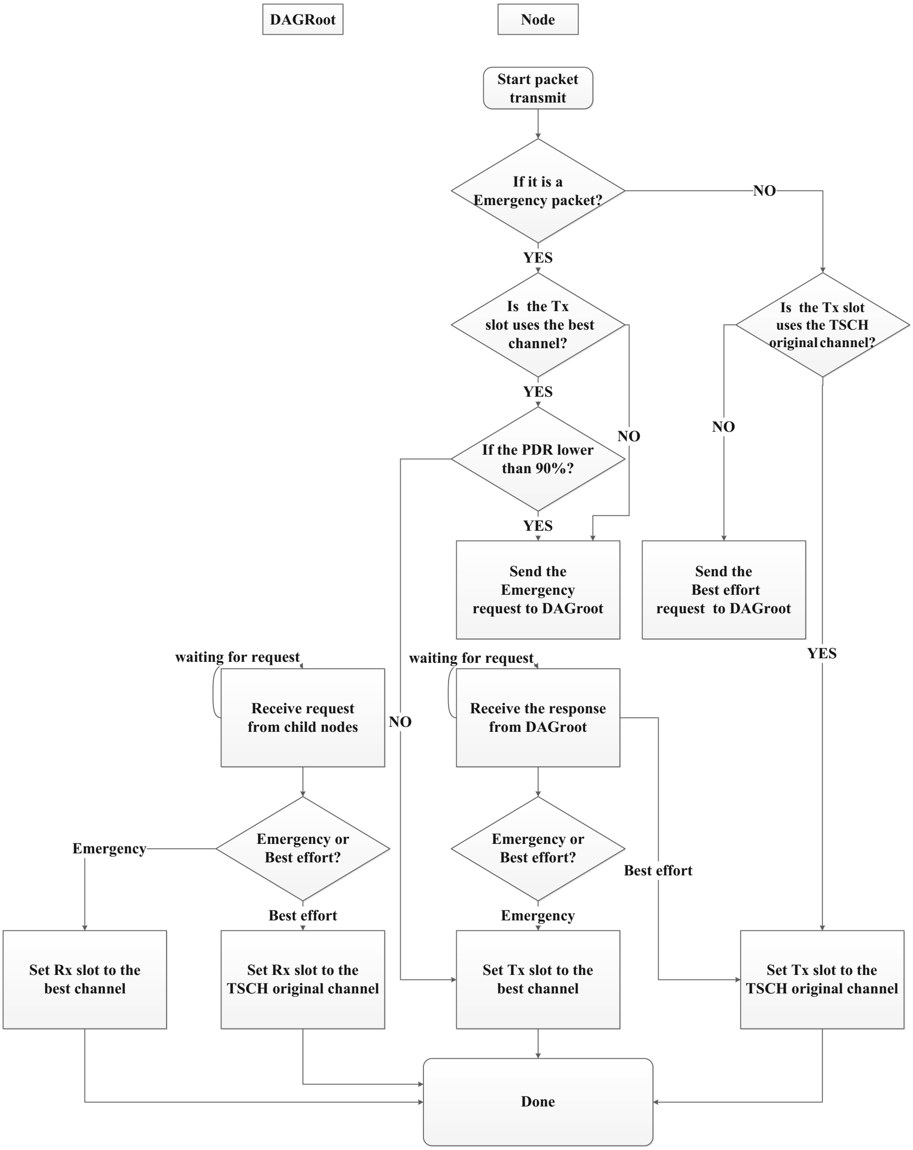

DAGRoot
Node
Start packet transmit
If it is a Emergency packet?
NO
YES
Is the Tx slot uses the best channel?
Is the Tx slot uses the TSCH original channel?
YES
NO
NO
If the PDR lower than 90%?
YES
Send the Emergency request to DAGroot
Send the Best effort request to DAGroot
YES
waiting for request
Receive request from child nodes
NO
waiting for request
Receive the response from DAGroot
Emergency
Emergency or Best effort?
Emergency or Best effort?
Best effort
Best effort
Emergency
Set Rx slot to the best channel
Set Rx slot to the TSCH original channel
Set Tx slot to the best channel
Set Tx slot to the TSCH original channel
Done

Fig. 1. The flowchart of proposed QoS Channel Allocation Scheme

4 Performance Evaluation

In this section, we first present the OpenWSN simulation configuration. Second, we show the results for both the proposed QCAS and the original 6TiSCH. We measure several parameters for performance evaluation as follows.

4.1 Simulation Configuration

Table 1 presents the simulation configuration. Figure 2 shows the network topology of the 5 testing node which contains a DAG root and four end nodes. Each end node will send a UDP packet to the DAGroot with one second interval. The length of slotframe is 11 cells, and each timeslot is 15 ms.

Table 1. Simulation configuration

Deployment example	
Number of sensor motes	5
Application data generation	
Period	1000 ms
Emergency packet generation	Packet transmission > 500
IEEE802.15.4e TSCH example	
Timeslot duration	15 ms
Length of a slotframe	11 cells
Max. MAC retries	0

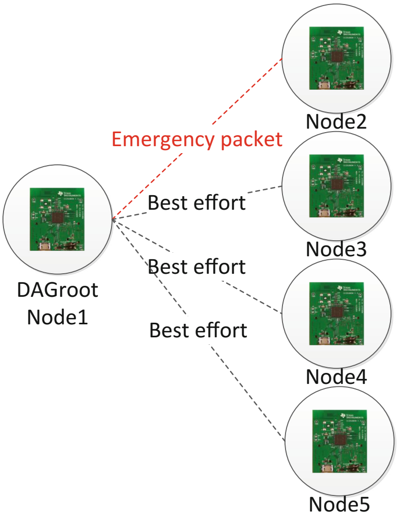


Fig. 2. Network topology

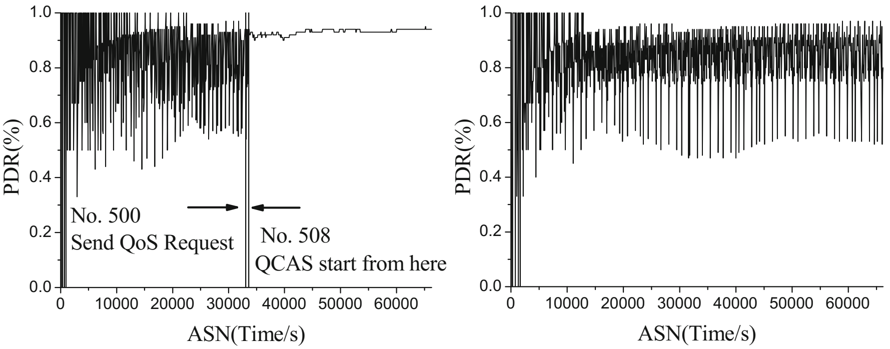
In this simulation, the first 500 packets are sent as best effort packets. The Table 2 is the value of PDR value for 16 channels.

Table 2. Simulation channel PDR parameters

Channel	11	12	13	14	15	16	17	18
PDR (%)	0.6	0.65	0.75	0.75	0.8	0.8	0.8	0.85
Channel	19	20	21	22	23	24	25	26
PDR (%)	0.85	0.85	0.85	0.9	0.9	0.9	0.9	0.95

4.2 The Average PDR for the Selected Channels

This section shows a comparison of simulation results from the 6TiSCH with QCAS and the original TSCH. Figure 3(a) shows the average PDR for the 6TiSCH with QCAS. The average PDR from packet no. 1 to no. 507 have a large variation. From packet no. 508 (ASN is 33606), the PDR is reduced. It is because, node 2 will generate a QoS request to node 1 (DAGroot) when emergency packets are inserted into the queue at packet no. 500 (ASN is 3303) and the best channel has been selected successfully at packet no. 508 with a delay of 8 packets transmission. Figure 3(b) shows the original TSCH channel hopping for the best effort packages. The unstable average PDR is affected by the random selected channel PDR.



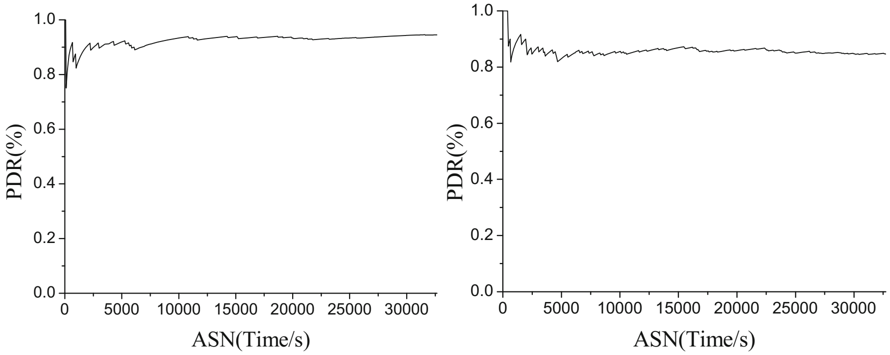
(a). The 6TiSCH with QCAS

(b). The original TSCH

Fig. 3. The average PDR for the selected channels

4.3 The Average PDR for the Selected Node 2

Figure 4 is the simulation result for the average PDR in the node2. We focus on the packet no. 508 to 1000. The average PDR of 6TiSCH with QCAS for node 2 is around 95%, which is very similar to the PDR value of channel 26 in Table 2. In Fig. 4(b), the average PDR of the original TSCH in node 2 is 83.5%. Thus, we can confirm that the QCAS can improve the link stability to provide better QoS for emergency packets.



(a) The 6TiSCH with QCAS

(b). The original TSCH

Fig. 4. The average PDR in the node 2

4.4 Throughput

In Fig. 5, shows the simulation results for the average throughput in both 6TiSCH with QCAS and original TSCH. The 6TiSCH with QCAS has higher average throughput than the original TSCH. The average throughput of 6TiSCH with QCAS is 0.966 kbps, and the original TSCH is 0.863 Kbps. In contrast, the QCAS approach improves the transmission efficiency and the link reliability.

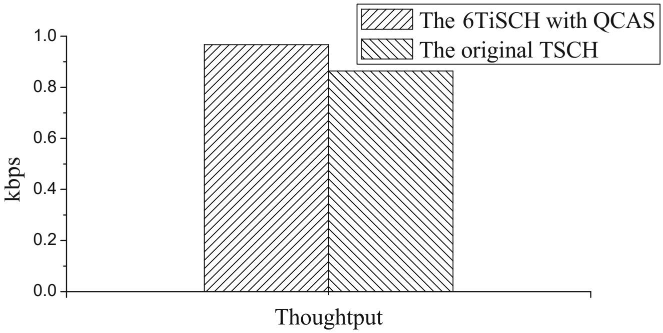


Fig. 5. The end-to-end throughput (kbps)

5 Conclusion

In this paper, we focus on the issue of 6TiSCH in the error-prone wireless channel condition. In this research, a QoS Channel Allocation Scheme is proposed to enhance the reliability of emergency packet transmission in 6TiSCH networks. Based on our simulation results, we have observed that the proposed QCAS can achieve higher reliability of transmission over selected channel to provide higher throughput and higher PDR than the original TSCH.

In the future works, we will extend this study to testing in more complex environment and established a multiple bandwidth modules based on QCAS to provide more efficient QoS support in 6TiSCH networks.

Acknowledgment. This research was supported by research grants from Ministry of Science and Technology, Taiwan 105-2221-E-142-002-MY2.

References

1. Salam HA, Khan BM (2016) IWSN - standards, challenges and future. IEEE Potentials 35(2): 9–16
2. IEEE 802.15 Work Group (2016) 802.15.4-2015 IEEE standard for low-rate wireless networks, April 2016
3. Thubert P (2016) An architecture for IPv6 over the TSCH mode of IEEE 802.15.4. draft-ietf-6tisch-architecture-10 (work in progress), June 2016
4. Watteyne T, Palattella M (2015) RFC7554 "Using IEEE 802.15.4e Time-Slotted Channel Hopping (TSCH) in the Internet of Things (IoT): problem statement", May 2015
5. IEEE standard for Information Technology (2012) IEEE std. 802.15.4e, Part. 15.4: Low-Rate Wireless Personal Area Networks (LR-WPANs) Amendment 1: MAC sublayer, April 2012
6. Wang Q, Vilajosana X (2015) 6TiSCH operation sublayer (6top) interface, IETF Standard draft-ietf-6tisch-6top-interface-04, July 2015
7. Phinney T, Thubert P, Assimiti R (2013) RPL applicability in industrial networks. Work in Progress, draft-ietf-roll-rplindustrial-applicability-02, October 2013
8. Wang Q, Vilajosana X (2016) 6top Protocol (6P). draft-ietf-6tisch-6top-protocol-03, October 2016
9. OpenWSN. https://openwsn.atlassian.net/wiki/
10. ITU Internet Reports (2005) The Internet of Things (ITU 2005), 7th edn
11. Shelby Z, Bormann C (2009) 6LoWPAN: the wireless embedded internet. Wiley series on communications networking & distributed systems. Wiley, Chichester
12. Palattella MR (2015) On-the-fly bandwidth reservation for 6TiSCH wireless industrial networks. IEEE Sens J
13. Muraoka K, Watteyne T, Accettura N (2016) Simple distributed scheduling with collision detection in TSCH networks. IEEE Sens J
14. Wang Q, Pister K (2016) Minimal 6TiSCH configuration. draft-ietf-6tisch-minimal-16, June 2016

An Adaptive User-Defined Traffic Control Mechanism for SDN

Hung-Chi Chu(✉) and Tzu-Hsuan Lin

Department of Information and Communication Engineering,
Chaoyang University of Technology, 168 Jifong E. Rd., Wufong District,
Taichung 41349, Taiwan
hcchu@cyut.edu.tw, hew28744b@gmail.com.tw

Abstract. In the traditional network architecture, must rely on network administrator to manually set up complex network parameters for a variety of QoS. Software-Defined Networking (SDN) divides networks into a three-tier architecture. The application layer that provides a centrally managed interface, the control layer that allows users to manage controller with software; and the controller sends command to the infrastructure layer, which runs a new network with settings to achieve the purpose of a centralized network control. This paper proposes a QoS mechanism that can determine the optimal traffic allocation based on network service requirements and adjust the overall network traffic allocation in real time to achieve the smooth operation of services. The experimental result showed that the proposed mechanism can be defined by the user to customize the required network services, and provides network resources to match requirements according to different priorities as well as dynamically adjusts the network parameters.

1 Introduction

The rapid development of the internet has created many new possibilities with many internet services having been achieved. With the emergence of new internet services, traditional distributed network architecture to address these emerging services faces many challenges. To handle increasing network traffic, network administrator must re-establish a network environment to maintain a stable network. Network administrators need to maintain all network devices individually and require a lot of time and effort, but these are limited to traditional networking features. To solve this issue, scholars have proposed a centralized network architecture, which we call Software-Defined Networking (SDN) [1–3]. SDN architecture was presented at the IEEE INFOCOM conference in 2009, but SDN came into the spotlight because of Google. Google spent two years converting their data center backbone network into an SDN architecture, this change made 30–40% usage of network bandwidth increased to 95%. This result allows many large-scale companies (such as Facebook or Microsoft) to research this technology. SDN architecture consists of three layers: t the application layer, the control layer, and the infrastructure Layer (Fig. 1). The authority responsible for managing an entire network will be located in the control layer. The interface between the control

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_67

layer and the infrastructure layer is also known as the southbound interface. This interface allows the infrastructure layer to successfully receive commands from the control layer to perform data forwarding tasks. The interface between the control layer and the application layer is also called the northbound interface, which provides a centralized management interface to allow user to control and manage it through software. Through such a network assignment method, network administrators can define the desired network settings and let the controller send commands to a switch in the infrastructure layer to apply the new network settings. This network architecture centrally manages network resources to effectively reduce the difficulty of network management and to enhance the flexibility of overall network traffic scheduling as well as increasing with processing speeds for unexpected situations. Even through the controller can implement a fully automated network management to reduce human error and management costs. The advantages of SDN are as follows: (1) A centralized network control and open network programming interface allows network administrator to monitor and manage a network through software or lets administrator develop programs they need; (2) A highly flexible network architecture allows network functions to be easily added or removed based on administrator need; and (3) In the network, the device configuration, management, and control can be completed via the controller to effectively reduce time and labor costs as well as providing high-performance network control and management.

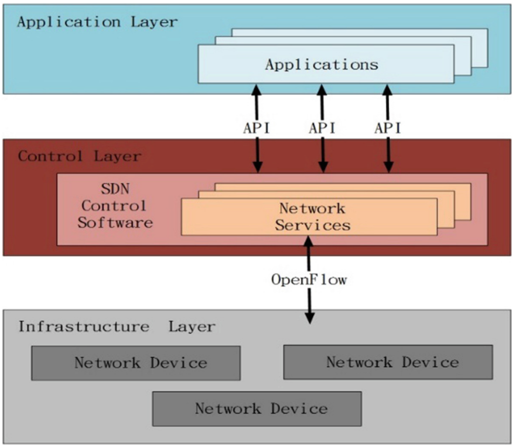


Fig. 1. SDN architecture

The communication protocol between the SDN controller and the switch was one of the projects that was part of the future web-based research program at Stanford University in 2008, i.e. OpenFlow [4]. OpenFlow is an open source protocol originally developed for a flow forward and centralized controller. In the SDN network, each switch has a flow table, so that when a packet arrives at the switch, it compares the information in the flow table to decide how to forward the packet. If it fails to find a match with the flow table, then the packet will be sent to the controller to plan the flow. After the planning is complete, the controller will inform the switch of the flow table update.

The communication between the switch and the controller is usually protected by SSL (Fig. 2). OpenFlow's performance is not good when it is launched, but it has become an important protocol after continuous improvements, i.e., it now supports IPv6, multiple controllers, QoS, and other functions.

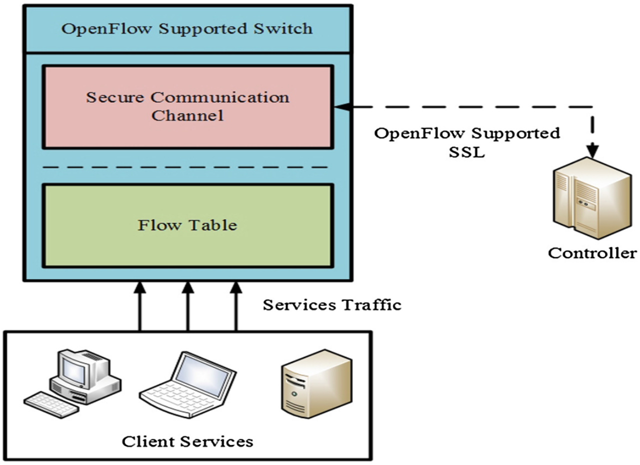


Fig. 2. OpenFlow communication architecture

According to the above description shows, SDN have many advantages compared to traditional networks. This paper proposes a traffic management mechanism based on SDN architecture, so that network administer can proactively define the flexible network traffic control mechanism. The controller automatically adjusts the traffic ratio of the service by priority to avoid problems associated with low network performance due to competing traffic between the various network services, which results in a high packet loss rate, network traffic shocks, and other issues.

In this paper, simulation experiments used Mininet [5], iPerf network speed test tool [6], and virtual switch to support the OpenFlow protocol: Open vSwitch [7] to demonstrate performance advantages from SDN architecture. Open vSwitch is a production quality, multilayer virtual switch designed to enable massive network automation through programmatic extensions. The advantages of Open vSwitch are as follows: (1) Highly flexible configuration that can establish up to hundreds of virtual switches in a single server (depending on server performance) and each virtual switch can have more detailed settings; and (2) Replaces a physical device with software simulation to effectively reduce the costs of building a network and virtual switch is not weakly than a more expensive physical device with the same functionality. Mininet is a virtual network simulator that supports the protocols and devices required by SDN architecture. This simulator, which only needs simple commands to simulate a simple network topology (such as a tree structure) and supports custom topologies written in Python, makes the situation closer to actual real-world network operations. iPerf allows for the analysis of the message throughput, jitter, packet loss, maximum transmission unit (MTU), and other statistics under physical or virtual networks. It is used to test the performance of

SDN and runs on common operating systems such as Linux, Windows, or Mac, among others. The operation is easy and only needs to create two IP-enabled computers on a custom network topology and sets them as a server and a client. We can begin experimenting with iPerf commands to test the performance of the network.

The rest of this paper is organized as follows: Sect. 2 presents the research related to the topic; Sect. 3 presents the traffic control mechanism and network topology; and Sect. 4 presents experimental results and comparison. Finally, Sect. 5 provides our conclusions.

2 Related Works

The biggest advantage of SDN is an ability to instantly monitor, configure, and reflect network changes. In the last few years of research, SDN are many major research projects. This paper, will introduce SDN and explore relevant QoS literature. [8] proposed the autonomous QoS management mechanisms for SDN, identify and tag packets in queue using OpenFlow Management and Configuration Protocol (OF-Config) and different QoS have different tags. So, it adds a Requirements DB, Context manager, Analysis and Rule decisions, and other modules to achieve the requirements of QoS in the controller. [9] proposed a dynamic traffic performance algorithm based on a single flow table and a group flow table. It uses a single flow meter to accurately monitor every client in the traffic and to use of group flow tables to effectively classify client hosts. The results showed that this algorithm can effectively avoid the use of too many flow table and solve flow table matching within the scope of too wide of defects. [10] compared IntServ [11] and DiffServ [12] of QoS. Intserv ensures that a single application can guarantee QoS, but the scalability is poor and all network devices within that network infrastructure must support IntServ, which increases the overall network construction costs. DiffServ allows the core router to no longer store the status of network traffic information but is only responsible for packet forwarding. From this, the network load is reduced, easier achieving, and improved scalability. But QoS cannot achieve the same results with IntServ. In addition, the paper analyzes the different routing algorithms in the SDN backbone network to dynamically establish the effectiveness of guaranteed data flow bandwidth and routing delays. In this research, the QoS performance indicator for the routing algorithm is using the not satisfied of bandwidth requirement rate and the results show that its performance is excellent, but its disadvantage is high computing time. [13] proposed to develop QoS for multimedia string data in the architecture of a decentralized control plane. Its features include: (1) Use topology aggregation and link summarization to effectively obtain information about network topology and status; and (2) By solving the communication between the controllers to perform secure QoS for an inter-domain.

3 Traffic Control Mechanism

In this paper, we propose a traffic control mechanism to establish a traffic allocation ratio by predefining the priority of various services. The definition of priority and the proportion of bandwidth allocation can be determined by a network administer to meet the needs of users. When the controller begins to detect that traffic affects overall network performance, then it will trigger the traffic control mechanism proposed. Figure 3 shows the operation flow. The controller is graded according to the predefined service to determine the proportion of bandwidth that can be allocated and checks whether the allocation matches the minimum bandwidth required for high priority services. If the requirements are matched, then the controller will begin sending commands to modify the overall network bandwidth to retain a route that meets the high priority service. Conversely, if the current network state cannot match the minimum bandwidth requirement for high priority services, then a temporary transfer of the service with a lower priority level to another link, so that the bandwidth of a single link is used in the transfer of high priority services. This mechanism ensures that high priority services can maintain quality until the end of the transmission and to minimize the impact of other low priority service traffic due to adjustments of bandwidth.

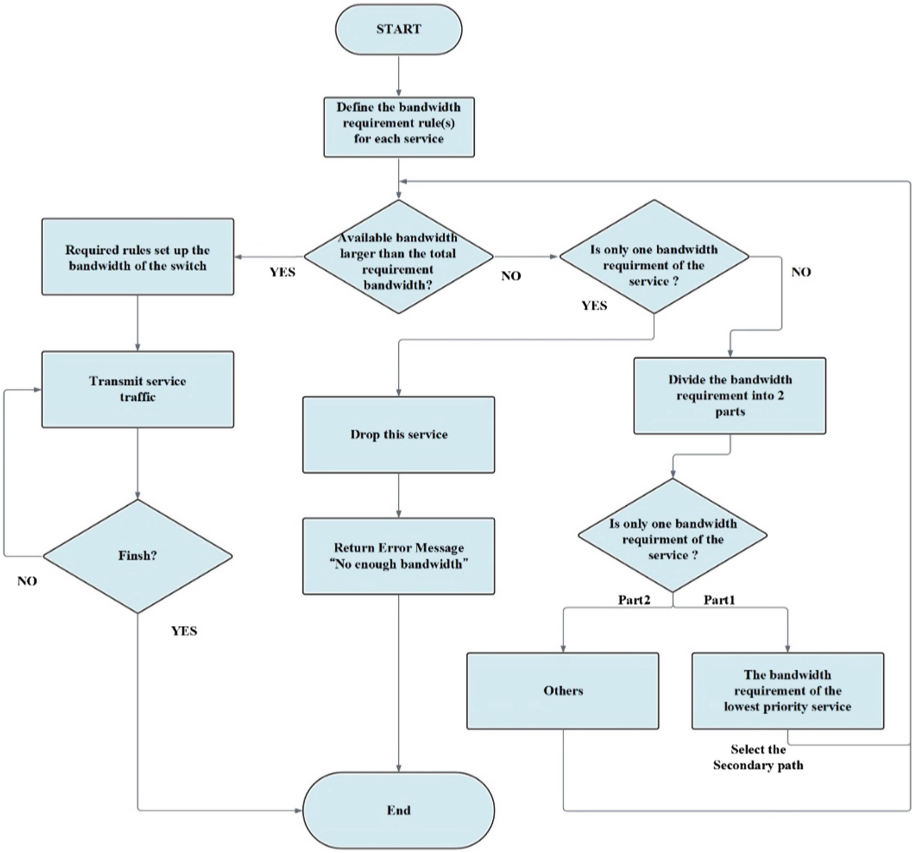

START
Define the bandwidth requirement rule(s) for each service
Available bandwidth larger than the total requirement bandwidth?
YES
NO
Required rules set up the bandwidth of the switch
Is only one bandwidth requirment of the service ?
NO
YES
Transmit service traffic
Drop this service
Divide the bandwidth requirement into 2 parts
Finsh?
NO
YES
Return Error Message "No enough bandwidth"
Is only one bandwidth requirment of the service ?
Part2
Part1
Others
The bandwidth requirement of the lowest priority service
Select the Secondary path
End

Fig. 3. Traffic control mechanism

To verify the effectiveness of the proposed mechanism, this paper will use a fully connected network topology (Fig. 4). The network has three kinds of data traffic: Traffic1, Traffic2, and Traffic3 to represent the different priority services. S represents the switch and a server is the destination of the data transmission. In a traditional network, if no special handling of traffic transmissions is necessary, the shortest path is prioritized; however, this approach is likely to cause each service to compete with traffic, which leads to low network performance, because the weight of each traffic is the same. If urgently changes packet routing, requires a lot of time to set the network parameters and restart network devices seriously reduces overall network performance. However, if the SDN architecture is used, then the controller will calculate the most suitable bandwidth allocation ratio and automatically set network parameters to achieve traffic control.

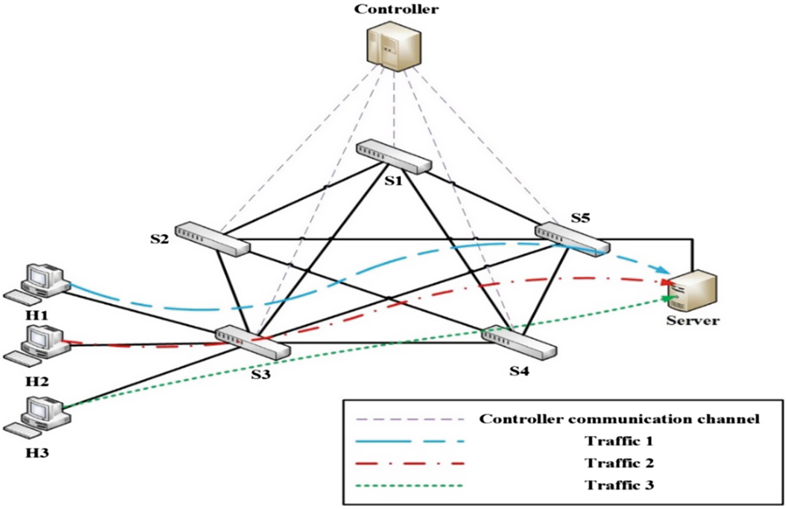


Fig. 4. Network topology

4 Experimental Results

This section will use the network topology in Fig. 4 and supports SDN architecture. It shows the difference between a traditional network and an SDN network for different network traffics. In the experiment, the routing algorithm of the traditional network will adopt the shortest path algorithm to carry out and set the links bandwidth between the various network devices as 10 Mbps. Table 1 shows the relevant experimental parameters using iPerf.

Table 1. EXPERIMENT PARAMETERS

Host	Command	Test time
Traffic	iperf –c Server IP –r X	100 s
Server	iperf –s	

The -r sets traffic, X is the bandwidth of the traffic requirement, and rate unit is bits per second.

We will perform two experiments to verify the effectiveness of the proposed traffic control mechanism as follows:

Experiment of user-defined traffic flow control: This experiment assumes that there are three streams at the same time: Traffic1, Traffic2, and Traffic3 (Fig. 4). In Traffic1, the transmission path is H1 → S3 → S5 → Server; in Traffic2, the transmission path is H2 → S3 → S5 → Server; and in Traffic 3, the transmission path is H3 → S3 → S5 → Server. These three traffics will be transmitted to the destination server using the UDP protocol. The entire experiment was simulated for 100 s, the traditional network uses the shortest path algorithm for data transmission from 0–50 s. In this case, there is no difference in the priority of each traffic. From 51–100 s, setting the priority of the three traffics to Traffic1, Traffic2, and Traffic3 from high to low; and the three traffics in 6:3:1 to configure the priority level to use the bandwidth ratio. For example, a link bandwidth of 10 Mbps, set Traffic1, Traffic2, and Traffic3 to 6, 3, and 1 Mbps. The user can configure the bandwidth ratio that can be used for each priority level on the controller.

Figure 5 shows the experimental results. From 0–50 s of the experiment indicates that each traffic is in a state of competing with the other. Because no priority level was set, so causes large magnitude of bandwidth usage change, i.e. behalf of the actual use of the traffic bandwidth instability. The average bandwidth used for from 0–50 s of these three streams is 3.46 Mbps, 2.95 Mbps, and 3.15 Mbps. It can be seen that in the state of competitive bandwidth, although the amount of available bandwidth changes is large but with the result of equalization for available bandwidth. When the experiment went to 50 s, Traffic1, Traffic2, and Traffic3 used bandwidth of 0.76, 2.74, and 6.42 Mbps, respectively. At this time the controller triggers the traffic control mechanism according to the set of bandwidth allocation to adjust the bandwidth to achieve traffic control. According priority setting, Traffic1 is allocated with more bandwidth and the controller begins to increase Traffic1's bandwidth to within the upper limit of 6 Mbps. The low priority of Traffic3 is assigned to the least bandwidth required and the controller reduces its use of bandwidth to within the upper limit of 1 Mbps. The medium priority of Traffic3's allocated bandwidth is close to the current bandwidth, so that the bandwidth is still within the upper limit of 3 Mbps. The simulation results clearly showed the bandwidth was adjusted after setting the priority level, the available bandwidth of the traffic became stable, and was close to the ratio of available bandwidth for each priority level. Therefore, the controller's network parameter settings can quickly impact devices on the network and achieve the set traffic control effects in a short period of time.

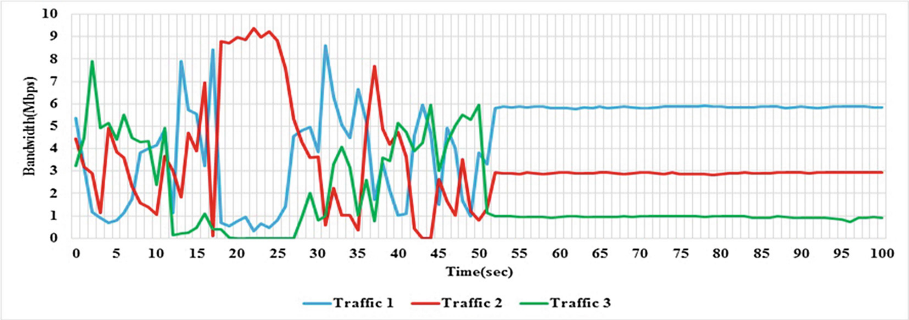


Fig. 5. Experiment of user-defined traffic flow control

Experiment of high priority traffic flow control: This experiment assumes that there are three different traffics based on different priorities and according to the priority level from high to low as in Traffic1, Traffic2, and Traffic3. The initial transmission path is the same as the previous experiment. The three traffics are configured with a bandwidth ratio of 6:3:1. These three traffics will be transmitted to the destination server using the UDP protocol. The experiment was simulated for 100 s, during 0–50 s, the three traffic configures bandwidth according to the priority level for data transfers. After 50 s, due to a special demand, the highest priority data traffic, i.e. Traffic1, can be used to upgrade the bandwidth to 7 Mbps, but the link bandwidth of only 10Mbps cannot match the three traffics required to use the complete bandwidth, i.e. 7 Mbps + 3 Mbps + 1 Mbps = 11 Mbps. In order to match the bandwidth required for the highest priority service, the available bandwidth for the lowest priority traffic Traffic3 is provided to the Traffic1, and Traffic3 is

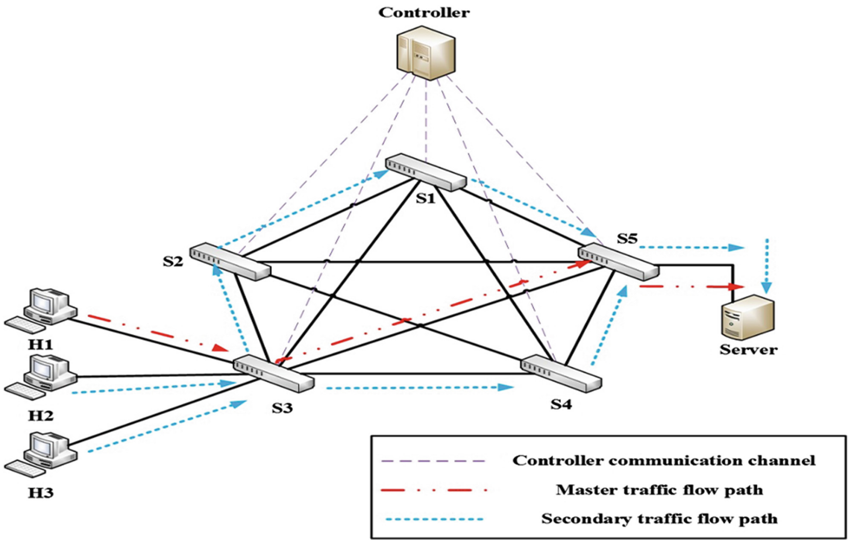
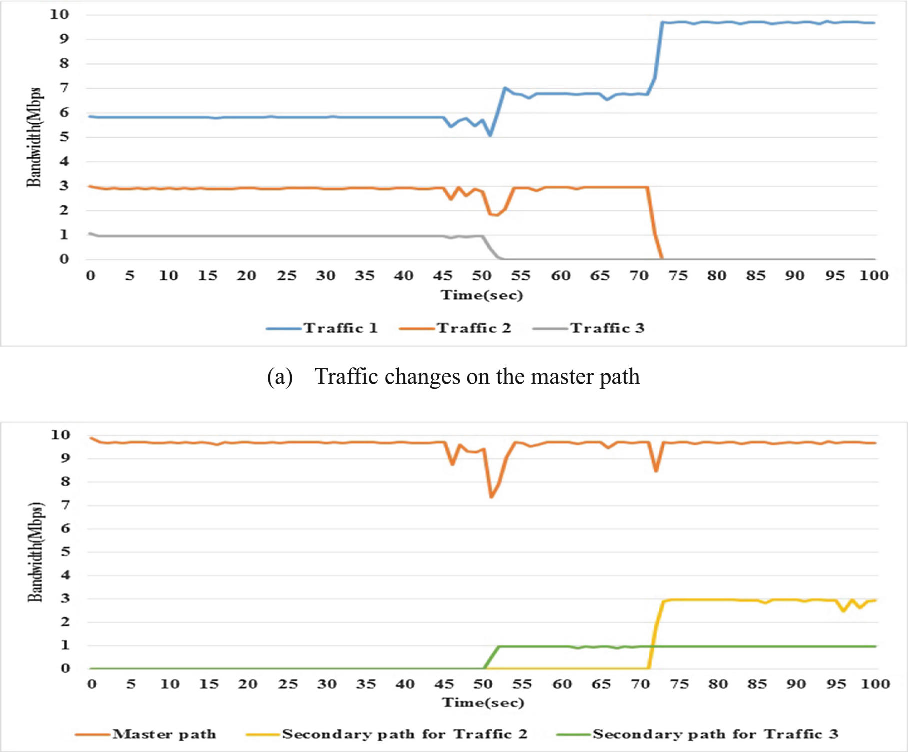


Fig. 6. An example of adaptive traffic flow

transferred to other links. This indicates that the lowest priority traffic Traffic3 is switched to a secondary transport path H3 → S3 → S4 → S5 → Server (Fig. 6). After 70 s, due to special demands again, the highest priority data traffic Traffic1 can use the bandwidth upgrade to 10Mbps, but the total bandwidth of the link can only match the required bandwidth for Traffic1. So, medium priority traffic, Traffic2 uses available bandwidth as provided to Traffic1; and Traffic2 will be transferred to other links. This indicates that the lowest priority traffic, Traffic2 is switched to the secondary transport path H2 → S3 → S1 → S5 → Server (Fig. 6).

Figure 7 shows the experimental results. From 0–50 s, according to priority settings, the available bandwidth of Traffic1, Traffic2, and Traffic3 is close to 6 Mbps, 3 Mbps, and 1 Mbp. After 50 s, because the available bandwidth of traffic1 increased to 7 Mbps. Therefore, Traffic3's available bandwidth is released on the original path (H3 → S3 → S5 → Server). Traffic3 changes to a secondary path (H3 → S3 → S4 → S5 → Server) for data transfer, which ensures that its available bandwidth is still 1 Mbps. After 70 s, because the available bandwidth of Traffic1 increases to 10 Mbps, Traffic2`s available bandwidth is released on the original path (H2 → S3 → S5 → Server), Traffic2 changes to a secondary path



(a) Traffic changes on the master path

(b) Traffic changes on a distributed path

Fig. 7. Experiment of high priority traffic flow control

(H3 → S3 → S1 → S5 → Server) for data transfers and ensures that its available bandwidth is still 3 Mbps.

The simulation results can be seen clearly, according to priority level to adjust the bandwidth. The available bandwidth of the traffic becomes stable as the available bandwidth of the highest priority service increases, the low priority traffic transfers to the secondary path and guarantees the available bandwidth requirements. The above experiment proves the traffic control mechanism of this paper adjusts the network parameters adaptively with a service requirements change to maintain the transmission quality for each priority level.

5 Conclusions

SDN architecture reduces the difficulty of network maintenance, update, and management. It also provides for a significant improvement in network performance. The controller has global network management functions and the network administrator can operate the controller remotely to manage all switches under its jurisdiction to reduce the consumption of human resources and time as well as offers the ability to respond quickly to emergencies. We use Mininet, iPerf, and Open vSwitch to achieve a QoS-support adaptive traffic control mechanism. Through the definition of the weight of the network parameters, the controller can automatically adjust network resources to match the requirements of different services. The experimental results show that our proposed traffic control mechanism can: (1) Allow the user to define the network service requirements needed; (2) Provide network resources that match user needs according to different priorities; (and 3) Automatic adjustment for dynamic network service requirement changes.

Acknowledgment. This work was supported in part by the Ministry of Science and Technology, Taiwan, ROC, under grant MOST 105-2221-E-324 -009 -MY2.

References

1. Open Networking Foundation (2012) Software-Defined Networking: The New Norm for Networks. White paper
2. Chowdhury SR, Bari MF, Ahmed R, Boutaba R (2014) PayLess: a low cost network monitoring framework for software defined networks. In: Network operations and management symposium (NOMS). IEEE, Krakow, pp 1–9
3. Liu CM (2014) The first book of the software-defined networking. Grandtech Information, Taiwan
4. McKeown N, Anderson T, Balakrishnan H, Parulkar G, Peterson L, Rexford J, Shenker S, Turner J (2008) OpenFlow: enabling innovation in campus networks. ACM SIGCOMM Comput Commun Rev 38(5):69–74
5. Mininet. http://mininet.org/. Accessed 1 Mar 2017
6. iPerf. https://iperf.fr/. Accessed 1 Mar 2017
7. Open vSwitch. http://openvswitch.org/. Accessed 1 Mar 2017

8. Wang W, Qi Q, Gong X, Hu Y, Que X (2014) Autonomic QoS management mechanism in software defined network. China Commun 11(7):13–23
9. Mao Q, Shen W (2015) A load balancing method based on SDN. In: International conference on measuring technology and mechatronics automation. IEEE, Nanchang, pp 18–21
10. Tomovic S, Radusinovic I, Prasad N (2015) Performance comparison of QoS routing algorithms applicable to large-scale SDN networks. In: IEEE international conference on computer as a tool (EUROCON). IEEE, Salamanca, pp 1–6
11. Braden R (1994) Integrated services in the internet architecture: an overview. RFC 1633
12. Blake S (1998) An architecture for differentiated services. RFC 2475
13. Egilmez HE, Tekalp AM (2014) Distributed QoS architectures for multimedia streaming over software defined networks. IEEE Trans Multimedia 6(6):1597–1609

Retrieval of 3D Trademark Based on Discrete Fourier Transform

Chu-Hui Lee(✉) and Liang-Hsiu Lai

Department of Information Management, Chaoyang University of Technology,
Wufong Township, Taichung County, Taiwan (R.O.C.)
{chlee,s10314630}@cyut.edu.tw

Abstract. With the diversity of the trademark, three-dimensional (3D) trademark has gradually been used by many companies. For the legitimate use, 3D trademark is required to be registered. In Taiwan, the registration application is proved by Intellectual Property Office under the Ministry of Economic Affairs. The 3D trademark is useful to identity characteristic products or company on the Internet or the physical store. At present, 3D trademark can be registered in many nations. However, it is must be made sure unique in the 3D trademark gallery before registration. Similarity measurement of 3D trademark is an important task and how to efficiently searching is an interesting issue. In this paper, an efficient searching method for 3D trademark gallery in Taiwan's Intellectual Property Office is provided. A 3D trademark is represented by multiple 2D images in the gallery. Discrete Fourier transform (DFT) is used to extract the feature, and the effective search mechanism is established by using the properties of multiple 2D images. The experiment shows the method is efficient.

Keywords: Image retrieval · Discrete Fourier transform (DFT) · Similarity measurement · 3D trademark

1 Introduction

In international economics and trade, 3D trademark images are crucial presence for maintaining international economic order. In many countries, 3D trademark application has to go through notary service agencies. In Taiwan, for example, it is required to apply for patents and registration with the IPO under the Ministry of Economic Affairs (MOEA) to receive legal protection. The state of 3D trademark application in recent years is shown in Table 1 [8]. According to the statistics published by four authorities that represent R.O.C., USA, South Korea and China, respectively. All of these countries have seen growth in the number of applications for invention patenting and trademark registration compared to the past and the growth also sets a new high record. More complex design models are required for 3D trademark application in comparison with trademark application in the past.

As multimedia databases (MMDBs) continue to increase, it is easier for users to transfer, store and create images on the internet and in multimedia devices, such as smart phones, digital cameras and 3C products. Therefore, how image-based retrieval retrieves

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_68

Table 1. Trademark and patent registration in recent years (2010-2015) [8]

	Number of registration					
	2010	2011	2012	2013	2014	2015
R.C.O.	66,496	67,620	74,537	74,031	75,933	78,523
USA	280,649	301,826	311,627	321,055	336,275	369,877
South Korea	121,312	124,000	132,611	147,667	150,226	172,512
China	1,072,187	1,416,785	1,648,316	1,881,546	2,285,358	2,800,000

similar images in extensive MMDBs effectively and rapidly after users have provided query images has become an important research topic and a popular issue. Many scholars today have proposed content-based image retrieval (CBIR) as one of image retrieval techniques. In most CBIR-based retrieval systems, the retrieval technique typically used is extraction of low-level features, such as color, texture, shape and spatial distribution as representations. These systems then calculate feature differences between images based on the extracted data and automatically retrieve more relevant images or achieve better retrieval results in MMDBs. Over the past few years, many approaches to CBIR have been introduced for the imaging domain [3, 4, 5, 7].

Today, many users use a set of 2D images to represent a 3D image. Such examples can be found in the 3D trademark database of the IPO, MOEA [8]. 3D trademarks are the important certificates in market trading order and 3D trademark registration has also significantly increased. A 3D trademark is defined as having fixed length, width and height, combining images that are in more than one direction and include shapes, text, graphics, symbols and colors and on each of which important information of the 3D trademark is contained at each angle.

Among CBIR-based techniques, some use the watershed algorithm for division of an image into regions, from which color and texture features are extracted [1]. Image retrieval also utilizes the reference feedback mechanism that uses feedback obtained via different interactive techniques based on implementation standards and makes weighting adjustment to similarities measurement in MMDBs [5]. Some scholars use the machine learning mechanism to first detect representative features of images, which are then used for retrieval, so selection of representative features can affect retrieval performance in image databases [3]. Some studies retrieve 3D trademarks using Harris corner detection for corner feature extraction in combination with contour point's distribution histogram (CPDH) for shape description and shape feature extraction [4].

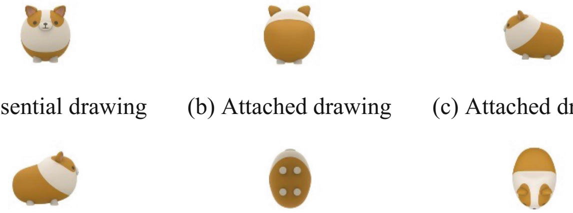
In this paper, the second section describes requirements for 3D trademark registration and the principle of DFT. The third section explains proposed methods and their steps. The fourth section gives the experiments of proposed method. The fifth section provides the conclusion.

2 Literature Review

With the rapid development of imaging technology in recent years, some flexible and efficient operations for query a 3D trademark for a large number of multimedia image databases are required to be established. Compared to traditional databases, a MMDB contains a wide range of content, such as shapes, text, graphics, symbols and colors, so the text-based retrieval method for traditional databases is not applicable.

2.1 3D Trademark

A 3D trademark is a trademark in a 3D shape formed by the three dimensions that are length, width and height and enables stakeholders to differentiate a product or service source. The form components for 3D trademark application include three parts, i.e. a statement, graphics and text description. For a 3D artwork, it is required to submit a master 3D artwork as the 'essential drawing' and maximum five other views of the artwork as the 'attached drawings', as shown in Fig. 1 [8]. If the filed 3D trademark contains text, graphics, symbols or colors, it is also required to provide clear and specific description [5, 8].



(a) Essential drawing (b) Attached drawing (c) Attached drawing

(d) Attached drawing (e) Attached drawing (f) Attached drawing

Fig. 1. A registered 3D trademark [8]

In Taiwan, the components of a 3D trademark include identifiable and non-functional ones and both are required for successful trademark registration. (1) Identifiable: a trademark should be identification that is sufficient to acquaint consumers with the products or services it represents and differentiate these products or services from others. The identification of a 3D trademark should take into account consumer perception and product features etc. (2) Non-functional: The 3D shape of a trademark or package that is essential for functionality may not be registered. Functional creations may receive patent rights protection in accordance with patent laws [8].

2.2 Discrete Fourier Transform

DFT is transform of spatial domain signals into frequency domain signals. As a type of Fourier transform, DFT is appropriate for sampling of continuous functions at uniform intervals. The computation time of DFT has complexity while fast and discrete Fourier computation can get faster results with complexity down to [2].

2.3 Image Retrieval

The performance of a CBIR-based system depends on extraction of specific features and similarity measurement. Some scholars propose feature extraction for a query image and tagging of extracted features in a database as an index table. In addition to this, different types of image databases naturally have different applicable features. Therefore, a mechanism that can independently adjust similarity measurement according to different types of database designs can increase the weights of appropriate features based on the characteristics of databases, thus improving image retrieval performance [1, 5].

3 Feature Extraction in 3D Trademark Retrieval

In future, as the time evolves and multimedia data continues to increase, retrieval of 3D trademark images will gain increasing attention. The DFT-based method presented in this paper for extraction of key features from images will benefit 3D trademark retrieval. We will introduce the feature extraction process and similarity measurement in this section.

3.1 Image Gray Scaling

As a 3D trademark consists of n images, represented by $P = (p_1, p_2, \ldots, p_n)$, the first step is to turn each image into a grayscale one. According to IPO's requirements, the maximum number of attached images is six.

3.2 Discrete Fourier Transform

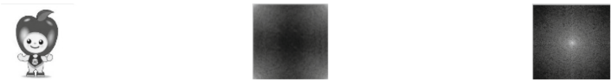
Multimedia images each consists of discrete signals. Transformation of these images is done via DFT. These images are two-dimensional $M \times N$, which is represented by the function $img(x, y)$, as shown in Eq. (1) [6]:

$$F(u, v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} img(x, y) e^{-j2\pi\left(\frac{ux}{M} + \frac{vy}{N}\right)} \tag{1}$$

Where, $u = 0, 1, 2, \ldots, M - 1$, $v = 0, 1, 2, \ldots, N - 1$ as a frequency variable. A spatial domain is represented by a coordinate formed by $img(x, y)$, which yields a frequency signal range that is the four corners of a spectrum (0, 0), (0, $N - 1$), ($M - 1$, 0), ($M - 1$, $N - 1$). For this study, let R and I represent the real and imaginary parts of F, as defined in Eq. (2):

$$|F(u, v)| = \left[R(u, v)^2 + I(u, v)^2\right] \tag{2}$$

Then the four corners of the spectrum are transformed and concentrated them to the center of the spectrum, as shown in Fig. 2.



(a) Original artwork (b) Original spectrum (c) transformed spectrum

Fig. 2. Post-DFT spectrum

Next, Fig. 2 (c) is the transformed spectrum as a feature image so that the data concentrates in the b × b range in the center and therefore the range for extraction of feature values is narrowed to b × b dimensions.

3.3 Similarity Measurement

Images of all 3D trademarks undergo DFT with the b × b spectrum in the center as feature block $B^k_{i,j}$, in which i and j represent different spectral coordinates and therefore $1 \leq i, j \leq b$. However, as each 3D trademark consists of multiple images, k represents number of images in the trademark. If the 3D trademark consists of 6 images, then k is set as 6. For the features of the 3D trademark, we use the minimum value of k images in each frequency as the feature, i.e. the feature of the 3D image is T, which is represented by Eq. (3) below and n is set as i × j:

$$T = [t_n] = \min\{B^1_{i,j}, B^2_{i,j}, \ldots, B^k_{i,j}\} \tag{3}$$

Next, the minimum value feature images are combined for matching a query image with all the images in the image database. The matching result is usually close to the query image, as shown in Eq. (4):

$$sim\left(T^1, T^2\right) = \frac{\sum_{i=1}^{n} T_i^1 T_i^2}{\sqrt{\sum_{i=1}^{n} \left(T_i^1\right)^2}\sqrt{\sum_{i=1}^{n} \left(T_i^2\right)^2}} \tag{4}$$

Where, T^1 represents the features of the image set for the query image while T^2 is the features of the image set for each 3D trademark in the multimedia database. The query result is shown by the similarity order.

3.4 Query Process

One set of images for a 3D trademark first undergoes DFT followed by feature extraction. The extracted features are then place into the feature database as the representative features for the 3D image database. During the subsequent query, the query image is compared with the feature data from the feature database for similarity measurement. After the measurement and matching, image features that are close matches to the query image are retrieved.

4 Experimental Results

In this paper, we used the DFT get the features to help retrieving the 3D trademark image. The proposed method can effectively and quickly retrieve the similar trademarks in a large image databases. The simulation experiments are programmed by MATLAB 2013a. In the hardware device, the processor is the Intel E5-2665, memory 32G, operating system Microsoft Windows 10. In the experiment, the MMDB is extracted from the 3D trademark database of the IPO under the MOEA, and contains in 131 3D trademarks that have total of 731 images. Each 3D trademark is rotated 0, 90, 180 and 270 degrees. The rotated 3D trademarks are treated as similar trademarks, which amount to a total of 5848 images. The retrieval result is shown in Fig. 3.

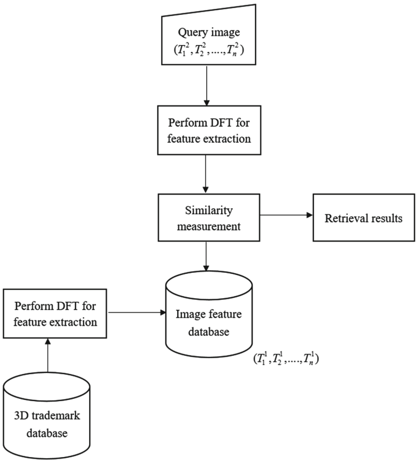

Query image
$(T_1^2, T_2^2, ..., T_n^2)$
Perform DFT for feature extraction
Similarity measurement
Retrieval results
Perform DFT for feature extraction
Image feature database
$(T_1^1, T_2^1, ..., T_n^1)$
3D trademark database

Fig. 3. Image query flowchart

The precision and recall are calculated to observe the retrieval performance. For a query process, let a be the number of retrieval images, b be the number of relevant images in the database, c be the number of retrieval and relevant images, the corresponding formulas are defined as follows (5)–(6):

$$\text{Precision} = \frac{c}{a} \tag{5}$$

$$Recall = \frac{c}{b} \tag{6}$$

There are five cases in the experiments, the size of feature blocks are $5 \times 5, 11 \times 11$, 15×15, 25×25, 50×50, respectively. The value of a is set as b, that is four, in our experiments. However, there are four similar trademarks in our database according to four different rotation degrees for every trademark. Hence, the system will output the top four similar trademarks. The retrieval results are shown in the Fig. 4. The best one is at case of 50×50. The precision and recall values reach 0.95 in the Fig. 5.

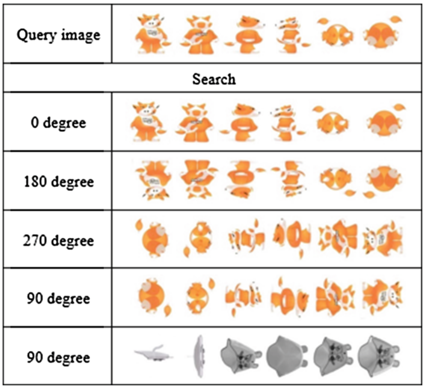


Fig. 4. The retrieval result

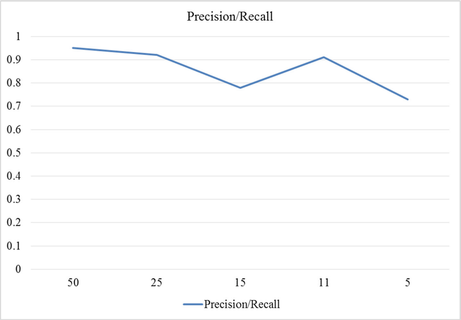


Fig. 5. the precision/recall values of five cases

5 Conclusion and Future Work

For effective retrieval of 3D trademarks, Discrete Fourier Transform was used for extraction of image features. Then, the extracted features were placed in the feature database for comparison with query images during the similarity measurement. Afterwards, the retrieved images were sorted in descending order of similarity values. The experiments were designed with different sizes of feature blocks to observe the precision and recall values. The experimental results proved that our method is efficient. In the real world, the objects will be captured in any angles. Hence, it disturbs for 3D trademark retrieval system. In future, we will explore how to retrieve the similar trademark with malevolence captured angle.

Acknowledgements. This research was supported by the Ministry of Science and Technology of Taiwan under Grant MOST-104-2221-E-324-025.

References

1. Chiang C-C et al (2009) Region-based image retrieval using color-size features of watershed regions. J Vis Commun Image Represent 20(3):167–177
2. Csuka B, Kollár S, Kollár Z, Kovács M (2016) Comparison of signal processing methods for calculating point-by-point discrete fourier transforms. In: 26th conference radioelektronika, 19–20 April, Košice, Slovak Republic, pp 217–221
3. Yang H-Y et al (2015) Color image representation using invariant exponent moments. Comput Electr Eng 46(August):273–287
4. Evgeniou T et al (2003) Image representations and feature selection for multimedia database search. IEEE Trans Knowl Data Eng 15(4):911–920
5. Lee C-H, Lin M-F (2010) Ego-similarity measurement for relevance feedback. Expert Syst Appl 37(1):871–877
6. Lee RCT, et al (2005) The divide and conquer strategy. In: Introduction to the design and analysis of algorithms, Flag Technology, pp 150–152
7. Zhang R, Zhang Z (2006) BALAS: empirical Bayesian learning in the relevance feedback for image retrieval. Image Vis Comput 24(3):211–223
8. Intellectual Property Office, the Ministry of Economic Affairs. https://www.tipo.gov.tw/mp.asp?mp=1

Performance Analysis of Video Transmission Over Wireless Multimedia Sensor Networks

Lin-huang Chang, Tzu-Chieh Lin, and Tsung-Han Lee(✉)

Department of Computer Science, National Taichung University of Education, Taichung, Taiwan
{lchang,thlee}@mail.ntcu.edu.tw, jakejakevbn1218@gmail.com

Abstract. Wireless multimedia sensing network, used to transmit multimedia files over zigbee wireless communication, faces a big challenge in terms of transmission bandwidth and node lifetime. In this paper, we will implement the testbed of 6LoWPAN transmission for video service. We used Contiki OS to implement the wireless zigbee Zigduino nodes. We first simulated the packet loss of video transmission to optimize the packet loss recovery mechanism. Then we conducted the testbed to analyze the performance and feasibility of the wireless zigbee communication for video transmission.

Keywords: WSN · Video transmission · IEEE802.15.4 · 6LoWPAN · GOP

1 Introduction

Wireless sensor networks (WSNs) are spatially distributed sensors to monitor physical or environmental conditions, such as pressure, temperature, sound, and so on. The sensing data are periodically transferred to certain node from different nodes in sparsely populated areas. The development of WSNs was originally motivated by military applications such as battlefield surveillance, they are further used in medical control or health monitoring, industrial automation, and forest fire detection or monitoring. Many wireless sensor devices follow the IEEE 802.15.4 standard [1] which requires smaller size, less power consumption, faster computing and Internet networking capabilities. The IEEE 802.15.4 standard defines the operation of low-rate wireless personal area networks (LR-WPANs). ZigBee [2] is an IEEE 802.15.4-based specification for a suite of high-level communication protocols.

The basic framework of IEEE 802.15.4 conceives a 10-meter communication range with the physical layer bandwidth of up to 250 Kbps. It offers the fundamental lower layers the characteristics which focuses on low-cost and low-power communication between nodes with little to no underlying infrastructure.

With the rapid development of WSN technology, the research of WSN has shifted from sensing and transmitting data, sound or still image to wireless multimedia sensing network (WMSN) which transmits video streaming or multimedia files over zigbee wireless communications. Some challenges, such as transmission bandwidth and node lifetime issues, for the realization of WMSN need to be faced as compared to the transmission of pure data or still images.

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_69

Besides the technologies on the WSN transmission and node design, the video encoding and decoding techniques as well as video recovery mechanisms for the playout are also important issues to evaluate the feasibility of deploying WMSN.

In this paper, we will evaluate the feasibility of video transmission over wireless multimedia sensor networks by implementing a zigbee communications testbed for video quality analysis. We first encode the video source into MPEG format video files followed by the analysis of video group of picture (GoP) for classification of I, P, B frame types. The video frames are then encapsulated into zigbee frame for wireless transmission. Upon receiving the video packets, the client node will conduct the assembly and recovery mechanism for video playout. The recovery mechanism is designed to handle the packet loss issue in WSNs. The video quality will be analyzed and evaluated using peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) index to compare the received video from sending video.

The organization of this paper is as follows: In Sect. 2, we present the related works with the scope of this study on the multimedia encoding and related skills as well as the development of video transmission nodes. The implementation of video transmission based on 6LoWPAN is proposed in Sect. 3. In Sect. 4, we will apply PSNR and SSIM to evaluate and discuss the video performance in WMSNs. Finally, we conclude this research in Sect. 5.

2 Related Standard and Mechanism

WMSNs have been used in medical surgery or healthcare monitoring [3], military applications such as battlefield surveillance [4], and building structure detection [5]. There are also some researches on the video transmission over WSNs [6, 7]. Some researches even focused on the video format analysis and tooling for video quality assessment [8]. In this section, we will review some video encoding format, and video quality measurement tools related to this research. Then, we will also briefly discuss the Contiki OS implemented in our WSN nodes.

2.1 Video Encoding Standard

The Xvid codec is used in this study to encode and decode MPEG-4 video files. In MPEG-4 video encoding, a group of picture (GOP) is defined as a collection of successive pictures within a coded video stream. Each coded video stream consists of successive GOPs which specify the order in which intra- and inter-frames are arranged. A GOP can contain the following picture types. I-frame (intra frame), is equivalent to a fixed video basemap which is coded independently of all other pictures. P-frame (predictive frame) contains motion compensated difference messages from the previous I or P-frame. B-frame (bipredictive frame) contains motion compensated difference information from the previous or later I or P-frame.

In general, the beginning of a GOP is an I frame which contains the full image of the video and does not require any additional information or frames to reconstruct it. The I frame is followed by several P and B frames. One example of GOP structure could be IBBPBBPBBPBBI. In general, the I frame is the most important frame in the video encoding file which can be more editable if there are more I frames in it. However, having more I frames will increase the video encoding rate substantially because I frame size is much larger than P or B frame size.

The I, P, B frames in MPEG-4 encoded video can be determined according to the frame header types. For instance, the I, P, B frame types in MPEG-4 are “00 00 01 B6” followed by “00”, “01”, and “10”, respectively. Therefore, the video packets can be classified into different I, P, B frames according to video header format in the ISO/IEC 14496-2 [9] standard.

2.2 Video Quality Measurement and Tools

The analysis of the quality of video transmission over zigbee wireless communications can be done by evaluating the video quality before and after transmission. That means we can compare the sourcing video with the receiving video signals to evaluate the video quality in WMSNs. Therefore, PSNR and SSIM measurements could be two approaches to achieve this goal.

PSNR is commonly used to measure the quality of reconstruction of image or video encoded files which can also be used to evaluate the video quality of reconstruction before and after transmission over WSNs. When comparing the receiving video (output video) file with the sending video (input video) file, PSNR can reach approximation to human perception of reconstruction quality.

The PSNR (in dB) can be defined as:

$$\mathrm{PSNR} = 10log\frac{(2^n - 1)^2}{MSE} \tag{1}$$

where MSE is the mean squared error of a monochrome image, however, it is the sum over all squared value differences divided by three and be image size for color images with three RGB values per pixel. And, n is the number of the bits per sample which representing one pixel of the image. So, 2^n-1 will be the maximum possible pixel value of the image. For example, when each pixel of the image is represented using 8 bits per sample, 2^n-1 will be 255. For wireless communications, the acceptable PSNR values is typically about 30 dB or above.

The SSIM index is one way to evaluate and predict the perceived quality of digital videos. That is, SSIM can be used to measure the similarity between two images such as receiving (output) and sourcing (input) images. SSIM was designed to improve the inconsistent with human visual perception, such as PSNR method.

The calculation of SSIM index on windows x and y (with common size NxN) is defined as:

$$\mathrm{SSIM(x,y)} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{2}$$

where μ_x and μ_y are the average of x and y, respectively; σ_x^2 and σ_y^2 are the variance of x and y, respectively; σ_{xy} is the covariance of x and y; c_1 and c_2 are two variables to stabilize the division with weak denominator. Typically, c_1 equals to $(0.01*L)^2$ and c_2 equals to $(0.03*L)^2$, where L is the dynamic range of the pixel-values, i.e. 2^n-1.

There are many tools that can be used to evaluate the video quality with PSNR and SSIM data, such as JSVM, SVEF, FFmpeg, etc. In this paper, we use FFmpeg to evaluate the PSNR and SSIM values regarding the video quality over WMSNs.

2.3 Contiki OS on Zigduino – Contiki 2.5 rc1

Contiki is an OpenSource operating system for memory-constrained systems focusing on low-power wireless networks and Internet of Things (IoTs) devices. It can be easily transplanted to any other computer operating system.

Contiki also provides a built-in Internet protocol suite, TCP/IP protocol stack, which only needs 10 kB of RAM and 30 kB of ROM in general. Even a full Contiki system, it only needs 30 kB of RAM to provide a graphical user interface.

In this study, Zigduino, being integrated with zigbee (IEEE 802.15.4) was used as WSN node for Contiki 2.5 rc1 operation and development. Zigduino is compatible with both the Arduino hardware and software. The Zigduino client is capable of sending UDP video packets to Zigduino servers via 6LoWPAN architecture. The designed Zigduino nodes also provides IPv6 header compression, packet fragmentation and reassembly functions.

3 Implementation of Video Transmission Based on 6LoWPAN

3.1 Network Set up

The implemented network setup testbed is shown in Fig. 1 for the video transmission implemented based on 6LoWPAN architecture.

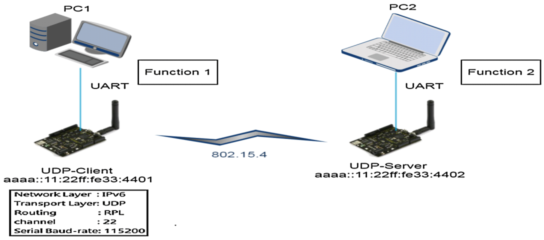


Fig. 1. Video transmission in WSN testbed

In this experiment, we use Akityo [10] as the source video. Xvid encoded is applied for video source coding techniques. We simulate source video from PC1 via Uart interface to Zigduino client by the assist of Function 1 for video processing. Function 1, with flow control comparable to Zigbee transmission speed, provides the capability of analyzing video for different video frame structures. Then, the Zigduino Client sends packets to the Zigduino Server using UDP transmission protocol. Upon receiving the packets, the Zigduino Server forwards the packet to PC2 via Uart interface. Finally, PC2 applies Function 2 to decode and analyze the video.

The wireless sensor network environment in the experiment is set with the IPv6 over Low power Wireless Personal Area Net-works (6LoWPAN) protocol combined with IEEE802.15.4 protocol. The Zigduino Client and Server are with Contiki OS. The related network parameters are shown in Table 1.

Table 1. Network parameters of testbed

Node	IPv6 address
UDP-Client	aaaa::11:22ff:fe33:4401
UDP-Server	aaaa::11:22ff:fe33:4402
Parameter	Value
Channel	22
Routing	RPL
Baud Rate(bps)	115200

In IEEE802.15.4 there are two methods to measure the packet transmission signal strength, received signal strength index (RSSI) and link quality indicator (LQI). The chip used by Zigduino in this research is Atmel 128rfa1. This chip has many RSSI registers. By applying the Eq. (3), we can obtain RSSI value with the register value. Equation (4) converts LQI to RSSI while Eq. (5) converts RSSI to LQI.

$$\mathrm{RSSI}_{\mathrm{dBm}} = \frac{RSSI_{dec}}{2} - RSSI_{offset} \tag{3}$$

$$\mathrm{RSSI} = -(81 - \frac{(\mathrm{LQI} \times 91)}{255}) \tag{4}$$

$$\mathrm{LQI} = 255 \times \frac{RSSI + 81}{91} \tag{5}$$

In our experiments, we conducted the measurements using low-power chips separated in 15 ms with −73 dBm RSSI and 22.65 LQI. We transmitted 1000 packets for five measurements and the average packet loss with 15 m separation is about 0.2%. Similarly, the average packet loss is about 3% for 20 m separation with −79 dBm RSSI and 4.74 LQI.

3.2 Video Codec and Transmission

As mentioned in the previous subsection, we use Akityo [10] as the source video for video transmission in this experiment. The video encoder uses open source video codec library Xvid which follows the MPEG-4 video coding standard. It uses MPEG-4 Part 2 Advanced Simple Profile features such as b-frames, global and quarter pixel motion compensation, and H.263, MPEG and custom quantization matrices. The video parameters using Xvid in this experiment is listed in Table 2.

Table 2. Video parameters using Xvid

Video	Akiyo
Enoding	MPEG-4
Frame rate	30 fps
Format	QCIF 176*144
Bit rate	32000 bps
Video frames	300
Size	36.7 kb

Each video frame has a start code value regarding the MEPG-4 file encoded by Xvid encoder. For example, the start code value of one P-frame in hexadecimal is 51 53 54 56 57 59 5A 5C 5D 68. Figure 2 shows a schematic diagram of the video sequence. In this paper, we call the frames sorted from #1 to #30 as a Block, and the frames sorted from #0 to #299 or #300 to #599 as a Cluster. Then, the whole video frames is a Group of picture (GoP). The Akiyo video, studied in this research, has a GoP with 300 frames.

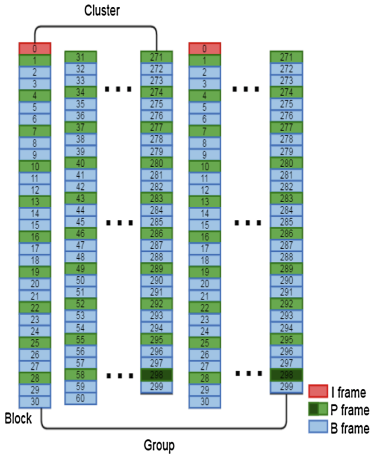


Fig. 2. Video sequence schematic diagram

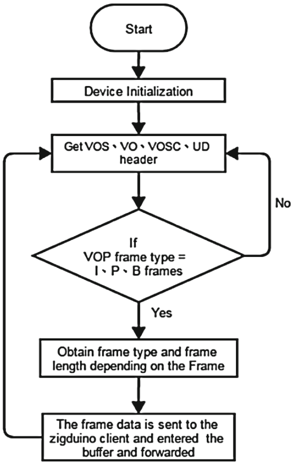

Start
Device Initialization
Get VOS、VO、VOSC、UD header
If VOP frame type = I、P、B frames
No
Yes
Obtain frame type and frame length depending on the Frame
The frame data is sent to the zigduino client and entered the buffer and forwarded

Fig. 3. Video analysis flow chart of Function 1

Function 1 and Function 2 algorithms are designed to encode/decode and processing video data and frames. Function 1 analyze and extract I, P or B frames from video headers and divide them into 100-Byte individual packet for zigbee wireless transmission. The video analysis flow chart of Function 1 is illustrated in Fig. 3.

Upon obtaining the video header, the algorithm determines the frame types and lengths, such as I, P or B format, of VOP. Calculate the frame length of each I, P or B frame and packetize them into 100-Byte Zigduino-Client buffer via the Uart interface.

Upon receiving packets from Zigduino-Client, Zigduino-Server sends the video packets to PC 2 via the Uart interface. The Zigduino operation flow chart is shown in Fig. 4.

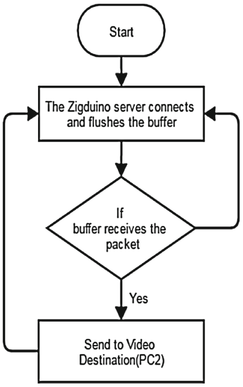

Start
The Zigduino server connects and flushes the buffer
If buffer receives the packet
No
Yes
Send to Video Destination(PC2)

Fig. 4. Zigduino operation flow chart

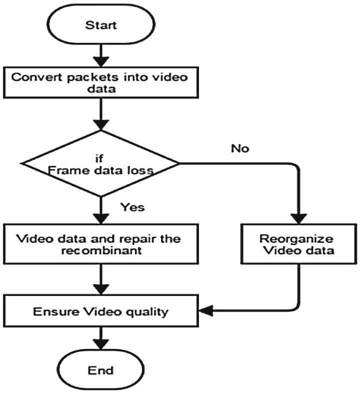


Fig. 5. Video packets recovery and assembly

Upon receiving the video packets from Zigduino-Server, Function 2, designed and embedded in PC 2, assembles and converts them for video playout, as shown in Fig. 5. In the meanwhile, Function 2 will reconstruct the lost frames by using the recovery mechanism which will be discussed in the next section.

The experimental result and analysis of the video quality and effectiveness of the video transmission over zigbee wireless communication will be evaluated in the next section.

4 Video Quality and Effectiveness Assessment

In this section, we first discuss the video quality in terms of PSNR and SSIM values for different packet loss rate. Then, we implement a testbed for video transmission over WSNs. We investigate the optimized recovery mechanism for video reconstruction due to the packet loss over wireless communications. By using the designed recovery mechanism, we will evaluate the feasibility of video transmission over WSNs.

4.1 Recovery from Different Packet Losses

For video transmission, packet loss problem may not be as significant as that of data transmission. However, large packet loss may also cause a problem for video reassembly and recovery. Due to the delay constraint, the retransmission of the lost packet for video playout is mostly infeasible. Therefore, the recovery from the packet loss of video frames is one important approach for this study.

First in this section we simulated the packet loss randomly with different rates from 1% to 20%. For Akityo video, each Block has 10 PBB frame sets and totally there are 30 Blocks. Table 3 shows one example of experimental result for different lost Block number regarding 1% packet loss case. As illustrated in Table 3 the PSNR value of Block number 1 frame loss is 31.8955 is much lower than those of Block number 5 or 10 frame loss. Similar results are found for different Block numbers as compared with Block number 1. The SSIM results also reach the same conclusion in which Block number 1 plays much important role in video encoding as compared with other Block numbers in the same Cluster.

Table 3. 1% packet loss rate simulation

Block Number	1	5	10
PSNR (dBm)	31.8955	42.2723	51.0091
SSIM (≒1,perfect)	0.96079	0.99653	0.99933

From the experimental results, we conclude that if the first PBB frame set is lost, we recover it from the second PBB frame set. When the lost packet is any one of the P-frame or B-frame except the first PBB frame set, we then recover the lost frames from the first PBB frame set.

This recovery mechanism is used for the rest of the experiments. The experimental results of the recovered video PSNR and SSIM values for different packet losses are shown in Table 4.

Table 4. 3%~20% packet loss rate simulation

Packet loss rate(%)	3	5	10	20
PSNR (dBm)	31.4558	30.4701	29.2637	27.1645
SSIM (≒1,perfect)	0.95685	0.94769	0.93293	0.89792

For 3% and 5% packet losses, the PSNR (SSIM) values are 31.4558 (0.95685) and 30.4701 (0.94769), respectively. These values are acceptable for video playout. However, for packet losses as high as 10% and 20%, the PSNR (SSIM) values are 29.2637 (0.93293) and 27.1645 (0.89792) which are below the acceptable level.

The experimental results conclude that the video transmission with less than 10% packet loss would be acceptable for the proposed frame recovery mechanism.

4.2 Video Quality Measurements for Video Transmission Testbed

In this section, we will conduct the real measurements for the video transmission over zigbee testbed as shown in Fig. 1. The Zigduino Client and Server nodes are separated in 15 m. The Akityo video, including I, P, and B frames, are packed into 370 100-Byte packets with corresponding frame header.

For the first test measurement, we again evaluate the recovery mechanism using different recovered frames. Figure 6 shows three different mechanisms for lost frames recovery. The white-color frames are assumed to be lost while the colored frames are received frames with red, green and blue colors corresponding to I, P, B frames, respectively. Figure 6(a) deploys mechanism 1 by using the first PBB frame set to recover from all other lost PBB frames. The PSNR (SSIM) result, listed in Table 5, for Mechanism 1 is 34.6492 (0.97858) which is acceptable even the packet loss rate for PBB frame set is as high as 90%.

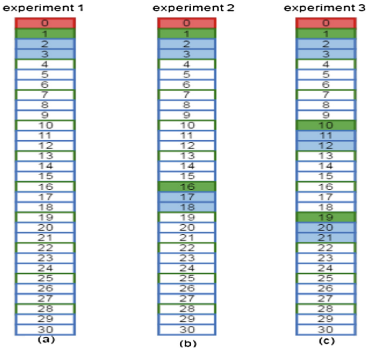


Fig. 6. Different mechanisms for lost video frames recovery

Table 5. Video quality result for different recovery mechanisms

Mechanism	1	2	3
PSNR (dBm)	34.6492	29.0809	31.8709
SSIM (≒1,perfect)	0.97858	0.92685	0.95711

Figure 6(b) deploys mechanism 2 by using two PBB frame sets to recover from all other lost PBB frames. The PSNR (SSIM) result, listed in Table 5, for Mechanism 2 is 29.0809 (0.92685) which is below the acceptable level. This result suggest that the first PBB frame set is the most significant PBB frame set in the same Block. This is also confirmed in Table 5 for Mechanism 3, illustrated in Fig. 6(c), which utilizes three PBB frame sets to recover from all other lost PBB frames.

The video playout results for different recovery schemes are demonstrated in Fig. 7. Figure 7(a) shows the recovery of Mechanism 1 using the first PBB frame set. This result demonstrates the overall integrity of the video and does not produce any problem of missing video or blur.

(a) (b)

Fig. 7. Different recovery schemes for video playout

As comparison, Fig. 7(b) shows the recovery of the lost PBB frame using the previous PBB frame set. As seen clearly that there are some missing blocks of video image, such as the white and blur portion of the lower right corner of the video which significantly affect the quality of the video playout.

From the simulation and testbed experimental results, the video quality is above acceptable for cases of overall packet loss less than 10% or PBB frames packet loss as high as 90% once the first PBB frame set is received. We have provided the recovery mechanism and showed the feasibility of the video transmission over zigbee wireless communication.

5 Conclusion

In this paper, we have implemented the testbed of 6LoWPAN transmission for video service. The Contiki OS was used to implement the wireless zigbee Zigduino node. We have simulated the packet loss of video transmission to optimize the packet loss recovery mechanism. From the simulation and testbed experimental results, we have shown that the video quality was above acceptable for cases of overall packet loss less than 10% or PBB frames packet loss as high as 90% once the first PBB frame set was received. The designed recovery mechanism has provided and shown the feasibility of the video transmission over zigbee wireless communication.

For future research, we will replace the PC nodes by embedded systems which are capable of processing all video encoding, frame type classification and assembly, as well as video recovery and reconstruction. Also, we will extend the testbed to multi-hopping WSNs. The complete performance analyses of WMSNs will also be conducted and analyzed.

References

1. IEEE standard for Information Technology, IEEE std. 802.15.4, Part. 15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (LR-WPANs), October 2003
2. ZigBee Standards Organization, ZigBee Specification, September 2012
3. Feller SD, Zheng Y, Cull E, Brady DJ (2002) Tracking and imaging humans on heterogeneous infrared sensor arrays for law enforcement applications. In: Proceedings of SPIE Aerosense, Orlando, USA, pp 212–221, April 2002
4. Bekmezci I, Alagoz F, (2005) New TDMA based sensor network for military monitoring (MIL-MON). In: Proceedings of IEEE military communications conference, Atlantic City, USA, pp 2238–2243, October 2005
5. Lee R, Chen K, Chiang S, Lai C, Liu H, Wei MS, (2006) A backup routing with wireless sensor network for bridge monitoring system. In: Proceedings of the communication networks and services research conference, Moncton, Canada, pp 161–165, May 2006
6. Angayarkanni V, Akshaya V, Radha S (2016) Distributed compressive video coding using enhanced side information for WSN. In: International conference on wireless communications, signal processing and networking (WiSPNET). IEEE

7. Abdollahzadeh M, Ghazijahani HA, Seyedarabi H Quality aware HEVC video transmission over wireless visual sensor networks
8. Baziz Y, Maimour M, Kechar B (2014) EvalVSN: a new tool for video quality evaluation in wireless sensor networks. In: 2014 International conference on multimedia computing and systems (ICMCS). IEEE
9. ISO/IEC JTC 1/SC 29/WG 11 Coding of Audio-Visual Objects - Part 2 (2001) Visual in ISO/ IEC 14496–2:2001, Geneva, Switzerland
10. Akiyo. https://media.xiph.org/video/derf/
11. El Dien ME, Abd Youssif AAA, Ghalwash AZ (2016) Energy aware and adaptive cross-layer scheme for video transmission over wireless sensor networks. IEEE Sens J 16(21):7792–7802
12. Ansari, AW et al (2014) ARM based real time video streaming using XBee for perimeter control in defense application. In: 2014 International conference on computing for sustainable global development (INDIACom). IEEE
13. de Dominguez HJO et al (2014) The H. 264 video coding standard. IEEE Potentials 33(2): 32–38
14. Farooq, MO, Kunz T, St-Hilaire M (2011) Cross layer architecture for supporting multiple applications in wireless multimedia sensor networks. In: 2011 7th international wireless communications and mobile computing conference (IWCMC). IEEE
15. Xiao, H et al (2011) A data collection system in wireless network integrated WSN and ZIGBEE for bridge health diagnosis. In: 2011 proceedings of SICE annual conference (SICE). IEEE
16. de Miranda R, Danilo C et al (2015) Design of objective video quality metrics using spatial and temporal informations. IEEE Latin America Transactions 13(3):790–795

Image Steganographic Method Based on Pencil-Shaped Pattern

Chin-Feng Lee(✉), Ying-Xiang Wang, and An-Tong Shih

Department of Information Management, Chaoyang University of Technology,
Taichung 41349, Taiwan, R.O.C.
lcf@cyut.edu.tw, ss6228g@yahoo.com.tw, ket12579@gmail.com

Abstract. Image Steganography is a covert communication to hide the existence of messages into images from a third party. High embedding capacity, good visual quality and security are three significant essentials. In the proposed method, every secret digit is embedded into each cover pixel pair based on a magic matrix with pencil-shaped patterns to obtain higher embedding capacity while preserving good image quality. Experimental results show that the proposed scheme ensures higher embedding capacity of 2 bits per pixel and PSNR of 44.7 dB on average compared with existing schemes, while ensuring security by scrambling the secret message with a secure key.

Keywords: Steganography · Data hiding · Embedding capacity · Visually quality

1 Introduction

Image steganography is a special branch of information hiding where a secret message is embedded in a cover image based on a shared stego-key, resulting in a stego-image. So, image steganography conceals the existence of a message for the purpose of secret communication. The receiver extracts the message if he/she has knowledge of the positions where secret data has been embedded. Since only minor modifications are made in the embedding process, the sender assumes that it is not aware of the message existence by an attacker. A successful steganographic method considers three aspects: embedding capacity, imperceptibility and security. Generally, a trade-off exists between embedding capacity and visually quality.

In 1989, Turner [1] introduced the method of least-significant bit (LSB) replacement that modifies the least-significant bit for data hiding. The LSB substitution [2] approach is very simple by performing the substitution operation by directly replacing the LSBs of the cover image with the message bits and generates only slight distortion but maintains high image quality. However, the LSB method can be easily detected by the steganography analytical methods. In 2006, Zhang and Wang [3] proposed an embedding scheme by exploiting the modification direction. This scheme also called EMD method, which first converts secret messages into a sequence of digits in an n-ary notational system. Then, each secret digit is embedded into a group of n cover pixels, and one of the cover pixels is at most increased or decreased by one. The best embedding capacity can achieve 1 bpp in the 5-ary notational system. In addition, the EMD embedding

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_70

scheme can provide the scalable degree of visual quality of stego-image along with the system parameter *n*.

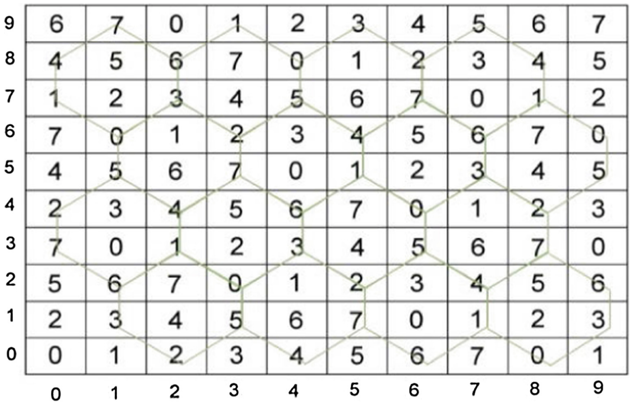
Inspired by the EMD method, Chang *et al.* [4] proposed a data hiding scheme based on Sudoku solutions to modify every cover pixel pairs with a help of a special magic matrix from a selected Sudoku solution to conceal secret data. The method requires less computational cost with acceptable visual quality of the stego-images. In 2014, a turtle-shaped method is proposed by Chang *et al.* [5] to enhance the embedding capacity of 1.5 bpp (bits per pixel) while preserving good visual perception of the stego-image with PSNR values of 53.3 dB on average.

This paper based on a pencil-shaped pattern is proposed to further enhance the embedding capacity based on the improvement of the turtle-shaped method. The pencil-shaped pattern contains 16 different digits from 0 to 15. Our method modifies pairs of cover pixels according to a magic matrix with each dimension ranging from 0 to 255 to achieve data hiding. The embedding capacity of proposed scheme achieves 4 bits for every pixel pairs while preserving high stego-image quality with PSNR values of 44.7 dB on average.

The rest of the paper is organized as follows. Section 2 gives a brief literature review. Section 3 describes the proposed data hiding scheme. Experimental results and the conclusions are discussed in Sects. 4 and 5, respectively.

2 Related Works

In this section, a data hiding scheme based on the concept of turtle-shaped shells is proposed by Chang *et al.* [5]. Assume that the secret data is represented by a binary stream. Each time, 3 secret bits are read from the secret message and converted a secret digit from 0 to 7. Figure 1 illustrates an example of the matrix *M* with the size of 256 × 256 based on turtle shells. The matrix *M* is designed according two rules: the value difference between the two adjacent elements in the same row of the matrix *M* is set to



9	6	7	0	1	2	3	4	5	6	7
8	4	5	6	7	0	1	2	3	4	5
7	1	2	3	4	5	6	7	0	1	2
6	7	0	1	2	3	4	5	6	7	0
5	4	5	6	7	0	1	2	3	4	5
4	2	3	4	5	6	7	0	1	2	3
3	7	0	1	2	3	4	5	6	7	0
2	5	6	7	0	1	2	3	4	5	6
1	2	3	4	5	6	7	0	1	2	3
0	0	1	2	3	4	5	6	7	0	1
	0	1	2	3	4	5	6	7	8	9

Fig. 1. An example of the magic matrix *M* based on turtle shells.

"1", and the value difference between the two adjacent elements in the same column is set alternately to "2" and "3".

Let $M(p_i, p_{i+1})$ and $M(p'_i, p'_{i+1})$, indicate the digits in the magic matrix M where (p_i, p_{i+1}) and (p'_i, p'_{i+1}) stand for the pixel pairs before/after secret embedding, respectively. A secret digit ranging from 0 to 7 is hidden in every pixel pair. If $M(p_i, p_{i+1})$ falls within a turtle shell, then the digit $M(p'_i, p'_{i+1})$ same as the to-be-embedded secret data s_i also can be found within the same shell. However, If $M(p_i, p_{i+1})$ is an edge digit, then the digit $M(p'_i, p'_{i+1})$ same as the secret data s_i can be found from the neighboring three turtle shells, with the minimum distance between (p_i, p_{i+1}) and (p'_i, p'_{i+1}). A special case occurs when the digit $M(p_i, p_{i+1})$ is located outside any shell. The solution also can be solved by finding the shortest distance between (p_i, p_{i+1}) and (p'_i, p'_{i+1}) under the condition that $M(p'_i, p'_{i+1})$ equals to the secret data s_i.

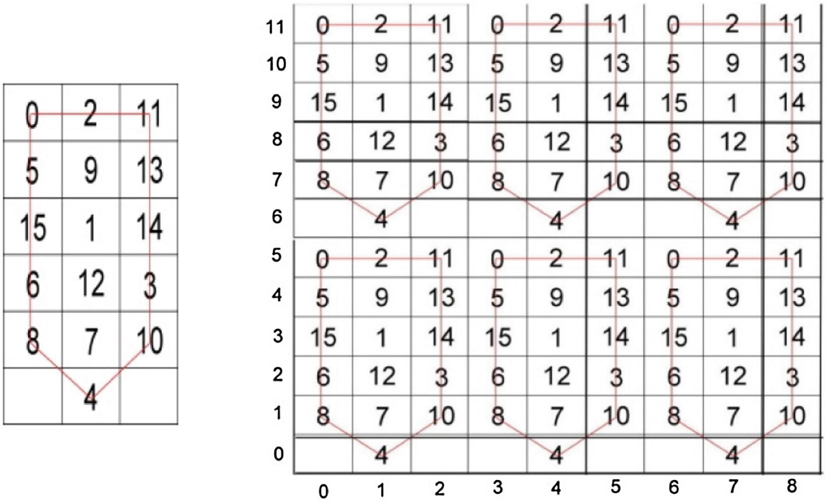
Method [5] is undoubtedly a great idea. However, there is a limitation in embedding payload. This paper based on a pencil-shaped pattern is proposed to further enhance the embedding capacity.

3 Proposed Scheme

A magic matrix based on pencil-shaped pattern of 16 different digits is first constructed. 4 secret bits, corresponding to one of the digits from 0 to 15, are read from the secret message and embedded into each cover pixel pair according to the matrix. Detailed description of the embedding and extraction procedures is provided as follows. To further improve the security of the method, the secret message is shuffled using a secure key SK1 created by a secure pseudo-random number generator (PRNG) before the embedding process. The scrambled secret message is then embedded onto the cover image using the magic matrix with pencil-shaped patterns. With the keys including the secure key and stego-keys, the secret message can be extracted from the exact stego-pixel pairs.

3.1 Embedding Procedure

(1) *Construction of magic matrix:* First, create a square of size 256×256. Second, use a stego-key SK2 to generate a pencil-shaped pattern contains 16 different digits, ranging from 0 to 15. There are 16! patterns and one of the pencil-shaped patterns is shown in Fig. 2(a). Third, use another key SK3 as a start location and repeat the pattern from left-to-right and top-to-bottom and in a non-overlapping matter to create a pattern collage like Fig. 2(b). For each remainder cell, use the random key SK4 to fill a digit from 0 to 15. Accordingly, a magic matrix M corresponding three stego-keys are constructed.



0	2	11
5	9	13
15	1	14
6	12	3
8	7	10
	4	

11	0	2	11	0	2	11	0	2	11
10	5	9	13	5	9	13	5	9	13
9	15	1	14	15	1	14	15	1	14
8	6	12	3	6	12	3	6	12	3
7	8	7	10	8	7	10	8	7	10
6		4			4			4	
5	0	2	11	0	2	11	0	2	11
4	5	9	13	5	9	13	5	9	13
3	15	1	14	15	1	14	15	1	14
2	6	12	3	6	12	3	6	12	3
1	8	7	10	8	7	10	8	7	10
0		4			4			4	
	0	1	2	3	4	5	6	7	8

Fig. 2. (a) (b) An example of magic matrix M based on pencil-shaped patterns.

Assume that the grayscale cover image P sized $H \times W$ is composed by $P = \{p_i | \, i = 1, 2, \ldots, (H \times W); 0 <= p_i <= 255$ and p_i is an integer$\}$. To embed the secret digits $S = \{s_j | j = 1, 2, \ldots, (H \times W)/2; 0 <= s_j <= 15$ and s_j is an integer$\}$, the location of each pixel pair (p_i, p_{i+1}) will be determined as $M(p_i, p_{i+1})$ in the magic matrix M, where p_i and p_{i+1} are the column value and row value, respectively.

(2) *Embedding Algorithm:* After the generation of the magic matrix M, all the elements in M can be classified as pattern/non-pattern elements, referring to those elements that fall inside/outside any of the pencil-shaped patterns. For a given to-be-embedded secret digit s_j, associated with a cover pixel pair (p_i, p_{i+1}), the stego-pixel pair (p'_i, p'_{i+1}) is generated by the embedding rules R1 and R2 described as follows.

R1: If $M(p_i, p_{i+1})$ belongs to the set of pattern elements, the stego-pixel pair can be obtained from $M(p'_i, p'_{i+1})$, where $M(p'_i, p'_{i+1}) = s_j$ and (p'_i, p'_{i+1}) and (p_i, p_{i+1}) belong to the same pencil-shaped pattern

R2: If $M(p_i, p_{i+1})$ does not involve in any pencil-shaped pattern, the stego-pixel pair comes from $M(p'_i, p'_{i+1})$, where $M(p'_i, p'_{i+1}) = s_j$ with the minimum Euclidean distance between (p'_i, p'_{i+1}) and (p_i, p_{i+1})

3.2 Extraction Procedure

Each stego-pixel pair (p'_i, p'_{i+1}) is mapped onto the element $M(p'_i, p'_{i+1})$ in the magic matrix M as used in the embedding procedure. The embedded secret message is extracted correctly from the stego-image. Since the secret data obtained will be a shuffled version of the original secret message, the receiver can use the key SK1 to de-shuffle the message to get the original one.

4 Experimental Results

Four 8-bit grayscale images, Barbara, Boat, Lena and Peppers, with sizes of 512 × 512 are taken as test images as shown in Fig. 3. In our experiments, the secret messages are randomly generated. The performance efficiency can be analyzed in terms of embedding capacity (EC) and peak-signal-to-noise-ratio (PSNR). A larger PSNR indicates that the quality of the stego-image is closer to the original one.

(a)Barbara (b)Boat (c)Lena (d)Peppers

Fig. 3. Four grayscale image of 512 × 512 size are taken as test images.

$$\text{PSNR} = 10log_{10}\frac{255^2}{MSE},$$

where MSE is the Mean Square Error between cover image and stego-image. The MSE is defined as follows:

$$\text{MSE} = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}(p'_{(i,j)} - p_{(i,j)})^2.$$

Here $p_{(i,j)}$ and $p'_{(i,j)}$ stand for the pixel values of cover image and stego-image, respectively. The embedding capacity (EC) is measured by the "bit per pixel" (bpp) which means the total number of secret bits that can be carried by an image pixel.

Table 1 shows the comparisons of the proposed scheme with two related works [4, 5] in terms of the embedding capacity as well as the visual quality of the stego-image. The proposed method has the embedding capacity of 2 *bpp* which is the highest among the three methods. However, the average PSNR value of proposed scheme is lower than the other two methods. In general, the PSNR value larger than 30 dB is acceptable, because the distortion on the stego-image is hard to be detected by human eyes.

Table 1. Comparison of image quality (PSNR) and embedding capacity (EC)

	Sudoku method [4]		Turtle-shaped method [5]		Proposed method	
	PSNR	EC	PSNR	EC	PSNR	EC
Barbara	46.25	1.5	53.34	1.5	44.73	2
Boat	46.24	1.5	53.29	1.5	44.69	2
Lena	46.24	1.5	53.31	1.5	44.72	2
Peppers	46.20	1.5	53.32	1.5	44.71	2

5 Conclusions

In this paper, a pencil-shaped pattern based scheme for data hiding is proposed to enhance the embedding capacity while guaranteeing good visual quality of the stego-image. Based on this magic matrix, each cover pixel pair can carry a four-bit secret data. Therefore, the embedding capacity can achieve 2 bits per pixel. By using the pencil-shaped pattern matrix, the embedded secret data can be extracted exactly from the stego-image. Experimental results demonstrate that the proposed scheme, in comparison with previous schemes, outperforms in embedding capacity and security, but might involve more distortion and complicated key management. In the future, a good design of magic matrix should be conducted to increase the PSNR without losing high embedding capacity.

Acknowledgment. This research was partially supported by the Ministry of Science and Technology of the Republic of China under the Grants MOST 105-2221-E-324-014.

References

1. Turner LF (1989) Digital data security system. Patent IPN, WO89/08915
2. Chan CK, Cheng LM (2004) Hiding data in images by simple LSB substitution. Pattern Recogn 37:469–474
3. Zhang X, Wang S (2006) Efficient steganographic embedding by exploiting modification direction. IEEE Communication Letters 10(11):781–783
4. Chang CC, Chou YC, Kieu TD (2008) An information hiding scheme using Sudoku. In: Proceedings of the computing, information and control (ICICIC 2008), Homburg, Germany
5. Chang CC, Liu Y, Nguyen TS (2014) A novel Turtle shell based scheme for data hiding. In: Proceedings of the tenth international conference on intelligent information hiding and multimedia signal processing, Kitakyushu, Japan

Wind Direction and Speed Estimation for Quadrotor Based Gas Tracking Robot

Kok Seng Eu(✉), Wei Zheng Chia, and Kian Meng Yap

Faculty of Science and Technology, Sunway University, Bandar Sunway, Selangor, Malaysia
{12058889,15013626,kmyap}@sunway.edu.my

Abstract. In gas extraction sites, the incidents of gas leaking poses a damage to workers on site. The protection for the workers is essential. However, due to the colorless and odorless nature of natural gas, it is difficult for humans to identify leaks. This paper proposes a quadrotor based gas tracking robot to be used in hazardous gas localization areas, and specifically, to detect methane leak from gas extraction sites. For the quadrotor to fulfill this purpose, it requires the ability to detect wind direction and speed, an endowment that commercial quadrotors lack. The need to detect wind direction and speed stems from the fact that gas plumes travel downwind, but the quadrotor needs to find the source of the leak, and hence, must determine the upwind direction to locate the source. In order to equip the quadrotor with the above skills, a wind direction and speed estimation algorithm based on Euler angles-velocity vectors has been proposed. For comparison purposes, we compared the proposed method with a generic ultrasonic wind sensor. We concluded that the proposed method achieves an error percentage as low as 10.73% for wind speed, and 9.09% for wind direction estimation. Thus, the algorithm is a significant addition to the quadrotors' capabilities, enabling the quadrotor to trace upwards, against the traversal of the gas plume, and carry out accurate calculations.

1 Introduction

According to Statista [1], the United States alone consumed 27.47 trillion cubic feet of natural gas in 2015. Oil and natural gas is an important industry that affects not only private corporations, but tens of millions of middle class Americans that own shares in oil and gas companies. However, as much money there is involved in the sector, there is something even more invaluable: human lives. According to Environmental Science and Technology, the United States of America loses 17 lives and 133 million USD in property damage annually, due to incidents involving natural gas pipelines. Natural gas, which consists of 75% methane, is hazardous and could cause health problems such as carbon monoxide poisoning. Methane leakage is so serious that the city of Boston alone had at least 1,868 documented unrepaired leaks in its gas lines as of March 2015, and the oldest has been leaking since 1985 [2]. The contribution of methane to global warming is also significant, as it absorbs more heat than carbon dioxide. On a timescale of 20 years, the effects on global warming are about 84 times more potent than carbon dioxide's.

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1_71

With environmental health and citizens' safety in mind, a quadrotor based gas tracking robot is proposed to solve this problem. The quadrotor is expected to determine the source of the leakage, so as not to expose a human environmental officer to the hazardous gas for long periods of time.

Gas leakage is usually attributed to pipeline leakages, but as shown in Fig. 1, gas leakage could be from the ground surface with some distance away from the gas extraction site. It is difficult to be detected by humans, even with mercaptan added to the gas to give it a distinct odor. Therefore, the best way to detect the source is not by sense of smell, but rather, through accurate calculations and measurements which could be carried out by the quadrotor [4].

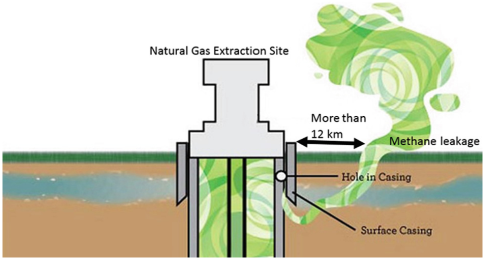


Fig. 1. Methane leaks could be certain distance away from gas extraction site [3]

The direction of gas leakage is usually downwind, along the direction of the wind. However, to be able to locate the source, the quadrotor has to trace against the gas plume and wind direction. To be able to determine which direction is upwind, one must first determine which direction is downwind. Intuitively, this is an easy task for a person who is standing in the wind and can easily determine in which direction the wind is flowing. However, for commercial quadrotors that are not blessed with intuition, they are not equipped with the ability to detect wind. Even if we were to attach a wind sensor onto the quadrotor, there are a few problems that may arise. There are two main problems. Firstly, the add-on wind sensor is unable to differentiate between environment wind and the disturbance caused by the quadrotors' propellers. The airflow generated by the propellers causes the wind sensor data to be inaccurate, and thus, the wind sensor is unable to tell the environments' wind direction and speed. Secondly, wind sensors are bulky and heavy, and for the quadrotor to carry it around while tracking the source of the leak is cumbersome and impractical. Therefore, the objective of the paper is to find a new way to enable the quadrotor to estimate wind direction and speed without the sensor.

In this paper, a wind direction and speed estimation algorithm is proposed. Based on Euler angles and velocity vectors obtained from the quadrotor, an algorithm can be derived to calculate the wind direction and speed. The data required for the algorithm to work is not collected from a wind sensor, but rather, from the quadrotors' built-in Inertial Measurement Unit (IMU) sensor. The IMU sensor is a combination of an accelerometer and a gyrometer. The purpose of the IMU sensor is to measure the quadrotors' orientation, roll, pitch, yaw, speed, and acceleration.

2 Related Work

Due to the fact that gas molecules are highly separated, their movement is highly dependent on the wind flow surrounding them. Therefore, in order to track the gas leak, collecting intel on the surrounding wind information is a crucial component. As mentioned above, since the odour plumes will be carried along the downwind direction, the sniffer robot will have to track along the upwind direction to find the source of the leak, as this will increase the chances of finding the location. This method is called the Anemotaxis-upwind approach. In the last decade, researchers used wheeled mobile robots to trace the gas plume. In order to obtain wind data, researchers mounted wind sensors on their wheeled based gas tracking robots. To the best of the authors' knowledge, Russell and Kennedy [5] were the first to propose a gas tracking robot with an add-on wind sensor.

The design of this anemometer is shown in Fig. 2 and it can work even in low airflow conditions by measuring the changes in the torque of the motor that was caused by the force exerted on the flat paddle by the airflow.

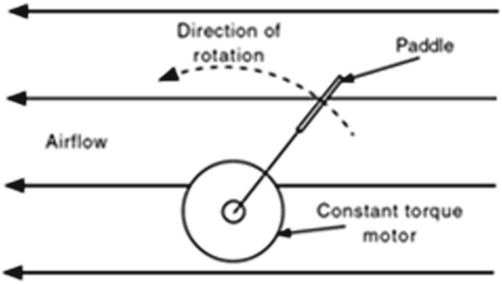


Fig. 2. Wind detection method proposed by Russell and Kennedy [5]

In recent years, researchers looked to improve the accuracy of wind sensing, thus, they proposed to mount a commercial wind sensor unto the wheeled based gas tracking robot as shown in Fig. 3 (ultrasonic wind sensor) and Fig. 4 (3-axis ultrasonic wind

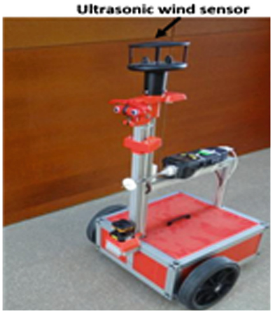


Fig. 3. Ultrasonic wind sensor mounted on wheeled base gas tracking robot [6]

sensor [6, 7]. This ultrasonic wind sensor has an accuracy of 2% @ 12 m/s, with a range of 0–60 m/s.

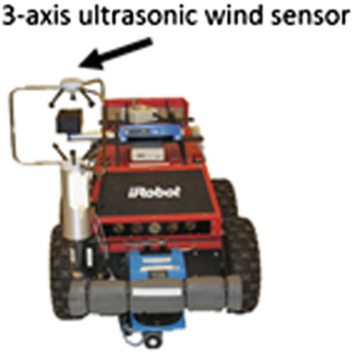


Fig. 4. Three-axis ultrasonic wind sensor mounted on wheeled base gas tracking robot [6]

Although the ultrasonic wind sensor has its merits, it cannot be applied to a quadrotor based gas tracking robot due to the reasons mentioned in the above section. Henceforth, the proposed method of this paper will be used to solve the mentioned problems.

3 Experiment Platform

For this experiment, the DJI Phantom 4 is used as the gas tracking robot. DJI Phantom 4 as shown in Fig. 5 is now the leading product in the quadrotor market, with DJI pulling in 1.47 billion USD in revenue as of December 2016. Meanwhile, their closest competitors, 3D Robotics, generated just 50 million USD in sales.

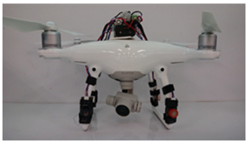


Fig. 5. DJI Phantom 4 armed with gas sensors.

In order to evaluate the performance and accuracy of the proposed method, an ultrasonic wind sensor from Gill Instruments Inc. as shown in Fig. 6 was used to measure the wind velocity up to 60 m/s, with an accuracy of 2% at 12 m/s and the wind direction in 360°. The results were used to compare between the performances of proposed method. The measurements will then be transmitted wirelessly to the central PC server.

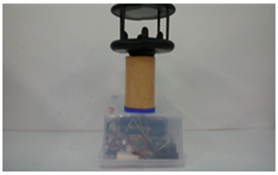


Fig. 6. Ultrasonic wind sensor with wireless transmission

From Fig. 7(a), the ultrasonic wind sensor is put at the center to measure wind from various angles and varying speed. After that, the ultrasonic wind sensor will be removed and taking its place will be the quadrotor as shown in Fig. 7(b).

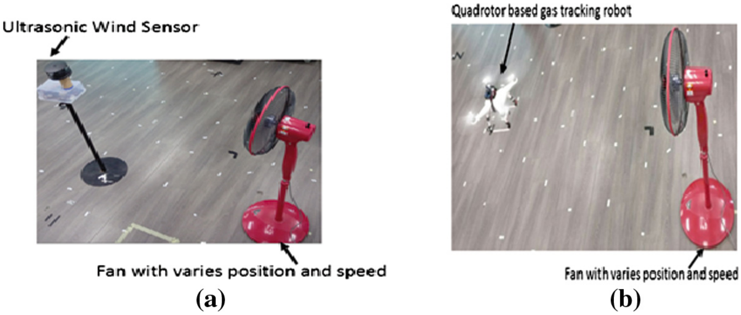

Ultrasonic Wind Sensor
Fan with varies position and speed
(a)
Quadrotor based gas tracking robot
Fan with varies position and speed
(b)

Fig. 7. Experiment setups, (a) the ultrasonic wind sensor at the center and (b) the quadrotor at the center, with varying angle degrees and positions and varying fan speed.

4 Proposed Method

The DJI phantom 4 has a very stable hover ability, maintaining its position firmly during flight despite disturbance of environment wind or drifting of IMU sensor. The hover function is based on an image processing technology: the optical flow algorithm. The DJI Phantom 4 has two pairs of vision sensors which were meant to aid in the execution of the optical flow algorithm for hovering purposes.

If there is wind blowing from a certain angle onto the DJI Phantom 4, it will tilt its roll (γ) and pitch (β) angles to compensate against the external disturbance from the wind. The stronger the force of the wind, the more the quadrotor will tilt the angle of roll and pitch. Hence, the response of the quadrotor towards wind force is a linear proportional response.

Based on the hypothesis above, we can derive the wind speed (w_s) formula by using the resultant vector of roll (γ) and pitch (β) angles, and then interpolate with parameters m and c as shown in Eq. (1). To derive the wind direction (w_{dir}) formula, we can apply four-quadrant inverse tangent (atan2) to compute the principal value of the arc tangent of β/γ, expressed in radians. We cannot apply inverse tangent function (tan^{-1}) only because inverse tangent function (tan^{-1}), as it is sign ambiguity where we cannot determine with certainty in which quadrant the angle lies. There is one condition for Eq. (1)

to be valid. According to the hypothesis, the quadrotor must be in hovering status, and therefore, the resultant velocity (v_r) of quadrotor must be equal to zero as shown in Eq. (2).

$$\begin{pmatrix} W_s \\ W_{dir} \end{pmatrix} = \begin{pmatrix} \left(\sqrt{\beta^2 + \gamma^2}\right) Xm + c \\ a\,tan2(\beta/\gamma) \end{pmatrix}, \; if \; v_r = 0 \tag{1}$$

$$v_r = \sqrt{v_x^2 + v_y^2 + v_z^2} \tag{2}$$

Where,

w_s = Wind Speed
w_{dir} = Wind Direction
β = pitch angle of quadrotor
γ = roll angle of quadrotor
v_r = resultant velocity of quadrotor
$v_{x,y,z}$ = velocity vector

5 Result and Discussion

Figure 8 shows how different experiment scenarios have been designed and conducted. Each experiment has a different distance, angle, and fan speed. The result is shown in Table 1. The maximum error percentage for wind speed as compared with the ultrasonic wind sensor and the proposed method is 10.73% whereas the maximum error percentage for wind direction is 9.09%. Wind direction is more important than wind speed when applying this information for gas tracking purposes. If a quadrotor has knowledge of the wind direction but not of the wind speed, it would still be able to continue its gas tracking tasks. Even though the data collected with the proposed method is not as accurate as the data collected from the ultrasonic wind sensor, it is sufficient for gas tracking tasks. The quadrotor based gas tracking robot is now able to detect wind speed and direction for its gas tracking purpose.

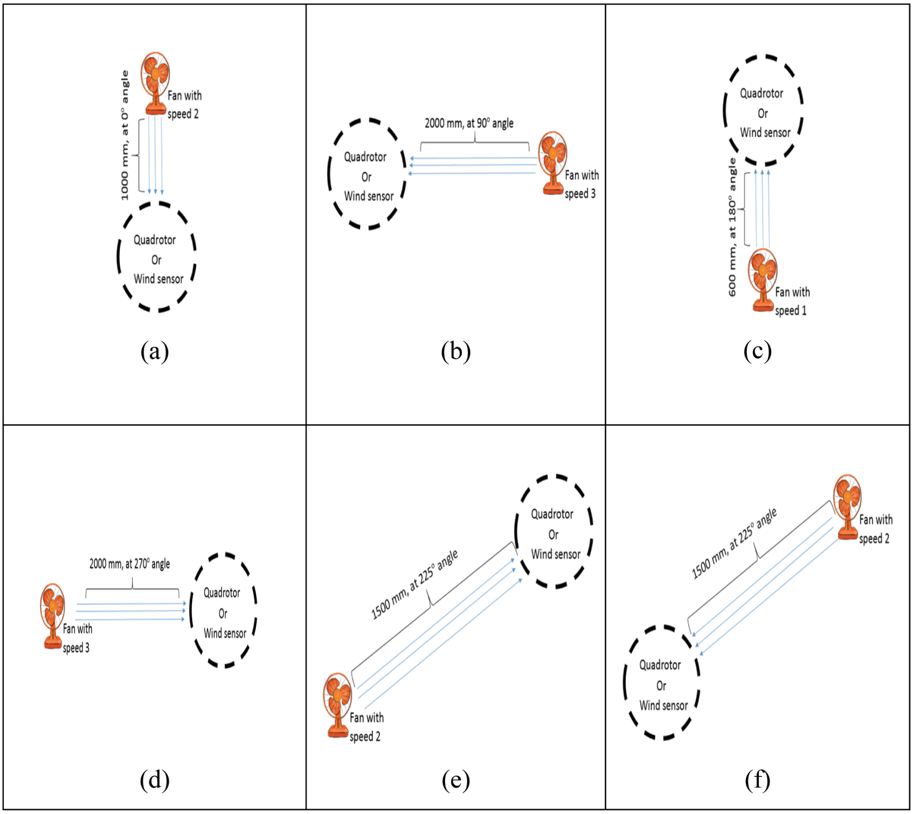


Fig. 8. Experiment scenarios with varies position and varies speeds of fan.

Table 1. Experiment result

Experiment	Ultrasonic wind sensor		Quadrotor wind estimation		Error (%)	
	Speed (m/s)	Direction (°)	Speed (m/s)	Direction (°)	Speed (%)	Direction (%)
(a)	3.4	359	3.38	351	0.59	2.23
(b)	1.9	91	1.7	84	10.53	7.69
(c)	2.89	183	2.58	187	10.73	2.19
(d)	1.94	271	1.82	263	6.19	2.95
(e)	2.55	223	2.48	235	2.75	5.38
(f)	2.39	44	2.15	48	10.04	9.09

6 Conclusion

We demonstrate that relying on the quadrotors' IMU alone is sufficient to calculate the surrounding wind speed and direction. The algorithm used enabled the quadrotor to obtain wind speed and wind information data with a minimal error percentage of 10.73% for the wind speed, and 9.09% for wind direction estimation. The quadrotor based gas tracking robot accomplished all this without attaching any add-on wind sensors. The experiments carried out covered a range of directions and speed, and therefore confirming that the quadrotor would be able to step up to the challenge of difficult and varying environments. This ability would prove useful in the oil and gas industry, where conditions could be marred by different factors, as oil and gas extraction sites are unpredictable by nature.

Acknowledgement. This work was supported by the Sunway Internal Grant scheme (Grant No: INT-FST-CIS-2016-03) at Sunway University, Malaysia.

References

1. Natural gas consumption in the United States from 1995 to 2015 (in trillion cubic feet). https://www.statista.com/statistics/184329/energy-consumption-from-natural-gas-in-the-us-from-1995/
2. City maps of gas leaks. http://www.heetma.org/squeaky-leak/natural-gas-leaks-maps/
3. Voiland A, Methane matters: scientists work to quantify the effects of a potent greenhouse gas. http://earthobservatory.nasa.gov/Features/MethaneMatters/
4. Eu KS, Yap KM, Tee TH (2014) An airflow analysis study of quadrotor based flying sniffer robot. In: Advanced development in industry and applied mechanics. Trans Tech Publications, pp 246–250
5. Russell A, Kennedy S (2000) A novel airflow sensor for miniature mobile robots. Mechatronics 10:935–942
6. Martinez D, Teixidó M, Font D, Moreno J, Tresanchez M, Marco S, Palacín J (2014) Ambient intelligence application based on environmental measurements performed with an assistant mobile robot. Sensors (Basel) 14:6045–6055
7. Hernandez Bennetts V, Lilienthal AJ, Neumann PP, Trincavelli M (2012) Mobile robots for localizing gas emission sources on landfill sites: is bio-inspiration the way to go? Front. Neuroeng. 4:1–12

Author Index

K.J. Kim and N. Joukov (eds.), *Mobile and Wireless Technologies 2017*,
Lecture Notes in Electrical Engineering 425, DOI 10.1007/978-981-10-5281-1