Computer Science & Information Technology 47

David C. Wyld
Natarajan Meghanathan (Eds)

Computer Science & Information Technology

Seventh International Conference on Networks & Communications
(NETCOM - 2015)
Sydney, Australia, December 26~27, 2015

AIRCC Publishing Corporation

Volume Editors

David C. Wyld,
Southeastern Louisiana University, USA
E-mail: David.Wyld@selu.edu

Natarajan Meghanathan,
Jackson State University, USA
E-mail: nmeghanathan@jsums.edu

ISSN: 2231 - 5403
ISBN: 978-1-921987-44-1
DOI : 10.5121/csit.2015.51601 - 10.5121/csit.2015.51620

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

Preface

The Seventh International Conference on Networks & Communications (NETCOM 2015) was held in Sydney, Australia, during December 26~27, 2015. The Seventh International Conference on Network and Communications Security (NCS 2015), The Seventh International Conference on Wireless & Mobile Networks (WiMoNe 2015), The Second International Conference on Computer Science, Engineering and Information Technology (CSEIT 2015) and The Second International Conference on Signal, Image Processing and Multimedia (SPM 2015) were collocated with the NETCOM-2015. The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The NETCOM-2015, NCS-2015, WiMoNe-2015, CSEIT-2015, SPM-2015 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in computer network and communications research.

In closing, NETCOM-2015, NCS-2015, WiMoNe-2015, CSEIT-2015, SPM-2015 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the NETCOM-2015, NCS-2015, WiMoNe-2015, CSEIT-2015, SPM-2015.

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

David C. Wyld
Natarajan Meghanathan

Organization

General Chair

Natarajan Meghanathan	Jackson State University, USA
Dhinaharan Nagamalai	Wireilla Net Solutions PTY LTD, Australia

Program Committee Members

Ahmed Asimi	Ibn Zohr University, Morocco
Alexandre Caminada	UTBM , France
Alexandre Cardoso	Federal University of Uberlandia, Brasil
Amalia Miliou	Aristotle University of Thessaloniki, Greece
Annamalai	Prairie View A&M University, USA
Antonio Ruiz Martinez	University of Murcia, Spain
Athanasios V.Vasilakos	Lulea University of Technology, Sweden
Bela Genge	Petru Maior University of Tg. Mures, Romania
Bhuyan Jay	Tuskegee University, United States
Ching-Hsien Hsu	Chung Hua University, Taiwan
Chunming Liu	T-Mobile, USA
Clement Nyirenda	University of Namibia, Namibia
Daqiang Zhang	Tongji University, China
Denivaldo Lopes	Federal University of Maranhao - UFMA, Brazil
Edward David Moreno	Federal University of Sergipe , Brazil
Elboukhari Mohamed	University Mohamed First, Morocco
Eman Shaaban	Ain Shams University, Egypt
Emmanouel (Manos) Varvarigos	University of Patras, Greece
Farzad Kiani	Istanbul S.Zaim University, Turkey
Francesca Lonetti	CNR-ISTI, Italy
Francine Krief	LaBRI Lab - University of Bordeaux, France
Gary Campbell	University of the West Indies, Jamaica
Geetha	VIT University, India
Hossam Afifi	Institut Mines Telecom, France
Hossein Jadidoleslamy	MUT University, Iran
Houcine Hassan	Univeridad Politecnica de Valencia, Spain
Ian Tan	Multimedia University, Malaysia
Ihab A.Ali	Helwan University, Egypt
Isa Maleki	Islamic Azad University, Iran
Islam Atef	Alexandria University, Egypt
Israa Sh.Tawfic	Ministry of science and technology, Iraq
Jae Kwang Lee	Hannam University, South Korea
Jani N.N	Kadi Sarva Viswavidyalaya University, India
Jianhong zhang	North China university of Technology, China
Junaid Ahsenali Chaudhry	Innopolis University, Russia
Kijun Han	Kyungpook National University, Korea
Kishore Bhamidipati	Manipal University, India
Mahdi Mazinani	Islamic Azad University, Iran

Technically Sponsored by

Networks & Communications Community (NCC)

Computer Science & Information Technology Community (CSITC)

Digital Signal & Image Processing Community (DSIPC)

Organized By

Academy & Industry Research Collaboration Center (AIRCC)

TABLE OF CONTENTS

The Seventh International Conference on Networks & Communications (NETCOM - 2015)

Implementation of Joint Network Channel Decoding Algorithm for Multiple Access Relay Channel Based on Convolutional Codes……………….……….... 01 - 08
Youssef Zid, Sonia Zaibi Ammar and Ridha Bouallègue

Spectrum Sensing in Cognitive Radio Networks : QoS Considerations……..... 09 - 19
Nabil Giweli, Seyed Shahrestani and Hon Cheung

A Method of Target Tracking and Prediction Based on Geomagnetic Sensor Technology……….…………………………………………………….……….... 21 - 37
Xinmin Tang, Shangfeng Gao, Songchen Han, Zhiyuan Shen, Liping Di and Binbin Liang

Medium Access Control with Scheduled Group Contention for Massive M2M Devices………………………………………………………….….………... 39 - 49
Joo Rak Kang and Tae-Jin Lee

Haarvested Energy-Adaptive MAC Protocol for Energy Harvesting IOT Networks……………………………………………………...….………... 51 - 58
Hyeong-Kyu Lee, MinGyu Lee and Tae-Jin Lee

Table-Based Identification Protocol of Computational RFID Tags……….…... 59 - 67
Yunmin Kim, Ji Hyoung Ahn and Tae-Jin Lee

SOTM : A Self Organized Trust Management System for VANET…………... 69 - 81
Amel Ltifi, Ahmed Zouinkhi and Mohamed Salim Bouhlel

Reducing Latency in African NRENS Using Performance-Based LISP/SDN Traffic Engineering…………………………………………….……..………... 83 - 98
Josiah Chavula, Melissa Densmore and Hussein Suleman

The Seventh International Conference on Network and Communications Security (NCS 2015)

Limiting Self-Propagating Malware Based on Connection Failure Behavior……………………………………………………………………….... 99 - 114
Yian Zhou, You Zhou, Shigang Chen and O. Patrick Kreidl

Selective Opening Secure Functional Encryption…………………..……….. 115 - 130
Yuanyuan Ji, Haixia Xu and Peili Li

Security and Privacy of Sensitive Data in Cloud Computing : A Survey of Recent Developments……………………………………………..…..………... 131 - 150
Ali Gholami and Erwin Laure

Security Analysis of Mobile Authentication Using QR-Codes……..……... 151 - 160
Siwon Sung, Joonghwan Lee, Jinmok Kim, Jongho Mun and Dongho Won

The Seventh International Conference on Wireless & Mobile Networks (WiMoNe 2015)

Application-Based QoS Evaluation of Heterogeneous Networks……..……... 161 - 172
Farnaz Farid, Seyed Shahrestani and Chun Ruan

BLE-Based Accurate Indoor Location Tracking for Home and Office…….. 173 - 181
Joonghong Park, Jaehoon Kim and Sungwon Kang

An Adaptive Remote Display Framework to Improve Power Efficiency…... 183 - 193
Dong Hyun, Jo and Dae Young, Kim

The Second International Conference on Computer Science, Engineering and Information Technology (CSEIT 2015)

Higher Educational Empowerment, Its Ideal Parameters and Their Co-Relation with Economy……………..…………………………………..……... 195 - 201
Mohit Satoskar and Anuprita Satoskar

Real Human Face Detection for Surveillance System Using Heterogeneous Sensors……..……………………………………………………………..…... 203 - 210
Yoon-Ki Kim, Doo-Hyun Hwang and Chang-Sung Jeong

Real-Time Pedestrian Detection Using Apache Storm in a Distributed Environment…………………………………………………………..……... 211 - 218
Du-Hyun Hwang, Yoon-Ki Kim and Chang-Sung Jeong

The Second International Conference on Signal, Image Processing and Multimedia (SPM 2015)

Robust Human Tracking Method Based on Appearance and Geometrical Features in Non-Overlapping Views…………….…………….………..…….. 219 - 235
Binbin Liang, Songchen Han, Yan Zhu and Liping Di

A Generalized Sampling Theorem Over Galois Field Domains for Experimental Design…………….………………...……………………..…….. 237 - 247
Yoshifumi Ukita

IMPLEMENTATION OF JOINT NETWORK CHANNEL DECODING ALGORITHM FOR MULTIPLE ACCESS RELAY CHANNEL BASED ON CONVOLUTIONAL CODES

Youssef Zid[1], Sonia Zaibi Ammar[1] and Ridha Bouallègue[2]

[1]SysCom Laboratory, National Engineering School of Tunis, Tunisia
[2]Innov'COM Laboratory, Higher School of Communications of Tunis, Tunisia

ABSTRACT

In this paper, we consider a Joint Network Channel Decoding (JNCD) algorithm applied to a wireless network consisting to M users. For this purpose M sources desire to send information to one receiver by the help of an intermediate node which is the relay. The Physical Layer Network Coding (PLNC) allows the relay to decode the combined information being sent from different transmitters. Then, it forwards additional information to the destination node which receives also signals from source nodes. An iterative JNCD algorithm is developed at the receiver to estimate the information being sent from each transmitter. Simulation results show that the Bit Error Rate (BER) can be decreased by using this concept comparing to the reference one which doesn't consider the network coding.

KEYWORDS

Joint network/channel decoding, network coding, wireless network.

1. INTRODUCTION

In the last years, researchers focused on cooperative communications [6][7] via relay. Then, by using this concept, the spatial diversity gain can be achieved. The principle task of the relay is firstly to combine information from different transmitters by using network coding [1]. Then, this mechanism, i.e. network coding, performs the gains on bandwidth especially [2][3]. In the next stage, the relay forwards the resulting signal to the destination node. In order to improve the throughput of wireless networks, the destination node receives the signals from different source nodes in the direct links, and eventually the one from the relay. In literature, This scheme is called Multiple Access Relay Channel [8] (MARC). The relay model in such scheme has been widely exploited. So, it was shown in [9][10] that several relaying architectures may be used like Amplify And Forward (AAF), Decode And Forward (DAF) and Compress And Forward (CAF). Indeed, authors proposed in [8] a joint network channel coding based on turbo codes for the MARC scheme with two transmitters.

David C. Wyld et al. (Eds) : NETCOM, NCS, WiMoNe, CSEIT, SPM - 2015
pp. 01–08, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51601

In a real case, the mentioned scheme cannot always be applied even it gives an improvement in throughput of wireless networks. So, practically ,a real network can contain more than two senders. Several algorithms are used at the relay to improve the performance of such scheme. In this case, the classical network coding at the relay, which consist to the XOR operation of signals from source nodes, cannot be applied. Authors presented in [11] a joint network channel coding for the MARC scheme with two senders based on distributed turbo code for the relay channel. They compared the proposed system to the one using separate network channel coding and proved that the use of joint network channel coding is necessary to exploit in a good way the redundant signal provided by the relay. Then, if the network and channel coding are treated separately, some performance loss is expected since the network decoder can't use the soft information computed by the channel decoder. Similarly, the channel code cannot exploit the redundant information of the network code. This concept is supported in [12][13][4] where authors proved that capacity can be achieved by conceiving channel and network coding as a single non-separated data processing stage. In addition, there are several joint channel network coding scheme like time division decode and forward MARC model presented in [5] in which authors proposed distributed regular LDPC codes as the joint network channel code at the relay. The same concept is proposed in [11] by using turbo-code-based joint channel network coding scheme.

In this paper, we propose a JNCD algorithm applied to the MARC scheme with large source nodes. The proposed scheme contains also one relay and one destination. We refer to [11] and we modify the joint network channel coding used at the relay. The redundant information provided at the relay is transmitted to the destination which receives eventually information from source nodes. An iterative decoding algorithm is presented at the destination to estimate the information being sent from each source.

This paper is organized as follows: the next section presents the considered MARC system model. In section III, a joint network channel coding algorithm is presented, then, a detailed description of the channel coding and network coding process is done. In section IV, an iterative joint network channel decoding algorithm is presented to estimate the source information's being sent from the transmitters. The section V presents the reference scheme. In section VI, the performance of the proposed decoding algorithm are evaluated and compared to the classic scheme which doesn't consider network coding.

2. SYSTEM MODEL

Fig. 1 depicts the system model of the proposed MARC scheme. It consists to wireless network that contains M Mobile Stations (MS), one Base Station (BS) and one Relay (R). We denote by $MS_1, MS_2,\ldots,MS_j,\ldots,MS_M$, $1 \le j \le M$, the set of M transmitters. Each MS_j generates a source information vector denoted $u_j, 1 \le j \le M$, of length equal to K bits. The source information's are firstly encoded with a channel coder in order to protect them from error transmission. Thus, the obtained sequence at the output of the j^{th} channel encoder is BPSK modulated to obtain x_j with length equal to N. In PLNC, there are two stages to recover the information being sent from each source at the destination. In the first one, the sources MS_j, $1 \le j \le M$, send their information simultaneously to the destination node and the relay over a Rayleigh channel. Then, the received message at BS, respectively R, is expressed as follows:

$$y_{D,j} = h_{D,j}x_j + n_{D,j} \qquad (1)$$

respectively

$$y_{R,j} = h_{R,j}x_j + n_{R,j} \qquad (2)$$

Where $h_{D,j}$ and $h_{R,j}$ are complex channel fading coefficient. $n_{D,j}$ and $n_{R,j}$ are zero mean and Gaussian noise with variance σ_D^2 and σ_R^2 .

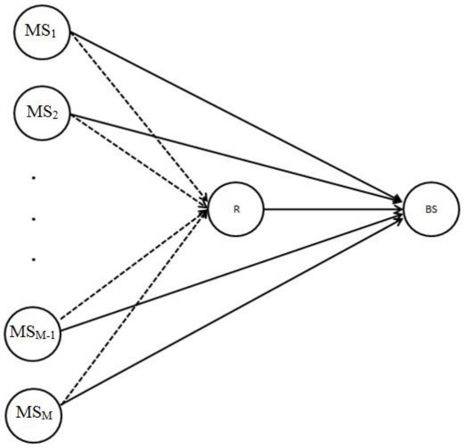


Fig. 1. System model for a network comprised of M users

In the next stage, a joint channel network coding is performed at the relay. It consists to the combination of signals being sent from the sources node $MS_j, 1 \le j \le M$. A detailed description of this process is done in the next section. The relay R provides itself redundant information which is sent to the destination node.

The main goal is to decode information being sent from the mobile stations at the destination node by using an iterative JNCD algorithm. It consists to *M* channel decoder and one network decoder. The last one provides information to each channel decoder in order to improve the decoding processes of the corresponding signal.

3. JOINT CHANNEL NETWORK CODING FOR M-USER NETWORK

3.1 Channel coding

As mentioned previously, each transmitter MS_j encode its source information in order to protect it from transmission error. The channel encoder used in this work is a convolutional code (37,21) with rate equal to 0.5 and constraint length 4. We assume that the length of each source information vector u_j, $1 \le j \le M$, is 1500 information bits. After channel encoding process, the jth encoder provides bit sequences of length equal to 3003 bits.

In order to increase the system rate, a puncturing method is used. So, the number of transmitted bits in each sequence is reduced. Then, we send only 2000 bits instead of 3003 from each channel coder. The choice of the transmitted bits must respect the following rule: we transmit only the third parity bit. Thus, we transmit only 500 from 1500 parity bits. As a result, the transmitted bit sequence from MS_j contains 500 parity bit and 1500 systematic bit. The proposed puncturing process is applied for all transmitters.

3.2 Network Coding

The network coding process consists to mixing information from different users. In the MARC scheme with two sources, it is done in a general way by computing the XOR of the corresponding sequences. The resulting sequence is used at the destination as additional information which improves the decoding process. This concept cannot be applied to network with large number of transmitters. We present in this section the network coding principle at the relay for the proposed M-users network.

Since the considered scheme has M transmitters, the network coding block contains M+1 sub-block: M channel decoders and one network encoder. The jth channel decoder has as input the sequence $y_{R,j}$ corresponding to the jth source node. The classic decoding process is performed for all channel decoders in order to get an estimation of the source information being sent from mobile stations. Then, we obtain at the output of the jth channel decoder a bit sequence denoted $\hat{u}_{j,4}$, $1 \leq j \leq M$. They are firstly interleaved and encoded by a convolutional encoder with the same characteristics as those used for mobile stations. This process is depicted in figure 2.

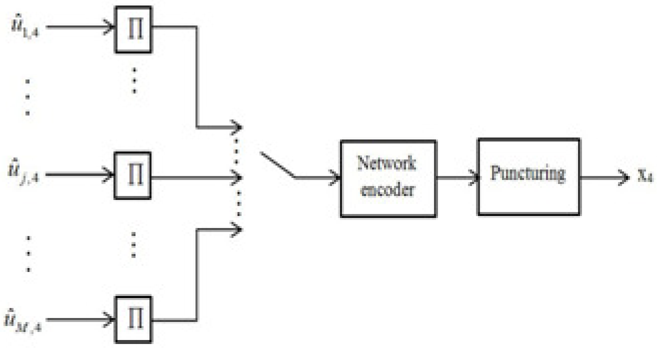


Fig. 2. Channel coding process in network codingblock

The network coding block contains M channel decoders, so, the network encoder has as input a bit sequence with length equal to $M \times 1500$.

At this stage, the network coding is performed by using a convolutional code (37,21) with rate equal to 0.5 and constraint length 4. Then, we obtain at the output of network coding block a bit sequence with length L equal to

$L = M \times 1500 \times 2 + 3$ bits, whose $M \times 1500 + 3$ parity bits.

The obtained bit sequence is punctured by using the same method being described in section 3.1 to obtain the signal x_4 So, we transmit only the parity bit at the third place. Consequently, in order to transmit $M \times 3003$ bits, only $M \times 2000$ bits ($M \times 1500$ systematic bits and $M \times 500$ parity bits) are transmitted to the destination node. Then, if we denote by R_R the system rate at the relay, and by N_R the number of transmitted parity bit, the network code rate can be written as follows:

$$R_R = MK / N_R = 2 \times M \times 1500 / M \times 2000 = 3000 / 2000 = 1.5 \qquad (3)$$

The sequence x_4 is BPSK modulated and transmitted over a Rayleigh channel to the destination node. We denote the received message by $y_{D,R}$.

4. JOINT NETWORK/CHANNEL DECODING

The notion of JNCD is illustrated when the system contains two or more decoders. It involves the combination of data provided by each decoder in order to ameliorate the system performances. For the MARC scheme with two transmitters, the JNCD block must contain three decoders: two channel decoders and one network decoder.

The proposed scheme in this paper contains *M* mobile stations, so, there are *M* channel decoders and one network decoder in the JNCD block. All these decoders are Soft Input Soft Output (SISO). Figure 3 presents the JNCD block for 4-users network.

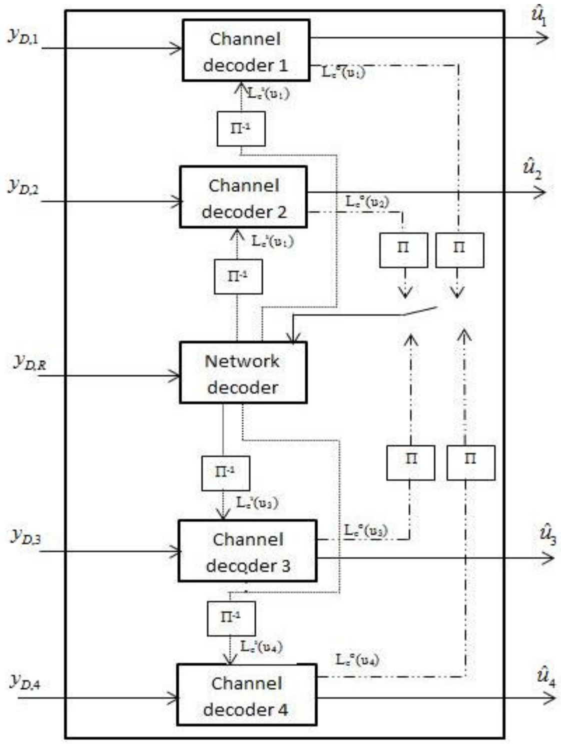


Fig. 3. joint network/channel decoding block for 4-users network

The JNCD block has as input the transmitted sequence from each mobile station and the combined sequence provided by the relay. Since all decoders are SISO, each channel decoder provides additional information called extrinsic information which is denoted by $L_e^o(u_j)$, $1 \le j \le M$. These quantities are firstly interleaved and mixed. Then, the resulting sequences will be the inputs to network decoder. It uses the quantities as additional information to decode the message $y_{D,R}$ transmitted from the relay. The network decoder provide itself en extrinsic information denoted $L_e^o(u_1,\ldots,u_M)$. this quantity contains a part related to the transmitted signal from MS_j.

The main goal is to extract the part corresponding to each channel decoder in the JNCD block. We denote by $L_e^j(u_j)$, $1 \le j \le M$, the desinterleaved information corresponding to the jth channel decoder. As a result, each channel decoder takes as input the signal $y_{D,j}$ and the quantity $L_e^j(u_j)$ provided by the network decoder. The decoding process is performed for each channel decoder, and a posteriori information is given from each one which will be the input for the network decoder.

This process is repeated until the total number of iterations is reached. An estimation of the transmitted information from the mobile stations is given.

5. REFERENCE CHAIN

In order to evaluate the performances of the JNCD algorithm, it is usually to fix a reference chain to which we compare the performances of the proposed scheme. Then, in the classic way, the network coding is not performed, so, the source information's are sent from mobile stations to the destination in a direct link without any additional information. Figure 4 depicts this reference chain. It contain *M* senders and one receiver, there is no relay.

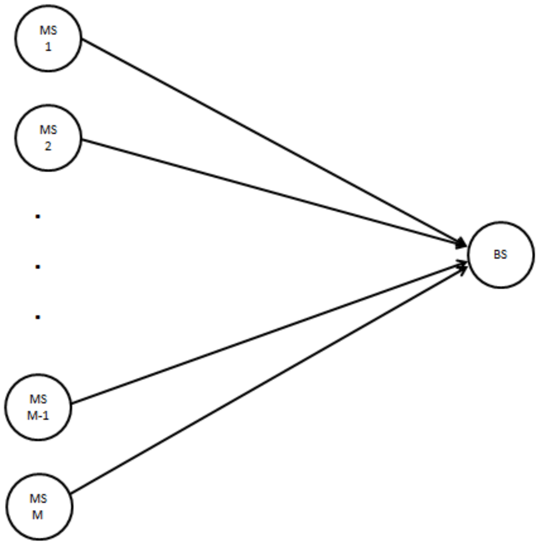


Fig. 4. Reference chain

The reference chain contains *M* transmitters and one receiver. Then, these mobile stations encoded its packets by using a channel decoder with the same characteristics as those using in the MARC scheme. The resulting quantities are sent in a direct link to the receiver over a Rayleigh channel. In addition, in order to get a meaningful comparison between the two chains, the same puncturing system used in the MARC scheme is adopted, so, from each transmitter, 2000 bits (1500 systematic bits and 500 parity bits) are sent instead of 3003. At the receiver, all punctured bits are replaced by 0.

6. SIMULATION RESULTS

In this section, we present the performances of joint network/channel decoding algorithm for the proposed MARC scheme which consists to *M* users. The number of mobile stations is set to 4. The characteristics of the convolutional code used in this work are presented in section II. In addition, we present the performances of the proposed scheme for the convolutional encoder (7,5). Then, we study the impact of the channel encoder on the performances of the JNCD algorithm. Thus, in order to evaluate the performances of the JNCD algorithm, we must make comparison between the conventional chain and MARC scheme.

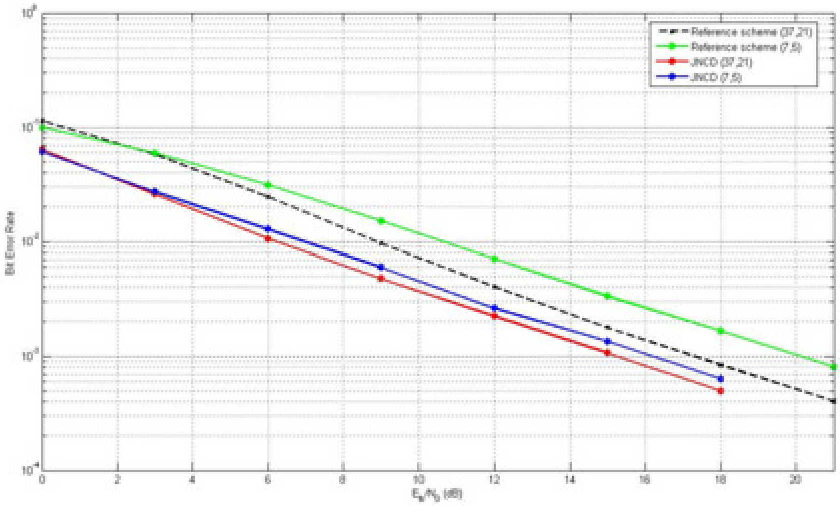


Fig. 5. Bit Error Rate of System applying joint network/channel decoding and reference chain

Figure 5. depicts the bit error rate (BER) for the two chains depending on the value of the ratio of Energy per Bit to the spectral noise density (Eb/N0) when using the two convolutional codes (37,21) and (7,5).. According to this figure, we can see that the BER decrease significantly comparing to the standard scheme in the two cases. Then, if the code (37,21) is used, the gain can achieve approximately 2.5 dB, and it achieves 3 dB for the code (7,5).

On the other hand, we can see that by using the JNCD decoding algorithm, we can achieve approximately the same performances for the two convolutional encoders. So, this algorithm corrects the errors in a good way unlike the standard scheme for which there is a difference of 2 dB between the standard scheme using the code (37,21) and the one using (7,5).

7. CONCLUSION

In this paper, we presented the joint network/channel decoding principle for MARC scheme that consists to *M* users, one relay and one receiver. A convolutional encoder is used as a channel code at the mobile stations. In order to increase the system rate, a puncturing method was presented. Indeed, a new design of the network decoding at the relay was presented. Finally, an iterative joint network channel decoding algorithm was developed to estimate the source information being sent from each mobile station. The implementation of this algorithm remains complex since the decoder contains *M*+1 decoders (always SISO). It has to take into account the information exchange between the channel decoders and the network decoder. Simulation results are presented for the two codes (37,21) and (7,5). It is shown that the JNCD algorithm give an improvement of gain mainly. So, a gain of approximately 2.5 dB is always achieved. In addition, simulation results shows that the JNCD algorithm can correct errors in good way even there is a significant difference in terms of gain between standard schemes when using two different convolutional codes.

REFERENCES

[1] R. Ahlswede, N. Cai, S-Y. R. Li, and R. W. Yeung. Network Information Flow. IEEE Trans. on Information Theory, 46(4):1204–1216, July 2000.

[2] D. Tuninetti and C. Fragouli. Processing Along the Way: Forwarding vs. Coding. In Proc. International Symposium on Information Theory and its Applications (ISITA), Oct. 2004.

[3] X. Bao and J. Li. Matching code-on-graph with network-on-graph : Adaptive network coding for wireless relay networks. In Proc. 43rdAllerton Conf. on Communication, Control, and Computing, Sept. 2005.

[4] C. Hausl and J. Hagenauer.Iterative Network and Channel Decoding for the Two-Way Relay Channel. In Proc. IEEE International Conference on Communications (ICC), June 2006.

[5] C. Hausl, F. Schreckenbach, I. Oikonomidis, and G. Bauch, Iterative network and channel decoding on a tanner graph, in Proceeding of 43rd Allerton Conference on Communication, Control and Computing, September 2005.

[6] J. N. Laneman and G. W. Wornell, Energy-efficient antenna sharing and relaying for wireless networks in Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC ’00), pp. 7–12, September 2000.

[7] A. Sendonaris, E. Erkip, and B. Aazhang, User cooperation diversity—part I: system description IEEE Transactions on Communications, vol. 51, no. 11, pp. 1927–1938, 2003.

[8] G. Kramer and A. J. van Wijngaarden. On the White Gaussian Multiple- Access Relay Channel. In Proc. IEEE International Symposium on Information Theory (ISIT), June 2000.

[9] J. Laneman, Cooperative diversity in wireless networks algorithms and architectures Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, Aug. 2002.

[10] E. Meulen, Three-Terminal Communication Channels Adv. App/. Prob., vol. 3, no. 1, pp. 120 - 154, 1971.

[11] C. Hausl and P. Dupraz, Joint Network-Channel Coding for the Multiple-Access Relay Channel, 2006 3rd Annual IEEE Communications Society on Sensor and Ad Hoc Communications and Networks, 2006. SECON '06.

[12] M. Effros, M. Medard, T. Ho, S. Ray, D. Karger, and R. Koetter, Linear Network Codes: A Unified Framework for Source, Channel and Network Coding in DIMACS03, 2003.

[13] N. Ratnakar and G. Kramer, The Multicast Capacity of Deterministic Relay Networks with No Interference, IEEE Trans. Information Theory, vol. 52, no. 6, pp. 2425–2432, June 2006.

SPECTRUM SENSING IN COGNITIVE RADIO NETWORKS: QOS CONSIDERATIONS

Nabil Giweli, Seyed Shahrestani and Hon Cheung

School of Computing, Engineering and Mathematics,
Western Sydney University, Sydney, Australia
ngiweli@scem.westernsydney.edu.au,
s.shahrestani@westernsydney.edu.au,
h.cheung@westernsydney.edu.au

ABSTRACT

The rapidly growing number of wireless communication devices has led to massive increases in radio traffic density, resulting in a noticeable shortage of available spectrum. To address this shortage, the Cognitive Radio (CR) technology offers promising solutions that aim to improve the spectrum utilization. The operation of CR relies on detecting the so-called spectrum holes, the frequency bands that remain unoccupied by their licensed operators. The unlicensed users are then allowed to communicate using these spectrum holes. As such, the performance of CR is highly dependent on the employed spectrum sensing methods. Several sensing methods are already available. However, no individual method can accommodate all potential CR operation scenarios. Hence, it is fair to ascertain that the performance of a CR device can be improved if it is capable of supporting several sensing methods. It should obviously also be able to select the most suitable method. In this paper, several spectrum sensing methods are compared and analyzed, aiming to identify their advantages and shortcomings in different CR operating conditions. Furthermore, it identifies the features that need to be considered while selecting a suitable sensing method from the catalog of available methods.

KEYWORDS

Cognitive Radio; Spectrum Sensing; Qos

1. INTRODUCTION

In general, the Radio Frequency Spectrum (RFS) is statically divided into licensed and unlicensed bands. While the use of the former is restricted to authorized operators, the unlicensed bands are available for use by the public, only subject to transmission constraints [1]. As such, the unlicensed bands may get heavily congested. On the other hand, several studies and measurements conducted around the world have indicated that the licensed RFS bands can be underutilized [2].

From a technical perspective, the Cognitive Radio (CR) concept is a promising technology to achieve an efficient utilization of RFS. The concepts for CR models were introduced in 1999 by Joseph Mitola [3]. In these models, the operators licensed to use some particular frequency bands are considered as the Primary Users (PUs). Whereas, the unlicensed participants are referred to as the Secondary Users (SUs). The CR model is based on the realization that a PU may not fully utilize its licensed bands, leaving parts of its spectrum unoccupied. These unoccupied white spaces, or holes, relate to use, or more correctly the lack of use, in terms of frequency, time, or

David C. Wyld et al. (Eds) : NETCOM, NCS, WiMoNe, CSEIT, SPM - 2015
pp. 09–19, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51602

space and location. An SU can utilize these holes in addition to the unlicensed bands that it may typically use.

To achieve their objectives, CR systems are dependent on the execution of a sequence of several functions, the so-called CR cycle. A typical CR cycle was proposed by Mitola [3]. This is illustrated in Figure 1.

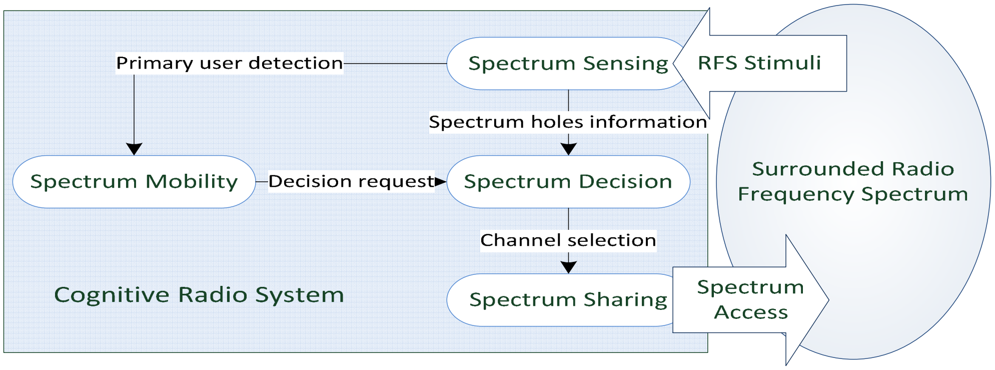


Figure 1. Basic CR cycle

The main functions of this CR cycle are spectrum sensing, spectrum decision, spectrum sharing, and spectrum mobility. More specifically, an SU should be able to perform the following [4]:

- Spectrum sensing: sense the surrounding RFS to determine spectrum holes and to detect the presence of the relevant PU.
- Spectrum decision: analyze and decide which spectrum hole is the most suitable for satisfying the application requirements.
- Spectrum sharing: share the available spectrum holes with other SUs as fairly as possible.
- Spectrum mobility: seamlessly switch to another suitable spectrum hole to avoid interference with a detected PU that may wish to start using its licensed spectrum.

Detecting the presence of a PU, or more precisely finding out whether the PU is using its allocated spectrum or not, is an essential task for a CR device. On one hand, this fundamental task requires improving sensing accuracy by avoiding false positive results while detecting the presence of a PU. On the other, the employed sensing technique should achieve a high detecting probability of the available spectrum holes. The nature of the electromagnetic signals makes accurate sensing a complicated process. More specifically, the Signal to Noise Ratio (SNR), the multipath fading of the PU signals, and the changing levels of noise can significantly affect the sensing accuracy [5, 6]. Moreover, imperfect spectrum sensing can result in increased transmission error rates, for both the PU and the SUs [7]. Such errors may contribute to the degradation of the quality of the services provided by a PU and SUs. Noticeably, any QoS degradation that can be attributed to the CR technology can potentially harm the progress of the CR-based solutions. In this paper, the main features and limitations of the prevalent spectrum sensing methods are examined. Furthermore, the key aspects that should be involved in selecting the appropriate sensing method are highlighted and discussed.

The remainder of this paper is organized as follows. Section II presents the background and motivation for this work. The effects of the sensing operation on the QoS of the applications running over CR networks are described in Section III. Several sensing approaches are discussed and evaluated in Section IV. Factors that may help in selecting the proper sensing techniques are outlined in Section V. The last section gives the conclusions and points to the potential future expansion of the reported work.

2. MOTIVATIONS

Most of the previous reviews of spectrum sensing techniques are mainly focused on the operation, accuracies, complexities, and implementation issues [8-11]. For instance, the relation between the sensing accuracy, and the speed, i.e., sensing time, and frequency, i.e., repeating the sensing, are the primary focus of the authors in [8]. They aim to achieve an optimal spectrum sensing performance with the capability of flexible tuning between the speed and frequency. However, they find that the available state-of-art sensing technologies do not offer a possible trade-off between complexity and accuracy. In contrast, other reviewers consider the characteristics of the PU signal as the main factor for selecting a proper sensing method [9]. Nevertheless, other factors should be considered for more adaptive sensing and improved performance.

In general, the dominant approach is about how to find an optimal sensing method for all possible CR operation requirements. However, none of the proposed sensing methods is suitable for all possible sensing situations, conditions and technologies of CR systems. This study shifts the focus to another approach where a CR device supports a range of various sensing methods. Thus, the proper sensing technique can be selected based on the real-time requirements. This approach implies the need for a real-time mechanism to select the most suitable sensing method. In this paper, various sensing methods are studied toward finding the relevant selection criteria that should be considered when designing such as real-time selection mechanism.

3. SENSING OPERATION IMPACT ON APPLICATIONS' QOS IN CR NETWORKS

As shown in Figure 2, the operation of a CR can be divided into repeated cycles of the sensing period. The sensing period T effectively represents the time interval where sensing is repeated. It also represents the communication frame in CR. During the sensing time t, a CR device obtains information from its environment. After the sensing time, the CR device can decide to transmit data on the same channel or in a new vacant channel, i.e., a spectrum hole.

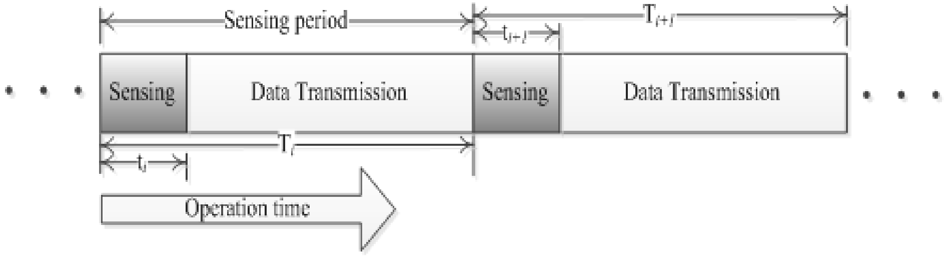


Figure 2. Simple structure of CR frames based on sensing operation

The decision taken by the CR device is based on the sensing outcome, i.e., the presence or absence of the PU. The transmission starts after the sensing time until the next sensing period, also called the CR communication frame. The transmission time (T-t) depends on the sensing time t and the frame time T that is based on the design of how frequent the sensing will be conducted. Thus, sensing frequency is 1/T. The sensing time t and the frame time T can be designed to be fixed for all frames or could be designed to vary based on the design goals [12]. Typically, the sensing operation should be limited and less frequent as much as possible without affecting the sensing accuracy [13].

Increasing the sensing time t and conducting the sensing more frequently, i.e. decreasing T, lead to an increase in the probability of correct detection of the PU's presence. In turn, this leads to more protection to the PU from interference by CR users and more utilization of the spectrum. On the other hand, this leads to less data transmission rate and hence to QoS degradation for SUs.

The degradation can be measured by several parameters such as throughput, delay and MAC layer process overhead [14]. Therefore, designing the sensing time and frequency of sensing operation should take into account the trade-off between protecting the PU's QoS and improving the QoS of SUs.

4. SENSING METHODS

The main challenge facing the sensing methods is how to improve the spectrum sensing performance by mainly increasing the positive detection probability and decreasing the false detection probability. A sensing technique with a higher positive detection probability provides more protection to PU. A CR user with a lower probability of false detection of the presence of the PU has more chance to use the available spectrum holes. Therefore, the user has more chance of achieving a higher throughput on the CR network. The design of a sensing technique is constrained by an acceptable level of false detection [15]. Additionally, improving sensing performance is challenged by a range of trade-offs and various constraints such as application requirements, hardware capability, complexity and required infrastructure [16].

In general, a sensing method that uses surrounding RFS information collected by the CR device only is called a local sensing. If the SUs do not exchange their surrounding RFS information gathered by local sensing, then this sensing is referred to as non-cooperative sensing. In this paper, the sensing methods are classified mainly into three categories: methods with no prior information required, based on prior information and based on SUs cooperation.

4.1. No Prior Information Required (Blind Sensing)

No prior information about the PUs' signal is necessary for the sensing methods under this category. However, prior information about the noise power of the targeted spectrum may be required for better performance. Otherwise, a reasonable estimation of the noise power is used instead. Two well-known blind sensing methods are energy detection and covariance-based detection.

4.1.1. Energy Detection

Also known as radiometry or periodogram, energy detection is the most common method for spectrum sensing because of its low implementation complexity and computational overhead [5]. In this method, the energy detector is used to detect a narrowband spectrum and then the observed signal energy level is compared with a predefined threshold. Thus, the channel is occupied by the PU if the detected signal energy is over the threshold. Otherwise, it is considered unoccupied, i.e., a spectrum hole. Because of this simplicity, this technique requires the shortest sensing duration t per frame compared to other common sensing technologies [17].

Generalizing the use of this method faces several challenges as a consequence of its simplicity. Firstly, selection of the threshold used for detection is an issue when the channel noise level is unknown or uncertain over time [18]. Secondly, under low SNR, it is hard to differentiate between modulated signals, including signals of other SUs, noise, and interference, resulting in poor detection performance [5]. Lastly, an energy detector is ineffective in detecting spread spectrum signals [19].

4.1.2. Covariance-based Detection

This method is based on comparing the covariance of the detected signal and the covariance of the noise where statistical covariance matrices of signal and noise are usually different [20]. The main improvement of this method is to overcome the energy detection shortcoming. In particular, it can distinguish between signal and noise in a low SNR, and without any prior information about the PU's signal and channel noise. This detection improvement is achieved at the expense of adding a computational overhead in computing the covariance matrix of the received signal samples [11]. In addition to increasing complexity, other drawbacks of the energy detection are still present in the covariance-based detection.

These sensing methods work with no prior information about the PU signals. They have a limited performance particularly for spread spectrum and in situations where other SUs are sharing the spectrum. Research is ongoing to improve the blind sensing approach in terms of performance and required sensing time, such as in [21, 22].

4.2. Prior Information Required

Methods belonging to this category rely on partial or full information about the PU's transmission signal to be able to differentiate it from other signals and noise.

4.2.1. Cyclostationarity Feature Detection

This method is based on distinguishing the PU signal from noise, interference, and other signals by identifying its cyclostationarity features [23, 24]. These cyclostationarity features are associated with the signal modulation type, carrier frequency, and data rate. Hence, the CR device needs sufficient prior information about these unique characteristics of the PU signal. Based on this information, it can perform a cyclostationarity analysis on the detected signal to identify matched features [9]. For this method to perform better than the energy detection method, an adequate number of real-time sample sets in the frequency domain need to be collected. As a consequence, better performance accrues more complexity and sensing time at the expense of the available throughput [9].

4.2.2. Correlation Detection

Sensing based on correlation is also known as waveform-based sensing or coherent sensing. In this method, the expected correlation or coherence between signal samples is identified to detect the PU signal based on previous knowledge about its waveform patterns [9]. The accuracy of the sensing increases when the length of the known signal pattern of the PU is increased [25]. The main drawback of this method is related to the large amount of information required for signal patterns of the PUs to achieve a high performance that is not practical for all CR systems.

4.2.3. Radio Identification Based Sensing

This method is based on having apriori information about the transmission technologies used by the PU. In the radio identification stage of the method, several features of the received signal are exploited and then classified to determine if the signal demonstrates the PU signal technology [26]. Fundamentally, the feature extraction and classification techniques are used in the context of European Transparent Ubiquitous Terminal (TRUST) project [27]. For collecting the signal features, the radio identification method may use one of the known sensing techniques, such as the energy detection method [9]. The radio identification improves the accuracy of the energy detection to some extent with complexity implication. The achieved precision is dependent on the signal features and classification techniques used to identify the presence of the PU.

4.2.4. Matched Filtering

The matched filtering method achieves a higher detection probability in a short detection time, compared to other methods that are similarly based on prior information [28, 29]. Hence, under this classification, this method is considered as the best sensing method. The collected signal is passed through a filter that will amplify the possible PU signal and attenuate any noise signal. The filter makes the detection of the presence of the PU signal more accurate [29]. The filter, which is known as a matched filter, has to be tuned based on some features of the PU signal. These characteristics include the required bandwidth, operating frequency, the modulation used and frame format [9]. One of the disadvantages of this method is in implementation where different PUs signal types require different dedicated hardware receivers. This requirement makes the method impractical to implement and also leads to higher power consumptions if the method is implemented based on current hardware technologies.

Figure 3 shows a comparison between non-cooperative sensing methods, based on accuracy and complexity metrics. Table 1 shows more comparison factors between local sensing methods.

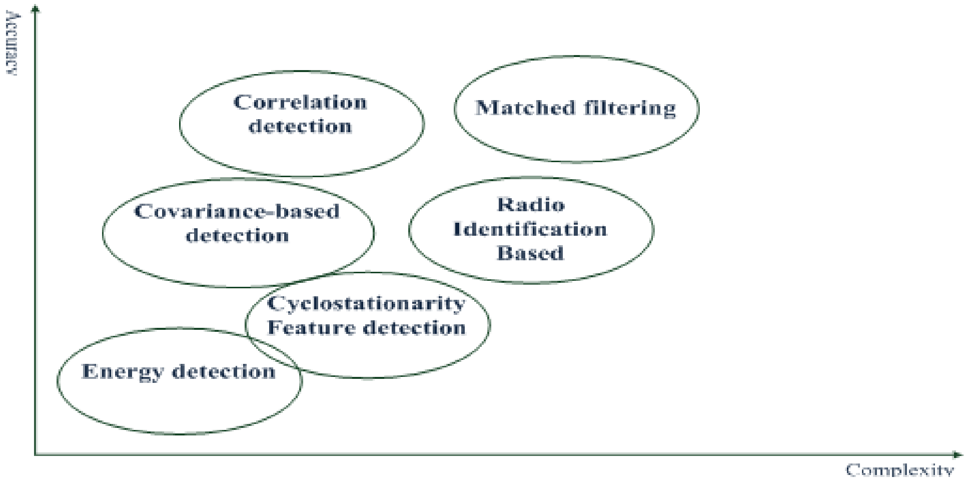
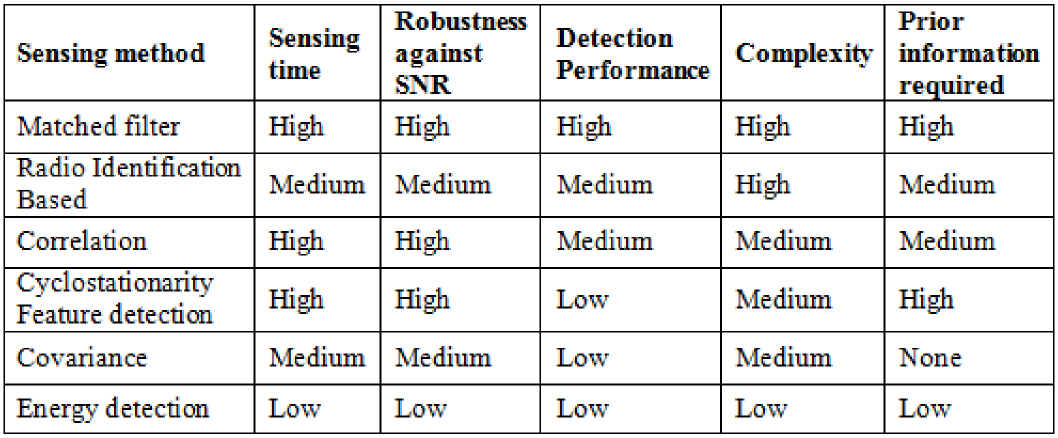

Accuracy
Correlation detection
Matched filtering
Covariance-based detection
Radio Identification Based
Cyclostationarity Feature detection
Energy detection
Complexity

Figure 3. Sensing method complexity versus accuracy

Table 1. Comparison between local sensing methods



Sensing method	Sensing time	Robustness against SNR	Detection Performance	Complexity	Prior information required
Matched filter	High	High	High	High	High
Radio Identification Based	Medium	Medium	Medium	High	Medium
Correlation	High	High	Medium	Medium	Medium
Cyclostationarity Feature detection	High	High	Low	Medium	High
Covariance	Medium	Medium	Low	Medium	None
Energy detection	Low	Low	Low	Low	Low

4.3. Based on SU Cooperation

The main principle of this approach is that SUs share their local sensed information of the spectrum. The use of sensed information from all SUs can produce a more accurate sensing outcome than relying solely on local sensing. The hidden transmitter problem is an example of the issues that may prohibit a CR from detecting the presence of a PU. The cause of this problem is the fading and shadowing of the signals from a PU, although it is within the transmission range of the CR [9]. However, when cooperated SUs are spatially distributed, it helps to overcome the hidden PU problem and other limitations of local sensing [30]. Sensing cooperation can also reduce the local sensing cost, e.g., sensing time duration and energy consumption while maintaining sensing quality by scheduling the sensing operation among cooperative SUs [31].

The sensing method used by an individual SU can be based on one of the sensing methods for local sensing, such as energy detection and cyclostationarity feature detection [10].

In some environments, cooperative sensing may lose its advantages as far as an individual SU is concerned. For instance, increasing the local sensing frequency in individual high mobility SUs is more efficient, in terms of sensing accuracy and overhead, than to cooperate with other SUs [19]. In cooperative sensing, the improvement of sensing is more noticeable when the number of cooperative SUs is increased. However, involvement of more SUs will increase the cooperation overhead in terms of the amount of data exchange and the time required for the exchange [32].

The cooperative approaches can only be used when SUs are able and willing to collaborate. Also, a SU may not always find other cooperative SUs within its transmission range. Therefore, the CR devices should not solely rely on cooperative sensing approaches. They should be able to use a fitting local sensing method and resort to cooperative sensing, only when an enhanced performance is possible.

5. Factors For Selecting The Fitting Sensing Method

Selecting the best sensing method for a particular cognitive radio operation condition depends on several factors. Based on the discussions in previous sections, notable factors are summarized below:

5.1. CR Device Capability

A CR device designed with limited hardware resources and power capacities will not be able to support a wider range of sensing methods. Some methods require sophisticated hardware components and higher power consumption, e.g. the matched filter method, compared to simple ones such as the energy detection method. An ideal CR device should be able to be reconfigured on-the-fly to support a broad range of sensing methods. In practice, a CR device's actual capability will limit the range of sensing methods that can be supported.

5.2. Qos Required for Applications Running on the CR Device

The QoS requirements differ based on the applications running on a CR device. The sensing delay and transmission throughput vary from one sensing method to another within the same conditions. As a result, the sensing operation used on a CR device has a direct impact on the QoS of an application running on the device, mainly in terms of the throughput and delay. As sensing is a repetitive operation, a CR device should be able to select a proper sensing method with the least impact on the QoS of the running application. Other operational requirements must also be taken into account. For example, the PU protection should have a higher priority than the QoS requirements of a CR user.

5.3. Apriori Information

The extent of information available about the characteristics of the PUs and the communications media is a major factor influencing the selection of a proper sensing method. For instance, insufficient information about the PU signals, excludes the use of matched filter method.
The CR device should be able to change the sensing method based on the information that becomes available about the PU signal or the SNR of the targeted spectrum by sensing.

5.4. Level of Protection Required for PU

The selection of the sensing method must be considered with regard to the degree of protection necessary for the PU. They may vary depending on available frequency bands and types of services. For instance, analog TV service is more robust against interference than digital TV service [16]. Hence, a sensing method that provides less protection, i.e., lower PU detection probability, should only be used when the PU is more tolerant of interference such as in analog TV services.

5.5. The CR Network Mode and Capability

The network mode and capability are important factors to CR systems to make a decision between cooperative and non-cooperative sensing approaches. In CR networks with infrastructure and centralized topology, a method based on cooperative sensing is more suitable than that based on local sensing only. Hence, the capability of such a CR network depends on how much management ability can provide for white space determination to its CR devices. Furthermore, the capacity of a CR network relies on how much information the network can gather and provide to its CR users about the PU signals and the ambient spectrum.

6. CONCLUSIONS

The work reported in this paper asserts that none of the available spectrum sensing techniques can achieve perfect solutions for all potential CR operating conditions. Therefore, to improve the performance of CR systems, the relevant devices must be capable of utilizing a catalog of sensing methods. The selection of the most suitable method from the catalog, which is an obvious necessity, is based on a number of factors that have also been identified and discussed in this paper. Our future works will focus on more exhaustive evaluations of these factors and how their fine-tunings can contribute to an improved CR performance.

REFERENCES

[1] L. Cui and M. B. Weiss, "Can unlicensed bands be used by unlicensed usage," Paper for the 41st annual TPRC, September, pp. 27-29, 2013.

[2] K. Patil, R. Prasad, and K. Skouby, "A Survey of Worldwide Spectrum Occupancy Measurement Campaigns for Cognitive Radio," in Devices and Communications (ICDeCom), 2011 International Conference on, 2011, pp. 1-5.

[3] J. Mitola and G. Q. Maguire Jr, "Cognitive radio: making software radios more personal," Personal Communications, IEEE, vol. 6, pp. 13-18, 1999.

[4] J. Marinho and E. Monteiro, "Cognitive radio: survey on communication protocols, spectrum decision issues, and future research directions," Wireless networks, vol. 18, pp. 147-164, 2012/02/01 2012.

[5] Y. Zeng, Y.-C. Liang, A. T. Hoang, and R. Zhang, "A review on spectrum sensing for cognitive radio: challenges and solutions," EURASIP J. Adv. Signal Process, vol. 2010, pp. 2-2, 2010.

[6] A. F. Molisch, L. J. Greenstein, and M. Shafi, "Propagation issues for cognitive radio," Proceedings of the IEEE, vol. 97, pp. 787-804, 2009.

[7] S. Haddadi, H. Saeedi, and K. Navaie, "Channel coding adoption versus increasing sensing time in secondary service to manage the effect of imperfect spectrum sensing in cognitive radio networks," in Communication and Information Theory (IWCIT), 2013 Iran Workshop on, 2013, pp. 1-5.

[8] D. Ariananda, M. Lakshmanan, and H. Nikookar, "A survey on spectrum sensing techniques for cognitive radio," in Cognitive Radio and Advanced Spectrum Management, 2009. CogART 2009. Second International Workshop on, 2009, pp. 74-79.

[9] T. Yucek and H. Arslan, "A survey of spectrum sensing algorithms for cognitive radio applications," Communications Surveys & Tutorials, IEEE, vol. 11, pp. 116-130, 2009.

[10] I. F. Akyildiz, B. F. Lo, and R. Balakrishnan, "Cooperative spectrum sensing in cognitive radio networks: A survey," Physical Communication, vol. 4, pp. 40-62, 2011.

[11] D. B. Rawat and G. Yan, "Spectrum sensing methods and dynamic spectrum sharing in cognitive radio networks: A survey," International Journal of Research and Reviews in Wireless Sensor Networks, vol. 1, pp. 1-13, 2011.

[12] P. Yiyang, L. Ying-Chang, K. C. Teh, and L. Kwok Hung, "How much time is needed for wideband spectrum sensing?," Wireless Communications, IEEE Transactions on, vol. 8, pp. 5466-5471, 2009.

[13] C. Kae Won, "Adaptive Sensing Technique to Maximize Spectrum Utilization in Cognitive Radio," Vehicular Technology, IEEE Transactions on, vol. 59, pp. 992-998, 2010.

[14] H. Xin-Lin, W. Gang, H. Fei, and S. Kumar, "The Impact of Spectrum Sensing Frequency and Packet-Loading Scheme on Multimedia Transmission Over Cognitive Radio Networks," Multimedia, IEEE Transactions on, vol. 13, pp. 748-761, 2011.

[15] L. Ying-Chang, Z. Yonghong, E. C. Y. Peh, and H. Anh Tuan, "Sensing-Throughput Tradeoff for Cognitive Radio Networks," Wireless Communications, IEEE Transactions on, vol. 7, pp. 1326-1337, 2008.

[16] A. Ghasemi and E. S. Sousa, "Spectrum sensing in cognitive radio networks: requirements, challenges and design trade-offs," Communications Magazine, IEEE, vol. 46, pp. 32-39, 2008.

[17] H. Kim and K. G. Shin, "In-band spectrum sensing in cognitive radio networks: energy detection or feature detection?," in Proceedings of the 14th ACM international conference on Mobile computing and networking, 2008, pp. 14-25.

[18] R. Tandra and A. Sahai, "SNR Walls for Signal Detection," Selected Topics in Signal Processing, IEEE Journal of, vol. 2, pp. 4-17, 2008.

[19] D. Cabric, S. M. Mishra, and R. W. Brodersen, "Implementation issues in spectrum sensing for cognitive radios," in Signals, Systems and Computers, 2004. Conference Record of the Thirty-Eighth Asilomar Conference on, 2004, pp. 772-776 Vol.1.

[20] Y. Zeng and Y.-C. Liang, "Spectrum-sensing algorithms for cognitive radio based on statistical covariances," Vehicular Technology, IEEE Transactions on, vol. 58, pp. 1804-1815, 2009.

[21] Y. Li, S. Shen, and Q. Wang, "A Blind Detection Algorithm Utilizing Statistical Covariance in Cognitive Radio," International Journal of Computer Science Issues (IJCSI), vol. 9, pp. 7-12, 2012.

[22] K. S. Kumar, R. Saravanan, and R. Muthaiah, "Cognitive Radio Spectrum Sensing Algorithms based on Eigenvalue and Covariance methods," International Journal of Engineering & Technology (0975-4024), vol. 5, 2013.

[23] J. Lundén, V. Koivunen, A. Huttunen, and H. V. Poor, "Spectrum sensing in cognitive radios based on multiple cyclic frequencies," in Cognitive Radio Oriented Wireless Networks and Communications, 2007. CrownCom 2007. 2nd International Conference on, 2007, pp. 37-43.

[24] S. Enserink and D. Cochran, "A cyclostationary feature detector," in Signals, Systems and Computers, 1994. 1994 Conference Record of the Twenty-Eighth Asilomar Conference on, 1994, pp. 806-810.

[25] H. Tang, "Some physical layer issues of wide-band cognitive radio systems," in New frontiers in dynamic spectrum access networks, 2005. DySPAN 2005. 2005 first IEEE international symposium on, 2005, pp. 151-159.

[26] T. Yucek and H. Arslan, "Spectrum characterization for opportunistic cognitive radio systems," in Military Communications Conference, 2006. MILCOM 2006. IEEE, 2006, pp. 1-6.

[27] T. Farnham, G. Clemo, R. Haines, E. Seidel, A. Benamar, S. Billington, et al., "Ist-trust: A perspective on the reconfiguration of future mobile terminals using software download," in Personal, Indoor and Mobile Radio Communications, 2000. PIMRC 2000. The 11th IEEE International Symposium on, 2000, pp. 1054-1059.

[28] S. Shobana, R. Saravanan, and R. Muthaiah, "Optimal Spectrum Sensing Approach on Cognitive Radio Systems," 2013.

[29] S. Shobana, R. Saravanan, and R. Muthaiah, "Matched Filter Based Spectrum Sensing on Cognitive Radio for OFDM WLANs," International Journal of Engineering and Technology (IJET), vol. 5, 2013.

[30] W. Chien-Min, S. Hui-Kai, L. Maw-Lin, L. Yi-Ching, and L. Chih-Pin, "Cooperative Power and Contention Control MAC Protocol in Multichannel Cognitive Radio Ad Hoc Networks," in Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2014 Eighth International Conference on, 2014, pp. 305-309.

[31] X. Dongyue, E. Ekici, and M. C. Vuran, "Cooperative Spectrum Sensing in Cognitive Radio Networks Using Multidimensional Correlations," Wireless Communications, IEEE Transactions on, vol. 13, pp. 1832-1843, 2014.

[32] Z. Yulong, Y. Yu-Dong, and Z. Baoyu, "Cooperative relay techniques for cognitive radio systems: Spectrum sensing and secondary user transmissions," Communications Magazine, IEEE, vol. 50, pp. 98-103, 2012.

AUTHORS

Nabil Giweli received the B.Sc. degree in Communication Engineering from Tripoli University, Libya, in 1997, the Master degree in Information and Communication Technology (with the dean medal award) from the Western Sydney University in 2011, and another M.Sc. form the same university in Cloud Security in 2013. Currently, he is a Ph.D. candidate and a casual teacher at the School of Computing, Engineering and Mathematics, Western Sydney University, Australia. His current research area is in Cognitive Radio Technologies.

Seyed Shahrestani completed his PhD degree in Electrical and Information Engineering at the University of Sydney. He joined Western Sydney University (Western) in 1999, where he is currently a Senior Lecturer. He is also the head of the Networking, Security and Cloud Research (NSCR) group at Western. His main teaching and research interests include: computer networking, management and security of networked systems, analysis, control and management of complex systems, artificial intelligence applications, and health ICT. He is also highly active in higher degree research training supervision, with successful results.

Dr Hon Cheung graduated from The University of Western Australia in 1984 with First Class Honours in Electrical Engineering. He received his PhD degree from the same university in 1988. He was a lecturer in the Department of Electronic Engineering, Hong Kong Polytechnic from 1988 to 1990. From 1990 to 1999, he was a lecturer in Computer Engineering at Edith Cowan University, Western Australia. He has been a senior lecturer in Computing at Western Sydney University since 2000. Dr Cheung has research experience in a number of areas, including conventional methods in artificial intelligence, fuzzy sets, artificial neural networks, digital signal processing, image processing, network security and forensics, and communications and networking. In the area of teaching, Dr Cheung has experience in development and delivery of a relative large number of subjects in computer science, electrical and electronic engineering, computer engineering and networking.

INTENTIONAL BLANK

A METHOD OF TARGET TRACKING AND PREDICTION BASED ON GEOMAGNETIC SENSOR TECHNOLOGY

Xinmin Tang[1], Shangfeng Gao[1], Songchen Han[2], Zhiyuan Shen[1],
Liping Di[2] and Binbin Liang[2]

[1]College of Civil Aviation,
Nanjing University of Aeronautics and Astronautics, Nanjing, China
tangxinmin@nuaa.edu.cn
[2]School of Aeronautics & Astronautics,
Si Chuan University, Chengdu, China
hansongchen@nuaa.edu.cn

ABSTRACT

In view of the inherent defects in current airport surface surveillance system, this paper proposes an asynchronous target-perceiving-event driven surface target surveillance scheme based on the geomagnetic sensor technology. Furthermore, a surface target tracking and prediction algorithm based on I-IMM is given, which is improved on the basis of IMM algorithm in the following aspects: Weighted sum is performed on the mean of residual errors and model probabilistic likelihood function is reconstructed, thus increasing the identification of a true motion model; Fixed model transition probability is updated with model posterior information, thus accelerating model switching as well as increasing the identification of a model. In the period when a target is non-perceptible, prediction of target trajectories can be implemented through the target motion model identified with I-IMM algorithm. Simulation results indicate that I-IMM algorithm is more effective and advantageous in comparison with the standard IMM algorithm.

KEYWORDS

surface surveillance, geomagnetic sensor technology, I-IMM, target trajectory, tracking and prediction

1. INTRODUCTION

Encountering increasingly serious problems regarding safety and efficiency on the airport surface, ICAO proposed an Advanced Surface Movement Guidance and Control System (A-SMGCS) [1]. The system performs surveillance, routing, guidance and control on a moving target using various sensor technologies, wherein surveillance is defined as the most important function in A-SMGCS[2].

At present, surface surveillance is mainly implemented through surface surveillance radar (SMR), automatic dependent surveillance (ADS) and multilateration (MLAT) and other surveillance

David C. Wyld et al. (Eds) : NETCOM, NCS, WiMoNe, CSEIT, SPM - 2015
pp. 21–37, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51603

devices. These three systems, however, have the following inherent defects: (1) SMR is susceptible to factors like building block, ground clutters and weather; (2) MALT and ADS can only monitor a target equipped with a transponder, but not a non-cooperative target on the surface; (3) These three surveillance approaches feature in low trajectory update rate, communication delay and high cost. The study of surface moving surveillance system based on event-driven non-cooperative can fundamentally solve the above-mentioned defects. Honeywell developed a dual infrared/magnetic sensor, and thousands of such sensors are equipped at airports for detection of the aircraft [3]. Chartier et al. proposed that the position of the aircraft could be determined though the information of coil sensor installed on the boundary of the airfield pavement segmentation[4]. K. Dimitropoulos et al. proposed to detect a magnetic target using a magnetic sensor network[5]. Schonefeld J et al. conducted comprehensive analysis on the performance of runway intrusion prevention system, XL-RIAS, based on distributed sensors, and testified that the response rate thereof is faster than that of ASDE-X [6].

Trajectory tracking and prediction of a target on the airport surface is a main function of the surface surveillance system. Two main trajectory tracking and prediction algorithms are studied. One is algorithm based on parameter identification in aircraft dynamics and kinemics models, wherein Gong studied taxing velocity and acceleration characteristics of the aircraft, and obtained kinematics trajectory model using regression analysis [7]; Capoozi et al. analyzed historical data of surface surveillance and excavated parameters of kinematics equation model [8]; Rabah W et al. employed high-gain observer and variable structure control method to perform output feedback tracking on nonlinear system, with effects of tracking uncertain system being undesirable[9]. Another is algorithm based on optimal estimation theory, wherein conventional Kalman filters like α-β and α-β-γ are single model tracking algorithms, which are not suitable for the variety and uncertainty of target motion on the surface[10]; Farina et al. applied the restricted information to IMM model set self-adaption in consideration of peculiarity of a target motion on the airport surface, thus improving the tracking precision [11]; Gong Shuli et al. applied VS-IMM algorithm to the surface target tracking in combination with the airport map [12].

In order to solve the inherent defects in SMR, ADS and MLAT, an asynchronous target-perceiving-event driven surface moving target surveillance scheme based on the geomagnetic sensor technology is proposed in this paper. In this scheme, geomagnetic detection nodes are deployed in the center of the runway/taxiway, thereby the target position can be accurately perceived as well as the real-time velocity being obtained as a target passes through the nodes. However, the node deployment density is low, causing the continuous motion state of a moving target in the adjacent nodes not to be perceived. Regarding such problems, this paper presents a new algorithm I-IMM, in which the likelihood function of IMM algorithm is improved to increase the identification of a true motion model. Furthermore, motion model switching is accelerated and model identification is improved through modification of state transition probability for self-adaption using posterior information. In the period when the motion state of a target is not perceptible, memory tracking and prediction on target trajectories can be implemented through the target motion model identified with I-IMM algorithm, combined with the final self-adapting state transition probability in the perceptible period.

2. SURFACE TARGET SURVEILLANCE SCHEME BASED ON GEO-MAGNETIC SENSOR TECHNOLOGY

2.1. Surface target surveillance scheme

In general, moving targets on the airport surface comprise aircraft and special vehicles, which are relatively large ferromagnetic objects, generating disturbance to the surrounding magnetic field during their moving, thereby targets can be detected by the geomagnetic sensor with an anisotropic magnetoresistance effect according to the disturbance[13]. Combination of geomagnetic sensor and event-driven wireless sensor network can achieve high precision, small volume, low cost, no need for wiring and deployment flexibility, without affecting the surface surveillance performance. The surface moving target surveillance scheme based on the magnetic sensor technology is as shown in Figure 1.

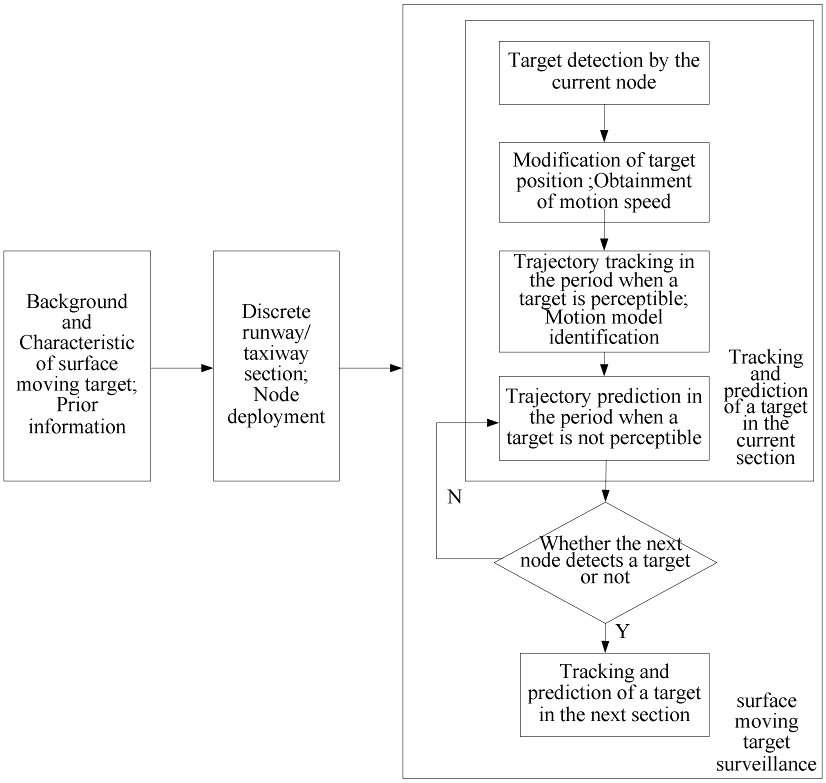

Background and Characteristic of surface moving target; Prior information
Discrete runway/ taxiway section; Node deployment
Target detection by the current node
Modification of target position ;Obtainment of motion speed
Trajectory tracking in the period when a target is perceptible; Motion model identification
Trajectory prediction in the period when a target is not perceptible
Tracking and prediction of a target in the current section
N
Whether the next node detects a target or not
Y
Tracking and prediction of a target in the next section
surface moving target surveillance

Figure1. Surveillance scheme for targets on the surface

2.2. Node deployment and runway section information

Due to the large surveillance area, the geomagnetic technology-based surveillance scheme needs to consider the way of deployment and quantity of geomagnetic detection nodes to reduce the cost of tracking and communication redundancy. From surface restrictions given by reference [14], it can be known that considering the restrictions on a moving target in different airport areas, the target motion characteristics can be transcendentally predicted. In this paper, nodes are deployed in combination with surface restrictions as is shown in Figure 2 (taking a taxiway section as an example).

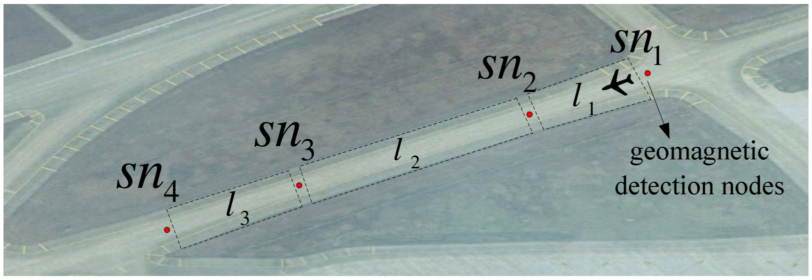


Figure 2. Node deployment

The taxiing route of a moving target on the surface is divided into different sections, $L=\{l_1,l_2,l_3\}$. A target mostly maintains single motion characteristics in different sections. For instance, the aircraft maintains accelerated motion during section l_1, constant motion during section l_2, and decelerated motion during section l_3. Geomagnetic detection nodes $SN=\{sn_1,sn_2,sn_3,sn_4\}$, are deployed at the cut-off rule of the adjacent sections. Each section comprises four parameters. For instance, section i can be defined as $(l_i,sn_i,sn_{i+1},long_i)$, where l_i denotes number, sn_i denotes start node, sn_{i+1} denotes terminal node, and $long_i$ denotes length. The section information is preserved in geomagnetic detection nodes for distributed computation after nodes perceive a target. In above-mentioned deployment, nodes can accurately perceive the target position as a target passes through them and modify the previous position information, and the velocity information can also be modified instantaneously via the target velocity obtained from nodes.

3. I-IMM-BASED SURFACE TARGET TRACKING AND PREDICTION ALGORITHM

In the surface surveillance scheme based on geomagnetic sensor technology, a target is in a perceptible state as it passes through nodes, which provide the velocity information. The data volume, however, is not large and can only be seen as small data samples. When a target completely detaches from nodes, it would be in an imperceptible state when moving in the section between nodes. Accordingly, the target tracking and prediction algorithm put forward in this paper needs to satisfy requirements as follows: When a target is perceptible, the real-time tracking is performed using the observed velocity information and the target motion model is accurately identified; When a target is not perceptible, extrapolated prediction is performed on trajectories thereof using the identified motion model.

3.1. I-IMM algorithm

I-IMM algorithm is improved based on IMM algorithm in the following two aspects: Weighted sum is performed on the mean of residual errors and model probabilistic likelihood function is reconstructed, thus increasing the identification of a true motion model; Model transition probability is updated for self-adaption using model posterior probability, thus accelerating model switching as well as increasing the identification of a model. The schematic diagram of I-IMM algorithm is as shown in Figure 3. This algorithm comprises the following 5 steps:Input interaction; Kalman filter; Model probability update; Model transition probability self-adaption; Output fusion.

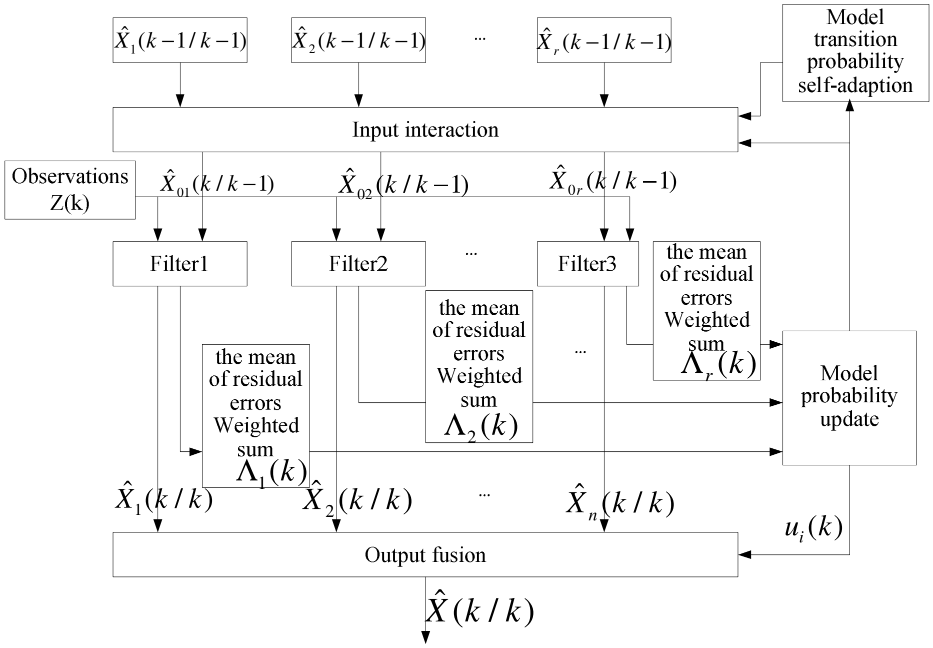


Figure 3. Schematic diagram of I-IMM algorithm

3.1.1. Input interaction

Assuming that a model set consists of r motion models, the state estimation value and covariance matrix of each model at time $k-1$ are respectively as follows: $\hat{X}_j(k-1|k-1)$ and $\hat{P}_j(k-1|k-1)$, $j=1,2,\cdots,r$.

After interaction, the input in model j at time k is expressed as follows:

$$\hat{X}_{0j}(k-1|k-1)=\sum_{i=1}^{r}\hat{X}_i(k-1|k-1)u_{ij}(k-1|k-1) \tag{1}$$

$$\begin{aligned}\hat{P}_{0j}(k-1|k-1)=\sum_{i=1}^{r}u_{ij}(k-1|k-1)\{\hat{P}_i(k-1|k-1)\\ +[\hat{X}_i(k-1|k-1)-\hat{X}_{0j}(k-1|k-1)]^T\}\end{aligned} \tag{2}$$

Where, the mixture probability after input interaction is defined as:

$$u_{ij}(k-1|k-1)=p_{ij}u_i(k-1)/\bar{c}_j \tag{3}$$

Where, $\overline{c}_j = \sum_{i=1}^{r} p_{ij} u_i(k-1)$, p_{ij} denotes model transition probability, and $u_i(k-1)$ denotes probability in model i at time $k-1$.

3.1.2. Kalman filter

Kalman filter consists of prediction process and update process. The prediction process is expressed by Eq. (4) and Eq. (5):

$$\hat{X}_j(k|k-1) = F_j \hat{X}_{0j}(k-1|k-1) \tag{4}$$

$$\hat{P}_j(k|k-1) = F_j \hat{P}_{0j}(k-1|k-1) F_j^{\mathrm{T}} + Q \tag{5}$$

In the above equations, F_j is the model state transition matrix; Q is the noise covariance in each model during the estimation.

Residual sequence and covariance matrix are:

$$r_j(k) = Z_j(k) - H\hat{X}_j(k|k-1) \tag{6}$$

$$S_j(k) = H\hat{P}_j(k|k-1)H^{\mathrm{T}} + R \tag{7}$$

In the above equations, $Z_j(k)$ is the observed value for the time k; H is the observation matrix; R is the noise covariance of observation.

Kalman filter gain matrix is:

$$K_j(k) = \hat{P}_j(k|k-1)H^T S_j^{-1}(k) \tag{8}$$

State estimate and covariance matrix update are expressed as follows:

$$\hat{X}_j(k|k) = \hat{X}_j(k|k-1) + K_j(k) r_j(k) \tag{9}$$

$$\hat{P}_j(k|k) = \hat{P}_j(k|k-1) - K_j(k) S_j(k|k-1) K_j^{\mathrm{T}}(k) \tag{10}$$

3.1.3. Model probability update

In IMM algorithm, maximum likelihood function in model j is as given in Eq. (11):

$$\Lambda_j(k) = \frac{1}{\sqrt{2\pi |S_j(k)|}} \exp\{-\frac{1}{2} r_j^T(k) S_j^{-1}(k) r_j(k)\} \tag{11}$$

As can be seen from the Eq. (11), it is assumed that the motion model set can contain all motion models of a target during the operation in IMM algorithm. However, due to the factors like uncertainty of the motion of a surface target, surface restrictions and spot dispatch, the target motion model may exceed the model set in the algorithm. Therefore, innovation information is no longer considered to obey Gaussian distribution, in which mean value is zero and variance is $S_j(k)$, and thus model probabilistic likelihood function is reconstructed.

Let assume the true motion model of a surface moving target to be as follows:

$$X_T(k) = F_T(k-1)X_T(k-1) + w_T(k-1) \tag{12}$$

$$Z(k) = HX_T(k) + v_T(k) \tag{13}$$

Define the model state transition matrix error as follows:

$$\Delta F_j = F_T - F_j \tag{14}$$

Define the state estimation error as follows:

$$e_j(k-1) = X_T(k-1 \mid k-1) - \hat{X}_j(k-1 \mid k-1) \tag{15}$$

Expression for the state estimation error after input interaction is obtained:

$$e_{0j}(k-1) = X_T(k-1 \mid k-1) - \hat{X}_{0j}(k-1 \mid k-1) = \sum_{i=1}^{r} u_{ij}(k-1 \mid k-1)e_i(k-1) \tag{16}$$

Given by Eq. (6) and Eq. (12), the residual error is obtained:

$$r_j(k) = HX_T(k) + v_T(k) - HF_j\hat{X}_{0j}(k-1 \mid k-1) \tag{17}$$

Given by Eq. (12) and Eq. (15), the residual error is obtained:

$$r_j(k) = HF_T e_{0j}(k\text{ -}1) + H\Delta F_j \hat{X}_{0j}(k\text{ -}1 \mid k\text{ -}1) + Hw_T(k) + v_T(k) \tag{18}$$

Mean value obtained from Eq. (18) can be expressed as follows:

$$\bar{r}_j(k) = HF_T\bar{e}_{0j}(k-1) + H\Delta F_j\hat{X}_{0j}(k\text{ -}1 \mid k\text{ -}1) \tag{19}$$

$$\text{Where,} \bar{e}_{0j}(k-1) = \sum_{i=1}^{r} u_{ij}(k-1 \mid k-1)\bar{e}_i(k-1) \tag{20}$$

Because of the uncertainty of the true motion model of a surface target, the quantization of F_T and ΔF_j in Eq. (19) cannot be performed, causing the mean of residual errors not to be obtained. To solve this problem, the true motion model of a target is assumed to be j, and weighted sum is performed on another model in the model set to obtain $\bar{r}_j(k)$.

$$\bar{r}_j(k) = \sum_{i=1}^{r}\left[HF_j\bar{e}_{0i}(k-1) + H\Delta F_i\hat{X}_{0j}(k\text{ -}1 \mid k\text{ -}1)_i\right]u_i\left(k-1 \mid k-1\right) \tag{21}$$

Where, $\Delta F_i = F_j - F_i$, and $\bar{e}_{0i}(k-1) = \sum_{n=1}^{r} u_{in}(k-1 \mid k-1)e_n(k-1)$.

Then, maximum likelihood function in model j at time k can be given as:

$$\Lambda_j(k) = \frac{1}{\sqrt{2\pi\left|S_j(k)\right|}}\exp\left\{-\frac{1}{2}\left[r_j(k) - \bar{r}_j(k)\right]^T S_j^{-1}(k)\left[r_j(k) - \bar{r}_j(k)\right]\right\} \tag{22}$$

$$u_j(k) = \frac{1}{c}\Lambda_j(k)\sum_{i=1}^{r} p_{ij}u_i(k-1) = \Lambda_j(k)\bar{c}_j / c \tag{23}$$

$$\text{Where,}\, c = \sum_{j=1}^{r}\Lambda_j(k)\bar{c}_j \tag{24}$$

3.1.4. Model transition probability self-adaption

In IMM algorithm, because of the uncertainty of the target maneuver and the distortion of the prior information, the fixed model transition probability p_{ij} fails to reflect the true motion model of a target, and switching velocity between models is also delayed during the target maneuver. Given that the observed velocity is small sample information, applying the fixed model transition probability p_{ij} may likely cause the target motion model hard to be identified or even not to be identified. Therefore, the model transition probability p_{ij} is updated using posterior information in I-IMM algorithm to solve this problem.

Assuming that the probability in model *j* at time $k-1$ is $u_j(k-1)$ and at time k is $u_j(k)$, the probability differential value of the same model at adjoining times reflects the change in the matching degree between model *j* and the true motion model. The rate of change of the posterior probability in model *j* can be defined as:

$$\Delta u_j(k) = u_j(k) - u_j(k-1) \tag{25}$$

Let the model transition probability from model i to model *j* at time $k-1$ be $p_{ij}(k-1)$, and update $p_{ij}(k-1)$ using $\Delta u_j(k)$, then the expression is obtained:

$$p'_{ij}(k) = \exp(\Delta u_j(k))\, p_{ij}(k-1) \tag{26}$$

Model transition probability needs to satisfy basic properties as follows:

$$\begin{cases} 0 < p_{ij} < 1, i, j = 1, 2, \cdots, r \\ \sum_{j=1}^{r} p_{ij} = 1 \end{cases} \tag{27}$$

Then, normalization needs to be performed on $p'_{ij}(k)$, and the transition probability $p_{ij}(k)$ can be obtained:

$$p_{ij}(k) = \frac{p'_{ij}(k)}{\sum_{j=1}^{r} p'_{ij}(k)} = \frac{\exp(\Delta u_j(k))\, p_{ij}(k-1)}{\sum_{j=1}^{r}\exp(\Delta u_j(k))\, p_{ij}(k-1)} \tag{28}$$

As can be seen from Eq. (28), updated $p_{ij}\left(k\right), i = 1, 2, \mathrm{L}\ \ , r$ increases as the transition of model from model i to model j , when the posterior information $\Delta u_j\left(k\right)$ increases, thus model j plays a critical role in input interaction at next time period.

3.1.5. Output fusion

Interactive output results at time k are expressed as follows:

$$\hat{X}\left(k \mid k\right) = \sum_{j=1}^{r} \hat{X}_j\left(k \mid k\right) u_j\left(k\right) \tag{29}$$

$$\hat{P}\left(k \mid k\right) = \sum_{j=1}^{r} u_j\left(k\right)\left\{P_j\left(k \mid k\right) + \left[\hat{X}_j\left(k \mid k\right) - \hat{X}\left(k \mid k\right)\right]\left[\hat{X}_j\left(k \mid k\right) - \hat{X}\left(k \mid k\right)\right]^{\mathrm{T}}\right\} \tag{30}$$

3.2. Trajectory prediction of targets not perceptible

A target would be not perceptible as moving in the section between two adjacent nodes, requiring memory tracking of target trajectories using extrapolated prediction.

At last moment of the period when a target is perceptible, I-IMM provides identification of each model in the model set, namely, model posterior probability $u_j\left(k\right), j = 1, 2, \mathrm{L}\ \ , r$. Then given by the self-adapting model transition probability $p_{ij}\left(k\right)$, the expression for prediction probability of each model in the model set when a target not perceptible at time $k+1$ can be obtained:

$$u'_j\left(k+1 \mid k\right) = \sum_{i=1}^{r} u_i(k) \mathrm{g} p_{ij}\left(k\right), j = 1, 2, \mathrm{L}\ \ , r \tag{31}$$

At the same time, posterior probability of each model needs to satisfy the following properties:

$$\begin{cases} 0 < u_j < 1, j = 1, 2, \mathrm{L}\ \ , r \\ \sum_{j=1}^{r} u_j = 1 \end{cases} \tag{32}$$

Then, normalization needs to be performed on posterior probability of each model predicted from Eq. (31), and model posterior probability at moment $k+1$ is obtained:

$$u_j\left(k+1 \mid k\right) = \frac{u'_j\left(k+1 \mid k\right)}{\sum_{j=1}^{r} u'_j\left(k+1 \mid k\right)} \tag{33}$$

After substituting state predicted value, $\hat{X}_j\left(k+1 \mid k\right)$ and prediction model probability, $u_j\left(k+1 \mid k\right)$ of each model into Eq. (29), state predicted value of the period when a target is not perceptible can be defined as :

$$\hat{X}\left(k+1|k\right)=\sum_{j=1}^{r}\hat{X}_j\left(k+1|k\right)u_j\left(k+1|k\right) \tag{34}$$

By performing extrapolated prediction on surface target trajectories using state predicted value obtained from Eq. (34), memory tracking and prediction can be implemented on a target not perceptible.

4. Simulation and Analysis

4.1. Preparation for simulation

This paper presents, taking aircraft passing through a certain geomagnetic detection node on the taxiway for an example, IMM algorithm and I-IMM algorithm are compared through Monte Carlo simulation, regarding the performance of trajectory tracking of the aircraft perceptible and trajectory prediction of the aircraft not perceptible.

According to the motion characteristics of the aircraft on the surface, the aircraft motion can be expressed by a model set comprising constant velocity (CV) model, constant acceleration (CA) model and constant jerk (CJ) model. State transition matrixes of three models are respectively expressions as follows:

$$F_{CV}=\begin{bmatrix}1 & T & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix},\quad F_{CA}=\begin{bmatrix}1 & T & \frac{T^2}{2} & 0\\ 0 & 1 & T & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0\end{bmatrix},\quad F_{CJ}=\begin{bmatrix}1 & T & \frac{T^2}{2} & \frac{T^3}{6}\\ 0 & 1 & T & \frac{T^2}{2}\\ 0 & 0 & 1 & T\\ 0 & 0 & 0 & 1\end{bmatrix}$$

Where, T is the interval of sampling. Motion state vector of the aircraft is $X=[x \quad \dot{x} \quad \ddot{x} \quad \dddot{x}]^{\mathrm{T}}$, and observation matrix is $H=\begin{bmatrix}0 & 1 & 0 & 0\end{bmatrix}$.

In the period when aircraft is perceptible, the process of the aircraft operation is as follows:(1) Performing CJ at $0.3\,m/s^3$ from 0 to 4.5 seconds; (2) Performing CA at $0.45\,m/s^2$ from 4.58 to 12 seconds; (3) Performing CV at the velocity obtained from step (2) from 12 to15 seconds.

In the period when aircraft is not perceptible, it maintains CV for about 30 seconds at the velocity obtained from step (3) and then operates to the next detection node.

The actual position of the aircraft according to the operation process is as shown in Figure 4.

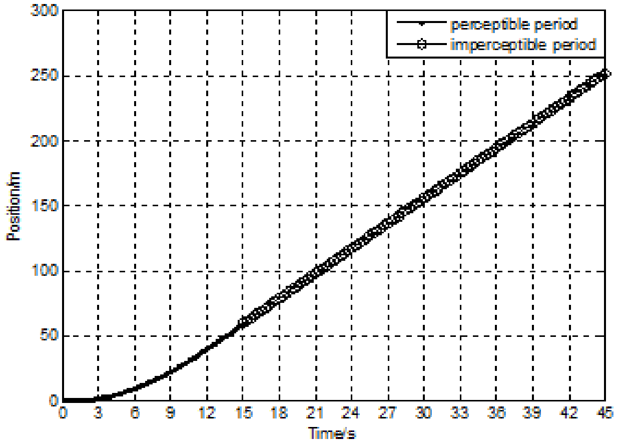


Figure 4. Actual position of the aircraft

The simulation parameters selection is as follows: Noise covariance in each model during the estimation is $Q=\begin{bmatrix} 0.01 & 0 & 0 & 0 \\ 0 & 0.01 & 0 & 0 \\ 0 & 0 & 0.01 & 0 \\ 0 & 0 & 0 & 0.01 \end{bmatrix}$; Noise covariance of velocity observation is $R=0.15$; the interval of sampling is $T=0.3s$; Initial model probabilities of CV model，CA model and CJ model are respectively 0.4，0.3 and 0.3; Initial model transition matrix can be given as: $P_{markov}=\begin{bmatrix} 0.9 & 0.05 & 0.05 \\ 0.05 & 0.9 & 0.05 \\ 0.05 & 0.05 & 0.9 \end{bmatrix}$.

4.2. Simulation results and analysis in perceptible period

Simulation results in the period when the aircraft is perceptible are as displayed in Figure 5 to 8.

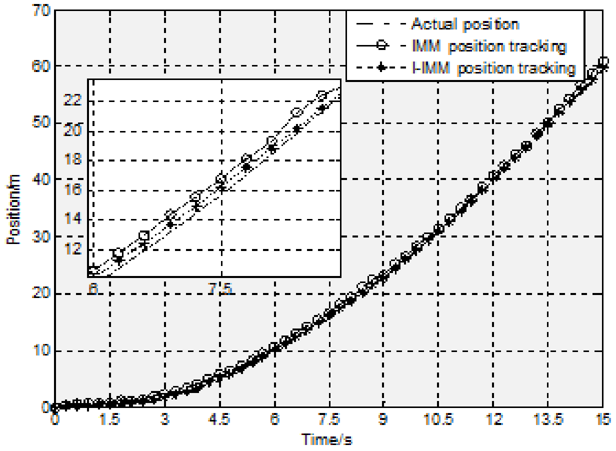


Figure 5. Position tracking

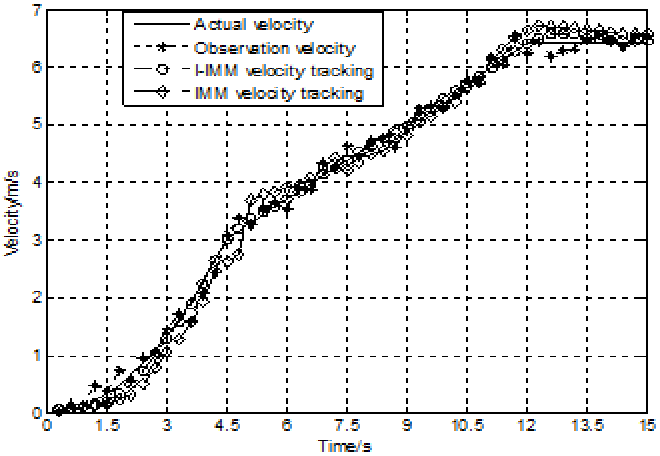


Figure 6. Velocity tracking

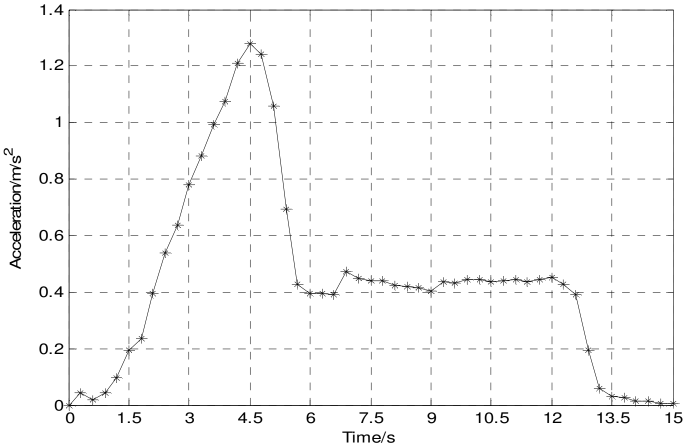


Figure 7. Acceleration tracking

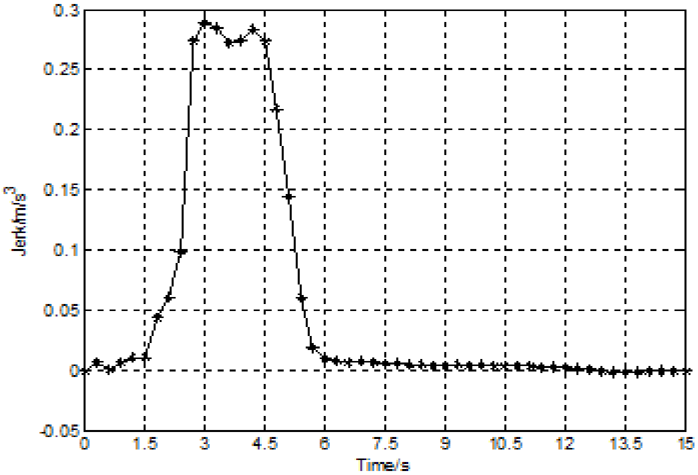


Figure 8. Jerk tracking

Figure 5 to 8 illustrate the excellence of I-IMM algorithm when used to track the motion state of the aircraft. Fig. 5 and 6 demonstrate that compared with IMM algorithm, the position and velocity tracked with I-IMM algorithm are more approximate to the actual position and velocity of the aircraft. To show the advantage of I-IMM more manifestly, the position error curve and velocity root-mean-square error(RMSE) curve of I-IMM and IMM algorithm are respectively plotted, as shown in Figure 9 and 10.

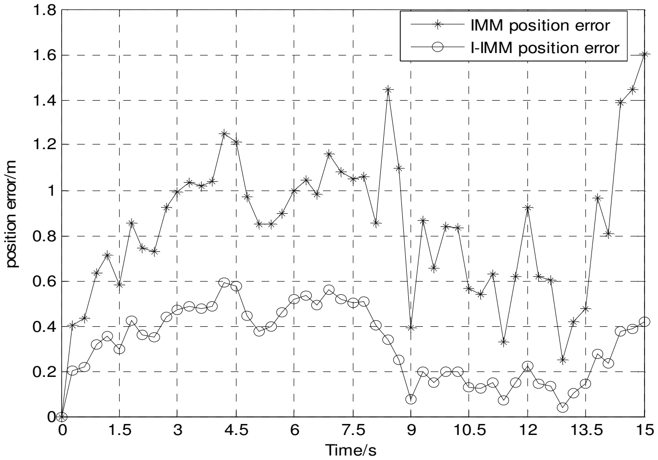


Figure 9. Position error curve

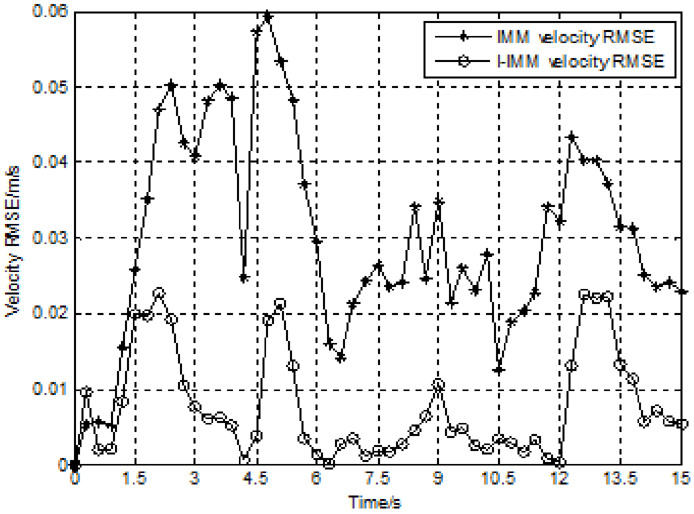


Figure 10. Velocity RMSE curve

Figure 9 shows maximum position error using IMM algorithm is 1.600m, while using I-IMM algorithm is only 0.600m. Figure 10 shows maximum RMSE error using IMM algorithm is 0.059m/s, while using I-IMM algorithm is 0.023m/s. As can be seen from the result, the tracking precision is well improved when using I-IMM algorithm.

Figure 11 presents the selection probability curves of CV, CA and CJ models when IMM and I-IMM algorithm are respectively employed. Figure 11 demonstrates that IMM algorithm cannot identify each motion model very clearly, and three selection probability curves intersect intensely. For instance, in constant accelerating phase, IMM algorithm's maximum identification degree of CA model is 0.710; Comparatively, I-IMM algorithm can largely improve the identification degree. In constant jerking phase, the maximum identification degree of CJ model is 0.987, while in constant accelerating phase, the maximum identification degree of CA model can reach to 0.987, and in constant velocity phase, the maximum identification degree of CV model is 0.987. As for model switching, the switching velocity in I-IMM algorithm is much faster than that in IMM algorithm. In IMM algorithm, it takes 2.4 seconds to switch from CJ model to CA model, and 1.8 seconds from CA model to CV model. In comparison, when employing I-IMM algorithm, it only takes 0.9 seconds to switch from CJ model to CA model, and only 0.9 seconds from CA model to CV model.

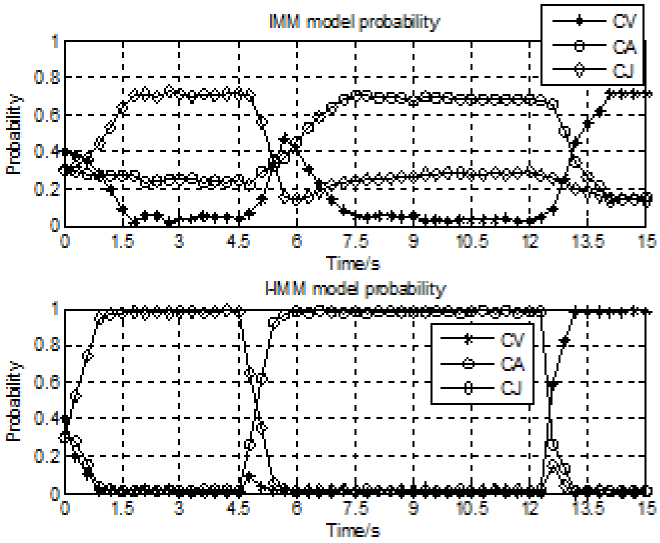


Figure 11. Selection probability in IMM and I-IMM

4.3. Simulation results and analysis in imperceptible period

Simulation results of trajectory prediction in the period when the aircraft is not perceptible are as displayed in Figure 12 and 13.

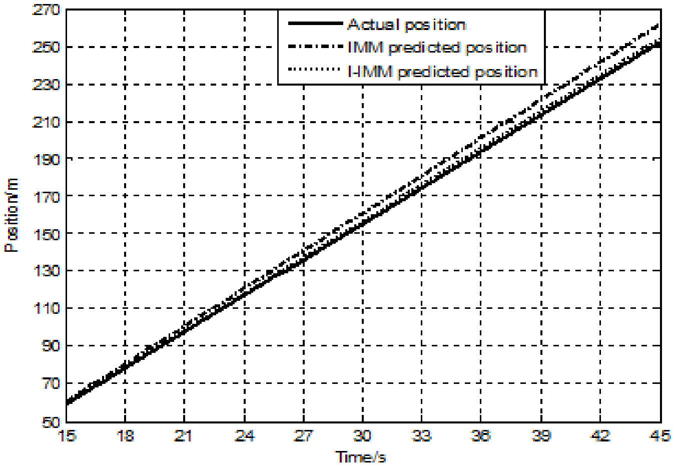


Figure 12. Position prediction

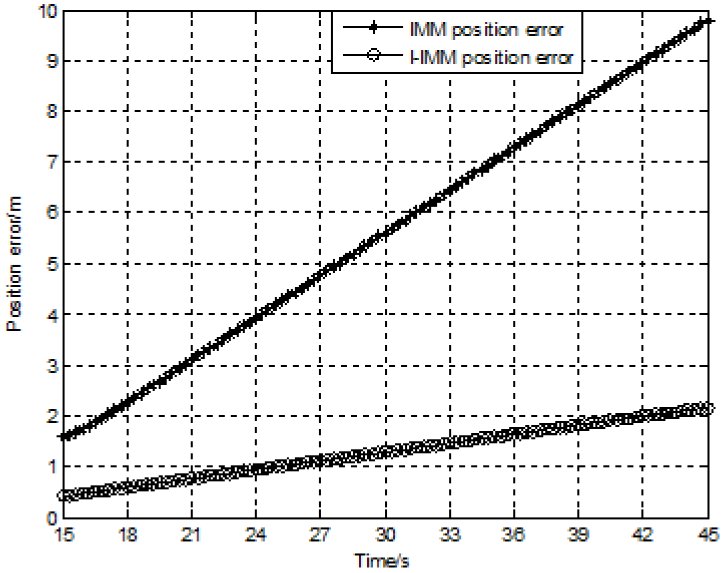


Figure 13. Position prediction error

Figure 12 illustrates that the deviation between the aircraft position predicted with either IMM or I-IMM algorithm and the real position increases with the increase of the running time. Figure 13 illustrates that at the last moment of position prediction, the position prediction error is accumulated to 9.790m when IMM algorithm is used, while only to 2.160m when I-IMM algorithm is used. It is apparent that I-IMM algorithm outperforms IMM algorithm in terms of trajectory extrapolated prediction, particularly in the period when the aircraft is not perceptible.

5. CONCLUSIONS

In view of the inherent defects in current surface surveillance system, this paper proposes a asynchronous target-perceiving-event driven surface moving target surveillance scheme based on the geomagnetic sensor technology. Furthermore, a surface moving target tracking and prediction algorithm is given based on I-IMM, which is improved on the basis of IMM algorithm in the following aspects: Weighted sum is performed on the mean of residual errors and model probabilistic likelihood function is reconstructed, thus increasing the identification of a true motion model; Model transition probability is updated for self-adaption with model posterior probability, thus accelerating model switching as well as increasing the identification of a model. At last, this paper presents simulation results of target tracking and prediction in both periods when a target is perceptible and not perceptible using two algorithms, demonstrating that the I-IMM algorithm is more effective than IMM algorithm, particularly when a target is not perceptible.

ACKNOWLEDGEMENTS

This work is supported by the National Science Foundation of China (grant no. U1433125), Civil Aviation Science and Technology Guidance Foundation (grant no. 14014J0340035), Science Foundation of Jiangsu Provence (grant no. BK20141413), Fundamental Research Funds for the Central Universities (grant no. NS2014065), and Startup Project for Introduced Talents of Sichuan University.

REFERENCES

[1] International Civil Aviation Organization. Doc. 9830-AN/452 Advanced surface movement guidance and control systems(A-SMGCS) manual. Montreal: ICAO, 2004.

[2] EUROCONTROL, Development and Validation of Improvement of Runway Safety Net (A-SMGCS Level 2) by Electronic Flight Strips–D16 Cost Benefit Analysis. Toulouse, EUROCONTROL, 2010.

[3] Stauffer D, French H & Lenz J. (1993) "A multi-sensor approach to airport surface traffic tracking", Digital Avionics Systems Conference, pp430-432.

[4] Chartier, E. & Hashemi, Z. (2001) "Surface surveillance systems using point sensors and segment-based tracking", Digital Avionics Systems, 2001. DASC. 20th Conference, Vol. 1, pp2E1/1 - 2E1/8.

[5] Dimitropoulos, K., Grammalidis, N., Gragopoulos, I., Gao, H., Heuer, T.& Weinmann, M., et al. (2006) "Detection, tracking and classification of vehicles and aircraft based on magnetic sensing technology", Proceedings of World Academy of Science Engineering & Technology, Vol. 4, No. 1, pp195-200.

[6] Schonefeld J, & Moller D P F. (2012) "Fast and robust detection of runway incursions using localized sensors", Multisensor Fusion and Integration for Intelligent Systems (MFI), pp33 - 39.

[7] Gong, C. (2009) "Kinematic airport surface trajectory model development", 9th AIAA Aviation Technology, Integration, and Operations Conference (ATIO) , pp175-179.

[8] Capozzi, B., Pledgie, S. & Kistler, M. (2010) "Surface Trajectory Characterization Report", NRA Contact NNA09DA89C, pp1-20.

[9] Aldhaheri, R. W. & Khalil, H. K. (1996) "Effect of unmodeled actuator dynamics on output feedback stabilization of nonlinear systems", Automatica, Vol. 32, No. 96, pp1323–1327.

[10] Zhou Hongren, Jing Zhongliang & Wang Peide. (1990) "Tracking Maneuvering Targets", Beijing: National Defense Industry Press.

[11] Farina A, Ferranti L & Golino G. (2003) "Constrained Tracking Filters For A-SMGCS", Information Fusion, 2003. Proceedings of the Sixth International Conference,Vol. 1, pp414-421.

[12] Gong Shuli, Tao Cheng & Wang Bangfeng. (2012) "A-SMGCS Surface Moving Target Tracking Based on VS-IMM Algorithm", Transactions of Nanjing University of Aeronautics & Astronautics, Vol. 44, No. 1, pp118-123.

[13] Klausner A, Tengg A & Rinner B. (2007)"Vehicle Classification on Multi-Sensor Smart Cameras Using Feature- and Decision-Fusion", First Acm/ieee International Conference on Distributed Smart Cameras, pp67 - 74.

[14] Herrero J G, Portas J A B. & Casar Corredera J R. (2003) "Use of map information for tracking targets on airport surface", IEEE Transactions on Aerospace and Electronic Systems, Vol. 39, No. 2, pp675 - 693.

AUTHORS

Dr. Xinmin Tang was born in 1979. He obtained his Ph.D. in Mechanical Engineering at the Harbin Institute of Technology in 2007. He is currently an Associate Professor in the College of Civil Aviation at Nanjing University of Aeronautics and Astronautics. His research interests include (1) Petri Net and Discrete Event Dynamic System theory; (2) Hybrid System Theory.

Shangfeng Gao was born in 1990. He holds a Master Degree in Engineering from the College of Civil Aviation at Nanjing University of Aeronautics and Astronautics. His major course is Transportation planning and management. His research interests include Advanced Surface Movement Guidance and Control System (A-SMGCS).

Dr. Songchen Han was born in 1964. He obtained his Ph.D. in Engineering at Harbin Institute of Technology. He is currently a professor in Sichuan University in China. His research interests include (1) next generation air traffic management system (2) air traffic planning and simulation.

Dr. Zhiyuan Shen is an assistant professor at college of civil aviation, Nanjing University of Aeronautics and Astronautics (NUAA). He received Ph.D in control science and engineering from the Harbin Institute of Technology, China. Between 2010 and 2012, he was a visiting scholar in Electrical and Computer Engineering at Georgia Institute of Technology, Atlanta. His current research interest includes ADS-B technique and 4D trajectory prediction.

Liping Di was born in 1964. She obtained his Bachelor Degree in Aircraft Control at the Harbin Institute of Technology in China. She is currently a teacher in Sichuan University. Her research interest is aircraft control.

Binbin Liang was born in 1990. He got a Master Degree in Engineering from the College of Civil Aviation at Nanjing University of Aeronautics and Astronautics. His research interests include civil aviation emergency management and computer vision.

INTENTIONAL BLANK

MEDIUM ACCESS CONTROL WITH SCHEDULED GROUP CONTENTION FOR MASSIVE M2M DEVICES

Joo Rak Kang[1] and Tae-Jin Lee[2]

[1]Samsung Electro-Mechanics Co., Ltd., Suwon, Korea
College of Information and Communication Engineering,
Sungkyunkwan University, Suwon, Korea
jrkang@skku.edu
[2]College of Information and Communication Engineering,
Sungkyunkwan University, Suwon, Korea
tjlee@skku.edu

ABSTRACT

In a dense Machine-to-Machine (M2M) network, a large number of stations contend to achieve transmission opportunity and it creates a critical congestion problem. To solve this issue, a group-based contention Medium Access Control (MAC) protocol is introduced. Stations are divided into small numbers of groups and only one station in each group will contend as a group leader to achieve the reserved time interval dedicated to a winner group. It can reduce the contention overhead and lessen the congestion problem. In this paper, we propose Scheduled Group Contention MAC (SGCMAC) protocol to enhance the group-based contention MAC. The proposed SGCMAC protocol divides groups based on the traffic categories of stations and schedules the contention groups to reduce the contention overhead. We also propose an efficient resource management mechanism in the group grant time to prevent the waste of time caused by idle stations. Simulations with IEEE 802.11ah parameters demonstrate that our proposed SGCMAC has performance gains over other group-based contention MAC protocols.

KEYWORDS

Medium Access Control (MAC), Machine-to-Machine, Grouping, Wireless LAN, Random Access Protocols

1. INTRODUCTION

Internet of Things (IoT) is a new paradigm for the future of the Internet. IoT will integrate a variety of sensors, actuators, smart devices and things through the Internet, and distributed smart services will be created to change our daily life. Machine-to-Machine (M2M) communications is considered as a basic communication technology for the realization of IoT. M2M communications involves information exchange among machines without any human interaction. M2M networks are expected to be widely utilized in many fields of various IoT applications such

David C. Wyld et al. (Eds) : NETCOM, NCS, WiMoNe, CSEIT, SPM - 2015
pp. 39–49, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51604

as public services, industrial automations, health care, intelligent transport systems, smart grids, and agricultural networks [1]-[5].

M2M networks have unique characteristics which are quite different from conventional networks. In M2M networks, tremendous number of devices may be involved in the service coverage and concurrent network access attempts may occur from the devices. Typically, the amount of data generated from each device may be small and data may be generated infrequently. If, however, many devices participate in and generate data at the same time, it may cause a critical congestion problem. Also, in M2M networks, high level of system automation, in which devices and systems exchange and share data automatically, is required. Therefore, congestion control is a fundamental issue in M2M networks. Especially, from the Medium Access Control (MAC) layer perspective, efficient access control and management of network resources is one of the most challenging issues in M2M networks [2]. In addition to congestion control and scalability, an M2M network has to support various traffic categories and quality of service (QoS) requirements. In this context, it is important how to distribute traffic and network access attempts subject to QoS requirements, which is the motivation to our work.

To address the issue, group-based access control schemes have been introduced in the literature [6]-[9]. The authors in [6] introduce a grouping-based radio resource management for 3GPP M2M communications. In [7], IEEE 802.11ah task group proposes Restricted Access Window (RAW) to allow contending channel only for a small number of stations. The group-based contention MAC protocols, a hybrid scheme to combine the advantages of contention-based and contention-free protocols, are presented, Group-based Medium Access Control (GMAC) [8] and Group Leader DCF-TDMA (DCFT) [9]. In GMAC and DCFT, all stations are divided into small numbers of groups and only one station in each group, called a group leader, involves in channel contention. If a group leader wins the competition, a certain time duration is reserved for its associated group and resources are allocated to the group members according to a contention-free protocol.

In this paper, we propose an enhanced group-based contention MAC protocol, Scheduled Group Contention MAC (SGCMAC) protocol, which is more suitable for M2M networks with heterogeneous M2M traffics. Our SGCMAC divides the machine type communications (MTC) stations into small numbers of groups according to their traffic categories, e.g., applications, trigger event types and periods of traffic. The proposed group-based resource allocation scheme is very efficient since the temporal correlation among the group members of a group is typically very high. The Access Point (AP) can schedule the contention time duration of each group based on its traffic characteristics. It can spread the contention groups over a wide range of time interval. Therefore, our proposed SGCMAC is not only efficient in reducing the contention overhead in massive M2M networks but also in being applicable to M2M networks with heterogeneous traffics. We develop an efficient mechanism to detect active stations and idle stations in the reserved time duration for a winner group, which removes the waste time caused by idle stations during the reserved time. The rest of this paper is organized as follows. In Section II and Section III, we describe the system model and the proposed SGCMAC protocol. Section IV presents simulation results to evaluate performance of the proposed protocol. Finally, Section V concludes the paper.

2. SYSTEM MODEL

We consider a single-hop wireless network consisting of one AP and n MTC stations. The network works based on the IEEE 802.11ah MAC and Physical (PHY) layer . Each device follows the Distributed Coordinate Function (DCF) to access the medium and operates in the sub-1GHz frequency band as described in [7], [10]. The MTC stations in a network have two types of operation status, i.e., active status and idle status. When a device is in the active status, it has upload data. We assume a single data buffer in an M2M device. If a active station successfully transmits data, this station goes to idle status. The events are assumed to arrive at each station by a Poisson process with the average arrival rate λ. Fig. 1 shows an example of DCF operation that active stations content the channel access according to DCF rules.

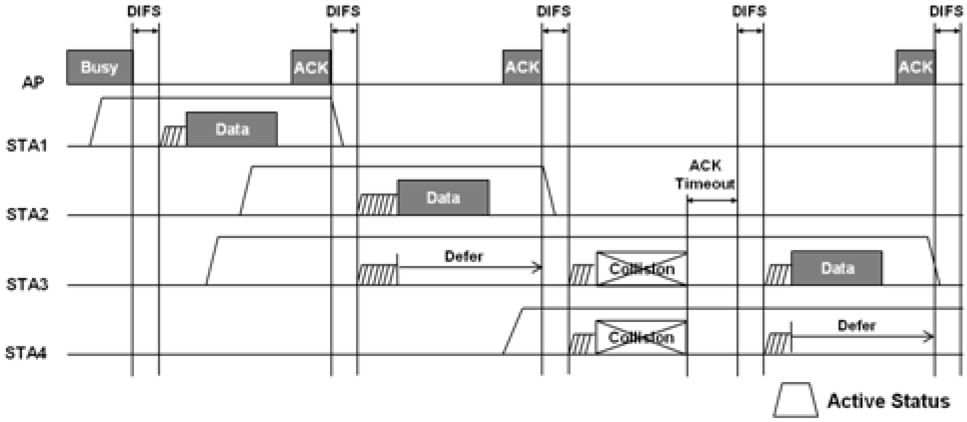


Fig. 1 An example of IEEE 802.11 Distributed Coordinate Function (DCF) operation.

3. PROPOSED SGCMAC PROTOCOL

In this section we describe our proposed SGCMAC protocol. The related grouping rule and the group leader selection algorithm are also introduced. To address the contention problem in a dense M2M network, we divide all MTC devices into several groups according to their traffic categories. The AP assigns the group leader role to a member in each group and schedules time intervals for some of the group leaders to coordinate the channel access. We call this specific time interval Group Access Window (GAW). The scheduled group leaders contend in a GAW by a modified DCF scheme. When a group leader wins the contention, the AP assigns time slots to the group members to allow data transmission by the group period granting procedure. The procedure for the SGCMAC protocol is shown in Fig. 2.

3.1. Group Access Window

The AP allocates a particular time period called GAW in a beacon interval for the group-based channel access. The information of a GAW such as the start time and the duration of the GAW and the scheduled groups for the GAW is announced through the Scheduled Group Indication Map (SGIM) Information Element in a beacon. The group leaders in the SGIM start contention to acquire a Group Granted Period (GGP) at the GAW start time.

The contention procedure among the group leaders is similar to the Request to Send (RTS)/Clear to Send (CTS) procedure based on the DCF scheme. After a random backoff, a group leader sends a Group Access Request (GAR) frame to the AP. If there is no collision, the AP responds by the Group Access Grant (GAG) frame. Since all the active devices in a network as well as the contending group leaders can hear the GAG frame, the winner group is allowed the GGP, defined in the duration field of the GAG frame. The AP checks the active group members in the winner group and assigns time slots to the active stations by Time Division Multiple Access (TDMA) scheme. If the total duration of the assigned time slots is less than GGP in the duration field of the GAG frame, the AP broadcasts the Group Access End (GAE) frame after the data exchange of the last active group member to terminate GGP. After the GGP, the group leaders resume contention again to access the channel until the end of GAW.

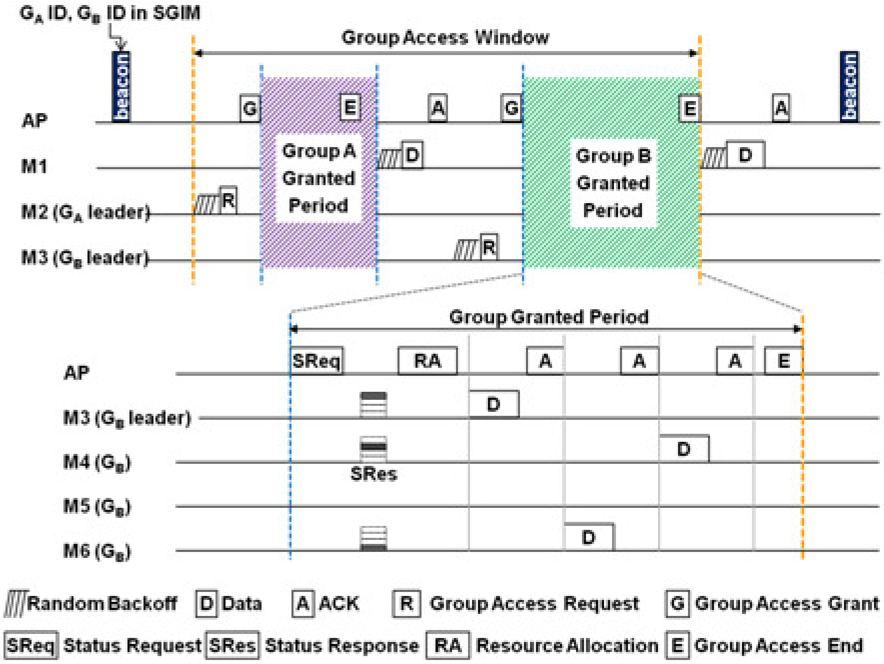
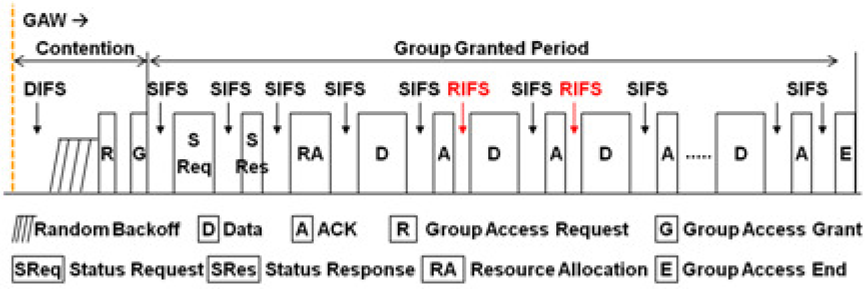
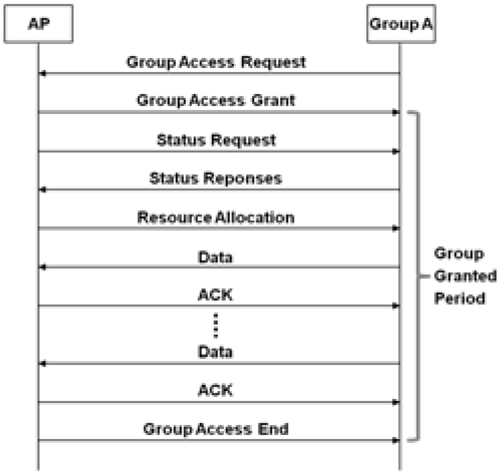


Fig. 2. Group Access Window (GAW) and Scheduled Group Contention MAC (SGCMAC) protocol.

The GAW has advantages. Because an AP can schedule a contention group at a particular GAW, M2M traffic and contention will be spread across beacon periods and QoS requirements of each group are able to be supported. Moreover, M2M devices can coexist with legacy devices since the mechanism of channel reservation for a group allows the protection mechanism.



(a) Transmission of stations in a GGP



(b) Frame exchange sequence during GGP

Fig. 2. An example of Group Granted Period (GGP).

3.2. Transmission of Stations in a GGP

When a group leader wins the contention, the winner group gets the dedicated channel access time called GGP. The transmission time for each station in a winner group is assigned by an AP using a TDMA scheme. However, if a station is not activated in GGP, the time slot for this station will be wasted. In order to prevent this waste, the AP has to check whether each station is in the active status or not before assigning time slots. This status check procedure may be an overhead. To minimize this overhead, we use a signal multiplexing scheme in the frequency domain as in [13]. When the GGP starts, the AP transmits the Status Request (SReq) frame to the group members. The SReq frame includes the sub-carrier assignment map for each group member. Active group members can hear the SReq frame and knows the sub-carrier assigned to itself. After Short Inter-Frame Space (SIFS), the active group members respond by sending Status Response (SRes) by using single tone signals assigned by the AP. Based on this SRes, the AP will be able to assign time slots to the active group members without wasting time. Since we assume that the group members have fixed data size, the AP can schedule the time slots for each of the active members by announcing the order of the active members. Then the AP transmits a Resource Allocation (RA) frame including the schedule information.

In our protocol, a time slot is assigned for a particular station. So, very small time interval larger than the propagation delay is enough to separate data exchange sequences of each active member. In this paper, this inter-time slot of a member is designed and utilized as in Reduced Inter-Frame Space (RIFS) defined in the IEEE 802.11 standard. Fig. 3 shows an example of GGP. The total channel occupation time for one group access consists of the time duration to check the active stations and the scheduled time slots for data transmission. Let T_C and T_S denote the time durations to confirm the active stations and the scheduled time slots, respectively.

$$T_C = T_{SREQ} + T_{SRES} + T_{RA} + SIFS \times 4 \tag{1}$$

where T_{SREQ}, T_{SRES}, and T_{RA} denote the duration for the SReq frame, the SRes, and the RA frame, respectively.

$$T_S = (T_D + SIFS + T_A + RIFS) \times N_{AGM}$$
$$+(SIFS - RIFS + T_E) \quad (2)$$

where T_D, T_A and T_E denote the duration for data, acknowledgment (ACK), and GAE frame, respectively. N_{AGM} denotes the number of active stations in a specific GGP. The total channel occupation time T_{Grant} for one group access is then calculated as follows.

$$T_{Grant} = T_C + T_S. \quad (3)$$

3.3. Grouping and Group Leader Selection Algorithm

We divide MTC stations into groups according to the traffic patterns of stations while GMAC and DCFT make groups based on the region or the coverage of nodes. After grouping, there is only a small number of stations involved in contention at a certain time duration and the hidden terminal issue may be reduced by the RTS/CTS mechanism as defined in [10]. If the temporal correlation between group member stations is low, it will adversely affect the channel efficiency of the group-based contention MAC.

A traffic pattern of each MTC station depends on its application or service type. The traffic of an M2M application or service is classified into Fixed-Scheduling (FS) or Event-Driven (ED) [11]. Furthermore, it can be categorized into periodic, trigger event, or random based on its application [6], [12]. When an MTC station associates with the AP, a station informs its traffic category and application type to the AP. The AP assigns this station to a proper group using the Group ID (GID) and the Group Member ID (GMID). In the proposed protocol, we limit the maximum number of group members to a number less than the number of sub-carriers for efficient group management. Accordingly, it can create multiple groups with the same traffic category.

One of the group members is selected as a group leader by the AP. The group leader consumes more power than the other group members due to the contention for its GGP. To prevent too early burning out of a group leader station, the AP may change the group leader every GGP. The AP chooses the next group leader among the active group members in a round-robin manner and sends its information in the RA frame.

4. SIMULATION RESULTS

In this section, we present the simulation results of the proposed SGCMAC protocol, GMAC [8] and DCFT [9]. We develop our own simulator which have been performed with the PHY and MAC layer characteristics based on the IEEE 802.11ah standard [7] as in Table 1. We assume that the number of stations in each group is fixed to 50. We measure the temporary reservation time for a winner group succeeded in the group contention and the length of group management frames such as the Polling frame in GMAC and the SReq frame and the RA frame in SGCMAC. The data rate of stations in the simulations is 650Kbps defined as the basic data rate in IEEE 802.11ah.

Table 1. Simulation Parameters

Parameter	Value	Description
aSlotTime	52 μs	Backoff slot time
aRIFSTime	20 μs	Reduced Inter-Frame Space
aSIFSTime	160 μs	Short Inter-Frame Space
aDIFSTime	264 μs	DCF Inter-Frame Space
CW_{min}	16	Minimum contention window size
CW_{max}	512	Maximum contention window size
N_{GM}	50	Number of group members
Data Rate	650 Kbps	S1G MCS0 for 2MHz channel
Payload	512 Bytes	Data payload size
MAC Overhead	22 Bytes	S1G short MAC header+FCS
T_{PHY}	280 μs	PHY header
T_{RTS}	560 μs	RTS, DCFT poll, GAR frames
T_{CTS}	480 μs	CTS, CFEND, GAG, GAE frames
T_{RA}	1120 μs	SReq, RA, GMAC polling frames
T_{SRES}	40 μs	SRes frame
T_{ACK}	240 μs	ACK frame

4.1. Performance in Homogeneous Traffic

Fig. 4 and Fig. 5 show the performance under the homogeneous traffic condition. The number of stations associated with the AP is 4000 and they are divided into 80 groups. All stations belong to one traffic category with the same event arrival rate. Our proposed SGCMAC shows gains in throughput and delay compared with the other group-based contention MAC schemes.

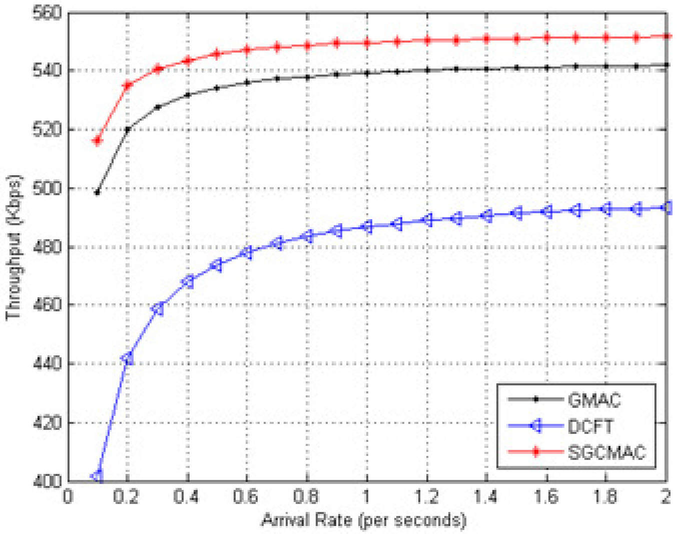


Fig. 3. Throughput for varying arrival rates under the homogeneous traffic condition.

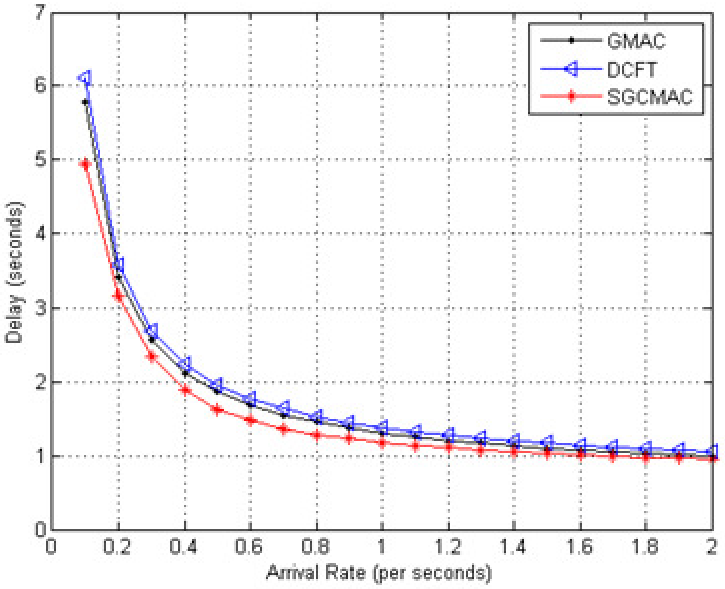


Fig. 5. Average delay for varying arrival rates under the homogeneous traffic condition.

Fig. 4 shows the throughput for varying event arrival rates from 0.1 to 2.0. As the arrival rate increases above 1.0, the network traffic reaches the saturation condition. Then almost all stations are in the active status whenever their groups achieve the reserved time slots for the group members. In this case, the contention overhead and the static overhead of each MAC protocol mainly affects the throughput. If the arrival rate decreases, stations stay in the idle status for a long time, and a large portion of the reserved time interval for a winner group will be wasted. In order to reduce this waste, it is important to efficiently detect idle stations in the reserved time interval for a winner group. Our proposed SGCMAC exhibits better performance than the others in terms of channel efficiency and throughput.

The average delay of stations is presented in Fig. 5. Our proposed SGCMAC has a gain of 850ms compared with the others at the arrival rate of 0.1. The proposed SGCMAC protocol has lower contention overhead and less waste of time in the granted group duration than the others. When the network traffic reaches its saturation condition, the average contention overhead and the static MAC and PHY overhead only affects the delay. In this case, our proposed SGCMAC also shows lower delay than the others.

4.2. Performance in Heterogeneous Traffic

In Fig. 6 and Fig. 7, the throughput and the delay are presented for varying numbers of stations under the heterogeneous traffic condition. There are 20 traffic categories which are classified according to their arrival rates from 0.01 to 0.2. All stations are grouped into 20 traffic categories such that equal number of stations is assigned to each of the categories.

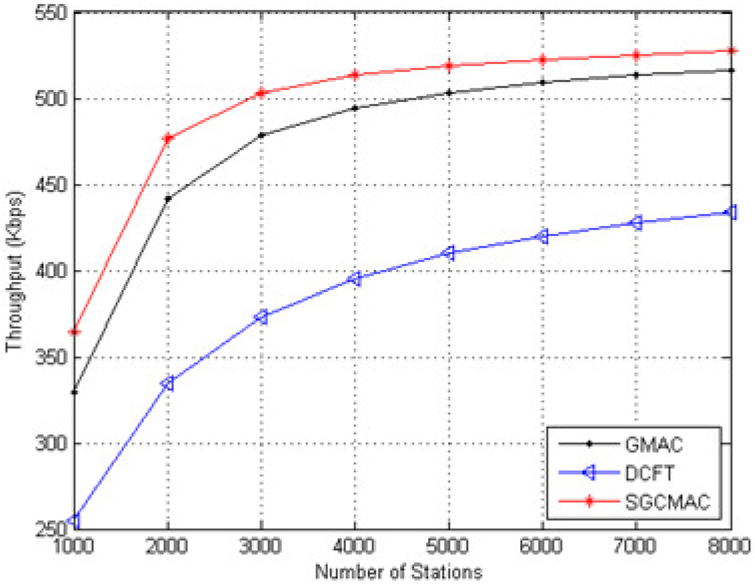


Fig. 4. Throughput for varying numbers of stations under the heterogeneous traffic condition.

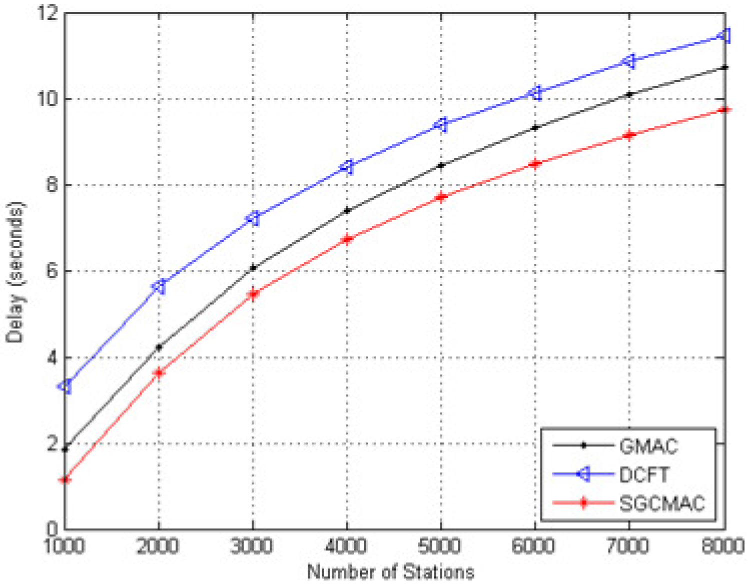


Fig. 5. Average delay for varying number of stations under heterogeneous traffic condition.

The throughput for varying number of stations is shown in Fig. 6. The throughput of our proposed SGCMAC is higher than those of GMAC and DCFT. The average gain of the throughput is 20 Kbps and 113 Kbps compared with those of GMAC and DCFT, respectively. If the number of groups increases due to varying number of stations, the contention overhead in DCF also increases, which then decreases throughput. In the simulation, the arrival rate of each group is lower than 0.2. So most stations stay in the idle status for a long time. However, if the channel access interval of each group gets longer due to increasing number of contention groups, the member stations in the active status will increase in the reserved time duration for a winner group. Then the network throughput increases as the number of stations increases as shown in Fig. 6.

In this traffic condition, the delay is a more important performance metric. The average delay of stations is presented in Fig. 7. The delay of SGCMAC is better than those of the other group-based contention MAC protocols. The delay of SGCMAC is 600 ms to 2.2 seconds lower than those of the others. In our proposed SGCMAC, the contention overhead does not change regardless of the number of stations since the AP is able to schedule a certain number of contention groups in GAW. In the simulation, the number of contention groups in every GAW is controlled to 10. Since the proposed SGCMAC keeps the low contention overhead and minimizes the waste time in the reserved time duration for a winner group, it shows lower delay over the whole range of the number of stations than the others.

5. CONCLUSION

In this paper, we have proposed a group-based contention MAC protocol for a dense M2M network. Our proposed SGCMAC divides groups based on the traffic categories of stations and schedules group contention time slots. It can reduce the contention overhead for channel access and solve a congestion problem in a dense M2M network. In addition, we propose an efficient mechanism to determine active stations in the reserved time duration for a winner group and resources are allocated to reside within the actual transmission time. We can reduce the waste time caused by idle stations in the group access time. Through simulations with the IEEE 802.11ah MAC and PHY parameters, we evaluate the throughput and delay performance of our proposed SGCMAC compared to GMAC and DCFT. The simulation results demonstrate that our proposed SGCMAC has performance gains compared with the other group-based contention MAC protocols. As a future work, we would like to evaluate our SGCMAC protocol performance based on real deployed M2M network traffic model.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (2014R1A5A1011478).

REFERENCES

[1] L. Atzori, A. Iera, and G. Morabito, “The Internet of Things: A Survey,” Computer Networks, vol. 54, no. 15, pp. 2787-2805, Oct. 2010.

[2] A. Rajandekar and B. Sikdar, “A Survey of MAC Layer Issues and Protocols for Machine-to-Machine Communications,” IEEE Internet of Things Journal, vol. 2, no. 2, pp. 175-186, Apr. 2015.

[3] J. Kim, J. Lee, J. Kim, and J. Yun, “M2M Service Platforms: Survey, Issues, and Enabling Technologies,” IEEE Communications Surveys & Tutorials, vol. 16, no. 1, Feb. 2014.

[4] L. Karim, A. Anpalagan, N. Nasser, and J. Almhana, "Sensor-based M2M Agriculture Monitoring Systems for Developing Countries: State and Challenges," Network Protocols and Algorithms, vol. 5, no. 3, pp. 68-86, 2013.

[5] A. Reche, S. Sendra, J. R. Diza, and J. Lloret, "A Smart M2M Deployment to Control the Agriculture Irrigation," Ad-hoc Networks and Wireless of the series Lecture Notes in Computer Science, vol. 8629, pp. 139-151, Feb. 2015.

[6] S.-Y. Lien, K.-C. Chen, and Y. Lin, “Toward Ubiquitous Massive Accesses in 3GPP Machine-to-Machine Communications,” IEEE Communication Magazine, vol. 49, no. 4, pp. 66-74, Apr. 2011.

[7] IEEE P802.11ah Draft Ver.5.0, IEEE Std. 802.11 TGah, Mar. 2015.

[8] Z. Abichar and J. M. Chang, “Group-Based Medium Access Control for IEEE 802.11n Wireless LANs,” IEEE Trans. Mobile Computing, vol. 12, no. 2, pp. 304-317, Feb. 2013.

[9] Y. Yang and S. Roy, “Grouping-Based MAC Protocols for EV Charging Data Transmission in Smart Metering Network,” IEEE J. Selected Areas in Communications, vol. 32, no. 7, pp. 1328-1343, Jul. 2014.

[10] IEEE Std 802.11-2012, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, IEEE, Mar. 2012.

[11] O. Al-Khatib, W. Hardjawana, and B. Vucetic, “Traffic Modeling for Machine-to-Machine (M2M) Last Mile Wireless Access Networks,” in Proc. of IEEE Global Communications Conference (GLOBECOM), pp. 1199-1204, Dec. 2014.

[12] R. Liu, W. Wu, H. Zhu, and D. Yang, “M2M-Oriented QoS Categorization in Cellular Network,” in Proc. of 7th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM), Sep. 2011.

[13] S. Sen, R. Roy Choudhury, and S. Nelakuditi, “No Time to Countdown: Migrating Backoff to the Frequency Domain,” in Proc. of 17th Annual International Conference on Mobile Computing and Networking (MobiCom), pp. 241-252, Sep. 2011.

INTENTIONAL BLANK

HARVESTED ENERGY-ADAPTIVE MAC PROTOCOL FOR ENERGY HARVESTING IOT NETWORKS

Hyeong-Kyu Lee, MinGyu Lee and Tae-Jin Lee

College of Information and Communication Engineering
Sungkyunkwan University, Suwon 16419, South Korea
lhkyu@skku.edu, lmk0218@skku.edu, tjlee@skku.edu

ABSTRACT

In energy harvesting IoT networks, an energy queue state of an IoT device will change dynamically and the number of IoT devices that transmit data to the IoT AP will vary in a frame. So we need a MAC protocol to adjust the frame length taking the amount of energy of IoT devices into consideration. Since the existing Framed slotted ALOHA (F-ALOHA) Medium Access Control (MAC) protocol utilizes the fixed frame size, the resource efficiency can be reduced. In this paper, we propose a Harvested Energy-adaptive Medium Access Control (HE-MAC) protocol where an IoT Access Point (AP) allocates slots in accordance with the number of IoT devices that try to transmit data in a frame. The proposed HE-MAC protocol improves the resource efficiency of the F-ALOHA MAC protocol. We show that the resource efficiency of the HE-MAC protocol is superior to those of the F-ALOHA MAC protocol through simulations.

KEYWORDS

Internet of Things, Energy Harvesting, Framed Slotted ALOHA, Medium Access Control.

1. INTRODUCTION

In Internet of Things (IoT) networks, IoT nodes have the limited battery lifetime and the batteries need to be replaced. To solve the problem, the energy harvesting technology draws much attention, since electric energy is created from the energy sources that exist in surroundings such as solar, heat, pressure, and electromagnetic wave energy [1]. The energy harvesting technology enables the sustainable use of the battery of IoT nodes in IoT networks [2]. However, in an energy harvesting IoT network, the energy states of IoT nodes change dynamically. A Computational Radio Frequency IDentification (CRFID) is an example of utilizing the energy harvesting technology. CRFID with sensing and computation functions is a new emerging technology which makes IoT devices to operate without battery [3].

According to the standard in ISO-18000-6C, the CRFID identification protocol utilizes the Framed Slotted ALOHA (F-ALOHA) Medium Access Control (MAC) protocol. CRFID tags randomly select one slot among the slots in a frame to transmit data packets [4]. The existing F-ALOHA MAC protocol utilizes the fixed frame length [5]. An IoT node selects one slot among the slots in a frame and transmits data to the IoT access point (AP) at the selected time slot. The number of the slots in a frame affects the throughput performance of the F-ALOHA MAC

David C. Wyld et al. (Eds) : NETCOM, NCS, WiMoNe, CSEIT, SPM - 2015
pp. 51–58, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51605

protocol. Since the amount of energy stored in a battery may change in a frame, some IoT nodes with insufficient energy to transmit data may exist. If the number of slots is more than the number of the IoT nodes in a frame, the resource efficiency is reduced due to the unused slots. If the number of slots is less than the number of the IoT nodes in a frame, the resource efficiency is also reduced due to the collided slots. Thus, a new MAC protocol to adjust the number of slots according to the energy state in a frame is required.

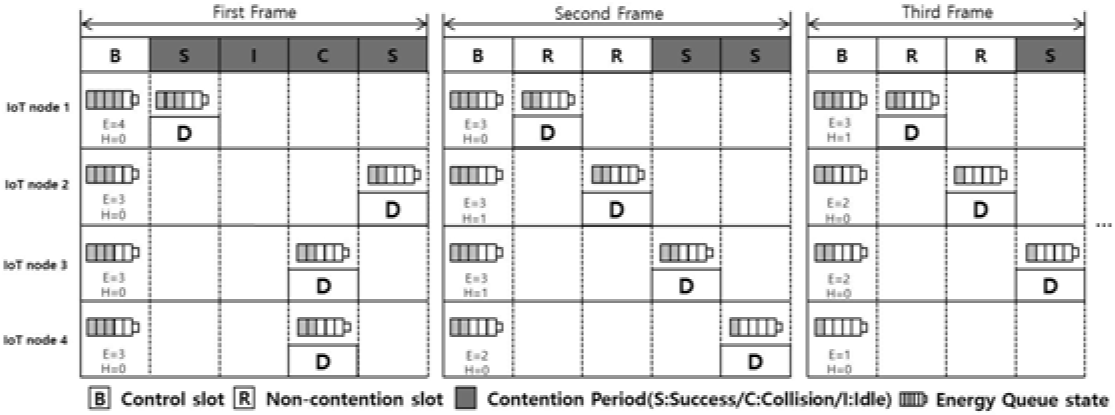


Figure 1. An example of the proposed the HE-MAC protocol with N (=4) and E_{min}.

In this paper, we propose a harvested energy-adaptive MAC (HE-MAC) protocol to adjust the frame length taking the energy states of IoT nodes into account. The proposed HE-MAC protocol alleviates the reduction of the resource efficiency of the F-ALOHA MAC protocol. The detailed explanation of the proposed HE-MAC protocol is presented in Section 2. In Section 3, we compare the throughput performance of the proposed HE-MAC with that of the existing F-ALOHA MAC protocol. Finally, we conclude in Section 4.

2. RELATED WORK

In [10], the S-MAC protocol is proposed to reduce energy consumption. The protocol utilizes periodic listen and sleep scheme, so each node does not need to listen at all the times. Also, the nodes exchange the Request To Send (RTS) and Clear To Send (CTS) before actual data transmission to avoid collision and overhearing. However, the protocol does not consider energy harvesting technology. The authors of [11] propose a duty-cycle optimization scheme by finding the minimum length of the non-active period satisfying Energy Neutral Operation (ENO) of energy harvesting nodes. Since the scheme utilizes F-ALOHA MAC protocol, it does not adapt the frame length according to the number of the nodes transmitting data packets during the active period.

The design of a MAC protocol such as TDMA, and F-ALOHA in an energy harvesting circumstance is presented in [12]. Also, the trade-off relationship between the delivery efficiency and the time efficiency is investigated. But the work does not consider variability in the number of nodes transmitting data packets due to fluctuation of the energy level of the individual nodes. In [13], the authors present a fair polling scheme for energy harvesting wireless sensor networks. The scheme improves the fairness by considering priority using the harvesting rate of the nodes. The scheme focuses on only fairness, but improvement of the throughput is not considered.

3. PROPOSED HE-MAC PROTOCOL

In this section, we present the proposed HE-MAC protocol. We consider an energy harvesting network which consists of an IoT AP and N IoT nodes. The IoT nodes transmit data with the energy queue states of IoT nodes. We assume that the energy stored in an energy queue is composed of energy blocks. The maximum number of energy blocks that can be stored in a battery is E_{max}. The energy queue state can be reduced with the unit of an energy block when an IoT node transmits a data packet. An IoT node harvests an energy block with the energy harvesting probability P_h in a slot. Let E_{min} denote the energy threshold. The energy queue state of IoT nodes has to exceed E_{min} in a frame to transmit data. For example, if E_{min}=1, the IoT node that the energy queue state is one does not transmit data.

A frame consists of the control slot, the non-contention slot, and the contention slot. An IoT AP broadcasts a control packet to the IoT nodes at the control slot. The control packet includes the number of non-contention slots, the number of contention slots, and the allocation information for the non-contention slots in the ith frame. A non-contention slot is allocated to only one IoT node that transmits data successfully in the (i-1)th frame and exceeds E_{min}. The contention slot is for the IoT nodes that collide with one another in the (i-1)th frame or for those not allocated non-contention slots. At the end of a frame, the contention slots can be classified as success slots, collision slots, and idle slots. The IoT AP receiving data packets with energy state information of the IoT nodes decides to allocate non-contention slots in the ith frame according to the information of the energy state of the IoT nodes. After deciding which IoT nodes transmit at non-contention slots, the IoT AP estimates the number of the IoT nodes trying to transmit at contention slots according to the information of the energy state of IoT nodes and the collision slots in the (i-1)th frame.

IoT nodes may receive control packets at the control slot. If IoT nodes succeed in transmission and the energy level exceeds E_{min} in the (i-1)th frame, they are allocated non-contention slots in the ith frame, they transmit their data packets at their allocated non-contention slots. If collision does not happen in the non-contention slot, the IoT nodes transmitting in the non-contention slots will be continuously allocated in the non-contention slots as long as the energy states of the IoT nodes exceed E_{min}. If the energy level of an IoT node does not exceed E_{min}, the IoT node harvests energy without transmitting of data packets. If IoT nodes collide in the (i-1)th frame and the energy levels exceed E_{min} at the start of the ith frame, they each randomly select one of the contention slots and try to transmit at the selected the contention slots in the ith frame.

Fig. 1 shows an example of the proposed HE-MAC protocol with N nodes and E_{min}(=1). Let E and H denote the energy queue state of an IoT node and the amount of harvested energy in the previous frame. In the first frame, all slots except the control slot are contention slots. The IoT terminals randomly select one of the four contention slots and transmit data packets to the IoT AP at the selected contention slot. The IoT node 1 and the IoT node 2 successfully transmit data packets in the first frame and consume one energy block. The IoT node 3 and the IoT node 4 collide with each other at the third slot and use energy blocks. After transmission, the amount of the remaining energy blocks of the IoT node 1 and the IoT node 2 is 3 and 2. The energy state of the IoT node 3 and the IoT 4 is 2 after collision at the third slot. Since the energy levels of the IoT nodes exceed E_{min}(=1), they can transmit data in the second frame.

In the second frame, the IoT nodes update energy states by adding the harvested energy during the first frame. The energy states of the IoT node 1, node 2, node 3, and node 4 are 3, 3, 3, and 2, respectively at the beginning of the second frame. The IoT node 1 and the IoT node 2 transmit data packets successfully in the allocated non-contention slots in the second frame since they succeed in transmission and the energy levels exceed E_{min} in the first frame. The IoT node 3 and the IoT node 4 succeed in transmission of the data packets by selecting different slots with each other. Since the IoT node 4 does not have enough energy after data transmission in the second frame, the energy queue state information indicates that it does not transmit its data packet in the next frame.

In the third frame, the IoT nodes update the energy states as in the second frame. If any IoT node does not transmit its data packet at the contention slot in the next frame, one of the IoT nodes that succeed in transmission in the previous frame transmits its data packet at the contention slot. The IoT node 1 and the IoT node 2 succeed in data transmission in the non-contention slots and the IoT node 3 transmits data successfully at the contention slot. However, the IoT node 4 does not transmit data, and harvests energy during the third frame. The IoT node 2 and the IoT node 3 do not transmit data in the next frame since the energy states do not exceed E_{min}. If the energy state of the IoT node 4 exceeds E_{min}, the IoT node 4 will transmit its data packet at a contention slot in the next frame.

Table 1. Parameters of Simulation.

Parameters	Value
Energy harvesting probability (P_h)	0.01 – 1
Number of IoT nodes (N)	10 – 150
Energy threshold (E_{min})	1
Size of the energy queue (E_{max})	5
Initial state of the energy queue (E_{init})	5

The proposed HE-MAC protocol adjusts the frame size according to the number of IoT nodes trying to transmit, while the F-ALOHA MAC protocol utilizes the fixed frame size. If the number of IoT nodes decreases, there may be unused slots in the F-ALOHA MAC protocol. However, the proposed HE-MAC protocol allocates slots according to the amount of stored energy states of the IoT nodes, and the number of wasted slots can be reduced.

4. SIMULATION

In this section, we present the resource efficiency of the proposed HE-MAC protocol. The F-ALOHA MAC protocol utilizes the fixed frame size in which the number of contention slots is the same as N. The parameters are summarized in Table 1. E_{init} denotes the initial state of the energy queue. We assume that the IoT AP knows all the energy queue states of the IoT nodes, and the number of IoT nodes does not change.

Fig. 2 presents throughput for varying harvesting probability with different N. The throughput increases as the harvesting probability increases. Since the harvested energy increases as the harvesting probability increases, the number of IoT nodes which have enough energy to transmit data packets increases. So, the IoT AP can allocate non-contention slots to the IoT nodes consistently. Since the success slots increases as the number of non-contention slots increases, the throughput also increases. However, the number of non-contention slots does not exceed N, thus the throughput is eventually saturated. When N=150, the throughput performances are similar for almost any harvesting probabilities since the maximum frame size when N=150 is larger than those of others, and the harvested energy of IoT nodes when N=150 is larger than those of others.

Therefore, the throughput performance when N=150 can be easily saturated even if the harvesting probability is low.

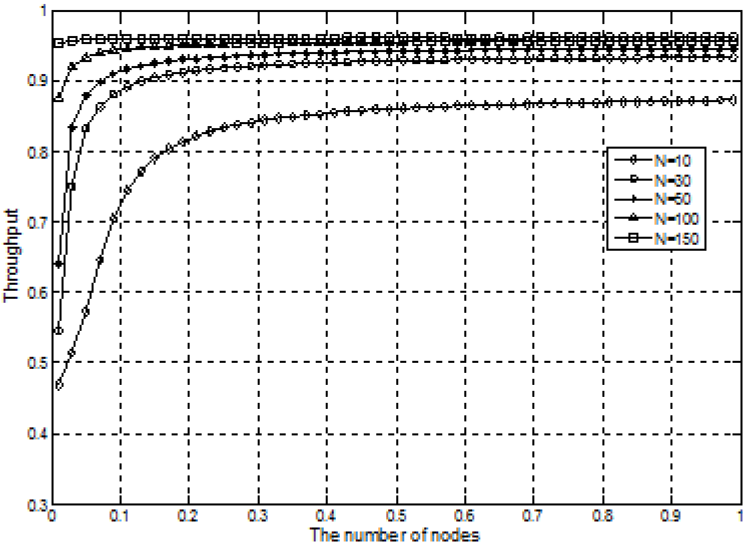


Figure 2. Throughput for varying harvesting probabilities with different N.

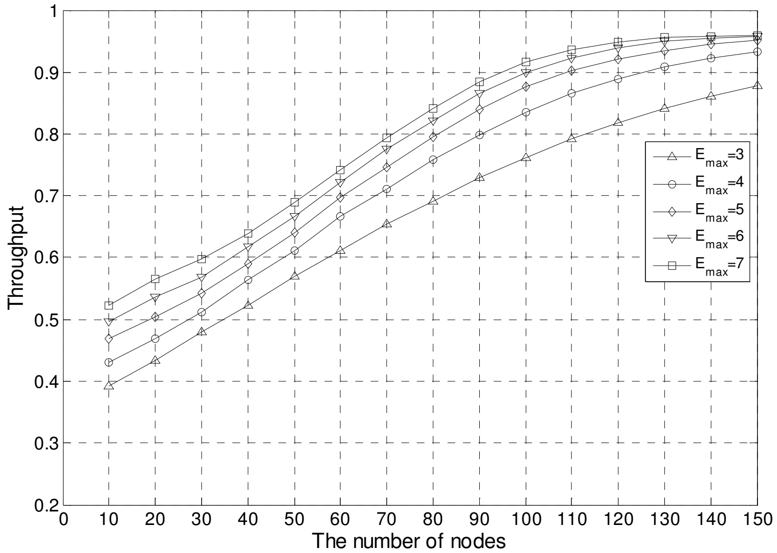


Figure 3. Throughput for varying number of IoT nodes with different E_{max} and P_h(=0.01).

Fig. 3 shows the throughput for varying number of nodes with different E_{max} and P_h. The battery of an IoT node can store more energy as the size of the energy queue increases. If an IoT node can store more energy, it can transmit data packets consistently at the non-contention slot until the energy state of the IoT node becomes lower than E_{min}. Since the number of IoT nodes that successfully transmits data packets consistently at the non-contention slot increases, the throughput increases according to the size of energy queue. In the case that the energy harvesting probability is high, the energy queue is charged quickly then the size of the energy queue does not much affect the throughput performance.

In Fig. 4, we compare the throughput performances of the HE-MAC protocol to those of the F-ALOHA MAC protocol by changing *N*. The throughputs of the two protocols increase as *N* increases. If the energy harvesting probability increases, the number of IoT nodes that harvested energy increases in a frame. Since the number of IoT nodes transmitting data packets at the non-contention slot increases, the throughputs increase. When *N*=10, as the energy harvesting probability increases from 0.01 to 0.09, the throughput increases from 0.47 to 0.84. Since the F-ALOHA MAC protocol utilizes fixed frame size, it does not adjust the number of slots according to the number of IoT nodes transmitting data packets. So, the throughputs of F-ALOHA MAC protocol are less than those of the HE-MAC protocol.

5. CONCLUSION

In this paper, we proposed the HE-MAC protocol that adjusts the frame length in accordance with the number of IoT nodes trying to transmit in a frame. Since the existing F-ALOHA MAC protocol utilizes the fixed frame size, it does not adjust the number of IoT nodes that try to transmit data packets according to the energy states of IoT nodes. The HE-MAC protocol can allocate the radio resources taking into account the energy states of IoT nodes. Thus, our proposed MAC protocol alleviates the reduction of the resource efficiency of the F-ALOHA MAC protocol. The simulation results show that our proposed HE-MAC protocol increases throughput performance than the F-ALOHA MAC protocol as the probability of energy harvesting and the number of IoT nodes increase.

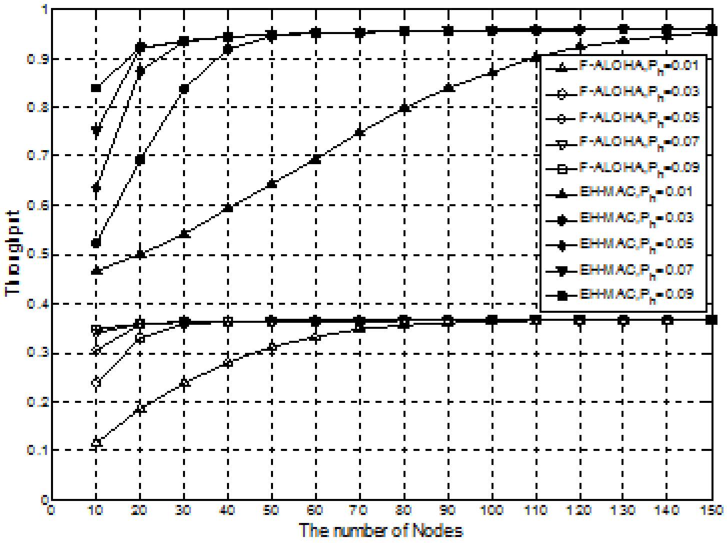


Figure 4. Throughput for varying number of IoT nodes.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP)(2014R1A5A1011478, 2015R1A2A2A01004067).

REFERENCES

[1] S. Sudevalayam and P. Kulkarni, “Energy Harvesting Sensor Nodes: Survey and Implications Energy Harvesting Sensor Nodes: Survey and Implications,” IEEE Communications Surveys & Tutorials, vol. 13, no. 3, pp. 443-461, Third Quarter, 2011.

[2] H. Li, J. Xu, R. Zhang, and S. Cui, “A General Utility Optimization Framework for Energy-Harvesting based Wireless Communications,” IEEE Communications Magazine, vol. 53, no. 4, pp. 79-85, Apr. 2015.

[3] W. Yang, D. Wu, M. J. Hussain, and L. Lu “Wireless Firmware Execution Control in Computational RFID Systems,” in Proc. of IEEE RFID, pp. 129-136, Apr. 2015.

[4] A. Wickramasinghe and D. C. Ranasinghe, “Ambulatory Monitoring Using Passive Computational RFID Sensors,” IEEE Sensors Journal, vol. 15, no. 10, pp. 5859-5869, Oct. 2015.

[5] J. E. Wieselthier, A. Ephremides, and L. A. Michaels, “An Exact Analysis and Performance Evaluation of Framed ALOHA with Capture,” IEEE Transactions on Communications, vol. 37, no. 2, pp. 125-137, Feb. 1989.

[6] J. Vales-Alonso, F. J. Parrado-Garcia, and J. J. Alcaraz, “Analytical Computation of the Mean Number of Tag Identifications during a Time Interval in FSA,” IEEE Communications Letters, vol. 18, no. 11, pp. 1923-1926, Nov. 2014.

[7] O. Omni, A. Wong, A. J. Burdett, and C. Toumazou, “Energy Efficient Medium Access Protocol for Wireless Medical Body Area Sensor Networks,” IEEE Transactions on Biomedical Circuits and Systems, vol. 2, no. 4, pp. 251-259, Dec. 2008.

[8] M. Gorlatova, P. Kinget, I. Kymissis, D. Rubenstein, X. Wang, and G. Zussaman, “Challenge: Ultra-low-power Energy-harvesting Active Neworked Tags(EnHANTs),” in Proc. of MobiCom, pp. 253-260, Sep. 2009.

[9] J. Liu, H. Dai, and W. Chen, “On Throughput Maximization of Time Division Multiple Access with Energy Harvesting Users,” IEEE Transactions on Vehicular Technology, to be published.

[10] W. Ye, J. Heidemann, D. Estrin, “An Energy-Efficient MAC Protocol for Wireless Sensor Networks ,”in Proc. of INFOCOM, pp. 1567-1576, Jun. 2002.

[11] O. Briante, A. M. Mandalari, A. Molinaro, G. Ruggeri, J. Alonso-Zarate, and F. Vazquez-Gallego, “Duty-Cycle Optimization for Machine-to-Machine Area Networks based on Frame Slotted-ALOHA with Energy Harvesting Capabilities,” in Proc. of EW, pp. 1-6, May. 2014.

[12] F. Iannello, O. Simeone, and U. Spagnolini, “Medium Access Control Protocols for Wireless Sensor Networks with Energy Harvesting,” IEEE Transactions on Communications, vol. 60, no. 5, pp. 1381-1389, May. 2012.

[13] M. Kunikawa, H. Yomo, K. Abe, T. Ito, “A Fair Polling Scheme for Energy Harvesting Wireless Sensor Networks,”in Proc. of VTC spring, pp. 1-5, May. 2015.

[14] Y. He, X. Cheng, W. Peng, and G. L. Stuber, “A Survey of Energy Harvesting Communications: Models and Offline Optimal Policies,” IEEE Communications Magazine, vol. 53, no. 6, pp. 79-85, Jun. 2015.

AUTHORS

Hyeong-Kyu Lee received B.S. degree in electronics communication engineering from Hanyang University, Korea in 2013. He is currently working toward his M.S. degree in the College of Information and Communication Engineering at Sungkyunkwan University, Suwon, Korea since March 2015. His research interests include energy harvesting, M2M communications, IoT and Medium Access Control (MAC).

Mingyu Lee received the B.S. degree in electronics engineering from Kwangwoon University, Korea in 2009 and the M.S. degree in mobile systems engineering from Sungkyunkwan University, Suwon, Korea in 2012. He is currently pursuing his Ph.D. degree in IT convergence at Sungkyunkwan University since March 2012. His research interests include cognitive radio networks, wireless LANs, ad hoc networks, vehicular networks, energy harvesting, M2M communications and IoT.

Tae-Jin Lee received his B.S. and M.S. in electronics engineering from Yonsei University, Korea in 1989 and 1991, respectively, and the M.S.E. degree in electrical engineering and computer science from University of Michigan, Ann Arbor, in 1995. He received the Ph.D. degree in electrical and computer engineering from the University of Texas, Austin, in May 1999. In 1999, he joined Corporate R & D Center, Samsung Electronics where he was a senior engineer. Since 2001, he has been Professor in the College of Information and Communication Engineering at Sungkyunkwan University, Korea. He was a visiting professor in Pennsylvania State University from 2007 to 2008. His research interests include performance evaluation, resource allocation, Medium Access Control (MAC), and design of communication networks and systems, wireless LANs/PANs, vehicular networks, energy-harvesting networks, IoT, ad hoc/sensor/RFID networks, and next generation wireless communication systems. He has been a voting member of IEEE 802.11 WLAN Working Group, and is a member of IEEE and IEICE.

TABLE-BASED IDENTIFICATION PROTOCOL OF COMPUTATIONAL RFID TAGS

Yunmin Kim, Ji Hyoung Ahn, and Tae-Jin Lee

College of Information and Communication Engineering
Sungkyunkwan University, Suwon 16419, South Korea
kym0413@skku.edu, beramode@skku.edu, tjlee@skku.edu

ABSTRACT

Computation RFID (CRFID) expands the limit of traditional RFID by granting computational capability to RFID tags. RFID tags are powered by Radio-Frequency (RF) energy harvesting. However, CRFID tags need extra energy and processing time than traditional RFID tags to sense the environment and generate sensing data. Therefore, Dynamic Framed Slotted ALOHA (DFSA) protocol for traditional RFID is no longer a best solution for CRFID systems. In this paper, we propose a table-based CRFID tag identification protocol considering CRFID operation. An RFID reader sends the message with the frame index. Using the frame index, CRFID tags calculate the energy requirement and processing time information to the reader along with data transmission. The RFID reader records the received information of tags into a table. After table recording is completed, the optimal frame size and proper time interval is provided based on the table. The proposed CRFID tag identification protocol is shown to enhance the identification rate and the delay via simulations.

KEYWORDS

Computational RFID, Energy harvesting, Sensing energy, Processing delay, Tag identification

1. INTRODUCTION

Internet of Things (IoT) is one of the emerging technologies that connects the physical world of things to the Internet [1]. Various applications such as metering, surveillance system, and factory maintenance systems are expected to be realized by IoT networks. There are many challenges to implement IoT networks [2]. IoT communications aim to support a massive number of deployed devices. To collect information of a large number of devices, efficient Medium Access Control (MAC) protocol is required. Moreover, there is an energy consumption issue due to energy-critical small IoT devices. Devices should be energy efficient to maximize the network lifetime to provide better network maintenance and reliability.

Radio Frequency IDentification (RFID) is a contact-less identification technology. An RFID network consists of a RFID reader and multiple tags [3]. RFID tags classified into passive tags and active tags. Passive RFID tags transmit data by backscattering the signal of the reader. Since the transmission relies on the signal from the reader, passive tags are battery-less and semi-

David C. Wyld et al. (Eds) : NETCOM, NCS, WiMoNe, CSEIT, SPM - 2015
pp. 59–67, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51606

permanent. Traditional RFID tags can only send the predefined information such as unique IDs of tags. Thus, the role of RFID tags may be limited in IoT networks which require smarter and more complex jobs of devices.

Computational RFID (CRFID) is an enhanced RFID that can overcome the limitation of the traditional RFID. CRFID incorporating the sensing capability into RFID expands the functionality of RFID [4]. Fig. 1 indicates the simple concept of the CRFID tag. In a sensor part, a surrounding environment is measured and data is produced by local computing process. Then, the generated data is delivered to the RFID part and tags transmit it by backscattering the signal of the reader. With this computational capability, potential applicability of CRFID can be widened with sufficiently low energy consumption [5]. In this sense, CRFID is treated as a viable solution of enabling IoT networks [6].

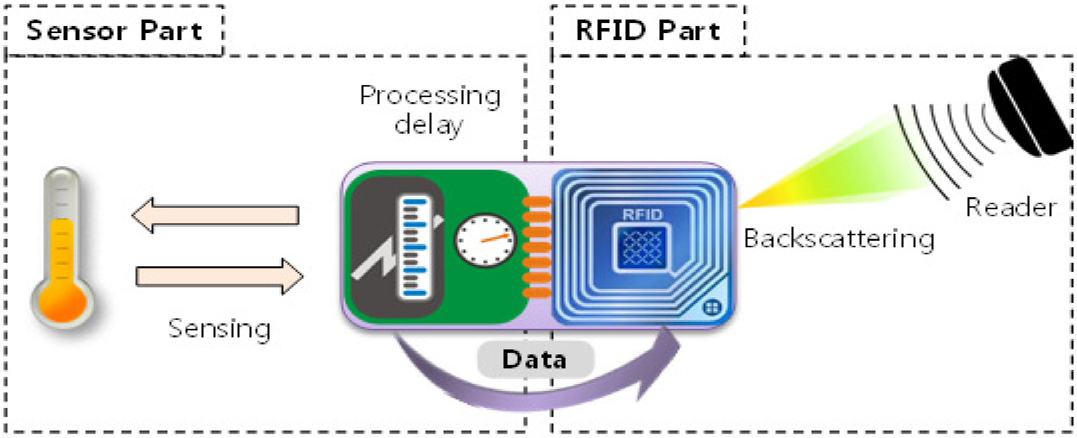


Figure 1. Structure of CRFID tag with sensing and processing capability.

There have been researches for CRFID systems. Yang et al. [7] proposed a switching mechanism of the firmware of a CRFID tag by the command from the reader. The inefficiency of firmware switching of a CRFID tag using wired interface is stated. To resolve the problem, the reader command for firmware switching is designed to be compatible with the EPC standard. The energy overhead and switching delay are simulated. In [8], an efficient way to transmit bulk data of a CRFID tag is studied. Authors stated the low performance of data stream of a CRFID tag caused by Cyclic Redundancy Check (CRC) calculation. Then, they proposed an efficient data transfer scheme for the CRFID tag by precomputing and exploiting the intermediate computations for CRC. Since the latency of CRFID tags is reduced with the proposed scheme, data transfer efficiency is improved. Wickramasinghe et al. [9] studied the ambulatory monitoring systems using CRFID sensors. Receiving data is segmented based on the natural activity boundaries and an algorithm to track the body movement transition is presented. In the experiment, CRFID in ambulatory monitoring shows high performance.

Existing studies do not tackle the most important difference of CRFID tags from traditional RFID tags. With the sensing operation, CRFID tags need extra energy and processing time before communicating with the reader. Dynamic Framed Slotted ALOHA (DFSA), which is used in the conventional RFID identification [10], may not be compatible with CRFID systems. Estimation of the number of tags in DFSA will no longer match with the actual number of active CRFID tags due to new characteristics of CRFID. Thus, the optimal frame size needs to be redesigned in DFSA.

In this paper, we propose a CRFID tag identification protocol considering the sensing capability. Initially, a CRFID reader collects the energy and processing time requirement of individual tags along with the identification process. When the reader collects all of the energy requirements and processing time of tags, reader estimates the exact number of active tags at every frames. Then, the optimal frame size is decided using the table to maximize the identification rate.

2. Proposed CRFID Tag Identification Method

2.1. DFSA Algorithm in CRFID Systems

One of the key differences of the CRFID tag is that it senses and generates data to report. With this sensing and processing capability, CRFID tag requires extra energy and processing time for local computation. Furthermore, energy and processing time requirement among CRFID tags may vary with one another. Depending on the type of sensing job, some CRFID tags may need more energy than others. Received energy from the reader will also vary depend on the distance between the reader and the tag. Another difference of CRFID tag is continuous data transfer. While conventional RFID tags stop operating after the successful identification, CRFID tags continuously respond to the reader when they have data to transmit.

For the conventional RFID, DFSA algorithm is used for tag identification. In DFSA, a reader assigns a frame size, the number of time slots in a frame. Then, tags randomly select their own time slots and transmit their IDs to the reader. In [11], it is shown that the optimal frame size is the same as the number of identifying tags. The number of identifying tags is estimated using the fraction of idle slots (zero estimation) [12]. Since the DFSA algorithm does not reflect the novel characteristics of the CRFID tag, it cannot simply adapt to CRFID systems.

Fig. 2 shows the CRFID tag identification process using DFSA. We assume that the reader provides extra energy for sensing operation at the beginning of a frame. Tags 1, 2, and 5 require charging twice for sensing, while tags 3 and 4 need charging only once. The reader assigns 5 slots in frame 1 since the number of tags is 5. However, only tags 3 and 4 respond to the reader. Then, the reader assigns 3 slots to frame 2 by zero estimation. In frame 2, all the tags respond to the reader and contend to transmit data. Frame 2 cannot accommodate all the active tags since the number of active CRFID tags is larger than the frame size. Furthermore, because of the processing delay, tags actually contend in 2 slots which causes further degradation of performance.

2.2. System Model

Now, we describe the proposed table-based CRFID tag identification protocol. A CRFID network consists of a reader and multiple tags. Since the CRFID network is likely to be installed for a purpose, we assume that the reader knows the number of tags. The tags are equipped with small capacitors that can store relatively small amount of energy. At the beginning of every frame, the reader sends a packet intended for charging tags to support sensing operation. The number of required charging for sensing is randomly distributed in [1, $e_{\max}$]. The main idea of the proposed MAC protocol is to collect the energy and processing time requirements of tags during the tag identification. After that, the reader can schedule the optimal frame size to identify tags based on the table. Also, the time duration for sensing operation is provided based on the table. The proposed scheme can be divided into table recording phase and normal phase.

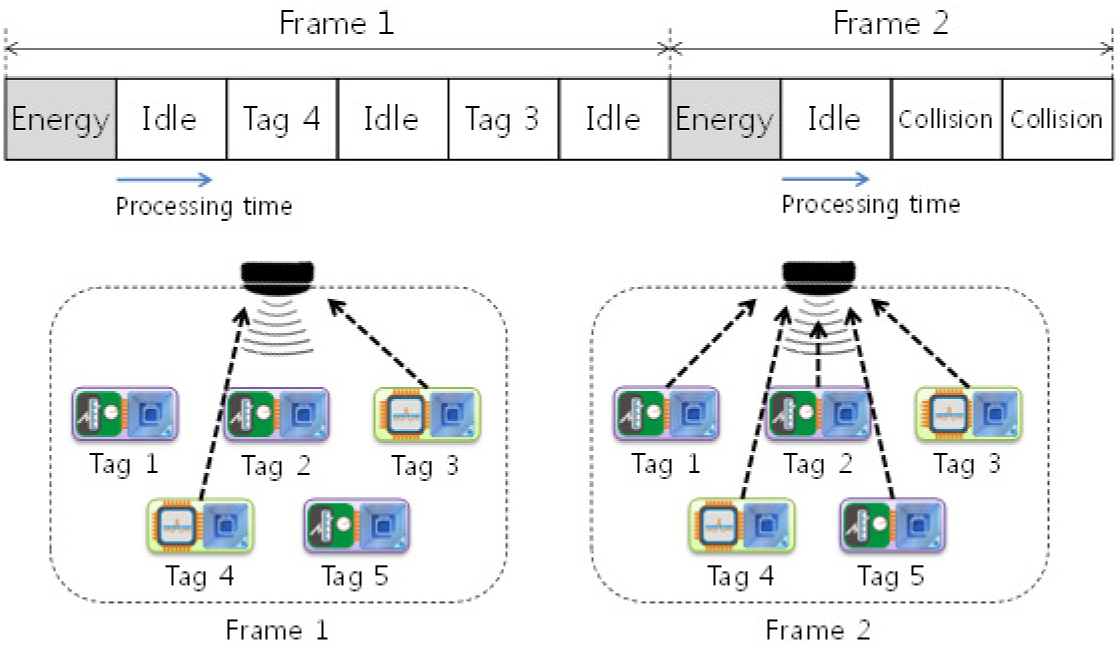
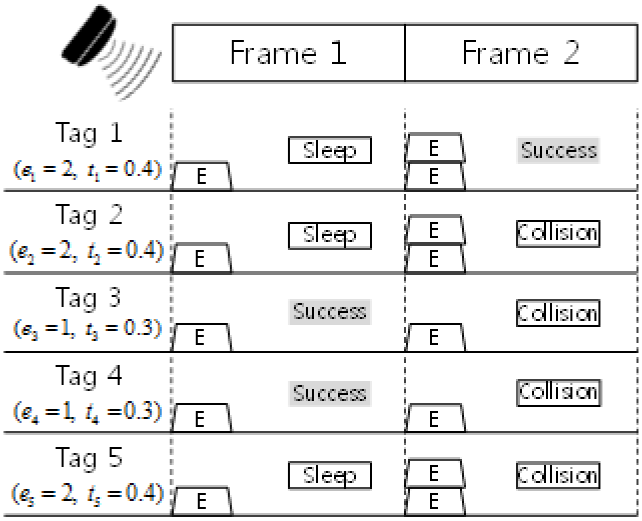


Figure 2. CRFID tag identification using DFSA algorithm.

2.3. Table Recording Phase

When the reader starts to receive data, it operates as the table recording phase. In the table recording phase, the reader collects the energy requirement (e) and processing delay (t) of individual tag along with receiving data. Let e_i and t_i be the energy and processing time requirements of tag i. Since energy and processing the information of tags is yet to be collected, the reader decides frame as the number of entire tags. The reader sends the querying message with a frame index. When the tag i charges sufficient enough amount of energy for sensing, it performs sensing and records the frame index as its e_i. Then, the tag transmits e_i and t_i along with the generated data to the reader. The reader records the received e_i and t_i information in the table.

Fig. 3 shows an example of collecting process for energy and processing information in the table recording phase. In frame 1, tags 3 and 4 respond to the reader and consume the charged energy. Since other tags need more energy to perform sensing, they sleep in frame 1. The reader records e_3, t_3, e_4, and t_4 of tags 3 and 4 in the table. In frame 2, all tags become active and respond to the reader. Since only tag 1 succeeds in frame 2, the reader obtains e_1 and t_1. Collided tags 2 and 5 may send their e and t information in the subsequent frames. The table recording phase ends when all the information of e and t are completely collected.



Table

Tag ID	1	2	3	4	5
e_i	-	-	1	1	-
t_i	-	-	0.3	0.3	-

Frame 1

Table

Tag ID	1	2	3	4	5
e_i	2	-	1	1	-
t_i	0.4	-	0.3	0.3	-

Frame 2

Figure 3. An example of the table recording process.

2.4. Normal Phase

Once the table is filled with all the information of e and t, the reader moves to the normal phase. In the normal phase, the reader can decide which tags try to transmit in the current frame. If e_i of tag i is a divisor of the current frame index, the tag will become active and transmit data. To maximize the identification rate, the reader should set the frame size to be the number of the active tags. By searching for the table, the reader evaluates the i-th frame size as

$$L_i = \sum_{j=1}^{N_{tag}} 1_{\{0\}}(i \bmod e_j), \tag{1}$$

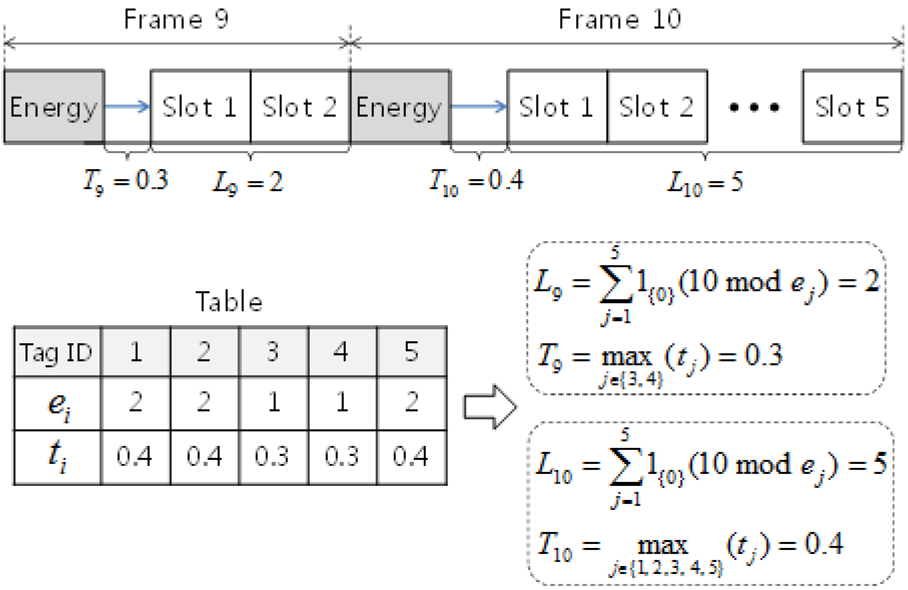
where $1_{\{0\}}$ and N_{tag} refer to the indicator function and the number of tags.

The reader can assign the time interval to mitigate the processing delay of tags. To ensure the processing delay of tags to be tolerable, the reader chooses the maximum t among those of the active tags in the frame. Then, the processing interval for the i-th frame is decided as

$$T_i = \max_{j\in\{l | 1_{\{0\}}(i \bmod e_j)=1\}} (t_j). \tag{2}$$

Fig. 4 shows the 9th and 10th frame structures in the normal phase. In the normal phase, the processing interval of length T_i is assigned after the energy transfer from the reader. So, all tags can sense and produce data before the slot contention. We assume that the information for energy and processing time of all tags are collected before frame 9. In frame 9, the reader finds out that tags 3 and 4 will be active tags by the table with the frame index 9. Then, the frame size is determined as $L_9 = 2$. The processing time interval for frame 9 is $T_9 = 0.3$ since $t_3 = t_4 = 0.3$ in the table. In frame 10, all tags will be active since 1 and 2 are divisors of the frame index 10.

The frame size L_{10} is decided as 5. Then, the processing time interval of frame 10 becomes $T_{10} = 0.4$.



Tag ID	1	2	3	4	5
e_i	2	2	1	1	2
t_i	0.4	0.4	0.3	0.3	0.4

$$L_9 = \sum_{j=1}^{5} 1_{\{0\}}(10 \bmod e_j) = 2$$

$$T_9 = \max_{j \in \{3,4\}} (t_j) = 0.3$$

$$L_{10} = \sum_{j=1}^{5} 1_{\{0\}}(10 \bmod e_j) = 5$$

$$T_{10} = \max_{j \in \{1,2,3,4,5\}} (t_j) = 0.4$$

Figure 4. The 9th and 10th frames in the normal phase.

3. PERFORMANCE EVALUATION

In this section, we evaluate the performance of our proposed CRFID tag identification method. The identification rate at frames and the delay are evaluated for varying number of CRFID tags. Monte Carlo method with MATLAB as a simulation tool is used. For each iteration, tags uniformly distributed in a square area with a reader located in the center. Then, tags perform random contention to send data to the reader using FSA or the proposed protocol. Since DFSA is not feasible as shown in Fig. 2, we compare the proposed method with the Framed Slotted ALOHA (FSA). The frame size of FSA is fixed as a half of the number of existing tags considering energy requirement. Energy requirement of CRFID tags is determined by the distance from the reader. During the simulation, the identification rate in frames and the delay of each tag are recorded. The parameters used in the simulations are shown in Table I.

Table 1. Simulation parameters.

Parameter	Value
Simulation area	5 m×5 m
Number of tags	200~300
Maximum energy requirement	3, 6
Processing delay	0.4 ms
Duration of time slot	1 ms
Data size	96 bits

Fig. 5 shows the identification rate as frames evolve. Since the number of active CRFID tags changes periodically, the identification rate fluctuates for a fixed frame size in FSA. In the proposed scheme, the identification rate fluctuates at first, then it converges to the value higher than that of the legacy FSA. The identification rate is maximized since the optimal frame size is allocated using the table. As the maximum energy requirement increases, the number of active tags at frames becomes smaller. Then, the identification rate decreases since the fixed frame size becomes too large. Also, the convergence of the identification time in the proposed method for larger minimum energy requirement becomes longer since the table information is collected more slowly. The proposed scheme shows better performance than the legacy FSA after about 60 frames.

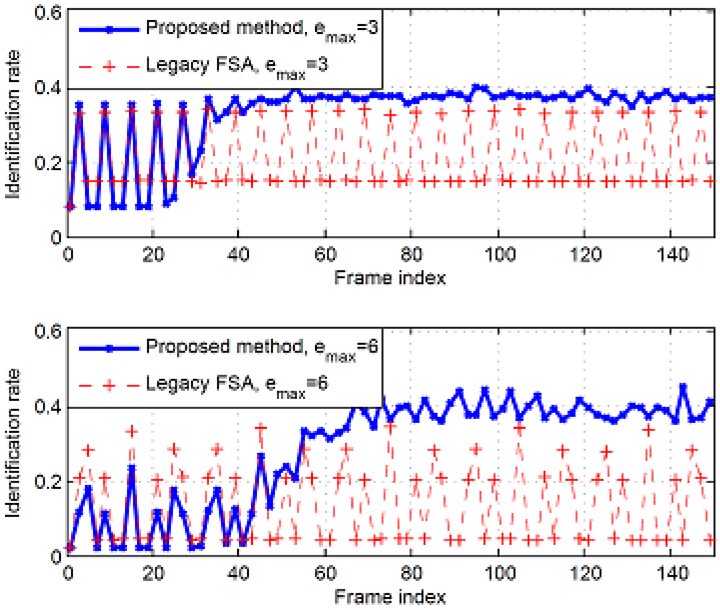


Figure 5. Identification rate of CRFID systems as frame evolve.

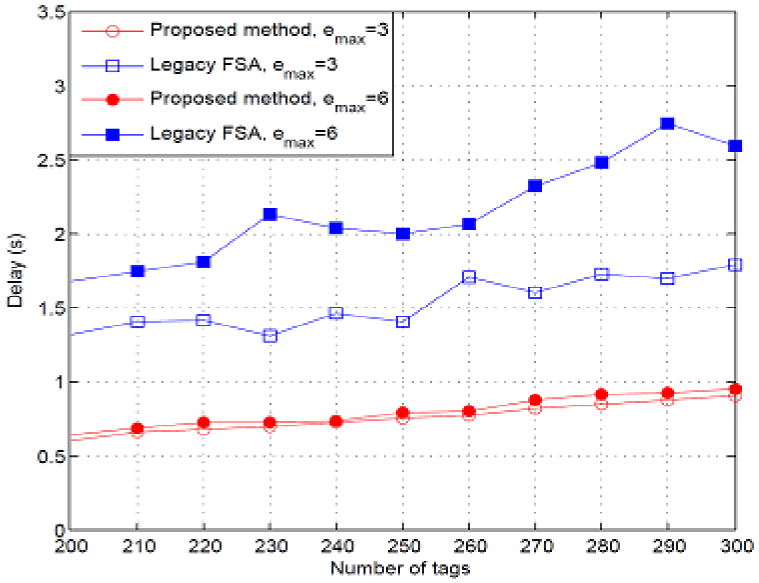


Figure 6. Delay of CRFID tags for varying number of tags.

Fig. 6 shows the delay of CRFID tags for varying number of tags. The average delay of tags increases as the number of tags increases. Since the chance of transmission reduces as the number of tags increases, the delay of tags becomes larger. The proposed method shows lower delay than the conventional FSA. By maximizing the identification rate, the delay between transmissions is reduced in the proposed scheme. When the maximum energy requirement is 6, the delay is larger than that of the case with the energy requirement of 3. Since tags sleep longer when more collisions occur, the delay increases. In the proposed method, the delay increment is negligible compared to that of the legacy FSA. Since the optimal frame size is reduced the chance of collisions, delay increment caused by sleep interval is minimized.

3. CONCLUSIONS

In this paper, we have proposed a CRFID tag identification protocol considering the features of CRFID tags. Especially, we consider CRFID tags which need energy for sensing process. The proposed MAC protocol collects the energy and processing time requirements of tags. After the requirement of all tags is collected, the reader can schedule the optimal frame size. We also provide the time duration to support the processing delay of CRFID tags in every frame. We have shown that the proposed method can improve the identification rate and delay compared to the conventional FSA scheme.

ACKNOWLEDGEMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP)(2014R1A5A1011478, 2015R1A2A2A01004067).

REFERENCES

[1] M. Bolic, M. Rostamian, and P. M. Djuric, “Proximity Detection with RFID: A Step Toward the Internet of Things,” IEEE transactions on Pervasive Computing, vol. 14, no. 2, pp. 70-76, Jun. 2015.

[2] H.-D. Ma, “Internet of Things: Objectives and Scientific Challenges,” Springer Journal of Computer Science and Technology, vol. 26, no. 6, pp. 919-924, Nov. 2011.

[3] C. W. Park, J. H. Ahn, and T.-J. Lee, “RFID Identification Protocol with Optimal Frame Size for Varying Slot Time,” International Journal of Information and Electronics Engineering, vol. 4, no. 2, pp. 87-91, Mar. 2014.

[4] A. P. Sample, D. J. Yeager, P. S. Powledge, A. V. Mamishev, and J. R. Smith, “Design of an RFID-Based Battery-Free Programmable Sensing Platform,” IEEE Transactions on Instrumentation and Measurement, vol. 57, no. 11, pp. 2608-2615, Nov. 2008.

[5] B. Sheng and C. C. Tan, “Group Authentication in Heterogeneous RFID Networks,” in Proc. IEEE Conference on Technologies for Homeland Security, pp. 167-172, Nov. 2015.

[6] L. Atzori, A. Iera, and G. Morabito, “The Internet of Things: A survey,” International Journal of Computer and Telecommunications Networking, vol. 54, no. 15, pp. 2787-2805, Oct. 2010.

[7] W. Yang, D. Wu, M. J. Hussain, and L. Liu, “Wireless Firmware Execution Control in Compuational RFID Systems,” in Proc. of IEEE International Conference on RFID, pp. 129-136, Apr. 2015.

[8] Y. Zheng and M. Li, "Read Bulk Data from Computational RFID," in Proc. of IEEE INFOCOM, pp. 495-503, Apr. 2014.

[9] A. Wickramasinghe and D. C. Ranasinghe, "Ambulatory Monitoring Using Passive Computational RFID Sensors," IEEE Transactions on Sensor Journals, vol. 15, no. 10, pp. 5859-5869, Oct. 2015.

[10] "EPC Radio-Frequency Identification Protocols Generation-2 UHF RFID Protocol for Communications at 860MHz-960MHz," Version 2.0.0, EPCglobal, Nov. 2013.

[11] J.-R. Cha and J.-H. Kim, "Novel Anti-collision Algorithms for Fast Object Identification in RFID Systems," in Proc. of International Conference on Parallel and Distributed Systems, pp. 63-67, Jul. 2005.

[12] G. Khandelwal, K. Lee, A. Yener, and S. Serbetli, "ASAP: A MAC Protocol for Dense and Time-constrained RFID Systems," EURASIP Journal on Wireless Communications and Networking, vol. 2007, no. 2, pp. 1-13, Jan. 2007.

AUTHORS

Yunmin Kim received the B.S. and M.S. degrees in electrical and computer engineering from Sungkyunkwan University, Korea, in 2012 and 2014, respectively. He is currently working toward the Ph.D. degree in electrical and computer engineering at Sungkyunkwan University since March 2014. His research interests include energy-harvesting networks, cognitive radio networks, vehicular networks, wireless LANs, RFID/NFC, and IoT, resource allocation and Medium Access Control (MAC) of wireless communication networks.

Ji Hyoung Ahn received the B.S. and M.S. degrees in electronic, electrical, and computer engineering from Sungkyunkwan University, Korea, in 2010 and 2012, respectively. He is currently working toward the Ph.D. degree in the College of Information and Communication Engineering at Sungkyunkwan University since March 2012. His research interests include routing protocols for ad hoc networks, wireless communication networks, wireless LAN, and wireless PAN.

Tae-Jin Lee received his B.S. and M.S. in electronics engineering from Yonsei University, Korea in 1989 and 1991, respectively, and the M.S.E. degree in electrical engineering and computer science from University of Michigan, Ann Arbor, in 1995. He received the Ph.D. degree in electrical and computer engineering from the University of Texas, Austin, in May 1999. In 1999, he joined Corporate R & D Center, Samsung Electronics where he was a senior engineer. Since 2001, he has been Professor in the College of Information and Communication Engineering at Sungkyunkwan University, Korea. He was a visiting professor in Pennsylvania State University from 2007 to 2008. His research interests include performance evaluation, resource allocation, Medium Access Control (MAC), and design of communication networks and systems, wireless LANs/PANs, vehicular networks, energy-harvesting networks, IoT, ad hoc/sensor/RFID networks, and next generation wireless communication systems. He has been a voting member of IEEE 802.11 WLAN Working Group, and is a member of IEEE and IEICE.

INTENTIONAL BLANK

SOTM: A SELF ORGANIZED TRUST MANAGEMENT SYSTEM FOR VANET

Amel Ltifi[1], Ahmed Zouinkhi[2] and Mohamed Salim Bouhlel[1]

[1]Research Unit: Sciences and Technologies of Image and Telecommunications, Higher Institute of Biotechnology of Sfax-Tunisia
altifi@gmail.com
[2]Research Unit: Modeling, Analysis and Control of Systems, National Engineering school of Gabes-Tunisia
Ahmed.zouinkhi@gmail.com

ABSTRACT

Security and trust management in Vehicular Adhoc NETworks (VANET) is a crucial research domain which is the scope of many researches and domains. Although, the majority of the proposed trust management systems for VANET are based on specific road infrastructure, which may not be present in all the roads. Therefore, road security should be managed by vehicles themselves. In this paper, we propose a new Self Organized Trust Management system (SOTM). This system has the responsibility to cut with the spread of false warnings in the network through four principal components: cooperation, trust management, communication and security.

KEYWORDS

Active vehicle, cooperation , trust management, VANET

1. INTRODUCTION

Road safety is the purpose of many researches and projects over the world, given the huge number of deaths and accidents [1]. VANET is a subclass of Mobile Ad-hoc Networks aiming at enabling a set of services for vehicles such as road security. It's a set of vehicles. Each one can communicate with other vehicles using DSRC (Dedicated Short Range Communication) technology (5.9 GHz) that supports ranges of up to 1 KM [2]. The vehicle is equipped with an On Board Unit (OBU). Each OBU is composed of a Global Positioning System (GPS) receiver, an Event Data Recorder (EDR), front radar, rear radar and a central computing system. EDR archives the sent/received messages to be available for use in emergency states. GPS receiver lends information about location, direction, speed and acceleration of the vehicle at a specific time. The computing system is used for data processing. Currently, VANET is the principal element in most current suggestions aimed at enhancing driving conditions. Intelligence ambient (AmI) and ubiquitous computing are new challenging technologies that can be used among VANET applications [3]. The concept of active object is a principal element in the AmI technology. It's recently introduced as an element of the active security in critical domains such as chemical industry [4].

David C. Wyld et al. (Eds) : NETCOM, NCS, WiMoNe, CSEIT, SPM - 2015
pp. 69–81, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51607

This paper illustrates a self organized trust management scheme for VANET. The nodes of this network are Active vehicles which can communicate with each other, they can decide about trustworthiness of received alerts messages, and they can manage their security states.

The organization of the paper is as follow: after an introduction, the second part presents some related works. The third part presents components of the proposed system SOTM. We dedicated the forth part for model evaluation. Finally, the fifth part concludes the paper.

2. RELATED WORK

Only a few trust models have recently been proposed for enhancement reliable information spreading in VANETs. For example, authors in [5], [6] have investigated in security and privacy on trust establishment in VANETs that relies on a security infrastructure and most often makes use of certificates. A survey on this kind of trust models can be found in [7]. Another different class of trust models is a set of systems which are independent from static infrastructure. In these models, cooperation between vehicles is the key to determine the trustworthiness of data transmitted between peers.

Golle et al. [8] present a technique that aims at addressing the problem of detecting and correcting malicious data in VANETs. Each vehicle maintains a model of VANET that contains all the knowledge that a particular vehicle has about the network. Data evaluation is done according to its coincidence to the peer's model of VANET.

A sociological trust model is proposed in [9] based on the principle of trust and confidence tagging. A new architecture for securing vehicular communication and a model for preserving location privacy of the vehicle are presented.

Dynamic Trust-Token (DTT) is an approach to strengthen cooperation in VANET [10]. The purpose of this mechanism is to detect and prevent misbehavior nodes intervention in the transmission of packets, and ensure the integrity of packets over the releases. DTT uses two cryptographic mechanisms: symmetric and asymmetric, to protect the integrity of packages. Thus, it applies "Neighborhood WatchDog" [11] to generate the trust token that based on instantaneous performance to verify the correctness of packets. Thus, many different solutions that rely on existing historical reputation or past records, DTT is based only on execution performance to implement instant reputation for each node, where no accumulation of information is necessary. With DTT, the packets containing incorrect information will not be propagated in VANET. In this approach, each node can play three logic roles: Predecessor, Relaying and successor in the process of transmission of the packet over time.

In our work, we established a trust management system based only on cooperation between vehicles. This work provides a new communication protocol between vehicles to be able to differentiate between trusted and non trusted messages transmitted in VANET.

3. SELF ORGANIZED TRUST MANAGEMENT SCHEME (SOTM)

The proposed scheme is based on the interaction and the communication between active vehicles supposed to manage by themselves their own security states. For this purpose, we have introduced the concept of active vehicle as a result of the integration of the ambient intelligence in the intelligent transport technologies. A new protocol of communication is defined between vehicles based on messages exchanging and aiming to have the ability to each vehicle to decide if a warning message received is correct or not. The SOTM system is composed of four principal components as depicted in figure 1. In this section, we will explain the roles of these four components.

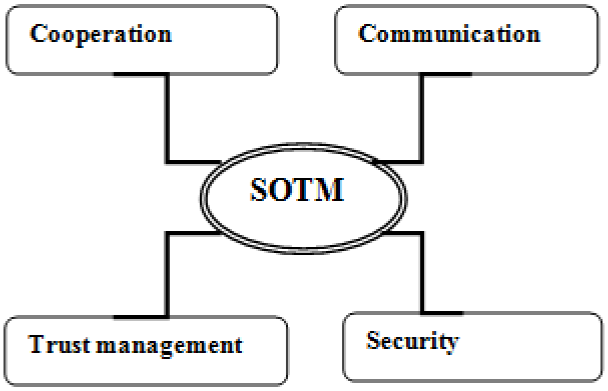


Fig.1 Component-based architecture for SOTM

3.1. Communication module

Generally, peers in VANET can communicate according to three modes of communication: Vehicle-to-Vehicle (V2V) among vehicles, Vehicle-to-Infrastructure (V2I), between vehicles and Road-Side Units (RSUs), and Vehicle-to-X (V2X), mixed V2V-V2I approach. In SOTM, vehicles are allowed to communicate only with V2V mode. For emergency message routing, the clustering model is applied. For each community of vehicles, there is a group leader that has the role of a trusted authority. There are two types of links between vehicles as depicted in figure 2: Unicast link and broadcast link.

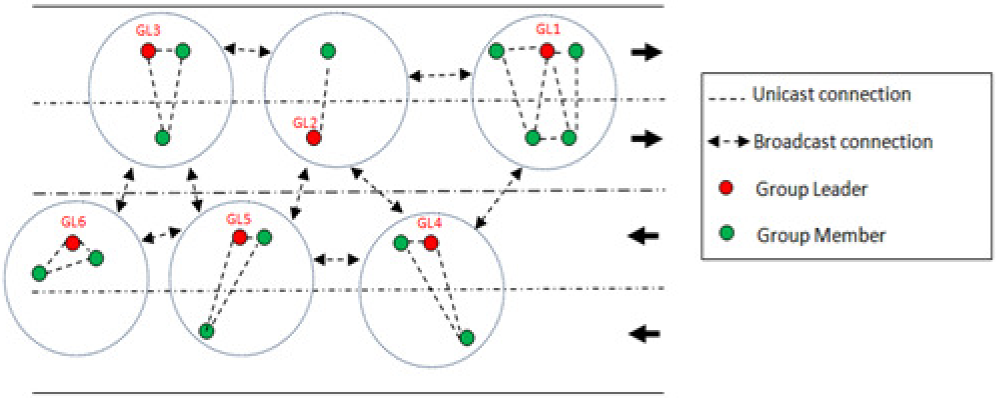


Fig.2 connection model between vehicles

This kind of application is very close to ad-hoc networks. In this situation, vehicles manage by themselves the traffic state. The V2V uses the standard IEEE 802.11p specification for network connection [12]. The 802.11p is an approved variant of the standard 802.11 used for Wi-Fi. The used band of spectrum is between 5.85GHz and 5.925GHz.

The vehicle-to-vehicle communication can be used alone on account of the existence of new wireless technologies and especially the IEEE 802.11p standard. The inter-vehicular communication gains benefit from wireless ad-hoc Networks and GPS to guarantee stable one hop and multi hop communications between vehicles [13].

Routing algorithm is the mainly challenging mission for VANET because of the strict requirements of VANET to high speed mobility and a rapidly changing topology [14]. For this reason, we opted to use a clustered architecture to create a network vision more stable and more reduced to each vehicle [15].

3.2 Trust management model

The aim of our work is to create a community of vehicles that is able to manage by itself its own active security state. It relies on the presence of communicant vehicles on the road. Each vehicle plays a specific role as a member of a disciplined community. To train vehicles facing their active security states, a new communication strategy is deployed by our trust management model. A set of rules is defined and should be applied concerning the collaboration way between vehicles. A knowledge base system is defined in the SOTM system to be integrated in the vehicle to be able to decide on received alarm messages trustworthiness.

3.2.1 Vehicles tasks in SOTM system

VANET is a sub-class of Ad-hoc networks. In such an environment, the trusted authorities couldn't be a part in the majority of security systems. In our case, the disciplined communication between vehicles is the key to create a stable community of vehicles that offers a number of services of road security. There are three main tasks for an Active Vehicle: announcement, communication and revocation. We will present in this part a description and the exchanged messages in each state.

3.2.1.1 Task 1: the announcement

During its driving life time, the vehicle may pass through different groups or it may create a new group. In order to announce its coming into a group, it should send a HELLO message on broadcast. The group leader GL should send an acknowledgement to the new coming vehicle that saves the address of the GL to be used during its transit through this group.

The "HELLO" message contains two fields. The first field is the identifier of the message sender. The vehicle identifier is a unique number aimed to distinguish between vehicles. And the second field is the public key generated by the On Board Unit (OBU) of the vehicle to be saved by the leader. We used the RSA method to achieve the authenticity and the integrity of messages. The "AckHello" has one field which is the leader Identifier.

3.2.1.2 Task 2: the communication

As is the case of any person in a human society, an active vehicle couldn't manage its security state without the interaction with other vehicles. It cooperates with its neighbourhood to be informed if there is an accident in the same space to react quickly. It should also transmit the received emergency messages to others. Before reacting and transmitting warning messages, the active vehicle should be sure about the trustworthiness of the received message. There is a set of messages exchanged between vehicles during their communication.

The “GRE” message is a periodic message sent by each vehicle member after the announcement step. It contains its type and the Id of the sender.

The “WARNING” message is sent by a vehicle when it detects an accident or an obstacle on the road. The destination of this message is the leader that verifies the trust level of the sender to decide whether to accept it or to ignore it. In the case of acceptance, first, the leader remunerates the sender by incrementing its Cooperation Counter, and second, it sends an “AckWARNING” to the sender to allow to it to transmit the warning to its successor. This message contains two fields: The Id of the sender and the warning number (numWrg) which is a unique number affected to each warning by the leader to distinguish between different warnings transmission sessions. A warning transmission session begins when the vehicle which triggered the warning receives an “AckWARNING” message from the leader. Consequently, it sends an “ALARM” message to its successor (the closest neighbor). The “ALARM” message contains the Id of the sender, the signature of the sender computed by its OBU based on the hashing method SHA-1 and the Data field containing the warning message.

When the vehicle successor receives the “ALARM” message, it should decide whether the warning is true or false. So, it begins the verification procedure by sending a “CONFIRM” message to the group leader to verify the trustworthiness of the sender. In this case, there are three possible statements:

- State 1: The sender is trustworthy. So, the leader sends a “VALIDATION” message to the vehicle successor to be able to transmit the “ALARM” message to another successor. The “VALIDATION” message contains the Id of the sender (IdS), the public key (KeyPb) used to verify the authentication of the sender and the warning number (NumWrg).

- State 2: The sender is untrustworthy. So the leader sends an “ERROR” message to the successor to stop the transmission session of the warning. The “ERROR” message contains the Id of the sender.

- State 3: The Data field is falsified by a malicious node. In this case, the leader sends a “CorVALIDATION” to the successor containing the original warning message received by the vehicle which triggered first the alarm. The structure of the “CorVALIDATION” message is similar to the "VALIDATION" message structure but it contains also the field Data describing the triggered alert.

At the end of each statement, the group leader updates the trust values (TV)s and the Cooperation Counters (CC)s of the vehicles which participated in the warning transmission session according to their behaviors.

This verification process is repeated by each successor receiving the “ALARM” message until the end of the transmission session when the last vehicle which received the “ALARM” message has no successor.

3.2.1.3 Task 3: Revocation

The revocation from a group can be a partial revocation or a total revocation. The first form handles the case of vehicles which pass through a group leader for many times. in this case the trust value of the correspondent vehicle saved by the GL will not be deleted. The exit of a vehicle can be explicit by sending an “EXIT” message to the GL or it can be implicit when the GL doesn’t receive message from the vehicle for a period of time. The exit time is saved by the GL into a timestamp to be used in the total revocation that is launched periodically for all the trust model items. For each entry in the trust model, the GL computes the duration between the timestamp saved for the last exit and the current time. If this duration exceeds a threshold, the item should be deleted.

3.2.2 Knowledge base

For registration purposes, we chose to apply a knowledge base system to be used to make appropriate decisions about received alert messages. This system is depicted in figure 3.

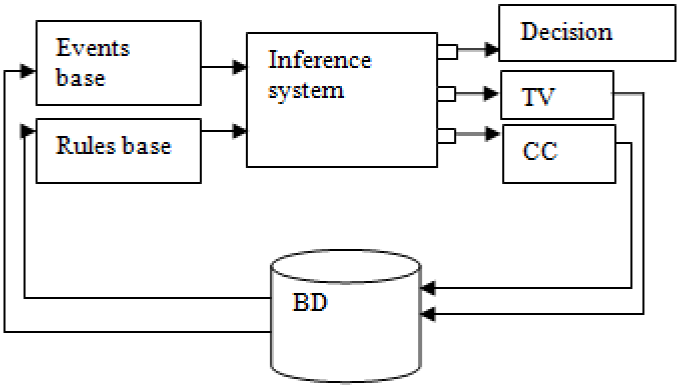


Fig. 3 knowledge base system

3.2.2.1 Events base

Events base contains vehicle properties (idVehicle, position, speed, acceleration …), the trust model structure (idVehicles of neighbors, TVs of neighbors, CCs of neighbors) and all possible road events (Accident (timeA, positionA), Obstacle (timeO, positionO).

3.2.2.2 Rules base

The rules base is a set of rules defining the action list of a vehicle after receiving a message. The vehicle behaviour depends on the message type and the parameters values registered in the Events base. The knowledge base process is the same as a traditional Inference system. It begins by the reception of a new message by a vehicle and it finishes by the generation of the decision about the message if it is accepted or not. Two others parameters are determined: the new trust value (TV) and the new Cooperation Counter of the vehicle source of the message.

3.3 Cooperation model

Our approach is totally autonomous with regard to the external infrastructure. It aims at detecting of malicious vehicles. Our approach guarantees the delivery of the authentic messages while messages containing incorrect information will not be propagated in the vehicular network. The proposed model is a secure and incentive model which has for objective to insure the cooperation encouragement between vehicles by various tools which are:

• The incentive mechanism: if a vehicle A behaves in a cooperative way, the GL modifies the Cooperation Counter "CC" of A by adding points.

• The system of punishment: if a vehicle A behaves in a not cooperative way, the GL modifies the Cooperation Counter "CC" of A by subtracting points.

• The isolation of malicious vehicles: if a vehicle reaches a threshold for the value of "CC", it will be eliminated from the group. So, it will not be covered by the community services.

• The evaluation of the trust level: it is the leading part of our system, the computing of the trust levels of vehicles is necessary to encourage them to cooperate. The GL updates the Trust Value (TV) of Active vehicles according to equation 1:

$$TV = TV + CC \times \propto \qquad (1)$$

Where :

- TV is the Trust Value,
- $TV \in [0,1]$
- CC : the value of the vehicle Cooperation Counter
- $CC \in [CC_{max}, CC_{min}]$
- $CC_{min} = -CC_{max}$
- $\propto \in [0, \frac{2}{CC_{max}}]$

3.4 Security model

Currently, because of its huge spread, wireless technology introduces many possible risks to its users. The security module, in our model, provides a solution for these possible risks. Our solution was inspired from the PGP (Pretty Good Privacy) algorithm that is used hugely in a self-organized network as VANET [16]. Social relationships between vehicles are close to those in the PGP system [17]. Unless, the very large amount of source of the complete PGP version makes from its comprehension and use a difficult task [18]. For this reason, we focused only on using the cryptographic and the hash methods used by PGP which are RSA and SHA. Our security module implies the algorithm SHA1-RSA [19]. RSA [20] is a public-key cryptosystem for both encryption and authentication. The public-key cryptography has many advantages [20] as providing the possibility to implement digital signatures. Many existing solutions for VANET security are using RSA [17][21-23]. We applied the SHA-1[24] function with the RSA encryption method. RSA is combined with the SHA1 hashing function to sign a message in this signature suite.

The group leader is in charge of the key distribution in its group. Each vehicle has a pair of public/private key generated by its OBU (On Board Unit). In the announcement step, each vehicle sends its public key to the leader to be used later in the communication step. When a vehicle A receives an ALARM message from its predecessor B, B sends a CONFIRM message to the leader to verify the trustworthiness of the message and to obtain the public key of A in order to verify the sender authenticity.

4. EVALUATION

We have evaluated our system with respect to two aspects: the number of peers integrated in the community and the average delay in the network.

Table 1 Simulation Parameters

Simulation parameter	Value
Speed Limit of Vehicles	30 Km/s
Acceleration/deceleration	0.5ms^{-1}/3ms^{-1}
Number of vehicles	8 to 40
Transmission power	9db, 12db, 15db, 18db, 21db
Simulation time	19s to 80s
Communication protocol	802.11a
Data rate	6Mb/s

.

4.1 Reliability of the suggested protocol

In order to evaluate the efficiency of the suggested model, it's important to start by studying the number of vehicles entering to the community according to a set of parameters such as the Id of the group leader and the transmission power of vehicles. In this section, the simulation time is equal to 60s and the number of vehicles is equal to 26.

4.1.1 Influence of the variation of the group leader identity

First, we have done a set of simulations with the same transmission power (21db). In each one, a different vehicle is designed to be the group leader. Figure 4 shows the percentage of vehicles which are entered in the group for each simulation. The group leader that accepted the great number of vehicles in its group is the vehicle V_0 (88% of the nodes number). The number of members in a group depends on the transmission power of vehicles and the number of vehicles in the leader surrounding.

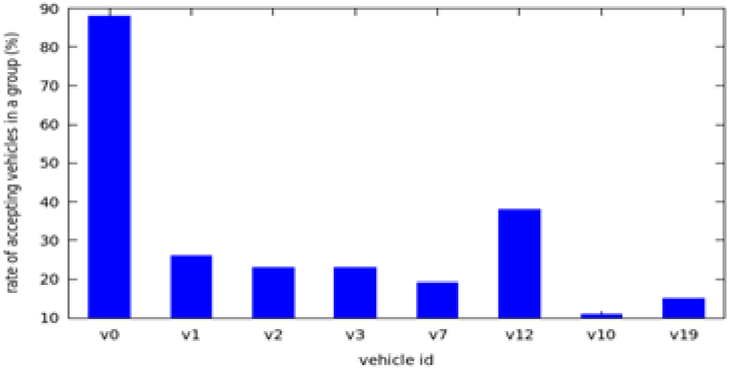


Fig. 4 Accepted vehicles percentage vs. group leader identity

4.1.2 Influence of the variation of the transmission power

In this case, we have launched a new set of simulations for different transmission powers. We have designed the vehicle V_0 as the group leader. As shown in figure 5, for a transmission power equal to 21db relevant to a transmission range of 250-300m, we have found that 88% of the total number of vehicles has undergone the announcement step.

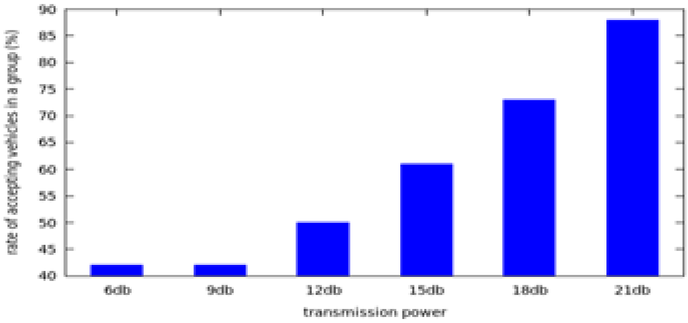


Fig.5 Accepted vehicles percentage vs. transmission power

4.2 Time overhead of SOTM

For evaluation purposes, we simulate our model with a transmission power equal to 21db and with the vehicle V_0 as the group leader. Figure 6 illustrates the end-to-end delay versus the number of nodes for these three speeds. Results obtained by the suggested communication model are compared to the end-to-end delay obtained by the simulation of another approach described in [25]. Authors in [25] proposed an infrastructure based authentication approach for VANET. The simulation results proved that the time overhead introduced by our suggestion is well under the overhead introduced by the approach [25] that ensures only the message authentication and doesn't include an algorithm for a complete trust management as it is the case of our approach. This improvement is due to the self organized approach adapted by our application.

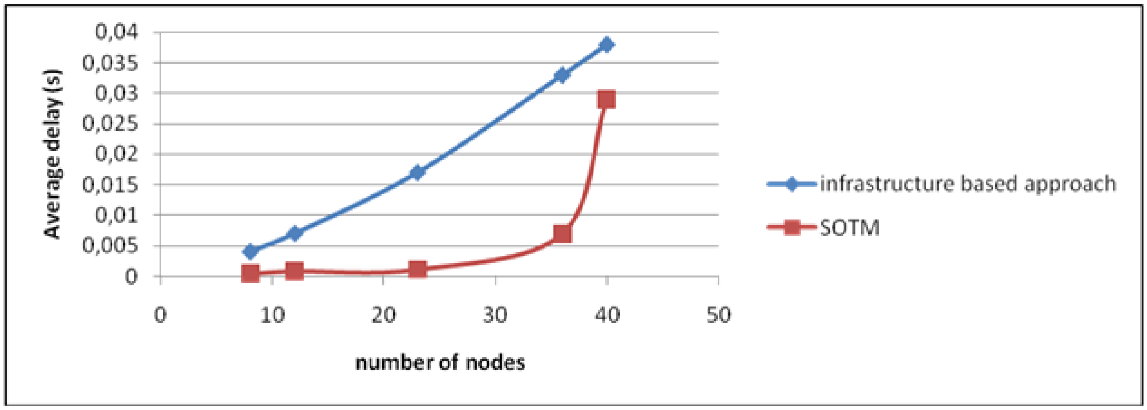


Fig. 6 SOTM vs. an infrastructure based approach

According to figure 6, the overhead introduced by our protocol is under the threshold fixed by the Dedicated Short Range Communications standard (DSRC) [12] that is 100 ms, although, this overhead is caused by messages sent periodically to maintain the linkability between vehicles (ex. the GRE packet). And, it can be reduced by studying and measuring the impact of the periodic time of such a control packet on the network delay in order to obtain the lowest overhead.

5. CONCLUSION

We presented in this paper the SOTM system as a new self organized trust management system for VANET. The SOTM system deals with the registration/updating of vehicles trust values based on historical and runtime vehicles behaviors. Indeed, our model allows the detection and the elimination of misbehaved nodes. And, it interrupts the spread of any false alert message transmitted between vehicles. In addition, simulation results show that the average delay of the proposed system is well under the tolerant delay constraint defined by the DSRC. In order to enhance the SOTM system performance, the privacy issue will be a priority task in the future works.

REFERENCES

[1] G. Samara, W. A. H. Al-Salihy, and R. Sures, “Security issues and challenges of vehicular ad hoc networks (VANET),” in Proceedings of the 4th International Conference on New Trends in Information Science and Service Science (NISS '10), pp. 393–398, Gyeongju-si, Republic of Korea, May 2010

[2] M. M. I. Taha and Y. M. Y. Hasan, “VANET-DSRC protocol for reliable broadcasting of life safety messages,” in Proceedings of the IEEE International Symposium on Signal Processing and Information Technology (ISSPIT '07), pp. 104–109, December 2007.

[3] Gillani, S., Khan, I., Qureshi, S., Qayyum, A.: Vehicular ad hoc network (VANET): enabling secure and efficient transportation system. Technical Journal, University of Engineering and Technology, Taxila, vol. 13 (2008)

[4] A. Zouinkhi, A. Ltifi, E. Bajic, E. Rondeau, M. B. Gayed and M. N. Abdelkrim, 'Simulation of active products cooperation for active security management', 8th International Conference of Modeling and Simulation, MOSIM'10, May 10-12, 2010, Hammamet, Tunisia, 2010

[5] J. J. Haas, Y.-C. Hu, and K. P. Laberteaux, "Design and analysis of a lightweight certificate revocation mechanism for VANET," in Proceedings of VANET, 2009, pp. 89–98.

[6] M. Raya, P. Papadimitratos, I. Aad, D. Jungels, and J.-P. Hubaux, "Eviction of misbehaving and faulty nodes in vehicular networks," IEEE Journal on Selected Areas in Communications, vol. 25, no. 8, pp. 1557– 1568, Oct. 2007.

[7] P. Wex, J. Breuer, A. Held, T. Leinmuller, and L. Delgrossi, "Trust issues for vehicular ad hoc networks," in Proceedings of the 67th IEEE Vehicular Technology Conference (VTC Spring), 2008.

[8] P. Golle, D. Greene, and J. Staddon, "Detecting and correcting malicious data in vanets," in Proceedings of VANET, 2004.

[9] M. Gerlach, "Trust for vehicular applications," in Proceedings of the International Symposium on Autonomous Decentralized Systems, 2007.

[10] Z. Wang and C. Chigan ,"Cooperation Enhancement for Message Transmission in VANETs", Wireless Personal Communications, Vol. 43, No.1, pp. 141-156, 2007.

[11] J. Hortelano, J.C. Ruiz and P. Manzoni, "Evaluating the Uselfusness of Watchdogs for Intrusion Detection in VANETs", IEEE International Conference on Communications Workshops, ICC Workshops, 2010.

[12] Jagdeep Kaur, Er.Parminder Singh,"Performance Comparison Between Unicast And Multicast Protocols In Vanets", International Journal of Advanced Technology & Engineering Research , Volume 3, Issue 1, Jan. 2013, pp109-115

[13] Adil Mudasir Malla, Ravi Kant Sahu, "A Review on Vehicle to Vehicle Communication Protocols in VANETs", IJARCSSE, Volume 3, Issue 2, February. 2013

[14] R. S. Shukla, I. A. Khan, N. Tyagi, "Performance of Modified Edge Based Greedy Routing Algorithm in VANET Using Real City Scenario",Advances in Mechanical Engineering and its , Applications (AMEA) 168 Vol. 2, No. 3, 2012, ISSN 2167-6380

[15] J. Y. Yu and P. H. J. Chong. "A Survey of Clustering Schemes for Mobile Ad Hoc Networks," IEEE Communications Surveys and Tutorials, Vol. 7. No. 1, 2005, pp. 32–48. doi:10.1109/ COMST.2005. 1423333

[16] Randhawa, Navdeep Kaur. "Design and Implementing PGP Algorithm in Vehicular Adhoc Networks (VANETs)," International Journal of Engineering Research and Applications, Vol. 2, Issue 3, May-Jun 2012, pp. 647-650

[17] Shafiullah Khan and Al-Sakib Khan Pathan, "Wireless Networks and Security: Issues, Challenges and Research Trends", Springer Series: Signals and Communication Technology, 2013, pp. 107-132, ISBN 978-3-642-36168-5

[18] Kurniawan, Y., Albone, A., & Rahyuwibowo, H. The design of mini PGP security. International Conference on the Electrical Engineering and Informatics (ICEEI), Indonesia, 17-19 July, 2011.

[19] Sophia A-J. A Score Based Trustworthy Declaration Scheme For Vanets, International Journal of Engineering Research and Applications, 2014; 4(3); 542-544.

[20] Rivest R, Shamir A, Adleman L. A Method for Obtaining Digital Signatures and Public-Key Cryptosystems. Communications of the ACM 1978; 21(2): 120–126.

[21] Serna J., Luna J. Medina M. Geolocation-based Trust for Vanet's Privacy. Journal of Information Assurance and Security 2009; 4(5):432-439.

[22] Alangudi B-N, Mahalakshmi R-S. Privacy Preserving Authentication for Security in VANET. International Journal of Advanced Research in Computer Science & Technology (IJARCST), 2014; 2(1): 200-203.

[23] Verma M, Dijiang H. SeGCom: secure group communication in VANETs. In Proceedings of 6th IEEE consumer communications and networking conference (CCNC 2009), Las Vegas, January 2009.

[24] Zhang J-P, Chen C, and Cohen R. Trust based decision making on message relay and local actions in VANET. Journal of Security Communication Networks, 2013; 6(1): 1-14.

[25] Chaurasia B-K, Verma S. Infrastructure based Authentication in VANETs. International Journal of Multimedia and Ubiquitous Engineering 2011; 6(2): 41-54.

AUTHORS

Amel Ltifi is a PhD student at the National Engineering School of Sfax (Tunisia) and a member of Sciences and Technologies of Image and Telecommunications (SETIT) laboratory. She received the National engineering Degree from the National School of Informatic sciences (ENSI), Tunisia in 2003 in computer sciences. She received the Master degree from the Higher School of Informatics and Multimedia of Gabes (ISIMG), Tunisia, in 2010. Her research activities are focused on Distributed Systems, Ambient Intelligence systems and architectures, VANET and Wireless Sensors Network Concepts.

Ahmed Zouinkhi is Associate Professor at the National Engineering School of Gabes (Tunisia) and a member of Modeling, Analysis and Control Systems (MACS) laboratory. He received the Notional engineering Degree from the National Engineering School of Monastir (ENIM), Tunisia in 1997 in industrial computing. He received the DEA degrees and the CESS (certificate high specialized electrical study) from the Higher School of Sciences and Techniques of Tunis (ESSTT), Tunisia, in 2001 and 2003, respectively. He received his PhD degree in 2011 in Automatic Control from the National Engineering School of Gabes (Tunisia) and a PhD degree in Computer Engineering from the Nancy University (France). His research activities are focused on Distributed Systems, Smart Objects theory and applications, Ambient Intelligence systems and architectures, RFID, VANET and Wireless Sensors Network Concepts and Applications in manufacturing and supply chain.

Mohamed-Salim BOUHLEL was born in Sfax (Tunisia) in December 1955. He received the engineering Diploma from the National Engineering School of Sfax (ENIS) in 1981, the DEA in Automatic and Informatic from the National Institute of Applied Sciences of Lyon in 1981, the degree of Doctor Engineer from the National Institute of Applied Sciences of Lyon in 1983. He has received in 1999 the golden medal with the special mention of jury in the first International Meeting of Invention, Innovation and Technology (Dubai). He was the Vice President of the Tunisian Association of the Specialists in Electronics. He is actually the Vice President of the Tunisian Association of the Experts in Imagery and President of the Tunisian Association of the Experts in Information technology and Telecommunication. He is the Editor in Chief of the International Journal of Electronic, Technology of Information and Telecommunication, Chairman of the international conference: Sciences of Electronic, Technologies of Information and Telecommunication: (SETIT 2003, SETIT 2004 ,SETIT 2005, SETIT 2007, SETIT 2009 and SETIT 2012) and member of the program committee of a lot of international conferences. In addition, he is an associate professor at the Department of Image and Information Technology in the Higher National School of Telecommunication ENST-Bretagne (France).

INTENTIONAL BLANK

REDUCING LATENCY IN AFRICAN NRENS USING PERFORMANCE-BASED LISP/SDN TRAFFIC ENGINEERING

Josiah Chavula, Melissa Densmore, and Hussein Suleman
Department of Computer Science, University of Cape Town, South Africa
{jchavula,mdensmore,Hussein}@cs.uct.ac.za

ABSTRACT

Active topology measurements on the African Internet have showed that over 75% of the intra-Africa traffic destined for Africa's National Research and Education Net- works (NRENs) uses intercontinental links, resulting in high latencies and data transmission costs. The goal of this work is to investigate how latency-based path selection using Locator/Identifier Separation Protocol (LISP) and Software Defined Networking (SDN) in NRENs can be used to reduce inter-NREN latencies. We present aspects of an experimental prototype implementation for real-time topology probes to discover lower-latency remote gateways and dynamic configuration of end-to-end Internet paths. Simulation results indicate that ranking remote ingress gateways, and dynamic configuration of end-to-end paths between gateways can lower the average latency for inter-NREN traffic exchange.

KEYWORDS

Software Defined Networking, Traffic Engineering, Latency, Internet exchange points, African National Research and Education Networks (NRENs)

1. INTRODUCTION

Research collaboration and e-resource sharing in sub-Saharan Africa continues to be hampered by the limited interconnectivity that is not only expensive, but fails to meet the quality of service (QoS) required for collaborative applications among the National Research and Education Networks (NRENs). Despite the ongoing interconnection efforts by African NRENs, over 75% of the inter-NREN traffic is being exchanged through circuitous routes traversing inter-continental links and Internet exchange points in Europe and North America, thereby experiencing high latencies [1].

Efforts have been made to improve the Internet traffic exchange among the NRENs in Africa. A key player in this effort is the UbuntuNet Alliance, an association of the NRENs in Eastern and Southern Africa. Between 2011 and 2014, the UbuntuNet Alliance has been implementing the Africa-Connect Project to create a regional research and education Internet network interconnecting the NRENs in the region. The project has aimed to interconnect Southern and Eastern African NRENs into a regional network through the use of terrestrial network facilities. The project involves establishment of Points of Presence (PoPs) in major cities in the region -

David C. Wyld et al. (Eds) : NETCOM, NCS, WiMoNe, CSEIT, SPM - 2015
pp. 83–98, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51608

notably in Mtunzini, Maputo, Dar es Salaam, Nairobi, Kampala and Kigali, and interconnecting them with broadband cross-border links to create a regional research network. So far, the intra-Africa interconnection serves six NRENs: TENET (South Africa), MoRENet (Mozambique), TERNET (Tanzania), KENET (Kenya), RENU (Uganda) and RwEdNet (Rwanda). Furthermore, transcontinental links between Nairobi and the Ubuntunet Alliance PoP in Amsterdam, as well as from Cape Town to London, have been established, thereby linking UbuntuNet Alliance with GEANT, the European research network (Figure 1).

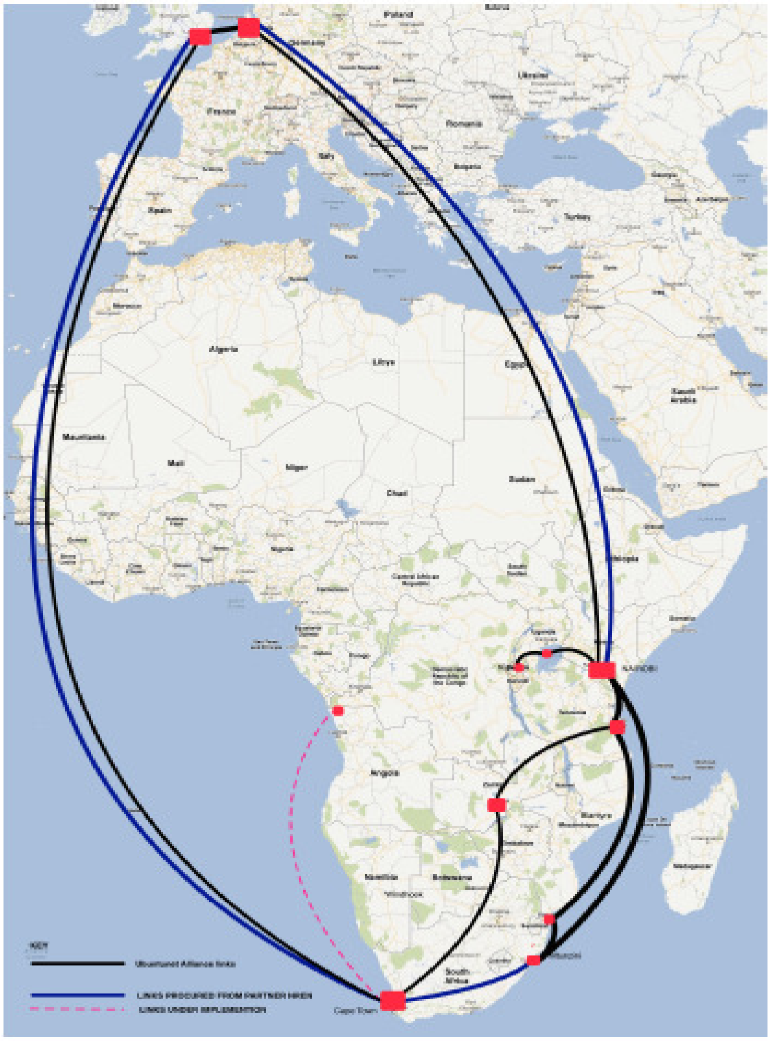


Fig. 1. The UbuntuNet Alliance regional network

The establishment of multiple PoPs, multiple intra-Africa Internet links, as well as multiple transcontinental links, provides new opportunities for implementing multipath routing and traffic engineering mechanisms with the aim of improving performance of traffic exchange among Africa's NRENs. NRENs could implement mechanisms that would allow them to announce to each other, and make use of multiple Internet attachment points to exchange traffic. One protocol that allows edge networks to exchange traffic over multiple Internet gateways is the Locator/Identifier Separation Protocol (LISP) [2, 3]. LISP allows edge networks to announce multiple Internet gateways, known as Route Locators (RLOCs), and to influence the selection of incoming paths. Through a mapping system, LISP allows networks to announce preferences for multiple RLOCs. The availability of multiple locators for the same destination increases path diversity[4] and enable multipath routing, as source networks can select among multiple gateways to reach a destination network.

This paper demonstrates the potential for performance improvement in the Pan-African NRENs by employing, at the network edge, traffic engineering techniques that are based on end-to-end multipath ranking. Using a Software Defined Network (SDN) experimental topology, and a LISP mapping system, the paper examines the potential for dynamically ranking egress and ingress links between multihomed NRENs based on end-to-end path metrics. The objective is to minimize latency for two-way delay sensitive traffic flows (e.g. real-time classroom streaming and video conferences between universities), and minimize intercontinental bandwidth utilization.

2. BACKGROUND AND RELATED WORK

Given the opportunities for traffic engineering provided by the multiple intra-continental and transcontinental links provided by the Africa Connect network, one way of improving the performance and optimization of traffic exchange across African NRENs is to enable dynamic selection of optimal traffic exchange routes based on application QoS needs. For example, path selection for delay sensitive applications can be made based on prevailing end-to-end latencies through either the intercontinental transit links or through the intra-Africa links.

2.1 LISP-based Multipath Traffic Engineering

The Locator/Identifier Separation Protocol (LISP) [3] simplifies multihoming. LISP divides the Internet's address space into two - locally routable Endpoint IDentifiers (EIDs) and global Route LOCators (RLOCs). By separating the host address space from the locator address space, LISP introduces a level of indirection that allows networks to specify preferences for multiple ingress gateways (locators). The availability of multiple locators for the same destination increases path diversity since the source networks are able to forward traffic for a particular destination through multiple remote locators (gateways).

Thus, additional end-to-end performance gains can be achieved with the ability to dynamically select the ingress link at the destination network. In Figure 2, for example, edge networks A and B are multi-homed to networks x,q and y,z respectively. Depending on how the routing is done in the Internet core, the choice of the egress link by network A, i.e. (A,x or A,q), has potential to influence selection of ingress link towards B i.e. 3,y or 4,z. Since each end-to-end path has its own unique path metrics in terms of bandwidth, delay, and loss, selection of particular egress and ingress links at A and B impacts the overall quality of the end-to-end path.

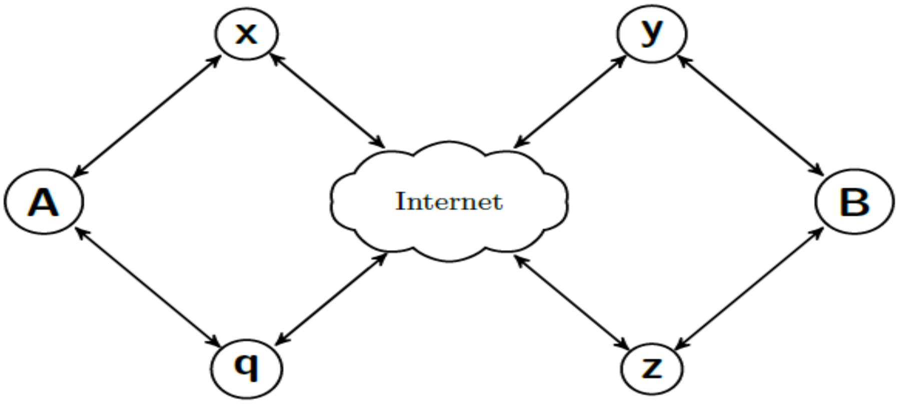


Fig. 2. Multihomed Networks A and B, multihomed through providers (x,q) and (y,z) respectively

One prominent work that uses LISP for traffic engineering is the ISP-Driven Informed Path Selection(IDIPS) [5]. IDIPs is a request/response service where centralized server nodes perform network measurements towards popular destinations, and clients request path rankings for a set of sources and destinations. The IDIPS server ranks the available paths based on a client's ranking preference and measured path metrics. In IDIPS implementation, path ranking is further influenced by the destination's preferences in the locator mapping. The selected paths therefore reflect not only the source network's ranking criteria, but also the destination's preferences for incoming traffic.

2.2 SDN-based Path Enforcement

One challenge with inter-domain multi-path routing and end-to-end traffic engineering is with regard to enforcement of paths across different domains. Software Defined Networking (SDN) provides new opportunities for flexible management of Internet routing and packet forwarding [6]. An SDN-based IXP [7] allows IXP participants to have access to an SDN controller and to write policies that override the default policies of the IXP's BGP route server. SDN has three important characteristics that are useful for interdomain traffic engineering [7]. Firstly, in contrast to traditional switches that forward traffic based only on the destination MAC address, SDN enables packet forwarding based on multiple header fields. Secondly, an SDN controller consolidates control messages from multiple remote networks, such that source and destination networks can remotely configure forwarding paths through a controller. Thirdly, the controller's direct control of the data plane enables dynamic/programmatic configuration of the forwarding tables. With these SDN opportunities, it is possible to allow edge networks some control over selection of inter-domain forwarding paths at Internet exchange points, thereby having more control on the end-to-end paths.

3. A MODEL FOR PERFORMANCE-BASED PATH SELECTION

The traffic engineering framework depicted in Figure 3 is based on the ability of traffic source gateways to select the destination's ingress link based on metrics of the edge-to-edge path. This requires that the source gateway should have a mechanism for learning the destination's multiple ingress links.

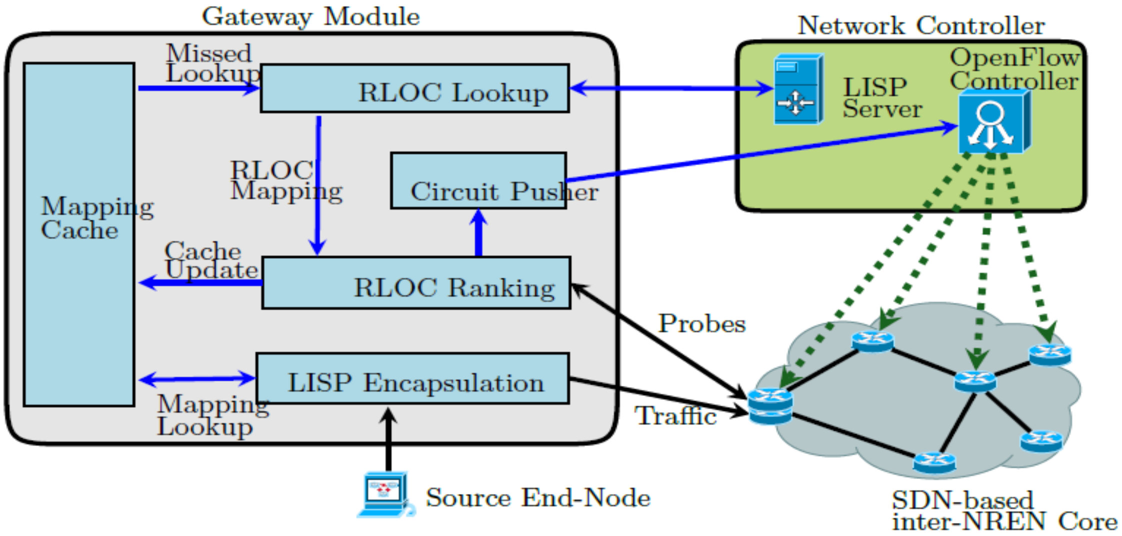


Fig. 3. LISP/SDN Multi-Path Traffic Engineering Model

In summary, the model works in the following manner: For each new traffic flow from a local source node to some remote network, the source network queries the LISP mapping system (through an RLOC lookup API) to obtain the destination network's RLOCs. After obtaining the remote RLOCs, a Locator Ranking module performs active measurements towards remote RLOCs. The Locator Ranking module uses the network metrics obtained from the active measurements to rank the local and remote gateways, after which it updates the local mapping cache. For each new traffic flow, the source network selects the egress and ingress gateways based on the rankings in the mapping cache. Once the egress and ingress links are selected, a Circuit Pusher module is invoked to configure, through an Open Flow SDN control, an end-to-end switching path between the source and destination RLOCs via the selected egress/ingress links. If no mapping exists in the mapping cache for a destination network, the RLOC Mapping module is invoked to perform the lookup, after which the RLOC Ranking module performs the ranking.

3.1 Path Performance Measurement

The key network metrics in this model are end-to-end latency, jitter and packet loss. Latency, measured as round trip time (RTT) for traffic to move from source to destination, and for the acknowledgement packet to be received by the sender, is an important characteristic that affects the performance and responsiveness of Internet applications. Jitter, on the other hand, is the variation in latency over time between a traffic source and destination node. Packet loss is a measure of the percentage of packets lost along the data path for each traffic flow. To obtain traffic characteristics of the network in terms of the key path metrics, active measurement techniques are used. In particular, a ping-based tool is used for sending probe packets, from each edge-network gateway, to the destination networks' RLOCs. Upon retrieving the destination network's RLOCs from the mapping server, the source gateway sends a ping probe, through each of its egress links, to each of the destination's RLOCs. By analysing the solicited responses, topological characteristics such as round-trip-times, jitter and packet loss are obtained. The values obtained from the responses are used to rank the different paths.

3.2 LISP-based Path Ranking

With LISP, multihomed edge networks are able to achieve some degree of path diversity, as multiple alternate gateways become visible between source and destination networks. Achieving optimal end-to-end performance in such environments requires that the source and destination networks should be able to evaluate the alternate paths, and to dynamically select both the source network's egress link and the destination network's ingress link. In particular, the source network needs a way of discovering and evaluating end-to-end links through alternate egress and ingress links.

The routing cost for an edge-to-edge path can be modelled as a vector comprising the measured performance metrics and the network RLOC preferences [8]. Let $P(A_{xy})$; $P(B_{yx})$ be the performance cost vectors for the two edge networks A and B, with respect to forwarding traffic through their access links x and y respectively. The performance cost P from each edge network comprises a set of end-to-end path metrics K, (eg. latency, packet loss, jitter) weighted by variable _, such that:

$$P(A_{xy}) = \sum_{i=1}^{n} \Lambda_i \cdot K_{xy_i}$$

$$P(B_{yx}) = \sum_{i=1}^{n} \Lambda_i \cdot K_{yx_i}$$

where $\sum_{i=1}^{n} \Lambda_i = 1.$

To calculate the total cost, $T(A_{xy})$; $T(B_{yx})$, the source preference cost ($\phi(x)$) is combined with the performance cost P, using a variable scaling factor λ:

$$T(A_{xy}) = \lambda P(A_{xy}) + (1 - \lambda)\phi(x);$$

$$T(B_{yx}) = \lambda P(B_{yx}) + (1 - \lambda)\psi(y);$$

where $0 \leq \lambda \leq 1.$

In LISP implementation, selection of the local egress RLOC as well as the remote ingress RLOC is determined by priority and weight values recorded in RLOC records that are retrieved from the mapping system and stored in a local cache. If multiple locators for the same destination exist, the priority values, ranging from 0 to 255, are used to select the locator that is most preferred. In this work, the calculated path costs are translated into RLOC priority values using a log function that scales the costs into values between 0 and 255. The egress-ingress RLOC pair that has the lowest resultant cost is the one that is used for the outgoing traffic flow between the locator pair. While the mapping remains valid, all subsequent matching flows between the locator pair uses the cached locator ranking. The cached locator mapping remains valid for a configurable TTL period of 60 seconds, after which the mapping and locator ranking process is repeated.

3.3 End-to-end Path Configuration

Each edge network gateway has a Circuit Pusher, an SDN module for setting up end-to-end circuits between locators. After the source gateway has selected both the local and remote locators, it invokes the SDN module to configure, through an SDN network controller, a unidirectional end-end circuit between locators. This is achieved by installing flow entries on all switches that are part of the shortest path between two selected locators. The installed path is unidirectional because each source gateway independently performs RLOC lookups and path ranking, and sets up a circuit toward the remote RLOC.

4. TESTBED IMPLEMENTATION

The overarching purpose of this study is to evaluate the extent to which LISP and SDN can support dynamic selection of end-to-end paths between multi-homed edge networks. The objective was therefore to assess network performance in a topology that uses a dynamic performance-based path selection, versus static default path routing. Figure 4 depicts the testbed implementation.

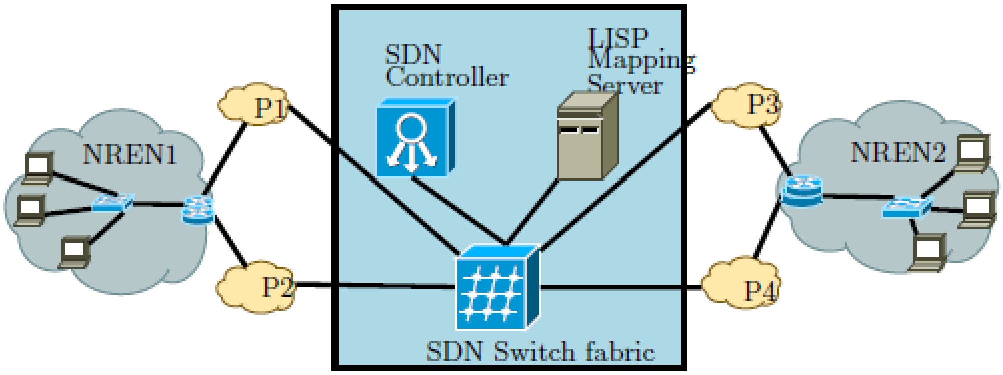


Fig. 4. SDN/LISP Traffic Engineering for dual homed NRENs

4.1 Topology

The testbed was built using a LISP-based SDN simulated network consisting of an SDN (Openflow) controller, an Open flow switch, and a LISP mapping server. The topology was built in a virtual environment, using the Mininet network emulator [9]. Mininet allows the creation of virtual hosts, switches, controllers and links. Furthermore, Mininet nodes run a standard Linux kernel and net-work stack, and can therefore run real network applications. The switches and network controllers are based on the standard SDN protocol Open Flow. An Internet-like topology was built in Mininet and was used together with a LISP mapping system and LISP transit routers (xTRs), which are based on an open source implementation of LISP called Open LISP[10]. The interconnection among the edge networks is through ISPs that interconnect at a common Internet exchange point. More specifically, the network was designed with the following features:

- 48 routers connected to the IXP switch; the routers represent participant networks at the IXP. The JINX has, as of December 2014, 48 participant networks.
- Each IXP participant has 10 edge networks connected to it, and each edge network is dual-homed and serves a total of 5 hosts, representing campus networks. In total, there were 1,200 end hosts as potential sources and destinations for traffic flowing through the IXP fabric.
- To simulate multi-homing, each edge network is connected to two provider networks present at the IXP. The primary link represents the intra-Africa link and is configured with lower latency, while the secondary link represents an transcontinental link and is configured with higher latency.
- The access links between the edge networks and the provider networks are configured with band-width evenly distributed between 2 Mbps and 4 Mbps. The links between the IXP participants (provider networks) and the IXP are equally provisioned with bandwidth of 100 Mbps.
- The end-to-end latencies are modelled on latencies measured for traffic exchanged between African NRENs. Round trip latencies in this experiment are distributed between 60 ms and 700 ms.

4.2 Network Traffic

A realistic evaluation of a network model requires emulating the network with traffic that characteristically resembles the traffic pattern of the emulated networks. This is important, as a major scalability issue with centralized network architectures, such as the Open flow controllers and LISP mapping systems, hinges on their ability to cope with the traffic flow characteristics in the network [11].

Researchers have characterized Internet traffic based on flow metrics such as byte volume, packet volume, flow duration, and flow inter-arrival time. The percentage of UDP traffic has increased with the advent of many UDP based P2P applications and streaming multimedia, which transport large volumes of data[12]. In 2009, a CAIDA survey showed that the ratio of UDP to TCP traffic was almost 0.21 in terms of packet numbers, 0.11 in terms of byte count, and 3.09 in terms of flows.

The length of data flows impacts the relative latency introduced at the controller and, furthermore, the number of active flows has implications for the size of the forwarding tables maintained at each Open Flow forwarding device. For example, a characterization of university campus network traffic [13] established that 21.4% of the traffic was carried by flows longer than 10 minutes, 12.6% by flows longer than 20 minutes, and nearly 2% was carried by flows longer than 100 minutes. Short flows are bursty and have flow speeds ranging from 1 Bps to over 10 kBps, while longer flows are slower, around 50 Bps for 40 min flows, [13].

The test traffic for the experiment is therefore based on the following Internet traffic characteristics:

1. Protocol flow: UDP to TCP ratio: 3:1
2. Flow Duration:
    - 0 - 2 sec : 45% of all the traffic
    - 2 sec - 5 mins : 55% of all the traffic

3. Flow inter-arrival time: 4 ms - 40 ms (or 25 to 250 new flows per sec)
4. Flow rate:
    - Short flows (0 - 60 sec) : 1 Bps - 10 kBps
    - Medium flows (1 min - 5 mins) : 100 Bps - 50 Bps

4.3 Test Traffic Generation

Some of the most widely used traffic generators include Iperf, PackETH, D-ITG, and Ostinato [14,15]. PackETH [16] is a stateless packet generator designed for Ethernet networks, and supports a number of protocols including UDP, TCP and ICMP. Iperf[17] is mostly used for evaluating topology parameters such as bandwidth, delay, window size and packet loss, for both TCP and UDP traffic. Iperf provides an estimation of received and transmitted data rates. Ostinato[14] is a user level traffic generator tool that supports UDP and TCP protocols at multiple rates. DITG (Distributed Internet Traffic Generator) [18] can generate Internet traffic with a user defined packet inter-departure times. This work made use of D-ITG for traffic generation and for obtaining performance metrics.

5. RESULTS

A key objective of the RLOC ranking was to discover and direct traffic flows through lower latency paths towards multi-homed remote networks. TCP traffic is particularly impacted by network round-trip-times, and this paper has considered performance of TCP traffic when RLOC ranking and dynamic path configuration is employed. Furthermore, the evaluation also considers how jitter is affected due to path ranking and circuit configuration.

5.1 Round Trip Times

The key results from a series of experiments suggest that in cases where the paths towards the different RLOCs of an edge network have significantly different RTTs, latency based ranking and selection of RLOCs does help to lower the overall latency in the network. Figure 5 shows the RTT dispersion and mean for TCP traffic in a LISP/SDN topology, with each flow lasting between 1sec and 300 sec. The vertical lines represent the dispersion of flow RTTs over the time interval, with each traffic flow RTT averaged over 2 sec intervals. The instantaneous average RTT for all the flows is indicated by the blue and red lines. The results show that RLOC ranking results in a 20 % lower overall latency compared to the default gateway forwarding.

The performance gains from RLOC ranking appear to diminish significantly with increased network load. As the network gets more congested, the observable gain from RLOC ranking is significantly reduced. This can be explained from the fact that as the network links get more congested, the otherwise shorter links begin to exhibit equally higher RTTs. With higher RTTs, many of the probe packets time-out, prompting the egress RLOC to forward packets towards the destination's default RLOC. Even when the probe packets generate responses, the RTT values of the otherwise shorter paths tend to be just as high, thereby being ranked lower and resulting in selection of the other paths for traffic forwarding. Figure 6 shows the dispersion and mean of the RTT at the point when the network is congested. Figure 7 and Figure 8 further illustrates the RLOC ranking effects in normal network operation and when there is congestion.

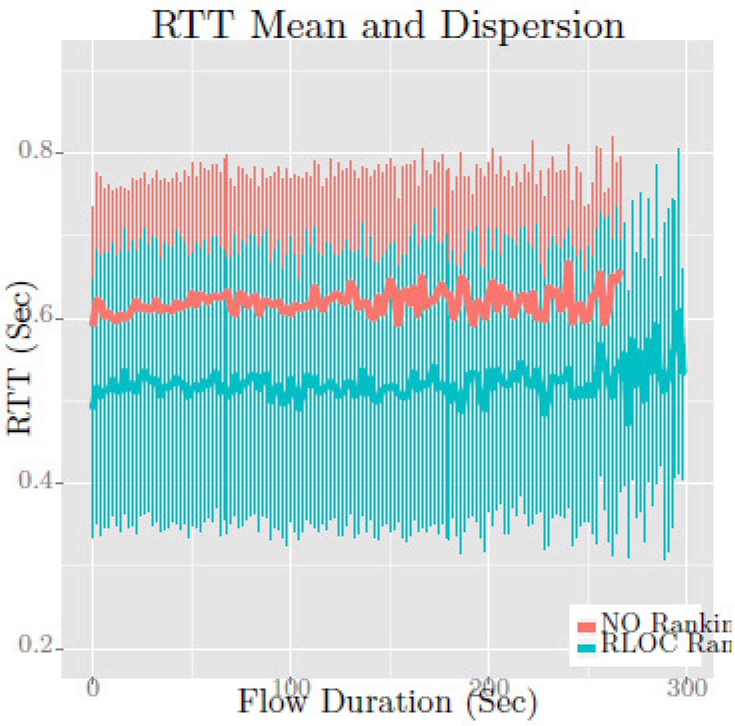


Fig. 5. Round-trip times for network operating with LISP gateway ranking vs non-path ranking LISP operation

5.2 Jitter

Apart from latency, jitter is another key metric that affects performance of interactive Internet applications. Some causes of Internet jitter include congestion in the core network as well as in the access links. Results from the simulation network suggest that RLOC ranking and dynamic path configuration does increase the overall jitter in the network. Given a network topology and traffic profile, RLOC ranking in a LISP network results in higher overall jitter than when no RLOC ranking is employed. Figure 9 shows the jitter for both RLOC ranking and the normal LISP operation. On average, RLOC probing is seen to increase jitter by 20 %. The jitter can be attributed to the delay experienced by some packets during the time the local gateway performs path measurement and ranking.

However, as the network reaches congestion point, both the RLOC ranking and non-ranking scenarios experience similarly higher jitter. This is illustrated in Figure 10, where jitter for both ranking and non-ranking LISP operations have their average jitter increase significantly. As congestion occurs, the probe engine fails to discover any lower latency RLOCs and resorts to using default paths.

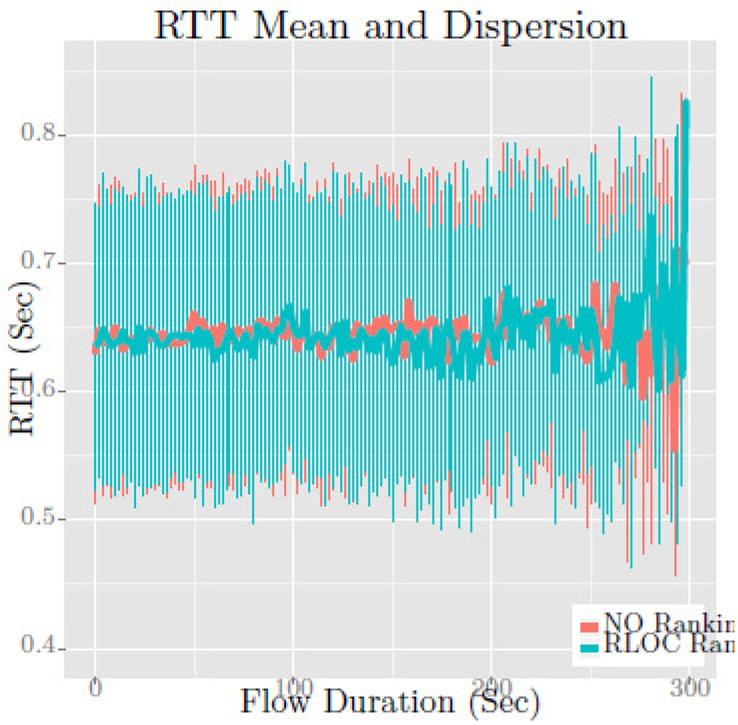


Fig. 6. Round trip times for path ranked and non-path ranked LISP operation at network congestion point

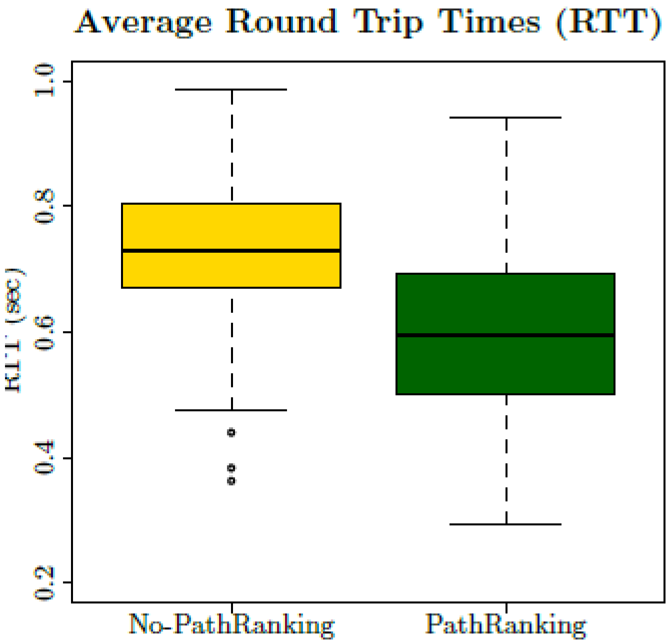


Fig. 7. RTT without congestion

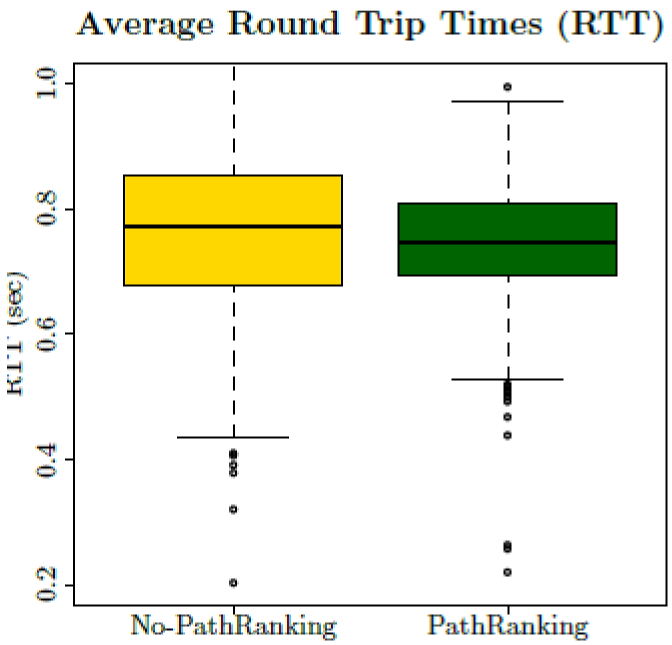


Fig. 8. RTT with congestion

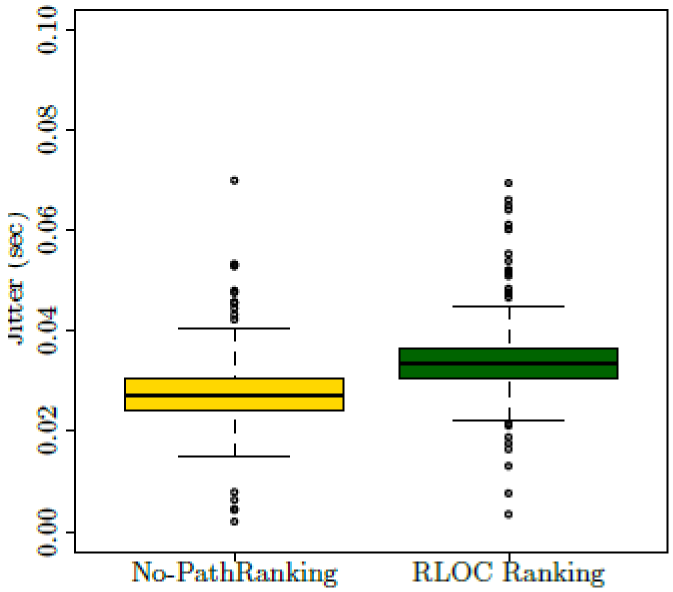


Fig. 9. Jitter with no congestion

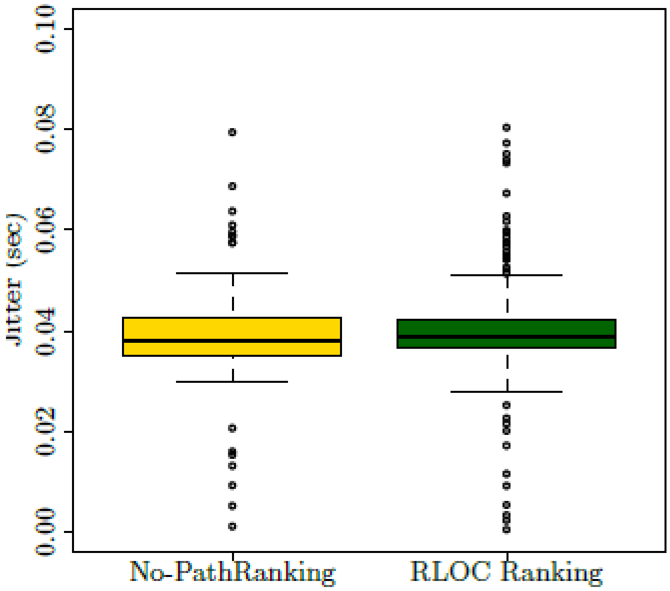


Fig. 10. Jitter with congestion

6. DISCUSSION

6.1 RLOC Ranking During Congestion

The RTT results show that the RLOC ranking mechanism fails to produce positive results under network congestion. In general, centralized systems are venerable to performance bottlenecks under system overload. For instance, inter-arrival times of traffic flows have implications on the performance of centralized network controllers [11]. In an Open Flow network architecture, for example, a scalability challenge stems from the fact that the first packet of each flow is forwarded to a central controller, which is responsible for determining and configuring the forwarding path for the flow. Similarly, for LISP, the egress gateway performs a lookup from a mapping server to determine each new flow's destination network's RLOC. In either case, the flow inter-arrival time has a scalability impact on the network, as higher rates for new flows result in bottlenecks at the SDN and LISP controllers, thereby introducing latency and jitter. Although the scalability and performance bottleneck would affect both the ranking and normal LISP operations, the RLOC ranking mechanism would experience more severe impact as it is dependent on receiving replies from probe packets, which take longer when there is congestion. One way of dealing with RLOC ranking during congestion is to reduce the amount of probing required by using historical performance information to select the RLOCs. Also, the amount of probing needs to be reduced by performing ranking only for critical flows (eg. delay intolerant applications).

6.2 Effects of RLOC Ranking on Jitter

The observed jitter in the experiments reveals that the process of RLOC ranking and path configuration does increase the overall jitter in the network. One way of minimising the jitter is to reduce the path setup time. In the presented model, a new end-to-end path is not setup until path probing/ranking and SDN path configuration is complete. Reducing jitter would require pre-ranking the RLOCs, so that new flows don't wait too long before circuits are set up. Pre-ranking of paths could be based on RLOC performance history as well as known traffic characteristics in the network.

6.3 Model Limitations

One challenge with the model presented in this paper is the assumption that NRENs are multihomed. This is true to a large extent as, in general, campus networks that are part of NRENs have more than one Internet attachment points; a specific network attachment point for traffic destined within the NRENs' network and another for traffic destined to non-NREN networks. Where an NREN has only one Internet attachment point, it can still appear multi-homed by performing prefix de-aggregation and announcing separate prefixes with different RLOCs. This would enable them to still benefit from multipath traffic engineering.

Another challenge is to do with independent selection of the remote RLOCs by the source network, which could result in violation of the destination network's preferences and policies. This could negatively impact on the destination's policies. For edge networks that have some form of cooperation, such as the case with NRENs within the UbuntuNet Alliance, a mutually beneficial approach would be to employ some level of coordination in selection of gateways. While each NREN would aim to optimize traffic cost and QoS performance (latency), collectively, the NRENs can optimize the performance of some common preferred applications. Balancing the individual NREN's optimization objectives with those of the peering domains requires coordination and routing cooperation among the peers. For UbuntuNet Alliance, benefits from this level of cooperation could include better performance of the network applications, reduction in usage of intercontinental links, as well as reduction in the cost of inter-NREN traffic exchange.

7. CONCLUSION AND FUTURE WORK

Results from this study show that through dynamic ranking of the local and remote LISP locators, a source network can perform traffic engineering towards a destination without requiring any form of cooperation in the network. It is evident that by leveraging LISP capabilities through integration with SDN, there is potential for improving traffic exchange performance. This however implies granting NRENs more exibility and control of routing and traffic engineering across Internet exchange points, such as to be able to dynamically select routing paths among multiple ingress and egress links. This could have important applications for NRENs that experience high bandwidth costs [19]. The ability to perform multipath routing has potential to over significant performance enhancements and cost savings. For example, delay-sensitive applications such as VoIP, can be mapped onto lower latency intra-continental links, while traffic for bandwidth intensive _le-sharing applications can be routed through higher capacity inter-continental links. Furthermore, by using Openow's ability to customize the packet forwarding rules, and by appropriately matching packets into ows using header tags, it is possible to set up application specific traffic engineering mechanisms. With this capability, a group of NRENs can jointly form traffic engineering strategies specifically for certain applications of common interest, e.g. inter-university video streaming, or access to e-library sites within the domain.

This paper has motivated for and described a performance based traffic engineering mechanism that involves path measurement, gateway ranking, and SDN-based edge-to-edge path configuration. While Internet infrastructure being implemented by the UbuntuNet Alliance through the Africa Connect project has great potential to improve traffic exchange among African NRENs, the inability of traditional protocols to fully take advantage of the available path diversity remains a challenge. For this reason, African NRENs should, apart from implementing the physical interconnectivity, also consider appropriate traffic engineering mechanisms to allow individual NRENs to discover and use optimal inter-NREN paths. However, a globally optimal solution requires coordination and collaboration among several domains that form part of the end-to-end path. Future work will investigate mechanisms that would enable African NRENs to perform collaborative performance based and application specific traffic engineering.

REFERENCES

[1] J. Chavula, N. Feamster, A. Bagula, and H. Suleman, \Quantifying the effects of circuitous routes on the latency of intra-africa internet traffic: A study of research and education networks," vol. 147, pp. 64{73, 2015.

[2] K. Li, S. Wang, and X. Wang, \Edge router selection and traffic engineering in lisp-capable networks,"Communications and Networks, Journal of, vol. 13, pp. 612{620, Dec 2011.

[3] D. Saucez, L. Iannone, O. Bonaventure, and D. Farinacci, \Designing a deployable internet: The locator/identifer separation protocol," Internet Computing, IEEE, vol. 16, pp. 14{21, Nov 2012.

[4] S. Secci, K. Liu, G. Rao, and B. Jabbari, \Resilient traffic engineering in a transit-edge separated internet routing," in Communications (ICC), 2011 IEEE International Conference on, pp. 1{6, June 2011.

[5] D. Saucez, B. Donnet, and O. Bonaventure, \Idips: Isp-driven informed path selection," 2008.

[6] C. E. Rothenberg, M. R. Nascimento, M. R. Salvador, C. N. A. Correa, S. Cunha de Lucena, and R. Raszuk, \Revisiting routing control platforms with the eyes and muscles of software-defined networking," pp. 13{18, ACM, 2012.

[7] A. Gupta, M. Shahbaz, L. Vanbever, H. Kim, R. Clark, N. Feamster, J. Rexford, and S. Shenker, \Sdx: A software defined internet exchange," 2013.

[8] S. Secci, K. Liu, and B. Jabbari, \Efficient inter-domain traffic engineering with transit-edge hierarchical routing," Computer Networks, vol. 57, no. 4, pp. 976{989, 2013.

[9] B. Heller, N. Handigol, V. Jeyakumar, B. Lantz, and N. McKeown, \Reproducible network experiments using container based emulation," in Proc. ACM CoNEXT, Dec. 2012.

[10] L. Iannone, D. Saucez, and O. Bonaventure, \Openlisp: an open source implementation of the locator/id separation protocol," IEEE INFOCOM, Demo paper, 2009.

[11] T. Benson, A. Akella, and D. A. Maltz, \Network traffic characteristics of data centers in the wild," in Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, pp. 267{280, ACM, 2010.

[12] M. Zhang, M. Dusi, W. John, and C. Chen, \Analysis of udp traffic usage on internet backbone links," in Applications and the Internet, 2009. SAINT'09. Ninth Annual International Symposium on, pp. 280{281, IEEE, 2009.

[13] L. Quan and J. Heidemann, \On the characteristics and reasons of long-lived internet
ows," in Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, pp. 444{450, ACM, 2010.

[14] A. Botta, A. Dainotti, and A. Pescap, \A tool for the generation of realistic network workload for emerging networking scenarios," Computer Networks, vol. 56, no. 15, pp. 3531 { 3547, 2012.

[15] S. S. Kolahi, S. Narayan, D. D. T. Nguyen, and Y. Sunarto, \Performance Monitoring of Various Network Traffic Generators," in International Conference on Computer Modeling and Simulation, 2011.

[16] M. Jemec, \packeth{ethernet packet generator," 2012.

[17] A. Tirumala, F. Qin, J. Dugan, J. Ferguson, and K. Gibbs, \Iperf: The tcp/udp bandwidth measurement tool," htt p://dast. nlanr. net/Projects, 2005.

[18] S. Avallone, S. Guadagno, D. Emma, A. Pescape, and G. Ventre, \D-itg distributed internet traffic generator," in Quantitative Evaluation of Systems, 2004. QEST 2004. Proceedings. First International Conference on the, pp. 316{317, IEEE, 2004.

[19] B. Barry, C. Barton, V. Chukwuma, L. Cottrell, U. Kalim, M. Petitdidier, and B. Rabiu, \egy-africa: better internet connectivity to reduce the digital divide," in IST-Africa, 2010, pp. 1{15, IEEE, 2010.

AUTHORS

Josiah Chavula is a PhD candidate at the Hasso-Plattner-Institute in Computer Science, at the University of Cape Town. His research focuses on Internet performance and traffic engineering in Africa's National Research and Education Networks (NRENs). He is interested in approaches for achieving flexible collaborative Internet traffic engineering using Software Defined Networking (SDN). He obtained a Master of Science degree (Networking and Internet Systems) from Lancaster University (UK), and a BSc (Computer Science) from University of Malawi.

Melissa Densmore is a Senior Lecturer in the Department of Computer Science at University of Cape Town. Melissa completed her PhD at University of California, Berkeley in Information Management and Systems, a 3 year ethnographic study of the use of Internet and mobile technologies by health practitioners and NGO staff in a health financing program in Uganda, has an MSc in Data Communications, Networks and Distributed Systems from University College London, and holds a BA in Computer Science from Cornell University.

Hussein Suleman is an Associate Professor in Computer Science at the University of Cape Town. His research is situated within the Centre for ICT for Development (ICT4D) and the Digital Libraries Laboratory. Hussein's main research interests are in digital libraries, ICT4D, information retrieval, cultural heritage preservation, Internet technology, high performance computing and computer science education. He completed his undergraduate degrees and MSc at the then University of Durban-Westville (now University of Kwazulu-Natal) and finished a PhD at Virginia Tech in 2002, in the area of component-based digital libraries.

LIMITING SELF-PROPAGATING MALWARE BASED ON CONNECTION FAILURE BEHAVIOR

Yian Zhou[1], You Zhou[1], Shigang Chen[1] and O. Patrick Kreidl[2]

[1]Department of Computer & Information Science & Engineering,
University of Florida, Gainesville, FL, USA 32611
yian,youzhou,sgchen@cise.ufl.edu
[2]Department of Electrical Engineering, University of North Florida,
Jacksonville, FL, USA 32224
patrick.kreidl@unf.edu

ABSTRACT

Self-propagating malware (e.g., an Internet worm) exploits security loopholes in software to infect servers and then use them to scan the Internet for more vulnerable servers. While the mechanisms of worm infection and their propagation models are well understood, defense against worms remains an open problem. One branch of defense research investigates the behavioral difference between worm-infected hosts and normal hosts to set them apart. One particular observation is that a worm-infected host, which scans the Internet with randomly selected addresses, has a much higher connection-failure rate than a normal host. Rate-limit algorithms have been proposed to control the spread of worms by traffic shaping based on connection failure rate. However, these rate-limit algorithms can work properly only if it is possible to measure failure rates of individual hosts efficiently and accurately. This paper points out a serious problem in the prior method and proposes a new solution based on a highly efficient double-bitmap data structure, which places only a small memory footprint on the routers, while providing good measurement of connection failure rates whose accuracy can be tuned by system parameters.

KEYWORDS

Self-propagating Malware, Connection Failure Behavior, Rate Limitation, Shared Bitmap

1. INTRODUCTION

Self-propagating malware (e.g., an Internet worm) exploits security loopholes in server software. It infects vulnerable servers and then uses them to scan the Internet for more vulnerable servers [1 - 3]. In the past two decades, we have witnessed a continuous stream of new worms raging across the Internet [4 - 7], sometimes infecting tens of thousands or even millions of computers and causing widespread service disruption or network congestion. The mechanisms of worm propagation have been well understood [8 - 11], and various propagation models were developed [12 - 15] to demonstrate analytically the properties of how worms spread among hosts across

David C. Wyld et al. (Eds) : NETCOM, NCS, WiMoNe, CSEIT, SPM - 2015
pp. 99–114, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51609

networks. Significant efforts have also been made to mitigate worms, with varying degrees of success and limitations. Worms remain a serious threat to the Internet.

Patching defects in software is the most common defense measure, not only to worms but also to other types of malware. However, it is a race for who (good guys or bad guys) will find the security defects first. Software is vulnerable and its hosts are subject to infection before the security problems are identified and patched. Moreover, not all users will patch their systems timely, leaving a window of vulnerability to the adversary that will try to exploit every opportunity. Moore et al. investigated worm containment technologies such as address blacklisting and content filtering, and such systems must interdict nearly all Internet paths in order to be successful [13]. Williamson proposed to modify the network stack to bound the rate of connection requests made to distinct destinations [16]. To be effective, it requires a majority of all Internet hosts are upgraded to the new network stack, which is difficult to realize. Similar Internet-wide upgrades are assumed by other host-based solutions in the literature, each employing intrusion detection and automatic control techniques whose supporting models must be calibrated for the specific machine that they will reside upon [17 - 20].

Avoiding the requirement of coordinated effort across the whole Internet, the distributed anti-worm architecture (DAW) [21] was designed for deployment on the edge routers of an Internet service provider (ISP) under a single administrative control. DAW observes a behavioral difference between worm-infected hosts and normal hosts: as an infected host scans random addresses for vulnerable hosts, it makes connection attempts but most will fail, whereas normal users's connection attempts to their familiar servers are mostly successful. By observing the failed connections made by the hosts, the edge routers are able to separate out hosts with large failure rates and contain the propagation of the worms. With a basic rate-limit algorithm, a temporal rate-limit algorithm and a spatial rate-limit algorithm, DAW offers the flexibility of tightly restricting the worm's scanning activity, while allowing the normal hosts to make successful connections at any rate.

However, for rate limit to work properly, we must be able to measure the connection failure rates of individual hosts accurately and efficiently. This paper points out that using Internet Control Message Protocol (ICMP) messages for this purpose [21] is flawed as they are widely blocked on today's Internet, and the total number of message packets in this big data [22] [23] [30] [31] and cloud computing [36 - 39] era is enormous. This paper designs a new measurement method that solves the problem with a highly efficient data structure based on bitmaps, which keeps record of connection attempts and results (success or fail) in bits, from which we can recover the connection failure rates, while removing the duplicate connection failures (which may cause bias against normal hosts). Our double-bitmap solution is highly efficient for online per-packet operations, and the simulation results show that not only does the data structure place a small memory footprint on the routers, but also it provides good measurement of connection failure rates whose accuracy can be tuned by system parameters.

The rest of the paper is organized as follows: Section 2 gives the propagation model of random-scanning worms and reviews the rate-limit algorithms based on connection failure rates. Section 3 explains the problem causing inaccurate failure rate measurement and provides a novel solution with double bitmaps. Section 4 presents simulation results. Section 5 draws the conclusion.

2. BACKGROUND

2.1 Propagation of Random-scanning Worms

This paper considers a type of common worms that replicates through random scanning of the Internet for vulnerable hosts. Their propagation can be roughly characterized by the classical simple epidemic model [26 - 28]:

$$\frac{di(t)}{d(t)} = \beta i(t)(1-i(t)), \tag{1}$$

where $i(t)$ is the percentage of vulnerable hosts that are infected with respect to time t, and β is the rate at which a worm-infected host detects other vulnerable hosts. More specifically, it has been derived [27] that the derivative formula of worm propagation is

$$\frac{di(t)}{dt} = r\frac{V}{N}i(t)(1-i(t)), \tag{2}$$

where r is the rate at which an infected host scans the address space, N is the size of the address space, and V is the total number of vulnerable hosts.

Solving the equation, the percentage of vulnerable hosts that are infected over time is

$$i(t) = \frac{e^{r\frac{V}{N}(t-T)}}{1+e^{r\frac{V}{N}(t-T)}}.$$

Let v be the number of initially infected hosts at time 0. Because $i(0) = v / V$, $T = -\frac{N}{r \cdot V}\ln\frac{v}{V-v}$. Solving this logistic growth equation for t, we know the time it takes for a percentage $\alpha(\geq v / V)$ of all vulnerable hosts to be infected is

$$t(\alpha) = \frac{N}{r \cdot V}(\ln\frac{\alpha}{1-\alpha} - \ln(\frac{v}{V-v})). \tag{3}$$

It is clear that $t(\alpha)$ is inversely proportional to the scanning rate r, which is the number of random addresses that an infected host attempts to contact (for finding and then infecting vulnerable hosts) in a certain measurement period. If we can limit the rate of worm scanning, we can slow down their propagation, buying time for system administrators across the Internet to take actions.

2.2. Behavior-based Rate-Limit Algorithms

In order to perform rate-limit, we need to identify hosts that are likely to be worm-infected. One way to do so is observing different behaviors exhibited from infected hosts and normal hosts. One important behavioral observation was made by [21], which argues that infected hosts have much larger failure rates in their initiated Transfer Contorl Protocol (TCP) connections than normal hosts. We can then apply rate limits to hosts with connection failure rates beyond a threshold and thus restrict the speed at which worms are spread to other vulnerable hosts. (Same as our work, the paper [21] studies worms that spread via TCP, which accounts for the majority of Internet traffic.) Below we briefly describe the host behavior difference in connection failure rate, which is defined as the number of failed TCP connection attempts made by a source host during a certain measurement period, where each attempt corresponds to a SYN packet and each SYN-ACK signals a successful attempt, while the absence of a SYN-ACK means a failure.

- Suppose a worm is designed to attack a software vulnerability in a certain version of web servers from a certain vendor. Consider an arbitrary infected host. Let N be the total number of possible IP addresses and N' be the number of addresses held by web servers, which listen to port 80. $N' = N$ because web servers only account for a small fraction of the accessible Internet. As the infected host picks a random IP address and sends a SYN packet to initiate a TCP connection to port 80 of that address, the connection only has a chance of N' / N to be successful. It has a chance of $1 - N' / N \approx 1$ to fail. The experiment in [21] shows that only 0.4% of all connections made to random addresses at TCP port 80 are successful. Together with a high scanning rate, the connection failure rate of an infect host will be high. Moreover, the measured connection failure rate is an approximation of the host's scanning rate.

- The connection failure rate of a normal host is generally low because a typical user accesses pre-configured servers (such as mail server and DNS server) that are known to be up for most of the time. An exception is web browsing, where the domain names of web servers are used, which again lead to successful connections for most of the time according to our experiences. Cases when the domain names are mistyped, it result in DNS lookup failure and no connection attempts will be made --- consequently no connection failure will occur.

By measuring the connection failure rates of individual hosts, the paper [21] proposes to limit the rate at which connection attempts are made by any host whose failure rate exceeds a certain threshold. By limiting the rate of connection attempts, it reduces the host's connection failure rate back under the threshold. An array of rate-limit algorithms were proposed. The basic algorithm rate-limits individual hosts with excessive failure rates. The temporal rate-limit algorithm can tolerate temporary high failure rates of normal hosts but make sure the long-term average failure rates are kept low. The spatial rate-limit algorithm can tolerate some hosts' high failure rates but make sure that the average failure rates in a network are kept low.

An important component that complements the rate-limit algorithms is the measurement of connection failure rates of individual hosts. This component is however not adequately addressed by [21]. As we will point out in the next section, its simple method does not provide accurate measurement on today's Internet. We will provide a new method that can efficiently solve this important problem with a novel data structure of double bitmaps.

3. A DOUBLE-BITMAP SOLUTION FOR LIMITING WORM PROPAGATION

In this section, we explain the problem that causes inaccurate measurement of connection failure rates and provide a new measurement solution that can work with existing rate-limit algorithms to limit worm propagation.

3.1. Failure Replies and the Problem of Blocked ICMP Messages

We first review the method of measuring the connection failure rates in [21]. After a source host sends a SYN packet to a destination host, the connection request fails if the destination host does not exist or does not listen on the port that the SYN is sent to. In the former case, an ICMP host-unreachable packet is returned to the source host; in the latter case, a TCP RESET packet is returned. The ICMP host-unreachable or TCP RESET packet is defined as a *failure reply*. The connection failure rate of a host s is measured as the rate of failure replies that are sent to s. The rationale behind this method [21] is that the rate of failure replies sent back to the source host should be close to the rate of failed connections initiated by the host. The underlying assumption is that, for each failed connection, a failure reply (either an ICMP host-unreachable packet or a TCP RESET packet) is for sure to be sent back to the source host.

However, this assumption may not be realistic. Today, many firewalls and domain gateways are configured to suppress failure replies. In particular, many organizations block outbound ICMP host-unreachable packets because attacks routinely use ICMP as a reconnaissance tool. When the ICMP host-unreachable packets are blocked, the rate of failure replies sent back to a source host will be essentially much lower than the rate of failed connections that the host has initiated. In other words, a potential worm-affected host may initiate many failed connections, but only a handful of failure replies will be sent back to it. Under these circumstances, the connection failure rate measured by failure replies will be far lower than the actual failure rate, which in turn misleads the rate-limit algorithms and makes them less effective.

To make the problem more complicated, when we measure the connection failure rates of individual hosts, all failed connections made from the same source host to the same destination host in each measurement period should be treated as duplicates and thus counted only once. We use an example to illustrate the reason: Suppose the mail server of a host is down and the email reader is configured to automatically attempt to connect to the server after each timeout period (e.g., one minute). In this case, a normal host will generate a lot of failed connections to the same destination, pushing its connection failure rate much higher than the usual value (when the server is not down) and falsely triggering the rate-limit algorithms to restrict the host's access to the Internet. Therefore, when we measure the connection failure rate of a source host, we want to remove the duplicates to the same destination and measure the rate of failed connections to distinct destinations.

3.2 SYN/SYN-ACK Solution and Problems of Duplicate Failures and Memory Consumption

We cannot use failure replies to measure the connection failure rates. Another simple solution is to use SYN and SYN-ACK packets. Each TCP connection begins with a SYN packet from the source host. If a SYN-ACK packet is received, we count the connection as a successful one; otherwise, we count it as a failed connection. (Technically speaking, a third packet of ACK from the source to the destination completes the establishing of the connection. For our anti-worm purpose, however, the returned SYN-ACK already shows that the destination host is reachable and listens to the port, which thus does not signal worm behavior --- random scanning likely hits unreachable hosts or hosts not listening to the port.)

Using SYN and SYN-ACK packets, a naive solution is for each edge router to maintain two counters, k_s and k_r, for each encountered source address, where k_s is the rate of SYN packets sent by the source (i.e., the number of SYN packets sent during a measurement period), and k_r is the rate of SYN-ACK packets received by the source (i.e., the number of SYN-ACKs received during a measurement period). The connection failure rate k is simply $k_s - k_r$.

This simple solution is memory efficient, as it only requires 64 bits per source host for failure rate measurement, assuming each counter takes 32 bits. However, this solution cannot address the problem of duplicate failures. As discussed in Section 3.1, when we measure the connection failure rate of a source host, we want to remove the duplicates to the same destination in the same measurement period, because measuring duplicate failures may cause bias against normal hosts. Maintaining two counters alone cannot achieve the goal of removing duplicate failures.

An alternative solution is to have the edge router store a list of distinct destination addresses for each source host. However, such per-source information consumes a large amount of memory. Suppose each address costs 32 bits. The memory required to store each source host's address list will grow linearly with the rate of distinct destination hosts that the source host initiates connection requests to. For example, the main gateway at our campus observes an average of more than 10 million distinct source-destination pairs per day. If the edge router keeps per-source address list, it will cost more than 320 megabits of memory, which soon exhausts the small on-die SRAM memory space of the edge router. Therefore, this solution is not feasible either.

The major goal of this paper is to accurately measure the connection failure rates with a small memory. However, tradeoffs must be made between measurement accuracy and memory consumption under the requirement of duplicate failure removal. Existing research uncovered the advantages of using Bloom filters [28] [29] or bitmaps [24] [25] [32 - 35] [40] to compress the connection information in limited memory space and automatically filter duplicates, which can be adopted to measure the connection failure rates. For example, the edge router can maintain two bitmaps for each source host, and map each SYN/SYN-ACK packet of the host into a bit in the host's corresponding bitmap, from which the rate of SYN/SYN-ACK packets of each host can be recovered. However, the measurement accuracy depends on setting the bitmap size for each source host properly in advance. In practice, it is difficult to pre-determine the values as different source hosts may initiate connection requests at unpredictable and different rates, which limits the practicability of this solution as well.

3.3. Double Bitmaps

In order to address the problems of duplicate failures and memory consumption, instead of using per-source address lists or bitmaps, we incorporate two shared bitmaps to store the SYN/SYN-ACK information of all source hosts. Our double-bitmap solution includes two phases: in the first phase, the edge router keeps recoding the SYN/SYN-ACK packets of all source hosts through setting bits in the bitmaps; in the second phase, the network management center will recover the connection failure rates from the two bitmaps based on maximum likelihood estimation (MLE), and notify the edge router to apply rate limit algorithms to limit the connection attempts made by any host whose failure rate exceeds some threshold. Below we will explain the two phases, and then mathematically derive an estimator to calculate the connection failure rate.

3.3.1. Phase I: SYN / SYN-ACK Encoding

In our solution, each edge router maintains two bitmaps B_s and B_r, which encode the distinct SYN packets and SYN-ACK packets of all source hosts within its network, respectively. Let m_s and m_r be the number of bits in B_s and B_r correspondingly. Below we will explain how an edge router encodes the distinct SYN packet information into B_s, which can later be used to estimate the SYN sending rate k_s for each source host. The way for the edge router to encode the distinct SYN-ACK packet information into B_r is quite similar, which we omit.

For each source host src, the edge router randomly selects l_s $(= m_s)$ bits from the bitmap B_s to form a logical bitmap src, which is denoted as $LB(src)$. The indices of the selected bits are $H(src \oplus R[0])$, $H(src \oplus R[1])$, $\cdots$, $H(src \oplus R[l_s - 1])$, where $\oplus$ is bitwise XOR, $H(\cdots)$ is a hash function whose range is $[0, m_s)$, and R is an integer array storing randomly chosen constants to arbitrarily alter the hash result. Similarly, the logical bitmap can be constructed from B_s for any other hosts. Essentially, we embed the bitmaps of all possible hosts in B_s. The bit-sharing relationship is dynamically determined on the fly as each new host src' will be allocated a logical bitmap $LB(src')$ from B_s to store its SYN packet information.

Given above notations and data structures, the online coding works as follows. At the beginning of each measurement period, all bits in B_s are reset to zeros. Suppose a SYN packet signatured with a $\langle src, dst \rangle$ host address pair is routed by the edge router. The router will randomly select a bit from the logical bitmap $LB(src)$ based on src and dst, and set this bit in B_s to be one. The index of the bit to be set for this SYN packet is given as follows:

$H(src \oplus R[H(dst \oplus K) \bmod l_s]).$

The second hash, $H(dst \oplus K)$, ensures that the bit is pseudo-randomly selected from $LB(src)$, and the private key K is introduced to prevent the hash collision attacks. Therefore, the overall effect to store the SYN packet information is :

$$B[H(src \oplus R[H(dst \oplus K) \bmod l_s])] = 1.$$

Similarly, the edge router only needs to set a bit in the bitmap B_r to be one for each SYN-ACK packet using the same mechanism. Note that in our solution, to store a SYN/SYN-ACK packet, the router only performs two hash operations and sets a single bit in its bitmap, which is quite efficient. In addition, duplicates of SYN and SYN-ACK information with same $\langle src, dst \rangle$

signature will mark the same bit in the shared bitmaps such that the duplicate information is filtered as desired.

3.3.2. Phase II: Failure Rate Measurement

At the end of each measurement period, the edge router will send the two bitmaps B_s and B_r to the network management center (NMC), which will estimate connection failure rate k for each source host src based on B_s and B_r, and notify the edge router to apply rate limit algorithms to limit the connection attempts made by any host whose failure rate exceeds some threshold. Since rate-limit algorithms have been fully studied in [21], we will focus on the measurement of connection failure rates based on the bitmaps. The measurement process is described in the following.

First, the NMC extracts the logical bitmaps $LB(src)$ and $LB'(src)$ of each source host src from the two bitmaps B_s and B_r, respectively. Second, the NMC counts the number of zeros in $LB(src)$, $LB'(src)$, B_s and B_r, which are denoted by U_s^l, U_r^l, U_s^m, and U_r^m, respectively. Then the NMC divides them by the corresponding bitmap size l_s, l_r, m_s, and m_r, and calculates the fraction of bits whose values are zeros in $LB(src)$, $LB'(src)$, B_s and B_r correspondingly. That is, $V_s^l = U_s^l / l_s$, $V_r^l = U_r^l / l_r$, $V_s^m = U_s^m / m_s$, and $V_r^m = U_r^m / m_r$. Finally, the NMC uses the following formula to estimate connection failure rate k for source host src:

$$\hat{k} = \frac{\ln V_s^l - \ln V_s^m}{\ln(1-\frac{1}{l_s}) - \ln(1-\frac{1}{m_s})} - \frac{\ln V_r^l - \ln V_r^m}{\ln(1-\frac{1}{l_r}) - \ln(1-\frac{1}{m_r})} \tag{4}$$

3.3.3. Derivation of the MLE estimator

Now we follow the standard MLE method to get the MLE estimators $\hat{k}_s$ and $\hat{k}_r$ of k_s and k_r, respectively, and then derive $\hat{k}$ given by (4). Since the way to derive the MLE estimator for k_s and k_r is quite similar, we will only derive the MLE estimator formula for $\hat{k}_s$, and directly give the result for $\hat{k}_r$. To derive $\hat{k}_s$, we first analyze the probability $q(k_s)$ for an arbitrary bit in $LB(src)$ to be '0', and use $q(k_s)$ to establish the likelihood function L to observer U_s^l '0' bits in $LB(src)$. Finally, maximizing L with respect to k_s will lead to the MLE estimator, $\hat{k}_s$.

Note that k_s is the actual rate of distinct SYN packets sent by a source host src, and n_s is the rate of distinct SYN packets sent by all hosts within the router's network. Consider an arbitrary bit b in $LB(src)$. A SYN packet sent by src has a probability of $1/l_s$ to set b to '1', and a SYN packet sent by any other host has a probability of $1/m_s$ to set b to '1'. Hence, the probability $q(k_s)$ for bit b to remain '0' at the end of the measurement period is

$$q(k_s) = \left(1 - \frac{1}{m_s}\right)^{n_s - k_s} \left(1 - \frac{1}{l_s}\right)^{k_s}. \tag{5}$$

Because the bits in any logical bit array are randomly selected from the bitmap B_s, each of the n_s SYN packets has about the same probability of $1/m_s$ to choose any bit in B_s. So for an arbitrary bit in B_s, the probability for it to be '0' after storing all n_s distinct SYN packets is

$$q(n_s) = \left(1 - \frac{1}{m_s}\right)^{n_s}. \tag{6}$$

In this sense, the number of zero bits in B_s follows a binomial distribution $U_s^m : B(m_s, q(n_s))$ $= B(m_s, (1 - 1/m_s)^{n_s})$. Therefore, the expected value for V_s^m is

$$E(V_s^m) = E\left(\frac{U_s^m}{m_s}\right) = \frac{m_s(1 - \frac{1}{m_s})^{n_s}}{m_s} = q(n_s). \tag{7}$$

Substituting (7) to (5), and replacing $E(V_s^m)$ by its instance value V_s^m, we have the following instance value for $q(k_s)$:

$$q(k_s) = V_s^m \times \left(\frac{1 - 1/l_s}{1 - 1/m_s}\right)^{k_s}. \tag{8}$$

Given the probability for each bit in $LB(src)$ to be '0' as $q(k_s)$, we can establish the likelihood function to observe U_s^l '0' bits in $LB(src)$ as follows:

$$L = q(k_s)^{U_s^l} (1 - q(k_s))^{l_s - U_s^l}. \tag{9}$$

The MLE estimator of k_s is the value of k_s that maximizes the above likelihood function. Namely,

$$\hat{k}_s = \arg\max_{k_s} \{L\}. \tag{10}$$

To find $\hat{k}_s$, we take logarithm on both sides, and then perform the first order derivative to obtain

$$\frac{\partial \ln(L)}{\partial k_s} = \left(\frac{U_s^l}{q(k_s)} - \frac{l_s - U_s^l}{1 - q(k_s)}\right) \times q'(k_s), \tag{11}$$

where $q'(k_s)$ is computed as

$$q'(k_s) = q(k_s) \times \ln\left(\frac{1-1/l_s}{1-1/m_s}\right). \tag{12}$$

Since $m_s > l_s \geq 1$ and $n_s > 0$, $q(k_s)$ and $q'(k_s)$ cannot be 0. Setting the right side of (11) be zero, we have

$$q(k_s) = \frac{U_s^l}{l_s} = V_s^l. \tag{13}$$

Substituting above equation to (8) and solving for k_s, we get the MLE estimator of k_s:

$$\hat{k}_s = \frac{\ln V_s^l - \ln V_s^m}{\ln(1-\frac{1}{l_s}) - \ln(1-\frac{1}{m_s})}. \tag{14}$$

Similarly, we can derive the MLE estimator of k_r:

$$\hat{k}_r = \frac{\ln V_r^l - \ln V_r^m}{\ln(1-\frac{1}{l_r}) - \ln(1-\frac{1}{m_r})}. \tag{15}$$

Since $k = k_s - k_r$, given the MLE estimators $\hat{k}_s$ and $\hat{k}_r$ of k_s and k_r, we can easily derive the estimator of k as

$$\hat{k} = \hat{k}_s - \hat{k}_r. \tag{16}$$

Substituting (14) and (15) to the above equation, we derive the estimator $\hat{k}$ as described in (4). Note that if the two bitmaps B_s and B_r have the same size, and the two logical bitmaps for each source host also have the same size, i.e., $m_s = m_r = m$ and $l_s = l_r = l$, then the estimator for the connection failure rate k will be in a more compact form:

$$\hat{k} = \frac{\ln V_s^l - \ln V_s^m - \ln V_r^l + \ln V_r^m}{\ln(1-\frac{1}{l}) - \ln(1-\frac{1}{m})}. \tag{17}$$

4. SIMULATION

We evaluate the measurement accuracy of our estimator for the connection failure rate through simulations. Recall that the major goal of this paper is to provide a good estimator for measuring the connection failure rates of individual hosts that can work well in a small memory. Hence, in our simulations, we purposely allocate memory with small sizes to encode the information of distinct SYN and SYN-ACK packets for all source hosts, such that the average memory size for each source host will be ranging from 10 bits to 40 bits only. As we explained in Section 3.2, the solution with per-source address lists or bitmaps will not work with this small memory size. Therefore, our solution outperforms in the aspect of greatly reducing the required online memory footprint for connection failure rate measurement while achieving duplicate failure removal.

4.1. Simulation Setup

Our simulations are conducted under the following setups. We simulate 50,000 distinct source hosts as normal hosts, and 100 distinct source hosts as worm-affected hosts. For the normal hosts, they will send distinct SYN packets to different destination hosts, with a rate following an exponential distribution whose mean is 5 distinct SYN packets per minute. For each distinct SYN packet that a normal host sends out, a corresponding SYN-ACK packet will be sent back to the host with a probability, which follows a uniform distribution in the range of [0.8, 1.0]. As for the worm-affected hosts, we simulate their aggressive scanning behavior by having them send distinct SYN packets to different destination hosts with a higher rate, which follows another exponential distribution whose mean is 10 distinct SYN packets per second. Since the worm-affected hosts will randomly scan the whole destination space, their failure rate is expected to be very high as we explained earlier. Therefore, in our simulations, no SYN-ACK packets will be sent back to them. Suppose each measurement period is 1 minute. Then each normal host will send 5 distinct SYN packets and receive 4.5 distinct SYN-ACK packets on average, and each worm-affected host will send 600 distinct SYN packets and 0 SYN-ACK packet on average, during each measurement period.

In our simulations, all the SYN and SYN-ACK packets are processed by a single simulated edge router and a simulated network management center according to our two-phase measurement scheme. First of all, the SYN and SYN-ACK packets are encoded into two *m*-bit bitmaps B_s and B_r of the edge router, respectively, as described in Section 3.3.1 (Phase I: SYN/SYN-ACK Encoding). After all packets are encoded into the two bitmaps B_s and B_r, the edge router will send B_s and B_r to the network management center, which will estimate the connection failure rate of each source host based on B_s and B_r offline, as described in Section 3.3.2 (Phase II: Failure Rate Measurement).

4.2. Simulation Results

We conduct three sets of simulations with three different sizes of memory allocated for the bitmaps B_s and B_r, $m_s = m_r = m = 0.5$Mb, 1Mb, and 2Mb, to observe the measurement accuracy under different memory constraints. The sizes of the logical bitmaps for each host is set to be $l_s = l_r = l = 300$. Figure. 1-3 present the simulation results when the allocated memory *m* equals 2Mb, 1Mb, and 0.5Mb, respectively. Since there are a total of 50,100 source hosts, the average memory consumption per source host will be about 40 bits, 20 bits, and 10 bits, accordingly. In each figure, each point represents a source host, with its x-coordinate showing the actual connection failure rate k (per minute) and y-coordinate showing the estimated connection failure rate $\hat{k}$ (per minute) measured by our scheme. The equality line *y* = *x* is also drawn for reference. Clearly, the closer a point is to the quality line, the better the measurement result.

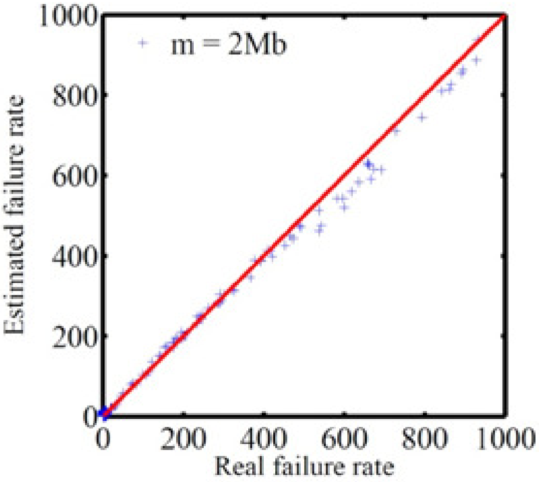


Figure 1. Measurement accuracy of connection failure rate per minute. m = 2Mb, l = 300.

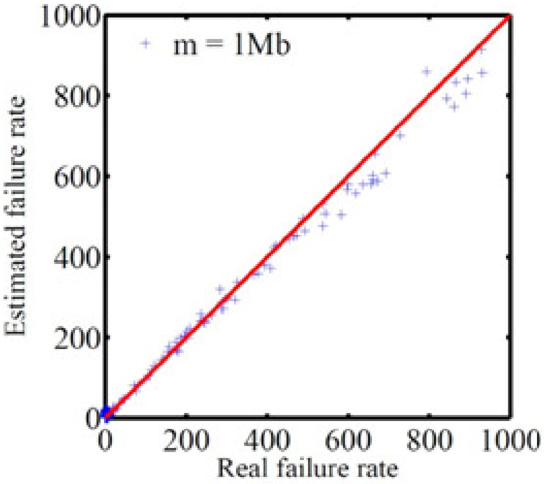


Figure 2. Measurement accuracy of connection failure rate per minute. m = 1Mb, l = 300.

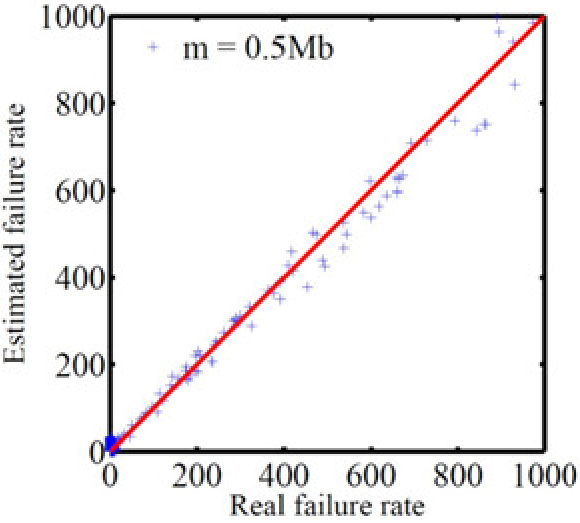


Figure 3. Measurement accuracy of connection failure rate per minute. m = 0.5Mb, l = 300.

From the three figures, one can observe that the measurement result for the connection failure rates of our scheme is quite accurate under all three different memory constraints. For almost every source host, the measured failure rate closely follows its real failure rate as shown in the figures. There is a tendency for the measurement result to be slightly more accurate with a larger memory size (compare Figure. 1 and Figure. 3). However, for our scheme, a small memory of size m = 0.5Mb (equivalent to 10 bits per source host on average) is adequate enough to generate sound measurement results as shown in Figure. 3. Recall that for the solution storing per-source address list, the destination address of every SYN packet must be stored for every source host. So for that solution, a normal source host initiating 5 connection requests (5 distinct SYN packets) per minute will require at least $32 \times 5 = 160$ bits to record its SYN packets, and a worm-affected host sending 10 SYN packets per second will require at least $32 \times 600 = 19200$ bits, for each measurement period of one minute. Clearly, through utilizing double bitmaps, our scheme outperforms the solution storing address lists, because it can work well with a much more strict memory constraint.

5. CONCLUSION

This paper proposes a new method of measuring connection failure rates of individual hosts, using a novel data structure based on double bitmaps. It addresses an important problem in rate-limiting worm propagation, where inaccurate failure rates will affect the performance of rate-limit algorithms. The past method relies on ICMP host-unreachable messages, which are however widely blocked on today's Internet. The new method makes the measurement based on SYN and SYN-ACK packets, which is more reliable and accurate. Its bitmap design helps significantly to reduce the memory footprint on the routers and eliminates the duplicate connection failures (another problem of the previous method).

ACKNOWLEDGEMENTS

This work is supported in part by a grant from Florida Cybersecurity Center.

REFERENCES

[1] M. Barwise, “What is an internet worm?,” http://www.bbc.co.uk/webwise/guides/internet-worms, 2010.

[2] D. M. Kienzle and M. C. Elder, “Recent Worms: A Survey and Trends,” Proc. of ACM Workshop on Rapid Malcode, pp. 1-10, 2003.

[3] J. Rochlis and M. Eichin, “With Microscope and Tweezers: The Worm from MIT's Perspective,” Communication of the ACM, vol. 32, no. 6, pp. 689-698, 1989.

[4] J. Maniscalchi, “Worm Evolution,” http://www.digitalthreat.net/2009/05/worm-evolution/, 2009.

[5] Computer Emergency Response Team, “CERT Advisory CA-2001-26 Nimda Worm,” http://www.cert.org/advisories/CA-2001-26.html, 2001.

[6] Computer Emergency Response Team, “CERT Advisory CA-2003-04 MS-SQL Server Worm,” http://www.cert.org/advisories/CA-2003-04.html, 2003.

[7] Computer Emergency Response Team, “CERT Alert Conficker Worm Targets Microsoft Windows Systems,” https://www.us-cert.gov/ncas/alerts/TA09-088A, 2013.

[8] Z. Chen and C. Ji, “Measuring Network-Aware Worm Spreading Ability,” Proc. of IEEE INFOCOM, pp. 116-124, 2007.

[9] Z. Li, L. Wang, Y. Chen, and Z. Fu, “Network-Based and Attack-Resilient Length Signature Generation for Zero-Day Polymorphic Worms,” Proc. of IEEE International Conference on Network Protocols (ICNP), pp. 164-173, 2007.

[10] P. K. Manna, S. Chen, and S. Ranka, “Inside the Permutation-Scanning Worms: Propagation Modeling and Analysis,” IEEE/ACM Transactions on Networking, vol. 18, no. 3, pp. 858-870, 2010.

[11] S. Stafford and J. Li, "Behavior-Based Worm Detectors Compared," Recent Advances in Intrusion Detection, pp. 38-57, 2010.

[12] M. Liljenstam, Y. Yuan, B. Premore, and D. Nicol, "A Mixed Abstraction Level Simulation Model of Large-Scale Internet Worm Infestations," Proc. of 10th IEEE/ACM Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS), 2002.

[13] D. Moore, C. Shannon, G. M. Voelker, and S. Savage, "Internet Quarantine: Requirements for Containing Self-Propagating Code," Proc. of IEEE INFOCOM, 2003.

[14] S. Staniford, V. Paxson, and N. Weaver, "How to Own the Internet in Your Spare Time," Proc. of 11th USENIX Security Symposium, 2002.

[15] C. C. Zou, W. Gong, and D. Towsley, "Code Red Worm Propagation Modeling and Analysis," Proc. of ACM CCS, 2002.

[16] M. M. Williamson, "Throttling Viruses: Restricting Propagation to Defeat Malicious Mobile Code," Proc. of Annual Computer Security Application Conference, 2002.

[17] R. Dantu, J. W. Cangussu, and S. Patwardhan, "Fast Worm Containment Using Feedback Control," IEEE Transactions on Dependable and Secure Computing, vol. 4, no. 2, pp. 119-136, 2007.

[18] O. P. Kreidl and T. M. Frazier, "Feedback Control Applied to Survivability: A Host-Based Autonomic Defense System," IEEE Transactions on Reliability, vol. 53, no. 1, pp. 148-166, 2004.

[19] X. Yan and Y. Zou, "Optimal Internet Worm Treatment Strategy Based on yhe Two-Factor Model," ETRI journal, vol. 30, no. 1, pp. 81-88, 2008.

[20] S. Zonouz, H. Khurana, W. H. Sanders, T. M. Yardley, et al., "RRE: A Game-Theoretic Intrusion Response and Recovery Engine," IEEE Transactions on Parallel and Distributed Systems, vol. 25, no. 2, pp. 395-406, 2014.

[21] S. Chen and Y. Tang, "DAW: A Distributed Antiworm System," IEEE Transactions on Parallel and Distributed Systems, vol. 18, no. 7, pp. 893-906, 2007.

[22] S. Sharma, U. S. Tim, J. Wong, S. Gadia, and S. Sharma, "A Brief Review on Leading Big Data Models," Data Science Journal, vol. 13, pp. 138-157, 2014.

[23] S. Sharma, R. Shandilya, S. Patnaik, and A. Mahapatra, "Leading NoSQL models for handling Big Data: a brief review," International Journal of Business Information Systems, 2015.

[24] Y. Zhou, Q. Xiao, Z. Mo, S. Chen, and Y. Yin, "Privacy-Preserving Point-to-Point Transportation Traffic Measurement through Bit Array Masking in Intelligent Cyber-physical Road Systems," Proc. of IEEE International Conference on Cyber, Physical and Social Computing (CPSCom), pp. 826-833, 2013.

[25] Y. Zhou, Z. Mo, Q. Xiao, S. Chen, and Y. Yin, "Privacy-Preserving Transportation Traffic Measurement in Intelligent Cyber-Physical Road Systems," IEEE Transactions on Vehicluar Technologies, 2015.

[26] H. W. Hethcote, "The Mathematics of Infectious Diseases," SIAM Review, vol. 42, no. 4, pp. 599-653, 2000.

[27] C. C. Zou, W. Gong, and D. Towsley, “Slowing Down Internet Worms,” Proc. of IEEE International Conference on Distributed Computing Systems (ICDCS), pp. 312-319, 2004.

[28] Y. Qiao, T. Li, and S. Chen, “Fast Bloom Filters and Their Generalization,” IEEE Transactions on Parallel and Distributed Systems, vol. 25, no. 1, pp. 93-103, 2014.

[29] O. Rottenstreich, Y. Kanizo, and I. Keslassy, “The Variable-Increment Counting Bloom Filter,” IEEE/ACM Transactions on Networking, vol. 22, no. 4, pp. 1092-1105, 2014.

[30] Y. Zhou, S. Chen, Z. Mo, and Q. Xiao, “Point-to-Point Traffic Volume Measurement through Variable-Length Bit Array Masking in Vehicular Cyber-Physical Systems,” Proc. of IEEE International Conference on Distributed Computing Systems (ICDCS), pp. 51-60, 2015.

[31] Y. Zhou, S. Chen, Y. Zhou, M. Chen, and Q. Xiao, “Privacy-Preserving Multi-Point Traffic Volume Measurement through Vehicle to Infrastructure Communications,” IEEE Transactions on Vehicular Technologies, 2015.

[32] S. Sharma, U. S. Tim, S. Gadia, R. Shandilya, and S. Peddoju, “Classification and Comparison of NoSQL Big Data Models,” International Journal of Big Data Intelligence (IJBDI), vol. 2, no. 3, 2015.

[33] S. Sharma, “Evolution of as-a-Service Era in Cloud,” arXiv preprint arXiv:1507.00939, 2015.

[34] T. Li and S. Chen, “Traffic Measurement on the Internet,” Springer Science & Business Media, 2012.

[35] T. Li, S. Chen, and Y. Ling, “Per-Flow Traffic Measurement Through Randomized Counter Sharing,” IEEE/ACM Transactions on Networking, vol. 20, no. 5, pp. 1622-1634, 2012.

[36] Z. Mo, Y. Zhou, S. Chen, and C. Xu, “Enabling Non-repudiable Data Possession Verification in Cloud Storage Systems,” Proc. of IEEE International Conference on Cloud Computing, pp. 232-239, 2014.

[37] Z. Mo, Q. Xiao, Y. Zhou,and S. Chen, “On Deletion of Outsourced Data in Cloud Computing,” IEEE International Conference on Cloud Computing, pp. 344-351, 2014.

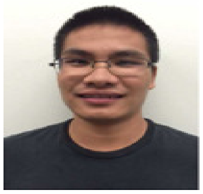
[38] Z. Mo, Y. Zhou, and S. Chen, “A Dynamic Proof of Retrievability (PoR) Scheme with O(logn) Complexity,” Proc. of IEEE International Conference on Communications, pp. 912-916, 2012.

[39] Z. Mo, Y. Zhou, and S. Chen, “An Efficient Dynamic Proof of Retrievability Scheme,” ZTE Communications, vol. 2, p. 008, 2013.

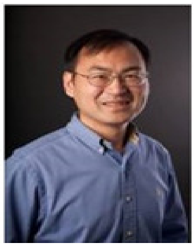
[40] Q. Xiao, M. Chen, S. Chen, and Y. Zhou, “Temporally or Spatially Dispersed Joint RFID Estimation Using Snapshots of Variable Lengths,” Proc. of ACM Mobihoc, 2015.

AUTHORS

Yian Zhou received her B.S. degree in computer science and B.S. degree in economics from the Peking University of China in 2010, and is currently pursuing her Ph.D. degree in computer and information science and engineering at the University of Florida, Gainesville, FL, USA. Her advisor is Prof. Shigang Chen. Her research interests include traffic flow measurement, cyber-physical systems, big network data, security and privacy, and cloud computing.



You Zhou received his B.S. degree in electronic information engineering from the University of Science and Technology of China, Hefei, China, in 2013, and is currently pursuing his Ph.D. degree in computer and information science and engineering at the University of Florida, Gainesville, FL, USA. His advisor is Prof. Shigang Chen. His research interests include network security and privacy, big network data, and Internet of Things.



Shigang Chen is a professor with Department of Computer and Information Science and Engineering at University of Florida. He received his B.S. degree in computer science from University of Science and Technology of China in 1993. He received M.S. and Ph.D. degrees in computer science from University of Illinois at Urbana-Champaign in 1996 and 1999, respectively. After graduation, he had worked with Cisco Systems for three years before joining University of Florida in 2002. He served on the technical advisory board for Protego Networks in 2002-2003. His research interests include computer networks, Internet security, wireless communications, and distributed computing. He published more than 100 peer-reviewed journal/conference papers. He received IEEE Communications Society Best Tutorial Paper Award in 1999 and NSF CAREER Award in 2007. He holds 11 US patents. He is an associate editor for IEEE/ACM Transactions on Networking, Elsevier Journal of Computer Networks, and IEEE Transactions on Vehicular Technology. He served in the steering committee of IEEE IWQoS from 2010 to 2013. He is a senior member of IEEE.

O. Patrick Kreidl has been an Assistant Professor of Electrical Engineering at the University of North Florida (UNF) since 2011, receiving his S.B. degree (with highest distinction) from George Mason University (GMU) in 1994 and his S.M. and Ph.D. degrees from the Massachusetts Institute of Technology (MIT) in 1996 and 2008, respectively. Past positions include Principal Research Engineer in the Cyber Operations and Networking Group within BAE Systems' Technology Solutions Directorate (via acquisition of Alphatech, Inc.), Research Affiliate in MIT's Laboratory for Information and Decision Systems, Adjunct Professor in GMU’s Department of Electrical & Computer Engineering as well as engineering positions in the Institute for Defense Analyses and the Naval Research Laboratory. His current research interests lie at the intersections of signal processing, stochastic control and optimization (particularly as they interface with algorithms, computation and statistics) with application to sensor networks, network security and distributed systems. He is a member of the IEEE.

SELECTIVE OPENING SECURE FUNCTIONAL ENCRYPTION

Yuanyuan Ji[1], Haixia Xu[2] and Peili Li[1]

[1]Chinese Academy of Sciences, Beijing, China
[2]State Key Laboratory of Information Security,
Institute of Information Engineering, CAS, Beijing, China
jiyuanyuan@iie.ac.cn, xuhaixia@iie.ac.cn, lipeili@iie.ac.cn

ABSTRACT

Functional encryption (FE) has more fine-grained control to encrypted data than traditional encryption schemes. The well-accepted security of FE is indistinguishability-based security (IND-FE) and simulation-based security (SIMFE), but the security is not sufficient. For example, if an adversary has the ability to access a vector of ciphertexts and can ask to open some information of the messages, such as coins used in the encryption or secret key in multi-key setting, whether the privacy of the unopened messages is guaranteed. This is called selective opening attack (SOA).

In this paper, we propose a stronger security of FE which is secure against SOA (we call SO-FE) and propose a concrete construction of SO-FE scheme in the standard model. Our scheme is a non-adaptive IND-FE which satisfies selective opening secure in the simulation sense. In addition, the scheme can encrypt messages of any bit length other than bitwise and it is secure against SOA-C and SOAK simultaneously while the two attacks were appeared in different model before. According to the different functionality f, our scheme can specialize as IBE, ABE and even PE schemes secure against SOA.

KEYWORDS

Functional encryption, Selective opening attack, Indistinguishability obfuscation, Deniable encryption

1. INTRODUCTION

Traditional encryption schemes provide rather coarse-grained access to encrypted data, because the receiver can get the message in its entirety if he possesses the right key or he can learn nothing without the secret key. Thus a new encryption scheme — functional encryption (FE), with much more fine-grained control, has been extensively studied. FE was introduced by Boneh, Sahai and Waters [13]. A FE scheme means one who owns SKf can decrypt the cipher of m to get the value of f(m). It requires that the user learns nothing other than f(m). There are two well-accepted security notions for FE: indistinguishable based security definition (IND-FE) and simulation based definition (SIM-FE) [13]. But the security can't

David C. Wyld et al. (Eds) : NETCOM, NCS, WiMoNe, CSEIT, SPM - 2015
pp. 115–130, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51610

satisfy people's needs because of the different modes of attack, here we consider selective opening attack.

Selective opening security had been first investigated to the traditional public key encryption field by Bellare, Hofheinz and Yilek [10] in 2009. In the public key encryption system, there are two kinds of selective opening attack (SOA). One is coin-revealing SOA (SOA-C), that is to say, if an adversary obtains a number of ciphertexts and then corrupts a subset of the senders, obtaining not only the corresponding messages but also the coins under which they were encrypted, then the unopened messages still remain privacy. The other is key-revealing SOA (SOA-K), which means if an adversary obtains a number of ciphertexts encrypted under different public keys, then the senders are asked to reveal a subset of the corresponding decryption keys, in this case it remains secure for the rest of the messages. Creating an encryption scheme secure against SOA has important practical meaning. Under the complex environment of cloud computing, distributed shares in a distributed file-system are allotted to different servers to perform a task, if a subset of the distributed servers are corrupted by an adversary who may get the encrypted messages as well as the randomness, then can messages under the other uncorrupted severs remain secure?

Achieving security against SOA is challenging but even so there has been some works to achieve the security goal ([5], [6], [8], [4], [9], [7]). There are two flavors of definitions to capture security under selective opening attacks: simulation-based selective opening security (SIM-SO) and indistinguishability-based selective opening security (IND-SO) [5]. Because IND-SO security notion requires that the joint plaintext distribution should be conditionally effective re-sampled, which restricts SOA security to limited setting, so we just concern SIM-SO security. SO secure PKE scheme had been investigated by Bellare et al. [5] in 2009. Bellare showed that any lossy encryption is able to achieve SO security. Later on, several other SOA secure PKE schemes had been constructed ([6],[9],[8]). In 2011, with the development of IBE, Bellare, Waters and Yilek [11] introduced SOA to IBE. In IBE, ciphertexts and secret keys SKID are generated according to the corresponding target identity ID, only the right SKID can open the ciphertexts and an adversary can make many key queries using the ID (different from the challenge ID) as input. Later, Junzuo Lai et al. [12] proposed a concrete CCA2 secure SO-IBE scheme. However, almost known SO-IBE schemes utilize the technology of one-side public openability which means these schemes have to encrypt bit by bit which is comparatively inefficient, and it is challenging to construct a SOA secure IBE scheme which is not bitwise.

FE schemes seems to be different from PKE or IBE, but it aims to keep the encrypted message secret even though the adversary can get some special information SKf. But if the adversary has more ability to open a part of the message and get the randomness used in the encryption, can the security of the unopened messages be kept?

[13] and [15] proved that the simulation secure FE can not be achieved in the standard model. So in this paper, we focus on the construction of IND-FE and simulation-based secure against SOA

1.1 Related Works

With the development of indistinguishability obfuscation (io), many difficult cryptography tasks can be achieved. In 2013, [16] proposed a concrete construction of functional encryption for all circuits. In their scheme, the SKf is generated by using indistinguishability obfuscation, at the same time, it uses double encryption of the same message as the ciphertext and statistical simulation soundness NIZK (SSS-NIZK) to get well-formed ciphertexts. With the help of io, their scheme can hide important process (decryption and compution) in the SKf. In 2014, Sahai and Waters [3] introduced a new technique: puncture programs. They proposed an effective method to transform the private key encryption to the public key encryption and they designed a deniable encryption scheme which had opened for 16 years [2]. In deniable encryption, if a sender is forced to reveal to an adversary both his message and the randomness under encryption, he should be able to provide a fake randomness and a fake message that will make the adversary believe the ciphertext is encryption of the fake message.

1.2 Our Contributions

The contribution of this work consists of the following two steps. We first propose a new security model of functional encryption secure against selective opening attacks (including coins and private keys), which we call SO-FE, and then propose a concrete construction of SO-FE scheme for general function without random oracle. In view of the impossiblility result of the SIM-FE in the standard model and the limitation of the IND-SO, the security of our scheme is indistinguishable based secure FE and simulation based secure against SOA.

In our scheme, we combine the coin-revealing selective opening security and key-revealing selective opening security owing to the special property of KeyGen process of FE. Before, SOA-C and SOA-K are mentioned in different scenes, specially, SOA-K is only used in the multi-key encryption, the feature of FE can make sure the key query even though ciphertexts are encrypted under the same public key.

The SO-FE scheme can be applied to the special situation, such as SO-IBE scheme, SO-ABE scheme, SO-PE scheme. Thus using io, we can get many encryption schemes secure against selective opening attacks. So far there are only SO-IBE schemes (ABE or PE scheme secure against SOA haven't be proposed). Moreover, all known SO-IBE schemes are bitwise, while our scheme can encrypt the message with any bit.

1.3 Our Technique

There are two difficult challenges in achieving this goal. The first is the corrupt query of coins in SOA-C process: when the adversary chooses a set I and asks to open the corresponding messages and randomness, how can the simulator provide the eligible randomness which is indistinguishable from the real one. The second is key queries in SOA-K process — a feature of FE security formalizations since [13], that allows the adversary to obtain the decryption key of any reasonable functionality f of his choice, but how to define reasonablity in SOA-based security model.

To solve the first problem, we adopt deniable encryption (DE, refer to section 2.2) which can output a fake random r_0 (satisfies DE_{Enc} (pk_{DE}, m_0, r_0) = C). The special property of DE can make sure the simulator generates a fake randomness to cheat the adversary that the opened coins match the opened ciphers and the opened messages.

To solve the second problem, we impose restrictions on the adversary's choice of functions that can be queried to the key generation. Here we define reasonable function.

Intuition. We start by giving an overview of the main ideas behind our SOA-based security definition. To convey the core ideas, it suffices to consider the simple case of X = m_1,m_2,f(m_1,m_2), ($m_i \in \{0,1\}$). Suppose that the adversary queries secret keys for function f. Now, recall that the IND-security definition guarantees that an adversary cannot differentiate between encryption of x_0 and x_1 as long as f(x_0) = f(x_1) for every f. It is the only restriction of IND-security definition, in SOA security model, the above restricting of f is not enough since an adversary can learn part information of message by making corrupt query of I. For example, an adversary can make I = {1} query and know m_1, by using key query to f, it can learn f(m_1,m_2). In particular, if f(m_1,0) ≠ f(m_1,1), it is easy to guess the unopened message m_2. Obviously, it makes no sense in SOA-based security definition. So we make the limitation of f: if the input of f contains the element of set m[I], which is opened in the corrupt query phase, thus except those messages in m[I], no matter what other input it is, the value of f is equal. That is to say, if ∃ i subject to $x_i \in$ m[I], the value of f(⋯ ,x_i,⋯) are equal (⋯ can be any value). Bellow, we present a unified definition of reasonable function.

Reasonable Function. Let M = {m_1,⋯ ,m_l} and X = {x_1,⋯ ,x_l} be any message of message space M, M is the challenge message, I = {i_1,⋯ ,i_t} ⊆ {1,⋯ ,l} is the query in the SOA-C process. Define:

$m[I] = \{m_{i_1}, \cdots, m_{i_t}\}$; $X|_I = \{x_i \in (m[I] \cap X)\}$; $X|_{\overline{I}} = X / X|_I$;

< y_1,y_2,⋯ ,y_l > denotes a permutation of the values y_1,⋯ ,y_l such that the value y_i is mapped to the k 'th location if y_i is the k' th input to f. Thus, .$X =< X|_{\bar{I}}, X|_I >$

Definition 1. (Reasonability). Let {f} be a set of functions f ∈ F. We say f is reasonable if $f < X|_{\bar{I}}, X|_I >= f < X'|_{\bar{I}}, X|_I >$ for ∀ X, X' ∈ M.

What we want to emphasize is that the key query and the corrupt query influence each other. The query of keys can increase the knowledge of the adversary, which can affect the choice of I; the corrupt query of I can make the adversary learn more about the message and can affect the choice of functionality f. In our scheme, we impose restrictions on the sequence of queries (the key queries of f must be made after the corrupt query of I) to remove the affect of the key queries, at the same time, on the KeyGen phase we limit the choice of f to remove the affect of the corrupt query on the basis of the opened messages in m[I], because an adversary may choose some special f in view of m[I] which can leak the information of unopened messages.

2. PRELIMINARIES

2.1 Functional encryption

A functional encryption scheme for a functionality f is a tuple of four algorithms: Setup. This is a PPT algorithm that takes the security parameter as input. It outputs a public and master secret key pair (PK,MSK).

Key Generation. This is a PPT algorithm that takes the functionality f as input, master secret key MSK. It outputs a decryption key SKf.

Encryption. This is a PPT algorithm that takes as input a message m and the public parameter PK. It outputs the ciphertext C.

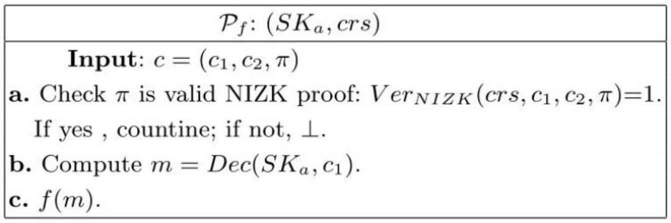
Decryption. This algorithm takes the ciphertext C and the decryption key SKf as input, and outputs f(m).

We utilize Garg et al.[16]'s construction of FE (dual system encryption):

Setup. Generate (PKa,SKa) ← SetupPKE, (PKb,SKb) ← Setup $_{\text{PKE}}$, crs ← Setup $_{\text{NIZK}}$ Key Generation(MSK,f). SKf = io(Pf) (refer to the following table).

Encryption(m). c = (c_1,c_2,π), where c_1 = Enc(PKa;m,r_1), c_2 = Enc(PKb;m,r_2), π is a NIZK proof of the fact that : ∃m,r_1,r_2 : c_1 = Enc(PKa;m,r_1) ∧ c_2 = Enc(PK_b;m,r_2).

Decryption. Compute SK_f(c).



$\mathcal{P}_f$: (SK_a, crs)
Input: $c = (c_1, c_2, \pi)$
a. Check π is valid NIZK proof: $Ver_{NIZK}(crs, c_1, c_2, \pi)$=1. If yes , countine; if not, $\perp$.
b. Compute $m = Dec(SK_a, c_1)$.
c. $f(m)$.

Table 1. Program $\mathcal{P}_f$

2.2 Deniable Encryption

An encryption scheme is deniable if the sender can generate fake randomness that will make the ciphertext looks like an encryption of a different plain message, thus to keep the real message private. A deniable encryption scheme contains the following algorithms:

Setup_{DE}. This is a PPT algorithm that takes the security parameter as input. It outputs a public and master secret key pair (pk_{DE}, sk_{DE}).

Enc_{DE}. This is a PPT algorithm that takes as input a message m and the public parameter pkDE, and outputs the ciphertext C.

Dec$_{DE}$. This algorithm takes C and the decryption key skDE as input, and outputs m. **Exp$_{DE}$.** This is a PPT algorithm that takes C,m0 as input. Output a fake random r_0 which satisfies EncDE(pk_{DE}, m_0, r_0) = C.

We utilize SW's [3] construction of DE:

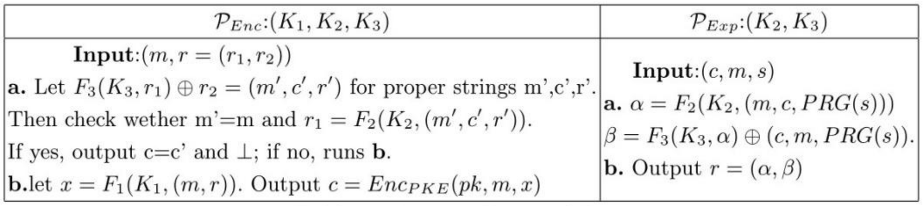
Bellare et al. [4] had proved no binding encryption scheme is simulator-based SOA security. That is why we use deniable encryption to realize our scheme. Specially, we use Sahai and Waters' scheme [3] which proposed a construction of deniable encryption. The scheme is proved to be IND-CPA secure and one-bit message encryption by using the technology of puncture, but it is not hard to generalize one-bit to a message string.

Setup$_{DE}$. (pk $_{PKE}$, sk $_{PKE}$) ← Setup $_{PKE}$. F1 is a puncturable extracting PRF, F2 is a puncturable statistically injective PRF, F3 is a puncturable PRF and (K1,K2,K3) is the corresponding puncturable PRFs' keys. pk_{DE} = (io(P_{Enc}),io(P_{Exp})), sk_{DE} = sk $_{PKE}$.

Enc$_{DE}$. c = io(P_{Enc}) (m,r)

Dec$_{DE}$. m = Dec $_{PKE}$ (sk $_{DE}$,c).

Exp$_{DE}$. r_0 ← io(P_{Exp}) (c, m_0, s): Enc$_{DE}$ (pk_{DE}, m_0, r_0) = c. (s is a randomness.)



$\mathcal{P}_{Enc}$:(K_1, K_2, K_3)	$\mathcal{P}_{Exp}$:(K_2, K_3)
Input:$(m, r = (r_1, r_2))$ **a.** Let $F_3(K_3, r_1) \oplus r_2 = (m', c', r')$ for proper strings m',c',r'. Then check wether m'=m and $r_1 = F_2(K_2, (m', c', r'))$. If yes, output c=c' and $\perp$; if no, runs **b**. **b.**let $x = F_1(K_1, (m, r))$. Output $c = Enc_{PKE}(pk, m, x)$	**Input**:(c, m, s) **a.** $\alpha = F_2(K_2, (m, c, PRG(s)))$ $\beta = F_3(K_3, \alpha) \oplus (c, m, PRG(s))$. **b.** Output $r = (\alpha, \beta)$

Table 2. Program $\mathcal{P}_{Enc}$ and $\mathcal{P}_{Exp}$

3. The Definition of SO-FE

We now propose the security model of a functional encryption secure against selective opening attacks, we call SO-FE.

Definition 2. We define two games GameREAL and GameSIM (refer to the following table).

GameREAL:

Setup. The challenger runs the Setup algorithm of FE, generates (PK,MSK) and gives the public parameters to the adversary.

Challenge. The adversary chooses a meessage distribution. The challenger chooses a message M from the distribution, and encrypts M . The ciphertext C is sent to the adversary. Corrupt query. The adversary makes one query to corrupt over a set of I (I ⊂ {1,2,⋯ ,l}), the challenger returns the messages m[I] and randomness r[I] used in challenge phase corresponding to I.

Key Query. The adversary is allowed to issue Key generation queries. That is to say the adversary outputs the function f to the challenger (f is reasonable), then the challenger runs KeyGen on f to generate the corresponding private key SKf and sends SKf to the adversary.

Final. The adversary guesses M.

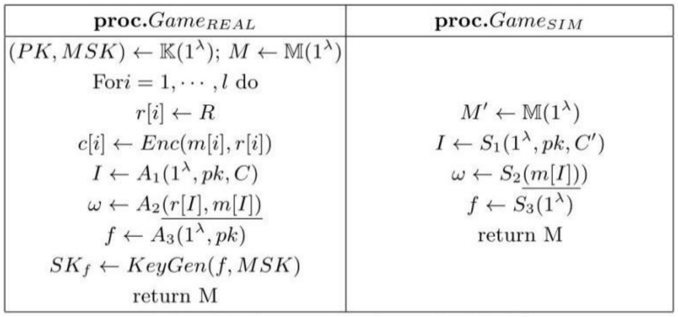
GameSIM:

Setup. The simulator generates (PK,MSK) and sends PK to the adversary.

Challenge. The simulator chooses a message M0 from the distribution, and encrypts M0 . The ciphertext C' is sent to the adversary which is indistinguishable with C in GameREAL. Corrupt query. The adversary makes one query to corrupt over a set of I, the simulator runs Oracle to get the messages m[I] ⊆ M in GameREAL and generates fake randomness r*[I] which satisfy C ‘[I] = EncFE(m[I],r*[I]).

Key Query. The simulator runs KeyGen on f to generate SKf and sends SKf to the adversary.

Final. The adversary guesses M.



proc.$Game_{REAL}$	**proc.**$Game_{SIM}$
$(PK, MSK) \leftarrow \mathbb{K}(1^\lambda)$; $M \leftarrow \mathbb{M}(1^\lambda)$	
For$i = 1, \cdots, l$ do	
$r[i] \leftarrow R$	$M' \leftarrow \mathbb{M}(1^\lambda)$
$c[i] \leftarrow Enc(m[i], r[i])$	$I \leftarrow S_1(1^\lambda, pk, C')$
$I \leftarrow A_1(1^\lambda, pk, C)$	$\omega \leftarrow S_2\underline{(m[I])})$
$\omega \leftarrow A_2\underline{(r[I], m[I])}$	$f \leftarrow S_3(1^\lambda)$
$f \leftarrow A_3(1^\lambda, pk)$	return M
$SK_f \leftarrow KeyGen(f, MSK)$	
return M	

Table 3. The security Games for SO-FE

We define the advantage of the adversary in this SO-FE Game:

AdvSO−FE(A) = |Pr[Gamereal ⇒ true] − Pr[GameSIM ⇒ true]|

A functional encryption scheme is secure against SOA if all polynomial time adversaries A have at most a negligible advantage in the Game.

Our scheme is post SO-FE, that is to say, the KeyGen queries of f must be made after the corrupt query of I. There are two reasons to explain why our scheme is asked to be post secure: one is to make sure the adversary choose the set of I without the help of the KeyGen queries. In the proof of the security, the simulator hope to run the adversary and utilize the rewind technology after the corrupt query hIi until the challenge cipher is not contain in I. The other is to make sure there is no leak about information of the challenge plaintext after the adversary receives SKf, because we restricy the choices of functions that can be queried based on I. The Specific reasons can refer to the proof of the security in section 5.

4. A CONSTRUCTION OF SO-FE

We now give our construction of SO-FE scheme. In fact, our construction is based on that of Garg et al.’ FE scheme, mixed with SW’ DE scheme. The dual public key encryption in FE is replaced with a dual DE.

Let M = $m_1, m_2, \cdots, m_l$ ($m_i \in \{0,1\}^n$), we have

Setup$_{\text{SO–FE}}$: The Setup algorithm first runs Setup $_{\text{NIZK}}$ to get crs and runs Setup $_{\text{DE}}$ twice to get $(pk^{\alpha}_{DE}, sk^{\alpha}_{DE})$ $(\alpha = a, b)$, where $pk^{\alpha}_{DE} = (io(P^{\alpha}_{Enc}), io(P^{\alpha}_{Exp}))$.

$$PK = (pk^{a}_{DE}, pk^{b}_{DE}, crs)\ ;\ MSK = (sk^{a}_{DE}, sk^{b}_{DE}, K^{a}_{1}, K^{b}_{1}, K^{a}_{2}, K^{b}_{2}, K^{a}_{3}, K^{b}_{3}).$$

(We utilize the SW’s DE scheme introduced in section 2, K_i^{α} (i = 1, 2, 3; α = a, b) are keys of F_1, F_2, F_3 in DE.)

Enc$_{\text{SO–FE}}$: ∀i = 1,⋯ ,l, α ∈ {a,b}, choose randomness $r^{\alpha}_{i} = (r^{\alpha}_{i,1}, r^{\alpha}_{i,2}) \leftarrow R$

Check if $r^{\alpha}_{i,1} = F_2(K^{\alpha}_{2}, F_3(K^{\alpha}_{3}, r^{\alpha}_{i,1}) \oplus r^{\alpha}_{i,2})$. If yes, choose randomness once again until the random does not satisfy the above condition.

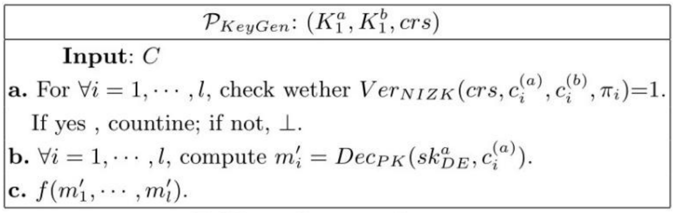
$$c_i^{(a)} = io(P^a_{Enc})(m_i, r_i^a)$$
$$c_i^{(b)} = io(P^b_{Enc})(m_i, r_i^b)$$

Creat a NIZK proof $\pi_i \leftarrow \Pr ove_{NIZK}(crs, (c_i^{(a)}, c_i^{(b)}), (r_i^a, r_i^b, m_i))$ to prove the fact that:

$$\exists m_i, r_i^a, r_i^b : c_i^{(a)} = Enc(PK_a; m_i, r_i^a) \wedge c_i^{(\ )} = Enc(PK_b; m_i, r_i^b).$$

$C_i = (c_i^{(a)}, c_i^{(b)}, \pi_i)$. Let the encryption of M is C = (C1,⋯ ,Cl).

KeyGen$_{\text{SO–FE}}$: Create an obfuscation of the program like the following Table 3, and output SKf = io(PKeyGen). DecSO–FE: Compute SKf (C).



$\mathcal{P}_{KeyGen}$: (K^a_1, K^b_1, crs)
Input: C
a. For $\forall i = 1, \cdots, l$, check wether $Ver_{NIZK}(crs, c_i^{(a)}, c_i^{(b)}, \pi_i)$=1. If yes , countine; if not, $\perp$.
b. $\forall i = 1, \cdots, l$, compute $m'_i = Dec_{PK}(sk^a_{DE}, c_i^{(a)})$.
c. $f(m'_1, \cdots, m'_l)$.

Table 4. Program $\mathcal{P}_{KeyGen}$

5. THE SECURITY OF SO-FE

The SO-FE scheme in section 4 is a SIM-SO FE scheme, the security model is given in section 3. Now we will give the security proof.

Theorem 1. If io is an indistinguishability obfuscator, DE is IND-CPA security and the NIZK is statistically simulation sound, the scheme is a no-adaptive secure SO-FE.

Proof. In order to prove the FE scheme is SIM-SO security, we need to construct a simulator which can run in the GameSIM to simulate all the possibility in the GameREAL. That is to say,

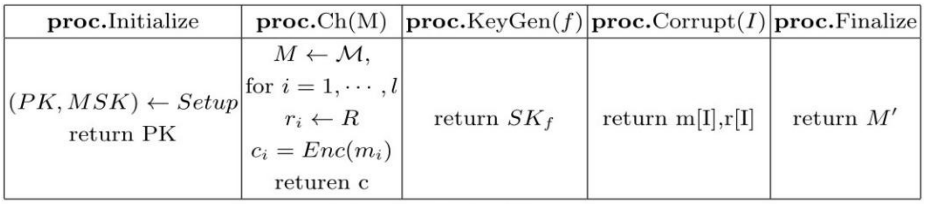
$$|\Pr(\text{GameREAL} \Rightarrow \text{true}) - \Pr(\text{GameSIM} \Rightarrow \text{true})| \leq \text{neg}(\cdot).$$

In short, the simulator needs to create equivocable ciphertexts as the challenge ciphertexts, then open them accordingly. Here, we must make sure the equivocable ciphertexts are indistinguishable from the real encryption of the messages in the REAL setting. In order to provide the environment of the adversary in GameREAL, on the corrupt phase, the simulator first gets the corrupt messages from the Oracle in the GameSIM and then outputs the fake randomness which is indistinguishable from the real random used in the encryption to the adversary (here we use the technology of DE).

we proof the theorem through a series of Hybrids:

Hybrid 0: Let A be an arbitrary adversary in GameREAL of the SO-FE security model. The challenger first generates (PK, MSK) and send the public key to to the adversary. Then the challenger chooses the message M from the message space M and encrypt the message running $\text{Enc}_{\text{SO-FE}}$. Later the adversary makes a corrupt query and some key generation queries, the challenger sends m[I],r[I] to A (r[I] is the real random used in encryption of m[I]). Finally, A give its guess of the message.

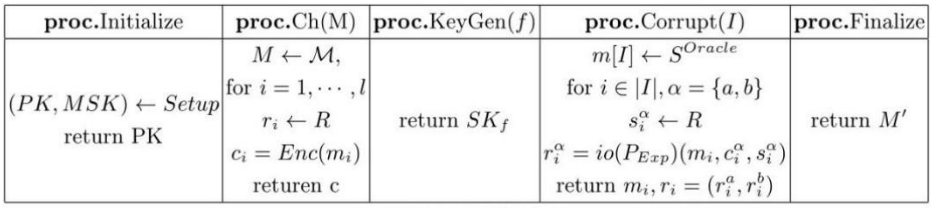
We can see Pr(Hybrid0 ⇒ true) = Pr(GameREAL ⇒ true)



proc. Initialize	**proc.**Ch(M)	**proc.**KeyGen(f)	**proc.**Corrupt(I)	**proc.**Finalize
$(PK, MSK) \leftarrow Setup$ return PK	$M \leftarrow \mathcal{M}$, for $i = 1, \cdots, l$ $r_i \leftarrow R$ $c_i = Enc(m_i)$ returen c	return SK_f	return m[I],r[I]	return M'

Table 5. Hybrid $0/Game_{REAL}$

Hybrid 1: We define Hybrid 1 to be the same as Hybrid 0, except that on the corrupt phase, the challenger first runs the Oracle in GameSIM to get the message m[I], for i ∈ |I|,α = {a,b}, set sαi ← R, riα = io(PExpα)(mi,ciα ,sαi). Output r[i] = (ria,rib). (cαi is the cipher generated by simulator, mi is the output of Oracle).



proc. Initialize	**proc.**Ch(M)	**proc.**KeyGen(f)	**proc.**Corrupt(I)	**proc.**Finalize
$(PK, MSK) \leftarrow Setup$ return PK	$M \leftarrow \mathcal{M}$, for $i = 1, \cdots, l$ $r_i \leftarrow R$ $c_i = Enc(m_i)$ returen c	return SK_f	$m[I] \leftarrow S^{Oracle}$ for $i \in \|I\|, \alpha = \{a, b\}$ $s_i^\alpha \leftarrow R$ $r_i^\alpha = io(P_{Exp})(m_i, c_i^\alpha, s_i^\alpha)$ return $m_i, r_i = (r_i^a, r_i^b)$	return M'

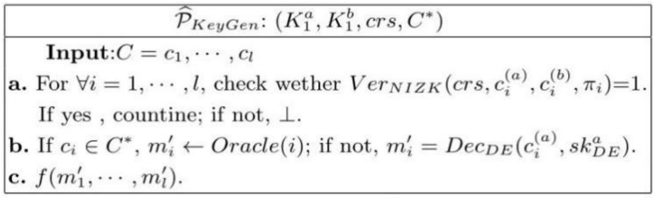
Table 6. Hybrid 1

We now say |Pr(Hybrid0 ⇒ true) − Pr(Hybrid1 ⇒ true)| ≤ neg(·), because the random returned in Hybrid 1 and Hybrid 0 are almost identically distributed in the view of A. The indistinguishability between Hybrid0 and Hybrid1 can reduce to the explainability of DE scheme.

In [3], Sahai and Waters had proved the explainability of deniable encryption: if the io is indistinguishable and F1 is a puncturable extracting PRF, F2 is a puncturable statistically injective PRF, F3 is a general puncturable PRF, then the generated pseudo-randomness is indistinguishable with the real random. While in Hybrid 0, the encrypted randomness is chosen from set {0,1}|r|/S,(S = {(a,b)|a = $F_2(K_2$, $F_3(K_3$,a) ⨁ b),a = {0,1}|r_1|,b = {0,1}|r_2|}). Now we can see the size of S: for any fixed a, there exist at most one preimage a0 because of F2 is a puncturable statistically injective PRF, thus b = a_0 ⨁ F3(K_3,a) is well-determined. So |S| = 2|r_1| and choose a random from S is negligible if r is large enough.

Hybrid 2: We define Hybrid 2 is the same with Hybrid 1 except that on the KeyGen query phase, the challenger returns $\widehat{SK}_f(\widehat{SK}_f = io(\widehat{P}_{KeyGen})$ is defined as follows). Our scheme is no-adaptive security, the KeyGen query is made after the challenge phase. It's easy to see SK[f and SKf is indistinguishable . So |Pr(Hybrid1 ⇒ true)−Pr(Hybrid2 ⇒ true)| ≤ neg(·).

The indistinguishability between Hybrid1 and Hybrid2 can reduce to the indistinguishability of io.

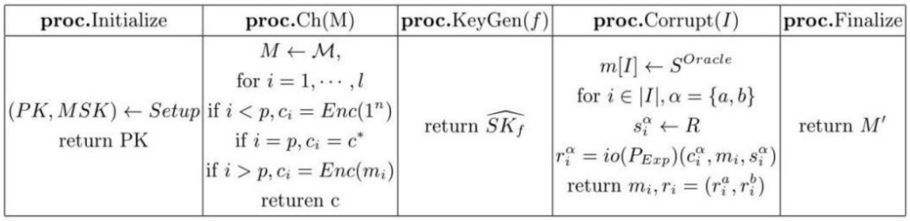


```
$\widehat{\mathcal{P}}_{KeyGen}$: $(K_1^a, K_1^b, crs, C^*)$
    Input:$C = c_1, \cdots, c_l$
a. For $\forall i = 1, \cdots, l$, check wether $Ver_{NIZK}(crs, c_i^{(a)}, c_i^{(b)}, \pi_i)$=1.
   If yes , countine; if not, $\perp$.
b. If $c_i \in C^*$, $m_i' \leftarrow Oracle(i)$; if not, $m_i' = Dec_{DE}(c_i^{(a)}, sk_{DE}^a)$.
c. $f(m_1', \cdots, m_l')$.
```

Table 7. Program $\widehat{\mathcal{P}}_{KeyGen}$

Hybrid 3−p:(0 ≤ p ≤ q) We define Hybrid 3−p is the same with Hybrid 2 except that on the challenge phase, if i ≤ p, we replace the real challenge cipher to new ones which are generate by simulater, (here specially the simulator choose messages mi = 1n and send the ciphers to A); If p < i ≤ q, the simulate sends the real challenge cipher to A.

We can see Pr(Hybrid3−0 ⇒true) =Pr(Hybrid2 ⇒true) and Pr(Hybrid3−q ⇒true) =Pr(GameSIM ⇒true). So our aim is to prove |Pr(Hybrid3−0 ⇒true) − Pr(Hybrid3−q ⇒true)| ≤ neg(·). We define the Hybrid3−p is like the following table 7.



proc. Initialize	**proc.**Ch(M)	**proc.**KeyGen(f)	**proc.**Corrupt(I)	**proc.**Finalize

$(PK, MSK) \leftarrow Setup$
return PK
	$M \leftarrow \mathcal{M}$,
	for $i = 1, \cdots, l$
	if $i < p, c_i = Enc(1^n)$
	if $i = p, c_i = c^*$
	if $i > p, c_i = Enc(m_i)$
	returen c
		return $\widehat{SK_f}$	$m[I] \leftarrow S^{Oracle}$
			for $i \in \|I\|, \alpha = \{a, b\}$
			$s_i^\alpha \leftarrow R$
			$r_i^\alpha = io(P_{Exp})(c_i^\alpha, m_i, s_i^\alpha)$
			return $m_i, r_i = (r_i^a, r_i^b)$
				return M'

Table 8. Hybrid $3_{-p}(0 \leq p \leq q$ and Hybrid $3_{-q}/Game_{SIM})$

Now we begin to explain the indistinguishability between Hybrid3−p and Hybrid3−(p−1). To prove the above problem, we first define the following hybrids and then reduce the indistinguishability to security of IND-CPA DE.

Hybrid3−(p−1)−(0): This hybrid is the same with Hybrid3−(p−1).

Hybrid3−(p−1)−(1): This hybrid uses the trapdoor in NIZK to generate an fake proof to make sure that the adversary can believe two ciphertexts in double system encryption is an encryption of the same message.

$$\pi^* \leftarrow Sim_{NIZK}(1^\lambda, \exists m, r_a, r_b : c_a^* = Enc(pk_{DE}^a, m, r_a) \wedge c_b^* = Enc(pk_{DE}^b, m, r_b)).$$

Hybrid3−(p−1)−(2): This hybrid change the pth ciphertext to $(c_p^a, c_p^b, \pi_p)^*$, where

$c_p^a* = Enc(pk_{DE}^a, m_p, r_a),\ c_p^b* = Enc(pk_{DE}^a, 1^n, r_b),\ \pi_p*$ is a fake proof generated by

Sim$_{\text{NIZK.}}$

Hybrid3−(p−1)−(3): This hybrid is the same with Hybrid3−p−(2) except that the pth ciphertext is $(c_p^a, c_p^b, \pi_p)^*$, where $c_p^a* = Enc(pk_{DE}^a, 1^n, r_a), c_p^b* = Enc(pk_{DE}^a, 1^n, r_b)$, and on the io of KeyGen query phase, we replace K_1^a to K_1^b and make sure we can use the key in the second part of the double encryption system. It's not hard to see Hybrid3−(p−1)−(3) ≈ Hybrid3−p.

If SSS-NIZK is computationally zero knowledge, then Hybrid3−(p−1)−(0), Hybrid3−(p−1)−(1) is indistinguish. For the indistinguishability between (1) and (2) or (2) and (3), we hope to reduce the problem to the IND-CPA secure DE. That is to say we hope to structure a simulator B who can run A, if there is an A who can distinguish (1) and (2) or (2) and (3), there is an adversary B who can distinguish the challenge cipher c□ in Game of IND-CPA DE. The reduction can refer to appendix. So

$$\left|Pr(Hybrid3_{-p} \Rightarrow true) - Pr(Hybrid3_{-(p-1)} \Rightarrow true)\right| \leq neg(\cdot)$$

$$\begin{aligned}
&|Pr(Hybrid3_{-0} \Rightarrow true) - Pr(Hybrid3_{-q} \Rightarrow true)| \\
&\leq |Pr(Hybrid3_{-0} \Rightarrow true) - Pr(Hybrid3_{-1} \Rightarrow true)| + \cdots \\
&\qquad + \left|Pr(Hybrid3_{-(q-1)} \Rightarrow true) - Pr(Hybrid3_{-q} \Rightarrow true)\right| \\
&\leq neg(\cdot).
\end{aligned}$$

6. CONCLUSION

Our paper proposed a stronger security of FE which is secure against SOA and proposed a concrete construction of SO-FE scheme. A lot of work is worth doing in the future, for example, how to concrete a SO-FE without indistinguishability obfuscation.

ACKNOWLEDGEMENTS

We would like to thank all workers who have helped us to make the paper better.

REFERENCES

[1] Mihir Bellare, Dennis Hofheinz, Scott Yilek: Possibility and impossibility results for encryption and commitment secure under selective opening. EUROCRYPT 2009. LNCS, vol. 5479, pp. 1-35. Springer, Heidelberg (2009)

[2] Ran Canetti, Cynthia Dwork, Moni Naor and Rafi Ostrovsky: Deniable Encryption. CRYPTO. Cryptology ePrint Archive, Report 1996/002. pp 90-104. (1997)

[3] Amit Sahai and Brent Waters: How to Use Indistinguishability Obfuscation: Deniable Encryption, and More. STOC 2014. Cryptology ePrint Archive, Report 2013/454. pp 475-484, (2014)

[4] Mihir Bellare, Rafael Dowsley, Brent Waters, Scott Yilek: Standard security does not imply security against selective-opening. EUROCRYPT 2012. LNCS, vol. 7237, pp. 645-662. Springer, Heidelberg (2012)

[5] Mihir Bellare, Dennis Hofheinz, Scott Yilek: Possibility and impossibility results for encryption and commitment secure under selective opening. EUROCRYPT 2009. LNCS, vol. 5479, pp. 1-35. Springer, Heidelberg (2009)

[6] Serge Fehr, Dennis Hofheinz, Eike Kiltz, Hoeteck Wee: Encryption schemes secure against chosen-ciphertext selective opening attacks. EUROCRYPT 2010. LNCS, vol. 6110, pp. 381-402. Springer, Heidelberg (2010)

[7] Zhengan Huang, Shengli Liu, Baodong Qin: Sender-equivocable encryption schemes secure against chosen-ciphertext attacks revisited. PKC2013. LNCS, vol. 7778, pp. 369-385. Springer, Heidelberg (2013)

[8] Brett Hemenway, Benoit Libert, Rafail Ostrovsky, Damien Vergnaud: Lossy encryption: Constructions from general assumptions and efficient selective opening chosen ciphertext security. ASIACRYPT 2011. LNCS, vol. 7073, pp. 70-88. Springer, Heidelberg (2011)

[9] Dennis Hofheinz: All-but-many lossy trapdoor functions. EUROCRYPT 2012. LNCS, vol. 7237, pp. 209-227. Springer, Heidelberg (2012)

[10] Mihir Bellare, Scott Yilek: Encryption schemes secure under selective opening attack. IACR Cryptology ePrint Archive, 2009:101 (2009)

[11] Mihir Bellare, Brent Waters, Scott Yilek: Identity-based encryption secure against selective opening attack. TCC 2011. LNCS, vol. 6597, pp. 235-252.Springer, Heidelberg (2011)

[12] Junzuo Lai, Robert H. Deng, Shengli Liu,Jian Weng, Yunlei Zhao : Identity-Based Encryption Secure against Selective Opening Chosen-Ciphertext Attack. EUROCRYPT 2014. LNCS, vol. 8441, pp 77-92. Springer, Heidelberg (2014)

[13] Dan Boneh, Amit Sahai, Brent Waters: Functional Encryption: Definitions and Challenges. LNCS, vol. 6597, pp 253-27 (2011)

[14] Florian B\ddot{o}hl, Dennis Hofheinz, Daniel Kraschewski: On definitions of selective opening security. PKC 2012. LNCS, vol. 7293, pp. 522-539. Springer, Heidelberg (2012)

[15] Mihir Bellare, Adam O'Neill: Semantically - secure functional encryption: Possibility results, impossibility results and the quest for a general definition. Cryptology ePrint Archive, Report 2012:515 (2012)

[16] Sanjam Garg, Craig Gentry, Shai Halevi, Mariana Raykova, Amit Sahai and Brent Waters: Candidate Indistinguishability Obfuscation and Functional Encryption for All Circuits. FOCS 2013, IEEE Computer Society. pp 40-49 (2013)

[17] Dan Boneh and Brent Waters: Constrained pseudorandom functions and their applications. IACR Cryptology ePrint Archive, 2013:352. (2013)

APPENDIX

A. Puncturable PRF

A puncturable family of PRFs F mapping ($\{0,1\}^{n(\cdot)} \rightarrow \{0,1\}^{m(\cdot)}$) is given by a triple of Turing Machines (Key_F,$\mathrm{Puncture}_F$,Eval_F) satisfying the following conditions:

Functionality preserved. For every PPT adversary A such that A(1λ) outputs a set $S \subseteq \{0,1\}^{n(\lambda)}$, then we have

$$Pr\left[Eval_F(K,x) = Eval_F(K_S,x) \middle| \begin{array}{l} K \leftarrow Key_F(1^\lambda) \\ K_S = Puncture_F(K,S) \\ \forall x \in \{0,1\}^{n(\lambda)} \wedge x \notin S \end{array} \right] = 1$$

Pseudorandom at punctured points.For every PPT adversary A such that $A(1^\lambda)$ outputs a set $S \subseteq \{0,1\}^{n(\lambda)}$ and state σ, consider an experiment where K ← KeyF(1^λ) and KS = PunctureF(K,S), for any PPT distinguisher D, we have

$$|\Pr[D(\sigma,KS,S,EvalF(K,S)) = 1] - \Pr[D(\sigma,KS,S,Um(\lambda)\cdot|S|) = 1]| \leq neg(\lambda)$$

Definition 3. A puncturable statistically injective PRF family with failure probability ε(·) is a family of PRFs F such that with probability 1 − ε(λ) over the random choice of key K ← KeyF(1λ), we have that F(K,) is injective.

Definition 4. A puncturable extracting PRF family with error ε(·) for min-entropy k(·) is a family of PRFs F mapping {0,1}n(λ) → {0,1}m(λ) such that for all λ, if X is any distribution over {0,1}m(λ) with min-entropy greater than k(λ), then the statistical distance between (K ← KeyF(1λ),F(K,X)) and (K ← KeyF(1λ),Um(λ)) is at most ε(λ).

B. Indistinguishability Obfuscator

A uniform PPT machine io is called an indistinguishability obfuscator (io) for a circuit family {Cλ} if the following conditions are satisfied:

Functionality preserved. For all security parameters λ ∈ N, for all C ∈ $\{C_\lambda\}$, for all input x, we have

Pr[C0(x) = C(x) : C0 ← io(λ,C)] = 1

Indistinguishability. For any PPT distinguisher D, for all security parameters λ ∈ N, for all pairs of circuits C0,C1 ∈ $\{C_\lambda\}$ which satisfies $\Pr[\forall x, C_0(x) = C_1(x)] > 1-neg(\cdot)$, then

$|Pr[D(io(\lambda, C_0)) = 1] - Pr[D(io(\lambda, C_1)) = 1]| \leq neg(\lambda)$

C.NIZK

A non-interactive zero-knowledge proof system (NIZK) contains three algorithms NIZK = (Setup,Prove,V er): crs ← Setup(1^k);π ← Prove(crs, stmt, ω);b ← V er(crs, stmt, π), where k is the security parameter, crs is the common reference string, stmt is the statement information, ω is a witness and π is the proof, moreover b is 0/1 means rejection or acceptance.

Completeness. Pr[crs ← Setup,π ← Prove(crs,stmt,ω): V er(crs,stmt,π) = 1] = 1

Soundness. Pr[crs ← Setup,∃(stmt,π) : (stmt /∈ L) ∧ V er(crs,stmt,π) = 1] ≤ neg(·)

Zero-knowledge. If there exists a simulator S=(SimSetup,SimProve),such that for all PPT adversary A, it holds that

$$\left| Pr\begin{bmatrix} crs \leftarrow Setup(1^k, m) \\ \forall i \in [m] : \pi_i \leftarrow Prove(crs, stmt_i, \omega_i) \\ \mathcal{A}(crs, \{stmt_i, \pi_i\}_{i\in[m]}) = 1 \end{bmatrix} - Pr\begin{bmatrix} (\widetilde{crs}, trap) \leftarrow SimSetup(1^k, \{stmt_i\}_{i\in[m]}) \\ \forall i \in [m] : \pi_i \leftarrow SimProve(crs, stmt_i, trap) \\ \mathcal{A}(\widetilde{crs}, \{stmt_i, \pi_i\}_{i\in[m]}) = 1 \end{bmatrix} \right|$$

is negligible.

In [16], the FE scheme used statistically simulation sound NIZK, which they called SSS-NIZK, and Garg et al. proposed a concrete construction of SSS-NIKZ. Informally, a NIZK system is statistically simulation sound, if under a simulated crs, there is no valid

proof for any false statement, except for the simulated proofs for statements fed into the SimSetup algorithm to generate crs. That is to say, f

$$Pr\begin{bmatrix} (\widetilde{crs}, trap) \leftarrow SimSetup(1^k, \{stmt_i\}_{i\in[m]}) \\ \forall i \in [m] : \pi_i \leftarrow SimProve(crs, stmt_i, trap) \\ \exists(stmt', \pi') s.t.(stmt' \notin \{stmt_i\}_{i\in[m]}) \wedge Ver(crs, stmt, \pi) = 1 \end{bmatrix} \leq neg(\cdot)$$

D. Reduct to IND-CPA DE

Here we will explain the indistinguishability between Hybrid3−(p−1)−(1) and Hybrid3−(p−1)−(2) or Hybrid3−(p−1)−(2) and Hybrid3−(p−1)−(3). We hope to structure a simulator B who can run A, if there is an A who can distinguish (1) and (2) or (2) and (3), there is an adversary B who can distinguish the challenge cipher c^* in Game of IND-CPA DE (refer to the following figures).

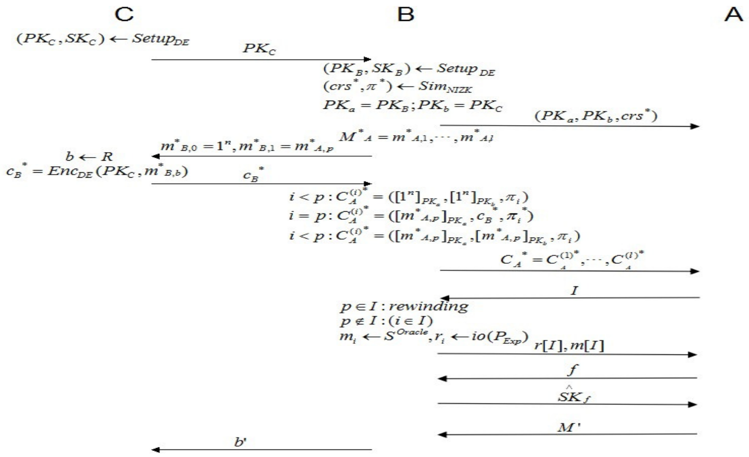

C
B
A
$(PK_C, SK_C) \leftarrow Setup_{DE}$
PK_C
$(PK_B, SK_B) \leftarrow Setup_{DE}$
$(crs^*, \pi^*) \leftarrow Sim_{NIZK}$
$PK_a = PK_B; PK_b = PK_C$
(PK_a, PK_b, crs^*)
$M^*_A = m^*_{A,1}, \cdots, m^*_{A,l}$
$m^*_{B,0} = 1^n, m^*_{B,1} = m^*_{A,p}$
$b \leftarrow R$
$c_B^* = Enc_{DE}(PK_C, m^*_{B,b})$
c_B^*
$i < p : C_A^{(i)*} = ([1^n]_{PK_a}, [1^n]_{PK_b}, \pi_i)$
$i = p : C_A^{(i)*} = ([m^*_{A,p}]_{PK_a}, c_B^*, \pi_i^*)$
$i < p : C_A^{(i)*} = ([m^*_{A,p}]_{PK_a}, [m^*_{A,p}]_{PK_b}, \pi_i)$
$C_A^* = C_A^{(1)*}, \cdots, C_A^{(l)*}$
I
$p \in I : rewinding$
$p \notin I : (i \in I)$
$m_i \leftarrow S^{Oracle}, r_i \leftarrow io(P_{Exp})$
$r[I], m[I]$
f
$\hat{SK}_f$
M'
b'

Fig.1.The reduction process: the indistinguishability between Hybrid3−(p−1)−(1) and Hybrid3−(p−1)−(2). [m]PK means encryption of m with public key PK.

Take Hybrid3−(p−1)−(2) and Hybrid3−(p−1)−(3) for example:

B gets PKC from the challenger from IND-CPA game of DE, then sets PKa = PKC and generates a pair of key (PK_b,SK_b). B sends (PK_a,PK_b) to adversary A in the SOAGame FE. B chooses message M*A = m*A,1,··· ,m*A,l from message space and makes the challenger's challenge message $m^*_{B,0} = 1^n, m^*_{B,1} = m^*_{A,p}$. . The challenger will return a

challenge cipher c*B. Then B hides the c*B into the challenge cipher $C^*_A = C^{(1)*}_A, \cdots, C^{(l)*}_A$ of A in the following way:

if $i < p, C^{(i)*}_A = (Enc^a_{DE}(1^n), Enc^b_{DE}(1^n), \pi_i)$;

if $i = p, C^{(i)*}_A = (c^*_B, Enc^b_{DE}(1^n), \pi^*_i)$;

if $i > p, C^{(i)*}_A = (Enc^a_{DE}(m_p), Enc^b_{DE}(m_p), \pi_i)$

When A makes corrupt query hIi: B first check whether p ∈ I, if yes, B uses the rewind technology to repeatedly run A until p ∉ I; if not, B makes pseudo randomness using io(P_{Exp}) after knowing the message m[I].

When A make key generate queries hfi(q-bounded): B replaces K^a_1 to K^b_1 in the SK[f to decrypt and make sure we can use the key in the second part of the double encryption system. Then send it to A.

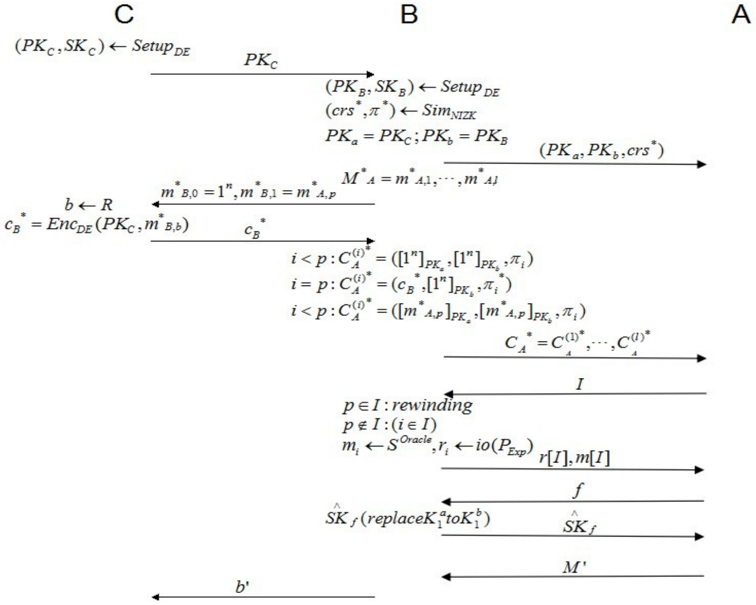

C
B
A
$(PK_C, SK_C) \leftarrow Setup_{DE}$
PK_C
$(PK_B, SK_B) \leftarrow Setup_{DE}$
$(crs^*, \pi^*) \leftarrow Sim_{NIZK}$
$PK_a = PK_C; PK_b = PK_B$
(PK_a, PK_b, crs^*)
$M^*_A = m^*_{A,1}, \cdots, m^*_{A,l}$
$m^*_{B,0} = 1^n, m^*_{B,1} = m^*_{A,p}$
$b \leftarrow R$
$c_B^* = Enc_{DE}(PK_C, m^*_{B,b})$
c_B^*
$i < p : C_A^{(i)^*} = ([1^n]_{PK_a}, [1^n]_{PK_b}, \pi_i)$
$i = p : C_A^{(i)^*} = (c_B^*, [1^n]_{PK_b}, \pi_i^*)$
$i < p : C_A^{(i)^*} = ([m^*_{A,p}]_{PK_a}, [m^*_{A,p}]_{PK_b}, \pi_i)$
$C_A^* = C_A^{(1)^*}, \cdots, C_A^{(l)^*}$
I
$p \in I : rewinding$
$p \notin I : (i \in I)$
$m_i \leftarrow S^{Oracle}, r_i \leftarrow io(P_{Exp})$
$r[I], m[I]$
f
$\hat{SK}_f (replace K_1^a to K_1^b)$
$\hat{SK}_f$
M'
b'

Fig.2. The reduction process: the indistinguishability betwee n Hybrid3−(p−1)−(2) and Hybrid3−(p−1)−(3)

Finally A will output its guess M0, then the B can utilize the pth guess to reply the challenger in Game DE. So if A can guess the message rightly, thus B can distinguish between the cipher of m0 or m1 with non-negligible advantage, which will break the INDCPA property of DE.

AUTHORS

Yuanyuan Ji, was born in henan, China, on Nov. 10, 1989. She is studying for a master’s degree at the university of Chinese academy of sciences, Beijing, China

SECURITY AND PRIVACY OF SENSITIVE DATA IN CLOUD COMPUTING: A SURVEY OF RECENT DEVELOPMENTS

Ali Gholami and Erwin Laure

HPCViz Dept., KTH- Royal Institute of Technology, Stockholm, Sweden
{gholami,erwinl}@pdc.kth.se

ABSTRACT

Cloud computing is revolutionizing many ecosystems by providing organizations with computing resources featuring easy deployment, connectivity, configuration, automation and scalability. This paradigm shift raises a broad range of security and privacy issues that must be taken into consideration. Multi-tenancy, loss of control, and trust are key challenges in cloud computing environments. This paper reviews the existing technologies and a wide array of both earlier and state-of-the-art projects on cloud security and privacy. We categorize the existing research according to the cloud reference architecture orchestration, resource control, physical resource, and cloud service management layers, in addition to reviewing the existing developments in privacy-preserving sensitive data approaches in cloud computing such as privacy threat modeling and privacy enhancing protocols and solutions.

KEYWORDS

Cloud Security, Privacy, Trust, Virtualization, Data Protection

1. INTRODUCTION

Cloud computing is revolutionizing many of our ecosystems, including healthcare. Compared with earlier methods of processing data, cloud computing environments provide significant benefits, such as the availability of automated tools to assemble, connect, configure and reconfigure virtualized resources on demand. These make it much easier to meet organizational goals as organizations can easily deploy cloud services. However, the shift in paradigm that accompanies the adoption of cloud computing is increasingly giving rise to security and privacy considerations relating to facets of cloud computing such as multi-tenancy, trust, loss of control and accountability [1]. Consequently cloud platforms that handle sensitive information are required to deploy technical measures and organizational safeguards to avoid data protection breakdowns that might result in enormous and costly damages.

Sensitive information in the context of cloud computing encompasses data from a wide range of different areas and disciplines. Data concerning health is a typical example of the type of sensitive information handled in cloud computing environments, and it is obvious that most individuals will want information related to their health to be secure. Hence, with the

David C. Wyld et al. (Eds) : NETCOM, NCS, WiMoNe, CSEIT, SPM - 2015
pp. 131–150, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51611

proliferation of these new cloud technologies in recent times, privacy and data protection requirements have been evolving to protect individuals against surveillance and database disclosure. Some examples of such protective legislation are the EU Data Protection Directive (DPD) [2] and the US Health Insurance Portability and Accountability Act (HIPAA) [3], both of which demand privacy preservation for handling personally identifiable information.

This paper presents an overview of the research on security and privacy of sensitive data in cloud computing environments. We identify new developments in the areas of orchestration, resource control, physical hardware, and cloud service management layers of a cloud provider. We also review the state-of-the-art in privacy-preserving sensitive data approaches for handling sensitive data in cloud computing such as privacy threat modeling and privacy enhancing protocols and solutions.

The rest of this paper is organized as follows. Section 2 gives an overview of cloud computing concepts and technologies. Section 3 describes the security and privacy issues that need to be solved in order to provide secure data management for cloud environments. Section 4, reviews the existing security solutions that are being used in the area of cloud computing. Section 5 describes research on privacy-preserving solutions for sensitive data. Finally, in Section 6, we present our findings and conclusions.

2. KEY CONCEPTS AND TECHNOLOGIES

Over the past few years, major IT vendors (such as Amazon, Microsoft and Google) have provided virtual machines (VMs), via their clouds, that customers could rent. These clouds utilize hardware resources and support live migration of VMs in addition to dynamic load-balancing and on-demand provisioning. This means that, by renting VMs via a cloud, the entire datacenter footprint of a modern enterprise can be reduced from thousands of physical servers to a few hundred (or even just dozens) of hosts.

While it is practical and cost effective to use cloud computing in this way, there can be issues with security when using systems that are not provided in-house. To look into these and find appropriate solutions, there are several key concepts and technologies that are widely used in cloud computing that need to be understood, such as virtualization mechanisms, varieties of cloud services, and "container" technologies.

2.1. Virtualization Mechanisms

A hypervisor or virtual machine monitor (VMM) is a key component that resides between VMs and hardware to control the virtualized resource [4]. It provides the means to run several isolated virtual machines on the same physical host. Hypervisors can be categorized into two groups [5]:

- ***Type I***: Here the hypervisor runs directly on the real system hardware, and there is no operating system (OS) under it. This approach is efficient as it eliminates any intermediary layers. Another benefit with this type of hypervisor is that security levels can be improved by isolating the guest VMs. That way, if a VM is compromised, it can only affect itself and will not interfere with the hypervisor or other guest VMs.

- ***Type II***: The second type of hypervisor runs on a hosted OS that provides virtualization services, such as input/output (IO) device support and memory management. All VM interactions, such as IO requests, network operations and interrupts, are handled by the hypervisor.

Xen[1] and kernel virtual machine (KVM)[2] are two popular open-source hypervisors (respectively of *Type I* and *Type II*). Xen runs directly on the underlying hardware and it inserts a virtualization layer between the system hardware and the virtual machines. The OSs running in the VMs interact with the virtual resources as if they were actually physical resources. KVM is a virtualization feature in the Linux Kernel that makes it possible to safely execute guest code directly on the host CPU.

2.2. Cloud Computing Characteristics

When considering cloud computing, we need to be aware of the types of services that are offered, the way those services are delivered to those using the services, and the different types of people and groups that are involved with cloud services.

Cloud computing delivers computing software, platforms and infrastructures as services based on pay-as-you go models. Cloud service models can be deployed for on-demand storage and computing power in various ways: *Software-as-a-Service* (*SaaS*), *Platform-as-a-Service (PaaS)* and *Infrastructure-as-a-Service (IaaS)*. Cloud computing service models have been evolved during the past few years within a variety of domains using the "*as-a-Service"* concept of cloud computing such as *Business Integration-as-a-Service, Cloud-Based Analytics-as-a-Service (CLAaaS), Data-as-a-Service (DaaS)* [61], [62]. This paper refers to the NIST cloud service models features [6] that are summarized in Table 1 that can be delivered to consumers using different models such as a private cloud, community cloud, public cloud, or hybrid cloud.

Table 1, Categorization of Cloud Service Models and Features

Service Model	Function	Example
SaaS	Allows consumers to run applications by virtualizing hardware on the resources of the cloud providers	Salesforce Customer Relationship Management (CRM)[3]
PaaS	Provides capability of deploying custom applications with their dependencies within an environment called a container.	Google App Engine[4], Heroku[5]
IaaS	Provides a hardware platform as a service such as virtual machines, processing, storage, networks and database services.	Amazon Elastic Compute Cloud (EC2)[6]

[1] Xen hypervisor, http://xen.org/products/xenhyp.html

[2] KVM, http://www.linux-kvm.org/

[3] Salesforce CRM, https://www.salesforce.com/crm/

[4] Google App Engine, https://appengine.google.com

[5] Heroku, https://www.heroku.com

[6] Amazon EC2, https://aws.amazon.com/ec2/

The NIST cloud computing reference architecture [7], defines five major actors in the cloud arena: cloud consumers, cloud providers, cloud carriers, cloud auditors and cloud brokers. Each of these actors is an entity (either a person or an organization) that participates in a cloud computing transaction or process, and/or performs cloud computing tasks.

A cloud consumer is a person or organization that uses services from cloud providers in the context of a business relationship. A cloud provider is an entity makes cloud services available to interested users. A cloud auditor conducts independent assessments of cloud services, operations, performance and security in relation to the cloud deployment. A cloud broker is an entity that manages the use, performance and delivery of cloud services, and also establishes relationships between cloud providers and cloud consumers. A cloud carrier is an entity that provides connectivity and transport of cloud services from cloud providers to cloud consumers through the physical networks.

The activities of cloud providers can be divided into five main categories: service deployment, resource abstraction, physical resources, service management, security and privacy [7]. Service deployment consists of delivering services to cloud consumers according to one of the service models (*SaaS, PaaS, IasS).* Resource abstraction refers to providing interfaces for interacting with networking, storage and compute resources. The physical resources layer includes the physical hardware and facilities that are accessible via the resource abstraction layer. Service management includes providing business support, resource provisioning, configuration management, portability and interoperability to other cloud providers or brokers. The security and privacy responsibilities of cloud providers include integrating solutions to ensure legitimate delivery of cloud services to the cloud consumers. The security and privacy features that are necessary for the activities of cloud providers are described in Table 2 [10].

Table 2, Security and Privacy Factors of the Cloud Providers

Security Context	Description
Authentication and Authorization	Authentication and authorization of cloud consumers using pre-defined identification schemes.
Identity and Access Management	Cloud consumer provisioning and deprovisioning via heterogeneous cloud service providers.
Confidentiality, Integrity, Availability	Assuring the confidentiality of the data objects, authorizing data modifications and ensuring that resources are available when needed.
Monitoring and Incident Response	Continuous monitoring of the cloud infrastructure to assure compliance with consumer security policies and auditing requirements.
Policy Management	Defining and enforcing rules for certain actions such as auditing or proof of compliance.
Privacy	Protect personally identifiable information (PII) within the cloud from adversarial attacks that aim to find out the identity of the person that PII relates to.

The majority of cloud computing infrastructures consist of reliable services delivered through data centers to achieve high availability through redundancy. A data center or computer center is a facility used to house computer systems and associated components, such as storage and network systems. It generally includes redundant or backup power units, redundant network connections, air conditioning, and fire safety controls.

2.3. Containers Technology

Clouds based on Linux container (LXC) technology are considered to be next-generation clouds, so LXCs has become an important part of the cloud computing infrastructures because of their ability to run several OS-level isolated VMs within a host with a very low overhead. LXCs are built on modern kernel features. An LXC resembles a light-weight execution environment within a host system that runs instructions native to the core CPU while eliminating the need for instruction level emulation or just-in-time compilation [8]. LXCs contain applications, configurations and the required storage dependencies, in a manner similar to the just enough OS (JeOS).

Using containers, several applications can share an OS, binaries or libraries, which results in significant increases in efficiency compared to using hypervisors. For example, the portability of applications and the provisioning time of VMs are very low with container technologies [9].

LXC technologies were introduced in the 1980s, starting with the chroot (change root) command, and evolving into to popular container managers such as Docker.

- **Chroot**: The Unix chroot system call, which was introduced as part of Unix version 7 in 1979, can be considered as the first step in the evolution of containerization. The chroot call changes the root directory of the calling process to a specified path, where the root directory is known by all children of the calling process. This feature is used by some containers for isolation and sharing the underlying file system. Chroot is often used when building system images by changing root to a temporary directory, downloading and installing packages in chroot or compressing chroot as a system root file system.

- **FreeBSD Jail**[7] extended chroot in 1998 to provide enhanced security. FreeBSD jail settings can explicitly restrict access outside the sandbox environment by files, processes, and user accounts (including accounts created by the jail definition). Jail can therefore define a new root user, who has full control inside the sandbox, but who cannot reach anything outside.

- **Namespaces** were introduced in 1992 [11] for process-based resource isolation. Namespaces provide tools for isolating the view of global resources such as details about file systems, processes, network interfaces, Inter Process Communication (IPC), host names, and user IDs. Processes in a particular namespace are invisible to other processes because they think that they are the only processes on the system and because "connectivity" is only permitted with the parent namespace

[7] https://www.freebsd.org/doc/handbook/jails.html

- **Control Groups (cgroups)**[8] are kernel mechanisms introduced by Google in 2007 to provide fine-grained control by grouping processes and their children into a tree structure for resource management. Each group can be assigned a task for operations related to CPU, memory, disk and network. For example, to isolate two groups such as applications resources and OS resources, two groups (group 1 and 2) can be created to assign resource profiles to each group.

- **Linux Security Modules (LSMs)** are kernel modules which provide a framework for mandatory access control (MAC) security implementations. In MAC implementations, the administrator (user or process) assigns access controls to subject / initiator. In discretionary access control (DAC), the resource owner (user) assigns access controls to the subject or initiator. Existing LSM implementations include AppArmor, SELinux and so forth to prevent virtual machines from attacking other virtual machines or the host. For this purpose, policies are used to define what actions a process can perform on a particular system.

- **Containers** are built on the hardware and operating system but they make use of kernel features called chroots, cgroups and namespaces to construct a contained environment without the need for a hypervisor. The most recent container technologies are Solaris Zones, OpenVZ and LXC.

  In 2004, Solaris version 10 used zones as facilities to provide protected virtualized environments within a single host. Every Solaris system includes a global zone for both system and system-wide administrative control, and may have one or more non-global zones. All processes run in the global zone if there is no non-global zone. The global zone is aware of all devices and all file systems, while non-global zones are not aware of the existence of any other zones. Zone-based containers provide isolation, security and virtualization. Zones are similar to jails with additional features such as snapshots and cloning that make it possible to clone efficiently or to duplicate a current zone into a new zone.

  In 2005 OpenVZ [9] containers were introduced using a modified Linux kernel with a set of extensions. OpenVZ is based on the namespace and control group concepts in contrast to jails, which were used in FreeBSD.

  Later in 2008, LXC[10] emerged as a container management tool and it combined namespaces and control groups to create a fully isolated environment. It provides libraries and command-line support to enable administrators to create new containers. LXC containers can be used in either privileged (as a root user) or unprivileged (as a non-root user) modes to easily customize kernel capabilities or configure cgroups to satisfy the particular requirements.

  Docker is another container management tool – it was introduced in 2013 and is based on namespaces, cgroups and SELinux. Docker provides automation for the deployment of containers through remote APIs and has additional features that make it possible to create

[8] Kernel CGroups, https://www.kernel.org/doc/Documentation/cgroups/

[9] OpenVZ, https://openvz.org/

[10] LXC Containers, https://linuxcontainers.org/

standardized environments for developing applications. This has made Docker a popular technology. Creating the standardized environments is achieved using a layered image format that enables users to add or remove applications and their dependencies to form a trusted image.

Docker adds portable deployment of LXCs across different machines. In cloud terms, one can think of LXC as the hypervisor and Docker as both the open virtualization appliance and the provision engine [12]. Docker images can run unchanged on any platform that supports Docker. In Docker, containers can be created from build files such as Web service management.

The use of containers in cloud computing is increasingly becoming popular amongst cloud providers such as Google[11] and Microsoft[12]. Significant improvements in performance and security are the main driving factors for employing containers compared to virtualization using hypervisors in cloud infrastructures.

3. Cloud Security and Privacy Challenges

Cloud computing has raised several security threats such as data breaches, data loss, denial of service, and malicious insiders that have been extensively studied in [67], [68]. These threats mainly originate from issues such as multi-tenancy, loss of control over data and trust. (Explanations of these issues follow in the next subsection.)

Consequently the majority of cloud providers – including Amazon's Simple Storage Service (S3)[13], the Google Compute Engine[14] and the Citrix Cloud Platform[15] - do not guarantee specific levels of security and privacy in their service level agreements (SLAs) as part of the contractual terms and conditions between cloud providers and consumers. This means that there are important concerns related to security and privacy that must be taken into consideration in using cloud computing by all parties involved in the cloud computing arena. These are discussed in the subsection 2.2.

3.1. Security Issues in Cloud Computing

- **Multi-tenancy:** Multi-tenancy refers to sharing physical devices and virtualized resources between multiple independent users. Using this kind of arrangement means that an attacker could be on the same physical machine as the target. Cloud providers use multi-tenancy features to build infrastructures that can efficiently scale to meet customers' needs, however the sharing of resources means that it can be easier for an attacker to gain access to the target's data.

- **Loss of Control:** Loss of control is another potential breach of security that can occur where consumers' data, applications, and resources are hosted at the cloud provider's owned premises. As the users do not have explicit control over their data, this makes it possible for cloud providers to perform data mining over the users' data, which can lead

[11] Google Container Engine, https://cloud.google.com/container-engine/
[12] Microsoft Azure Container, https://azure.microsoft.com/en-us/blog/azure-container-service-now-and-the-future/
[13] Amazon S3 SLA, https://aws.amazon.com/s3/sla/
[14] Google Compute Engine SLA, https://cloud.google.com/compute/sla
[15] Citrix Cloud Platform SLA, https://www.citrix.se/products/cloudplatform/overview.html

to security issues. In addition, when the cloud providers backup data at different data centers, the consumers cannot be sure that their data is completely erased everywhere when they delete their data. This has the potential to lead to misuse of the unerased data. In these types of situations where the consumers lose control over their data, they see the cloud provider as a black-box where they cannot directly monitor the resources transparently.

- **Trust Chain in Clouds:** Trust plays an important role in attracting more consumers by assuring on cloud providers. Due to loss of control (as discussed earlier), cloud users rely on the cloud providers using trust mechanisms as an alternative to giving users transparent control over their data and cloud resources. Therefore cloud providers build confidence amongst their customers by assuring them that the provider's operations are certified in compliance with organizational safeguards and standards.

3.2. Privacy Considerations of Processing Sensitive Data

The security issues in cloud computing lead to a number of privacy concerns. Privacy is a complex topic that has different interpretations depending on contexts, cultures and communities, and it has been recognized as a fundamental human right by the United Nations [13]. It worth nothing that privacy and security are two distinct topics although security is generally necessary for providing privacy [1], [59].

Several efforts have been made to conceptualize privacy by jurists, philosophers, researchers, psychologists, and sociologists in order to give us a better understanding of privacy – for example, Alan Westin's research in 1960 is considered to be the first significant work on the problem of consumer data privacy and data protection. Westin [14] defined privacy as follow.

"*Privacy is the claim of individuals, groups, or institutions to determine for themselves
when, how, and to what extent information about them is communicated to others.*"

The International Association of Privacy Professionals (IAPP)[16] glossary 27 refers to privacy as the appropriate use of information under the circumstances. The notion of what constitutes appropriate handling of data handling varies depending on several factors such as individual preferences, the context of the situation, law, collection, how the data would be used and what information would be disclosed.

In jurisdictions such as the US, "privacy" is the term that is used to encompass the relevant laws, policies and regulations, while in the EU the term "data protection" is more commonly used when referring to privacy laws and regulations. Legislation that aims to protect the privacy of individuals – such as the European Union (EU) DPD [1], the Gramm-Leach-Bliley Act (GLBA) [15], the Right to Financial Privacy Act (RFPA) [16], and the HIPAA [2] – can become very complicated and have a variety of specific requirements. Organizations collecting and storing data in clouds that are subject to data protection regulations must ensure that the privacy of the data is preserved appropriately to lay the foundations for legal access to sensitive personal data.
The development of a legal definition for cybercrime, the issue of jurisdiction (who is responsible for what information and where are they held responsible for it) and the regulation of data

[16] IAPP Glossary, https://iapp.org/resources/glossary

transfers to third countries [17] are among other challenging issues when it comes to security in cloud computing. For example, the DPD, which is the EU's initial attempt at privacy protection, presents 72 recitals and 34 articles to harmonize the regulations for information flow within the EU Member States.

The DPD highlights the demand for cross-border transfer of data through non-legislative measures and self-control. One example of where these types of privacy principles are being used is the Safe Harbor Agreement (SHA) which makes it possible transfer data to US-based cloud providers that are assumed to have appropriate data protection mechanisms. However, cloud carriers are not subject to the SHA, which leads to complexity in respect to international laws.

There is an ongoing effort [18] to replace the EU DPD with a new a data protection regulation containing 91 articles that aims to lay out a data protection framework in Europe. The proposed regulations expand the definition of personal data protection to cover any information related to the people who are the subjects of the data, irrespective of whether the information is private, public or professional in nature. The regulations also include definitions of new roles related to handling data (such as data transfer officers) and propose restricting the transfer of data to third-party countries that do not guarantee adequate levels of protection. Currently Argentina, Canada, Guernsey, Jersey, the Isle of Man, Israel, Switzerland, the US Safe Harbor Privacy Program, and the US Transfer of Air Passenger Name Data Record are considered to offer adequate protection. The new regulations consider imposing significant penalties for privacy breaches that result from violations of the regulations, for example, such a penalty could be 0.5 percent of the worldwide annual turnover of the offending enterprise.

4. SECURITY SOLUTIONS

This section reviews the research on security solution such as authentication, authorization, and identity management that were identified in Table 2.2 [10] as being necessary so that the activities of cloud providers are sufficiently secure.

4.1 Authentication and Authorization

In [19] the authors propose a credential classification and a framework for analyzing and developing solutions for credential management that include strategies to evaluate the complexity of cloud ecosystems. This study identifies a set of categories relevant for authentication and authorization for the cloud focusing on infrastructural organization which include classifications for credentials, and adapt those categories to the cloud context. The study also summarizes important factors that need to be taken into consideration when adopting or developing a solution for authentication and authorization – for example, identifying the appropriate requirements, categories, services, deployment models, lifecycle, and entities. In other work, a design model for multi-factor authentication in cloud computing environments is proposed in [20], and this model includes an analysis of the potential security threats in the proposed model. Another authentication solution is seen with MiLAMob [21], which provides a SaaS authentication middleware for mobile consumers of IaaS cloud applications. MiLAMob is a middleware-layer that handles the real-time authentication events on behalf of consumer devices with minimal HTTP traffic. The middleware currently supports mobile consumption of data on IaaS clouds such as Amazon's S3.

FermiCloud [22] uses another approach for authentication and authorization - it utilizes public key infrastructure (PKI) X.509 certificates for user identification and authentication. FermiCloud is built in OpenNebula1[17] and it develops both X.509 authentication in Sunstone OpenNebula – a Web interface intended for user management – and X.509 authentication via command-line interfaces. To avoid the limitations of OpenNebula access control lists that are used for authorization after successful authentication of users, authors integrated an existing local credential mapping service. This solution has also been extended in cloud federations to authorize users across different cloud providers that have established trust relationships through trusted certification authorities.

Tang et al. [23] introduce collaborative access control properties such as centralized facilities, agility, homogeneity, and outsourcing trust. They have introduced an authorization-as-a-service (AaaS) approach using a formalized multi-tenancy authorization system, and providing administrative control over enhanced fine-grained trust models. Integrating trust with cryptographic role-based access control (RBAC) [24] is another solution that ensures trust for secure sharing of data in the cloud. The authors propose using cryptographic RBAC to enforce authorization policies regarding the trustworthiness of roles that are evaluated by the data owner. Another feature of the authorization system in this solution is that it develops a new concept using role inheritance for evaluating the trustworthiness of the system. In another study, Sendo et al. [25] propose a user-centric approach for platform-level authorization of cloud services using the OAuth2 protocol to allow services to act on behalf of users when interacting with other services in order to avoid sharing usernames and passwords across service.

4.2 Identity and Access Management

The important functionalities of identity management systems for the success of clouds in relation to consumer satisfaction is discussed in [26]. The authors also present an authorization system for cloud federation using Shibboleth - an open source implementation of the security assertion markup language (SAML) for single sign-on with different cloud providers. This solution demonstrates how organizations can outsource authentication and authorization to third-party clouds using an identity management system. Stihler et al. [27] also propose an integral federated identity management for cloud computing. A trust relationship between a given user and SaaS domains is required so that SaaS users can access the application and resources that are provided. In a PaaS domain, there is an interceptor that acts as a proxy to accept the user’s requests and execute them. The interceptor interacts with the secure token service (STS), and requests the security token using the WS-Trust specification.

IBHMCC [28] is another solution that contains identity-based encryption (IBE) and identity-based signature (IBS) schemes. Based on the IBE and IBS schemes, an identity-based authentication for cloud computing has been proposed. The idea is based on the identity-based hierarchical model for cloud computing along with the corresponding encryption and signature schemes without using certificates for simplified key management.

Contrail [29] is another approach that aims to enhance integration among heterogeneous clouds both vertically and horizontally. Vertical integration provides a unified platform for the different kinds of resources while horizontal integration abstracts the interaction models of different cloud providers. In [29] the authors develop a horizontal federation scheme as a requirement for vertical integration. The proposed federation architecture contains several layers, such as users’ identities,

[17] http://opennebula.org/

business logic and a federation manager to support APIs for resources, storage, and networking across different providers.

E-ID authentication and uniform access to cloud storage service providers [30] is an effort to build identity management systems for authenticating Portuguese citizens using national e-identification cards for cloud storage systems. In this approach, the OAuth protocol is integrated for authorizing the cloud users. The e-ID cards contain PKI certificates that are signed by several levels of governmental departments. A certification authority is responsible for issuing the e-ID cards and verifying them. The e-ID cards enable users for identity-based encryption of data in cloud storage.

In [31], the authors consider the issues related to inter-cloud federation and the proposed ICEMAN identity management architecture. ICEMAN discusses identity life cycle, self-service, key management, provisioning and deprovisioning functionalities that need to be included in an appropriate intercloud identity management system.

The EGI delivered a hybrid federated cloud [32] as a collaboration of communities developing, innovating, operating and using clouds for research and education. The EGI federated cloud provides IaaS, persistent block storage attached to VMs, and object-level storage for transparent data sharing. The EGI controls access to resources using X.509 certificates and the concept of "Virtual Organization" (VO). VO refers to a dynamic set of users or institutions using resource-sharing rules and conditions. The authorization attributes are issued through a VO management system that can be integrated with SAML for federation.

4.3 Confidentiality, Integrity, and Availability

Santos et al. [33] extend the Terra [34] design that enables users to verify the integrity of VMs in the cloud. The proposed solution is called the trusted cloud computing platform (TCCP), and the whole IaaS is considered to be a single system instead of granular hosts in Terra. In this approach, all nodes run a trusted virtual machine monitor to isolate and protect virtual machines. Users are given access to cloud services through the cloud manager component. The external trusted entity (ETE) is another component that provides a trust coordinator service in order to keep track of the trusted VMs in a cluster. The ETE can be used to attest the security of the VMs. A TCCP guarantees confidentiality and integrity in data and computation and it also enables users to attest to the cloud service provider to ensure whether the services are secure prior to setting up their VMs. These features are based on the trusted platform module (TPM) chip. The TPM contains a private endorsement key that uniquely identifies the TPM and some cryptographic functions that cannot be altered.

In 2011, Popa et al. proposed CloudProof [35] as a secure storage system to guarantee confidentiality, integrity and write-serializability using verifiable proofs of violation by external third parties. Confidentiality is ensured by private keys that are known only to the owner of the data that is to be encrypted. The main idea behind CloudProof is the use of the attestation mechanism. Attestations provide proof of sanity of users, data owners and cloud service providers. Data owners use a block identifier to acquire the content of a block. This mechanism enables users to store data by putting a block identifier and the contents of the block in the cloud. The attestation structure implements a solution called "block hash" for performing integrity checks through signature verification. The block hash provides proof for write-serializabilty using a forked sequence of the attestations while a chain hash is used for a broken chain of attestations which are not sequenced correctly.

Fuzzy authorization (FA) for cloud storage [36] is another flexible and scalable approach to enable data to be shared securely among cloud participants. FA ensures confidentiality, integrity and secure access control by utilizing secret sharing schemes for users with smartphones who are using the cloud services.

In [37] the authors define threats to cloud server hypervisors thorough analysis of the codebase of two popular open-source hypervisors: Xen and KVM. In addition, they discuss the vulnerabilities reports associated with them. As a result, a model is proposed for characterization of hypervisor vulnerabilities in three dimensions: the trigger source, the attack vector and the attack target. The attack vector consists of the Hypervisor functionality that makes security breaches possible - for example, virtual CPUs, symmetric multiprocessing, soft memory management units, interrupt and timer mechanisms, IO and networking, paravirtualized IO, VM exits, hypercalls, VM management (configure, start, pause and stop VMs), remote management, and software hypervisor add-ons. Successful exploitation of a vulnerability in these functionalities enables an attacker to compromise the confidentiality, integrity, or availability of the Hypervisor or one of its guest VMs.

The vulnerability reports in [37] show 59 vulnerability cases for Xen and 38 cases for KVM. Approximately 50 percent of these vulnerabilities are the same for Xen a dKVM and consist of issues relating to confidentiality, integrity and availability. The remote management software of Xen contributes to 15.3 percent of the vulnerabilities that demonstrates the increase attack surface by non-essential services. The VM management component contributes to 11.9 percent of the vulnerabilities in Xen compared to 5.3 percent in KVM. The lower vulnerability rate in KVM is due to the libvirt toolkit inside the hypervisor, whereas Xen's decision to allocate an entire privileged is done in Dom0. Other factors that have been studied in [20] are trigger sources and likely attack targets, including the overall network, the guest VM's user-space, the guest VM's kernel-space, the Dom0/host OS, and the hypervisor. The most common trigger source is the guest VM user-space, which gives rise to 39.0 percent of Xen's and 34.2 percent of KVM's vulnerabilities. This makes it possible for any user-space guest VM to be a threat to the hypervisor. The guest VM kernel-space has around 32 percent of the total vulnerabilities in both cases. The authors show Dom0 to be a more common target than the hypervisor in Xen, whereas the host OS in KVM is a less common target compared to the hypervisor. The location of the IO device emulation back-end drivers plays an important factor in this difference. The IO and network device emulation functionalities cause one third of the 15 vulnerabilities in both.

In [38] the authors propose Swap and Play as a new approach for live updating of hypervisors without the need to reboot the VM for high availability. The proposed design is scalable, usable and applicable in cloud environments and it has been implemented in Xen as one of the most popular hypervisors. Swap and Play provides methods to transfer the in-memory state of the running hypervisor to the updating state, in addition to updating the underlying host. Swap and Play consists of three independent phases: preparation, distribution and update. In the preparation phase information for the later state transfer is collected. The distribution phase deploys the update package on the target host for updating. In the last step, the update package is patched to individual hosts in the cloud. Each host applies the update package independently of the others and does not require any network resources. The Xen implementation of the Swap and Play solution is called SwapVisor. SwapVisor introduces a new hypercall in the Xen architecture. A hypercall is a trap from a domain to the hypervisor (similar to a syscall from an application to the kernel). Hypercalls are used by domains to request privileged operations such as updating page tables. The experiments show that updating from Xen version 4.2.0 to version 4.2.1 is fulfilled within approximately 45 ms which seems to be intangible and have almost zero effect on the network performance.

Klein et al. [39] improve cloud service resilience using a load-balancing mechanism called brownout. The idea behind this solution is to maximize the optional contents to provide a solution that is resilient to volatility in terms of flash crowds and capacity shortages (through load-balancing over replicas) when compared to other approaches that are implemented using response-time or queue length. In another effort [40] the authors proposed a synchronization mechanism for cloud accounting systems that are distributed. The run time resource usage generated from different clusters is synchronized to maintain a single cloud-wide view of the data so that a single bill can be created. The authors also proposed a set of accounting system requirements and an evaluation method which verifies that the solution fulfills these requirements.

4.4 Security Monitoring and Incident Response

Anand [41] presents a centralized monitoring solution for cloud applications consisting of monitoring the server, monitors, agents, configuration files and notification components. Redundancy, automatic healing, and multi-level notifications are other benefits of the proposed solution which are designed to avoid the typical drawbacks of a centralized monitoring system, such as limited scalability, low performance and single point of failure.

Brinkmann et al. [42] present a scalable distributed monitoring system for clouds using a distributed management tree that covers all the protocol-specific parameters for data collection. Data acquisition is done through specific handler implementations for each infrastructure-level data supplier. Data suppliers provide interoperability with cloud software, virtualization libraries and OS-level monitoring tools. The authors review the limitations of existing intrusion detection systems and discuss VM-level intrusion detection as an emerging area for securing VMs in cloud environments. The requirements for an efficient intrusion detection system for cloud infrastructures – including multi-tenancy, scalability and availability – are identified and a VM introspection detection mechanism via a hypervisor is proposed.

Hypervisor-based cloud intrusion detection systems are a new approach (compared to existing host-based and network-based intrusion detection systems) that is discussed in [43]. The idea is to use hypervisor capabilities to improve performance over data residing in a VM. Performance metrics are defined as networking transmitted and received data, read/write over data blocks, and CPU utilization. These metrics are retrieved in near real-time intervals by endpoint agents that are connected directly to a controller that analyzes the collected data using signatures to find any malicious activity. The controller component sends an alert to a notification service in case there is any potential attack.

4.5 Security Policy Management

In [44] the authors propose a generic security management framework allowing providers of cloud data management systems to define and enforce complex security policies through a policy management module. The user activities are stored and monitored for each storage system, and are made available to the policy management module. Users' actions are evaluated by a trust management module based on their past activities and are grouped as "fair" or "malicious". An appropriate architecture for security management which satisfies the requirements of policy definitions (such as flexibility, expressiveness, extendibility and correctness) has been implemented. The authors evaluated the proposed system on a data management system that is built on data storage.

Takabi et al. [45] introduce policy management as a service (PMaaS) to provide users with a unified control point for managing access policies in order to control access to cloud resources independently of the physical location of cloud providers. PMaaS is designed specifically to solve the issue of having multiple access control authorization mechanisms employed by cloud service providers that restrict the flexibility of applying custom access control to a particular service. For this purpose, the PMaaS architecture includes a policy management service provider that is the entry point for cloud users to define and manage the policies. The cloud service provider imports the user-defined policies and acts a policy decision point to enforce the user policies.

The challenges associated with policy enforcement in heterogeneous distributed environments are discussed in [46]. The authors propose a framework to support flexible policy enforcement and a feedback system using rule- and context-based access control to inform cloud users about the effect of defined policies. There are three main requirements for building a general policy enforcement framework. First it must support various data types such as image, structured and textual data. Secondly, in a distributed environment there need to be several compute engines such as Map/Reduce, relational database management systems or clusters. Finally, access policy requirements in terms of access control policies, data sharing policies, and privacy policies need to be integrated with the general policy management framework. Several policy enforcement mechanisms (such as extensible access control markup language or inline-reference monitors to enforce user-centric policies in accord with cloud provider approval) were also discussed.

In [47] the authors describe A4Cloud with the aim of developing solutions to ensure accountability and transparency in cloud environments. Users need to be able to track their data usage to know how the cloud provider satisfies their expectations for data protection. For this purpose cloud providers must employ solutions that provide users with appropriate control and transparency over their data, e.g. tools to define policies for compliance with regulatory frameworks. In another effort [48] the authors discuss the issue of usable transparent data processing in cloud computing and also consider how to enable users to define transparency policies over their data. They identify the requirements for transparent policy management in the cloud based on two aspects: user demands and legal aspects of transparent data processing.

5. Privacy-Preservation For Sensitive Data in Cloud Computing

Over the time, organizations have collected valuable information about the individuals in our societies that contain sensitive information, e.g. medical data. Researchers need to access and analyze such data using big data technologies [63], [64], [65] in cloud computing, while organizations are required to enforce data protection compliance (subsection 3.2).

There has been considerable progress on privacy preservation for sensitive data in both industry and academia, e.g., solutions that develop protocols and tools for anonymization or encryption of data for confidentiality purposes. This section categorizes work related to this area according to different privacy protection requirements. However, these solutions have not yet been widely adopted by cloud service providers or organizations.

Pearson [1] discusses a range of security and privacy challenges that are raised by cloud computing. Lack of user control, lack of training and expertise, unauthorized secondary usage, complexity of regulatory compliance, transborder data flow restrictions and litigation are among the challenges faced in cloud computing environments. In [66], the authors describe the privacy challenges of genomic data in the cloud including terms of services of cloud providers that are not developed with a healthcare mindset, awareness of patient to upload their data into the cloud without their consent, multi-tenancy, data monitoring, data security and accountability. The authors also provide recommendations for data owners when aiming to use cloud provider services.

In [49] the authors discussed several privacy issues associated with genomic sequencing. This study also described several open research problems (such as outsourcing to cloud providers, genomic data encryption, replication, integrity, and removal of genomic data) along with giving suggestions to improve privacy through collaboration between different entities and organizations. In another effort [50], raw genomic data storage through encrypted short reads is proposed.

Outsourcing privacy is another topic that is discussed in [51]. The authors define the concept of "outsourcing privacy" where a database owner updates the database over time on untrusted servers. This definition assumes that database clients and the untrusted servers are not able to learn anything about the contents of the databases without authorized access. The authors implements a server-side indexing structure to produce a system that allows a single database owner to privately and efficiently write data to, and multiple database clients to privately read data from, an outsourced database.

Homomorphic encryption is another privacy-preserving solution that is based on the idea of computing over encrypted data without knowing the keys belonging to different parties. To ensure confidentiality, the data owner may encrypt data with a public key and store data in the cloud. When the process engine reads the data, there is no need to have the DP's private key to decrypt the data. In private computation on encrypted genomic data [52], the authors proposed a privacy-preserving model for genomic data processing using homomorphic encryption on genome-wide association studies.

Anonymization is another approach to ensure the privacy of sensitive data. SAIL [53] provides individual-level information on the availability of data types within a collection. Researchers are not able to cross-link (which is similar to an equality join in SQL) data from different outside studies, as the identities of the samples are anonymized. In another effort [57] the authors propose an integration architecture to make it possible to perform aggregated queries over anonymized medical data sets from different data providers. In this solution, data providers remove the data subjects' identifiers and apply a two-level encryption using hashing and PKI certificates. The sensitive information will then be anonymized using an open-source toolkit and will be encrypted granularly using the cloud provider's public key. ScaBIA [60] is another solution for processing and storing anonymized brain imaging data in cloud. This approach provides PKI authentication for administrator roles to deploy a PaaS middleware and defines researchers as users in the in Microsoft Azure cloud. Researchers are allowed to login by username/password to run statistical parametric mapping workflows within isolated generic worker containers. The brain imaging datasets and related results can be shared by the researchers using a RBAC model over secure HTTPS connections.

In [54], the design and implementation of a security framework for BiobankCloud, a platform that supports the secure storage and processing of genomic data in cloud computing environments, has been discussed. The proposed framework is built on the cloud privacy threat modeling approach [55], [56] which is used to define the privacy threat model for processing next-generation sequencing data according to the DPD [2]. This solution includes a flexible two-factor authentication and an RBAC access control mechanism, in addition to auditing mechanisms to ensure that the requirements of the DPD are fulfilled.

6. CONCLUSIONS

This paper surveyed recent advances in cloud computing security and privacy research. It described several cloud computing key concepts and technologies, such as virtualization, and containers. We also discussed several security challenges that are raised by existing or forthcoming privacy legislation, such as the EU DPD and the HIPAA.

The results that are presented in the area of cloud security and privacy are based on cloud provider activities, such as providing orchestration, resource abstraction, physical resource and cloud service management layers. Security and privacy factors that affect the activities of cloud providers in relation to the legal processioning of consumer data were identified and a review of existing research was conducted to summarize the state-of-the-art in the field.

ACKNOWLEDGEMENTS

This work is funded by the EU FP7 project "Scalable, Secure Storage and Analysis of Biobank Data" under Grant Agreement no. 317871.

REFERENCES

[1] S. Pearson, "Privacy, security and trust in cloud computing," in Privacy and Security for Cloud Computing (S. Pearson and G. Yee, eds.), Computer Communications and Networks, pp. 3–42, Springer London, 2013.

[2] E. U. Directive, “95/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of such Data,” Official Journal of the EC, vol. 23, 1995.

[3] U. States., “Health insurance portability and accountability act of 1996 [micro form] : conference report (to accompany h.r. 3103).” http://nla.gov.au/nla.catvn4117366, 1996.

[4] “Hypervisors, virtualization, and the cloud: Learn about hypervisors, system virtualization, and how it works in a cloud environment.” Retrieved June 2015.

[5] M. Portnoy, Virtualization Essentials. 1st ed., 2012.Alameda, CA, USA: SYBEX Inc.,

[6] P. Mell and T. Grance, “The NIST Definition of Cloud Computing,” tech. rep., July 2009.

[7] F. Liu, J. Tong, J. Mao, R. Bohn, J. Messina, L. Badger, and D. Leaf, NIST Cloud Computing Reference Architecture: Recommendations of the National Institute of Standards and Technology (Special Publication 500-292). USA: CreateSpace Independent Publishing Platform, 2012.

[8] R. Dua, A. Raja, and D. Kakadia, “Virtualization vs containerization to support paas,” in Cloud Engineering (IC2E), 2014 IEEE International Conference on, pp. 610–614, March 2014.

[9] D. Bernstein, "Containers and Cloud: From LXC to Docker to Kubernetes," IEEE Cloud Computing, vol. 1, no. 3, pp. 81-84, 2014.

[10] NIST Special Publication 500–291 version 2, NIST Cloud Computing Standards Roadmap, July 2013, Available at http://www.nist.gov/itl/cloud/publications.cfm.

[11] R. Pike, D. Presotto, K. Thompson, H. Trickey, and P. Winterbottom, “The use of name spaces in plan 9,” SIGOPS Oper. Syst. Rev., vol. 27, pp. 72–76, Apr. 1993.

[12] B. Russell, “Realizing Linux Containers (LXC).” http://www.slideshare.net/BodenRussell/linux-containers-next-gen- virtualization-for-cloud-atl-summit-ar4-3-copy. Retrieved October 2015.

[13] United Nations, “The Universal Declaration of Human Rights.” http://www.un.org/en/documents/udhr/index.shtml, 1948. Retrieved August 2015.

[14] A. Westin, Privacy and Freedom. New Jork Atheneum, 1967.

[15] U. States., “Gramm-leach-bliley act.” http://www.gpo.gov/fdsys/pkg/PLAW-106publ102/pdf/PLAW-106publ102.pdf, November 1999.

[16] U. S. F. Law, “Right to financial https://epic.org/privacy/rfpa/, 1978. privacy act of 1978.”

[17] D. Bigo, G. Boulet, C. Bowden, S. Carrera, J. Jeandesboz, and A. Scherrer, “Fighting cyber crime and protecting privacy in the cloud.” European Parliament, Policy Department C: Citizens’ Rights and Constitutional Affairs, October 2012.

[18] S. Stalla-Bourdillon, “Liability exemptions wanted! internet intermediaries’ liability under uk law,” Journal of International Commercial Law and Technology, vol. 7, no. 4, 2012.

[19] N. Mimura Gonzalez, M. Torrez Rojas, M. Maciel da Silva, F. Redigolo, T. Melo de Brito Carvalho, C. Miers, M. Naslund, and A. Ahmed, “A framework for authentication and authorization credentials in cloud computing,” in Trust, Security and Privacy in Computing and Communications (TrustCom), 2013 12th IEEE International Conference on, pp. 509–516, July 2013.

[20] R. Banyal, P. Jain, and V. Jain, “Multi-factor authentication framework for cloud computing,” in Computational Intelligence, Modelling and Simulation (CIMSim), 2013 Fifth International Conference on, pp. 105–110, Sept 2013.

[21] R. Lomotey and R. Deters, “Saas authentication middleware for mobile consumers of iaas cloud,” in Services (SERVICES), 2013 IEEE Ninth World Congress on, pp. 448–455, June 2013.

[22] H. Kim and S. Timm, “X.509 authentication and authorization in fermi cloud,” in Utility and Cloud Computing (UCC), 2014 IEEE/ACM 7th International Conference on, pp. 732–737, Dec 2014.

[23] B. Tang, R. Sandhu, and Q. Li, “Multi-tenancy authorization models for collaborative cloud services,” in Collaboration Technologies and Systems (CTS), 2013 International Conference on, pp. 132–138, May 2013.

[24] L. Zhou, V. Varadharajan, and M. Hitchens, “Integrating trust with cryptographic role-based access control for secure cloud data storage,” in Trust, Security and Privacy in Computing and Communications (TrustCom), 2013 12th IEEE International Conference on, pp. 560–569, July 2013.

[25] J. Sendor, Y. Lehmann, G. Serme, and A. Santana de Oliveira, “Platform level support for authorization in cloud services with oauth 2,” in Proceedings of the 2014 IEEE International

Conference on Cloud Engineering, IC2E '14, (Washington, DC, USA), pp. 458–465, IEEE Computer Society, 2014.

[26] M. A. Leandro, T. J. Nascimento, D. R. dos Santos, C. M. Westphall, and C. B. Westphall, "Multi-tenancy authorization system with federated identity for cloud-based environments using shibboleth," in Proceedings of the 11th International Conference on Networks, ICN 2012, pp. 88–93, 2012.

[27] M. Stihler, A. Santin, A. Marcon, and J. Fraga, "Integral federated identity management for cloud computing," in New Technologies, Mobility and Security (NTMS), 2012 5th International Conference on, pp. 1–5, May 2012.

[28] H. Li, Y. Dai, L. Tian, and H. Yang, "Identity-based authentication for cloud computing," in Cloud Computing (M. Jaatun, G. Zhao, and C. Rong, eds.), vol. 5931 of Lecture Notes in Computer Science, pp. 157–166, Springer Berlin Heidelberg, 2009.

[29] E. Carlini, M. Coppola, P. Dazzi, L. Ricci, and G. Righetti, "Cloud federations in contrail," in Euro-Par 2011: Parallel Processing Workshops (M. Alexander,P. D'Ambra, A. Belloum, G. Bosilca, M. Cannataro, M. Danelutto, B. Di Mar tino, M. Gerndt, E. Jeannot, R. Namyst, J. Roman, S. Scott, J. Traff, G. Vallée, and J. Weidendorfer, eds.), vol. 7155 of Lecture Notes in Computer Science, pp. 159–168, Springer Berlin Heidelberg, 2012.

[30] J. Gouveia, P. Crocker, S. Melo De Sousa, and R. Azevedo, "E-id authentication and uniform access to cloud storage service providers," in Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference on, vol. 1, pp. 487–492, Dec 2013.

[31] G. Dreo, M. Golling, W. Hommel, and F. Tietze, "Iceman: An architecture for secure federated inter-cloud identity management," in Integrated Network Management (IM 2013), 2013 IFIP/IEEE International Symposium on,pp. 1207–1210, May 2013.

[32] G. Sipos, D. Scardaci, D. Wallom, and Y. Chen, "The user support programme and the training infrastructure of the egi federated cloud," in High Performance Computing Simulation (HPCS), 2015 International Conference on, pp. 9–18, July 2015.

[33] N. Santos, K. P. Gummadi, and R. Rodrigues, "Towards trusted cloud computing," in Proceedings of the 2009 Conference on Hot Topics in Cloud Computing, HotCloud'09, (Berkeley, CA, USA), USENIX Association, 2009.

[34] T. Garfinkel, B. Pfaff, J. Chow, M. Rosenblum, and D. Boneh, "Terra: A virtual machine-based platform for trusted computing," in Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles, SOSP '03, (New York, NY, USA), pp. 193–206, ACM, 2003.

[35] R. A. Popa, J. R. Lorch, D. Molnar, H. J. Wang, and L. Zhuang, "Enabling security in cloud storage slas with cloudproof," in Proceedings of the 2011 USENIX Conference on USENIX Annual Technical Conference, USENIX ATC'11, (Berkeley, CA, USA), pp. 31–31, USENIX Association, 2011.

[36] S. Zhu and G. Gong, "Fuzzy authorization for cloud storage," Cloud Computing, IEEE Transactions on, vol. 2, pp. 422–435, Oct 2014.

[37] D. Perez-Botero, J. Szefer, and R. B. Lee, "Characterizing hypervisor vulnerabilities in cloud computing servers," in Proceedings of the 2013 International Workshop on Security in Cloud Computing, Cloud Computing '13, (New York, NY, USA), pp. 3–10, ACM, 2013.

[38] F. F. Brasser, M. Bucicoiu, and A.-R. Sadeghi, "Swap and play: Live updating hypervisors and its application to xen," in Proceedings of the 6th Edition of the ACM Workshop on Cloud Computing Security, CCSW '14, (New York, NY, USA), pp. 33–44, ACM, 2014.

[39] C. Klein, A. Papadopoulos, M. Dellkrantz, J. Durango, M. Maggio, K.-E. Arzen, F. Hernandez-Rodriguez, and E. Elmroth, "Improving cloud service re silience using brownout-aware load-balancing," in Reliable Distributed Systems (SRDS), 2014 IEEE 33rd International Symposium on, pp. 31–40, Oct 2014.

[40] E. Lakew, L. Xu, F. Hernandez-Rodriguez, E. Elmroth, and C. Pahl, "A synchronization mechanism for cloud accounting systems," in Cloud and Autonomic Computing (ICCAC), 2014 International Conference on, pp. 111–120, Sept 2014.

[41] M. Anand, "Cloud monitor: Monitoring applications in cloud," in Cloud Computing in Emerging Markets (CCEM), 2012 IEEE International Conference on, pp. 1–4, Oct 2012.

[42] A. Brinkmann, C. Fiehe, A. Litvina, I. Lück, L. Nagel, K. Narayanan, F. Ostermair, and W. Thronicke, “Scalable monitoring system for clouds,” in Proceedings of the 2013 IEEE/ACM 6th International Conference on Utility and Cloud Computing, UCC ’13, (Washington, DC, USA), pp. 351–356, IEEE Computer Society, 2013.

[43] J. Nikolai and Y. Wang, “Hypervisor-based cloud intrusion detection system,” in Computing, Networking and Communications (ICNC), 2014 International Conference on, pp. 989–993, Feb 2014.

[44] C. Basescu, A. Carpen-Amarie, C. Leordeanu, A. Costan, and G. Antoniu, “Managing data access on clouds: A generic framework for enforcing security policies,” in Advanced Information Networking and Applications (AINA), 2011 IEEE International Conference on, pp. 459–466, March 2011.

[45] H. Takabi and J. Joshi, “Policy management as a service: An approach to manage policy heterogeneity in cloud computing environment,” in System Science (HICSS), 2012 45th Hawaii International Conference on, pp. 5500–5508, Jan 2012.

[46] K. W. Hamlen, L. Kagal, and M. Kantarcioglu, “Policy enforcement framework for cloud data management.,” IEEE Data Eng. Bull., vol. 35, no. 4, pp. 39–45, 2012.

[47] S. Pearson, V. Tountopoulos, D. Catteddu, M. Sudholt, R. Molva, C. Reich, S. Fischer-Hubner, C. Millard, V. Lotz, M. Jaatun, R. Leenes, C. Rong, and J. Lopez, “Accountability for cloud and other future internet services,” in Cloud Computing Technology and Science (CloudCom), 2012 IEEE 4th International Conference on, pp. 629–632, Dec 2012.

[48] S. Fischer-Hubner, J. Angulo, and T. Pulls, “How can cloud users be supported in deciding on, tracking and controlling how their data are used?,” in Privacy and Identity Management for Emerging Services and Technologies (M. Hansen, J.-H. Hoepman, R. Leenes, and D. Whitehouse, eds.), vol. 421 of IFIP Advances in Information and Communication Technology, pp. 77–92, Springer Berlin Heidelberg, 2014.

[49] E. Ayday, J. Raisaro, U. Hengartner, A. Molyneaux, and J.-P. Hubaux, “Privacy-preserving processing of raw genomic data,” in Data Privacy Management and Autonomous Spontaneous Security (J. Garcia-Alfaro, G. Lioudakis, N. Cuppens-Boulahia, S. Foley, and W. M. Fitzgerald, eds.), vol. 8247 of Lecture Notes in Computer Science, pp. 133147, Springer Berlin Heidelberg, 2014.

[50] E. Ayday, E. D. Cristofaro, J.-P. Hubaux and G. Tsudik "The chills and thrills of whole genome sequencing", Computer, vol. 99, pp.1, 2013.

[51] Y. Huang and I. Goldberg, “Outsourced private information retrieval,” in Proceedings of the 12th ACM Workshop on Workshop on Privacy in the Electronic Society, WPES ’13, (New York, NY, USA), pp. 119–130, ACM, 2013.

[52] K. Lauter, A. Lopez-Alt, and M. Naehrig, “Private computation on encrypted genomic data,” Tech. Rep. MSR-TR-2014-93, June 2014.

[53] M. Gostev, J. Fernandez-Banet, J. Rung, J. Dietrich, I. Prokopenko, S. Ripatti, M. I. McCarthy, A. Brazma, and M. Krestyaninova, “SAIL - a software system for sample and phenotype availability across biobanks and cohorts,” Bioinformatics , vol. 27, no. 4, pp. 589591, 2011.

[54] A. Gholami and E. Laure, “Advanced cloud privacy threat modeling,” The Fourth International Conference on Software Engineering and Applications (SEAS-2015), to be published in Computer Science Conference Proceedings in Computer Science and Information Technology (CS/IT) series.

[55] A. Gholami, J. Dowling, and E. Laure, “A security framework for population-scale genomics analysis,” in High Performance Computing Simulation (HPCS), 2015 International Conference on, pp. 106–114, July 2015.

[56] A. Gholami, A.-S. Lind, J. Reichel, J.-E. Litton, A. Edlund, and E. Laure, “Privacy threat modeling for emerging biobankclouds,” Procedia Computer Science, vol. 37, no. 0, pp. 489 – 496, 2014. The 5th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN-2014)/The 4th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH 2014)/ Affiliated Work- shops.

[57] A. Gholami, E. Laure, P. Somogyi, O. Spjuth, S. Niazi, and J. Dowling, “Privacy-preservation for publishing sample availability data with personal identifiers,” Journal of Medical and Bioengineering, vol. 4, pp. 117–125, April 2014.

[58] C. Wang, Q. Wang, K. Ren, and W. Lou, “Privacy-preserving public auditing for data storage security in cloud computing,” in Proceedings of the 29th Conference on Information Communications, INFOCOM’10, (Piscataway, NJ, USA), pp. 525–533, IEEE Press, 2010.

[59] A. Cavoukian, The Security-Privacy Paradox: Issues, misconceptions, and Strategies. https://www.ipc.on.ca/images/Resources/sec-priv.pdf, Retrieved November 2015.

[60] A. Gholami, G. Svensson, E. Laure, M. Eickhoff, and G. Brasche, “Scabia: Scalable Brain Image Analysis in the Cloud,” in CLOSER 2013 - Proceedings of the 3rd International Conference on Cloud Computing and Services Science, Aachen, Germany, 8-10 May, 2013, pp. 329–336, 2013.

[61] S. Sharma, “Evolution of as-a-service era in cloud,” CoRR, vol. abs/1507.00939, 2015.

[62] S. Sharma, U. S. Tim, J. Wong, S. Gadia, “Proliferating Cloud Density through Big Data Ecosystem, Novel XCLOUDX Classification and Emergence of as-a-Service Era,” 2015

[63] S. Sharma, U. S. Tim, J. Wong, S. Gadia, S. Sharma, “A Brief Review on Leading Big Data Models,” Data Science Journal, 13(0), 138-157. 2014.

[64] S. Sharma, U. S. Tim, J. Wong, S. Gadia, R. Shandilya, S. K. Peddoju, “Classification and comparison of NoSQL big data models,” International Journal of Big Data Intelligence (IJBDI), Vol. 2, No. 3, 2015.

[65] S. Sharma, R. Shandilya, S. Patnaik, A. Mahapatra, “Leading NoSQL models for handling Big Data: a brief review,” International Journal of Business Information Systems, Inderscience, 2015.

[66] Dove, E. S, Y. Joly, A.-M. Tassé, P. P. P. in Genomics, S. P. I. S. Committee, I. C. G. C. I. Ethics, P. Committee, and B. M Knoppers, “Genomic cloud computing: legal and ethical points to consider,” European Journal of Human Genetics, August 2014.

[67] Cloud Security Alliance (CSA), “Security Guidance for Critical Areas of Focus in Cloud Computing” version 3, 2011. Available at: https://cloudsecurityalliance.org/guidance/csaguide.v3.0.pdf

[68] Cloud Security Alliance (CSA), “The Notorious Nine: Cloud Computing Top Threats in 2013”. Available at: https://cloudsecurityalliance.org.

AUTHORS

Ali Gholami is a PhD student at the KTH Royal Institute of Technology. His research interests include the use of data structures and algorithms to build adaptive data management systems. Another area of his research focuses on the security concerns associated with cloud computing. He is currently exploring strong and usable security factors to enable researchers to process sensitive data in the cloud.

Professor Erwin Laure is Director of the PDC - Center for High Performance Computing Center at KTH, Stockholm. He is the Coordinator of the EC-funded "EPiGRAM" and "ExaFLOW" projects as well as of the HPC Centre of Excellence for Bio-molecular Research "BioExcel" and actively involved in major e-infrastructure projects (EGI, PRACE, EUDAT) as well as exascale computing projects. His research interests include programming environments, languages, compilers and runtime systems for parallel and distributed computing, with a focus on exascale computing.

SECURITY ANALYSIS OF MOBILE AUTHENTICATION USING QR-CODES

Siwon Sung[1,2], Joonghwan Lee[1,2], Jinmok Kim[1,2], Jongho Mun[2] and
Dongho Won[2]

[1]Samsung Electronics, Suwon, Korea
{siwon.sung, joonghwan.lee, jinmok.kim}@samsung.com
[2]College of Information and Communication Engineering,
Sungkyunkwan University, Suwon, Korea
{jhmoon, dhwon}@security.re.kr

ABSTRACT

The QR-Code authentication system using mobile application is easily implemented in a mobile device with high recognition rate without short distance wireless communication support such as NFC. This system has been widely used for physical authentication system does not require a strong level of security. The system also can be implemented at a low cost. However, the system has a vulnerability of tampering or counterfeiting, because of the nature of the mobile application that should be installed on the user's smart device. In this paper we analyze the vulnerabilities about each type of architectures of the system and discuss the concerns about the implementation aspect to reduce these vulnerabilities.

KEYWORDS

Authentication, Security, QR-Code, Mobile

1. INTRODUCTION

The one-dimensional barcodes have been used in limited areas such as identifier of goods due to the limitation of information capacity. The two-dimensional barcode has been widely used in various fields because of great increase of the information capacity. Thus the user authentication system using the QR-Code becomes possible. Therefore the widely use of smart devices has promoted to spread the QR-Code authentication system. Users could generate the QR-Code image for authentication easily through mobile application on their smart devices. The system is mainly used in a low cost authentication system such as gate access control in buildings and unmanned book rental system in libraries. However, this system is vulnerable to attempts to disguise access to someone else's identity because of the nature of easily replicable QR-Code. In case of mobile device, it is very easy to duplicate through screen capture and be transferred to other user via instant messenger. Thus it is very difficult to eradicate these user violations because of the limitations by those characteristics. In order to overcome these vulnerabilities, various considerations such as security policy, authentication protocols, network security are required. In addition, the system needs also to consider reverse engineering because the vulnerable attributes of the mobile application. The reverse engineering can occur on the various components to configure the system. Particularly the mobile application running on the smart

David C. Wyld et al. (Eds) : NETCOM, NCS, WiMoNe, CSEIT, SPM - 2015
pp. 151–160, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51612

device platform is extremely vulnerable to reverse engineering. Even if an application use very secure authentication scheme, the security might not meet the security levels by lower implementation maturity. Whereas the system uses vulnerable methods, fine implementation recommendations help to prevent those vulnerabilities. In this paper, we suggest some recommendations in implementation points of view to prevent those threats. The remainder of the paper is organized as follows. In Section 2, related works are presented. We introduce the general architecture of the QR-Code authentication system using mobile application in Section 3. The attacker models are described in Section 4. The vulnerabilities in the software implementation point of the view are discussed in Section 5. The result of the forgery attack on real system is reported in Section 6. Then the countermeasure to reduce this software threats is given in Section 7. Finally, we conclude this paper in Section 8.

2. RELATED WORK

QR-TANs is a transaction authentication technique based on QR-Code, allow the user to directly validate the content of a transaction within a trusted device [1]. *Liao et al.* [2] proposed a scheme based on two-dimensional barcode not only eliminates the usage of the password verification table, but also is a cost effective solution. *Lee et al.* [3] proposed an authentication system used Mobile OTP with the combination of QR-Code. *Oh et al.* [4] suggests a technology to authorize users three kinds of QR-Code. *Kao et al.* [5] try to develop a safe and efficient authentication way by using mobile device and implement it for access control system that enhances the security of physical access control systems. *2CAuth* is a two factor authentication scheme that enhances secure usage of application information and preserves usability, without sacrificing user's privacy [6]. *Lee et al.* designed secured *QR-Login* user verification protocol for smart devices that are ready to communicate with QR-Code and proposed a way to keep critical data safe when using the Internet [7]. *Kale et al.* proposed a anti phishing *single sign-on* authentication model using QR-Code [8]. *Malik et al.* introduced the idea of a *one-time password*, which makes unauthorized access difficult for unauthorized users [9].

3. ARCHITECTURE OF AUTHENTICATION SYSTEM

The QR-Code authentication system using mobile application (*BAS-MA*) consists of five participants: the users U, the mobile application *MA*, QR-Code generator *QG*, the service provider *SP* and authentication center *AC*. We summarize the notations and acronyms used in this paper in Table 1.

Table 1. Notations and Acronyms

U_i	User
C_i	User credentials for U_i
B_i	QR-Code for authentication issued to U_i
QG	QR-Code generator
SP	Service provider
AC	Authentication center
BAS-MA	QR-Code authentication system using mobile application
AC-AR	Authentication center based architecture

MA-AR	Mobile application center based architecture
A_{it}	Internal attacker
A_{ex}	External attacker

BAS-MA is classified into two kinds of architectures based on the location of the *QG* :

- Authentication Center based Architecture
- Mobile Application based Architecture

In the *AC-AR*, *QG* is located in *AC* that could be regarded as a secure server in the remote location. The basic configuration of *AC-AR* is shown in Figure 1. The procedure through which a U_i can get a valid access to *SP* is the following:

1. U_i is willing to use the *SP*. U_i get authentication through the user authentication scheme of the system.
2. *MA* requires B_i for authentication to *AC*.
3. *AC* retrieves user data from *User Data Storage* then calculates C_i. *AC* asks the *QG* a B_i carrying C_i.
4. *QG* transform the C_i to B_i without any modification on C_i and returns a B_i.
5. *AC* passes through the B_i to *MA* via secure channel.
6. *MA* displays the B_i on its screen and submits to *SP*.
7. *SP* decodes the B_i and extracts C_i. *SP* asks the *AC* the verification carrying the C_i.
8. *AC* verifies C_i and return with an authentication callback.

The B_i is generated in the *AC* on remote. The *MA* has just downloaded the image B_i from *AC* and displays B_i on its screen. The *MA* does not have any logic for generating B_i or processing C_i and is just provided with a bitmap image. If the *MA* is authorized to download the B_i, the C_i can be secured to the reversing attack because they are created and managed on *AC* in remote.

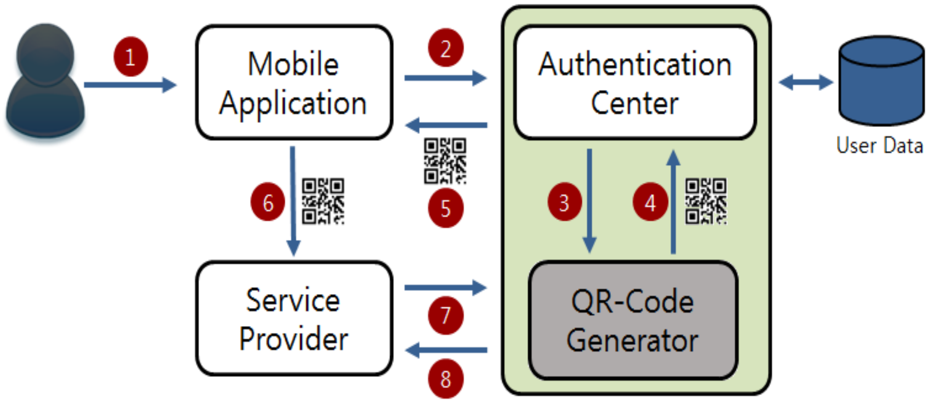


Figure 1. Basic procedure of Authentication Center based Architecture

Otherwise, the *QG* is associated with *MA* tightly in *MA-AR*. The *MA* includes *QG* itself or is able to interact with the external *QG* in the same device. Figure 2 shows the operation flows that the

MA is granted the permission for *SP*. The procedure on *MA-AR* is different to that of the *AC-AR*, yet it has the following differences:

1. U_i try to get a permission to access *SP*.
2. *MA* requests a C_i to generate B_i.
3. *AC* assemble C_i using the *User Data Storage* and return to *MA*.
4. *MA* askes *QG* existed in mobile device a B_i carrying C_i.
5. *QG* generate B_i containing C_i then pass it to *MA*.
6. *MA* gives the B_i to *SP*.
7. *SP* open the B_i and find C_i then *SP* request a verification to the *AC* with the C_i.
8. *AC* checks the equality of the C_i and response.

The *MA* is responsible to manage the C_i in secure and to generate the B_i itself. The *QG* is exposed to threat to get to be attempt software attack.

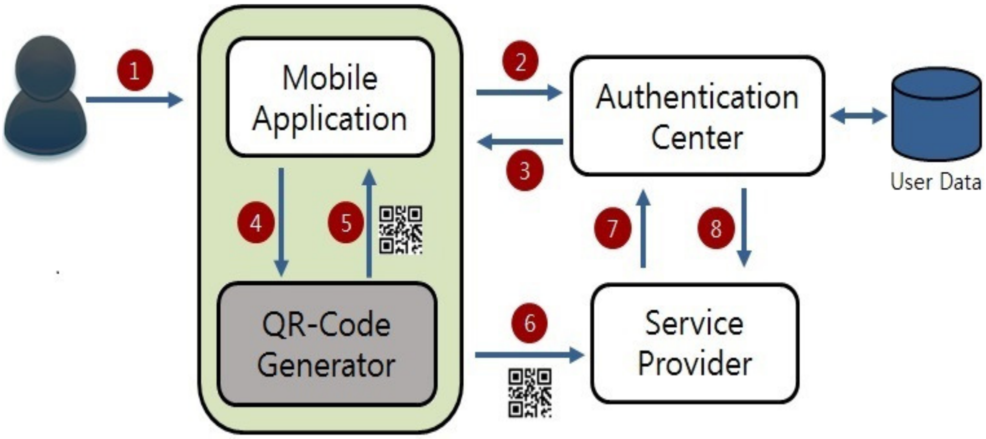


Figure 2. Basic procedure of Mobile Application based Architecture

4. ATTACK MOTELS

The attackers who threaten the *BAS-MA* can be divided into two groups:

4.1. The Internal Attacker

The internal attacker A_{it} is a member of the organization provided services by *SP* and has the correct permissions can access the *SP*. A_{it} may be a malicious user himself willing to access to *SP* by the identity of the other user U_j . The A_{it} also tries to let an unauthorized person to get the authentication to the system by sharing his identity. The A_{it} can try various attacks on the system because they know very well about the *BAS-MA*, and can examine the *MA* on their mobile devices. A_{it} is able to perform repeated incorrect attempts to authentication system with less doubt of the system administrator.

4.2. The External Attacker

Otherwise, the external attacker A_{ex} is not participated in the organization. A_{ex} is the attacker would like to acquire authentication to the system even if they don't have any right permissions. Depending on the mobile application distribution policy, the A_{ex} may have difficulties to obtain the installable packages or binary of the mobile applications. For that reason, A_{ex} distribute a malware to find vulnerabilities in order to steal someone's C_i and B_i.

5. VULNERABILITY ANALYSIS

5.1. QR-Code Cloning

In the *BAS-MA*, the B_i is the key object to achieve the authentication. However, the B_i could be cloned very easily on *MA* through screen capture. The A_{it} might cause the masquerade attack transferring their image to others via an instant messenger. The A_{ex} also try to get the copies of the B_i. In fact, it is impossible to completely prevent such cloning because of the nature of the QR-Code that should be exposed on the screen. The A_{ex} could even take a picture of the B_i using a separate camera.

5.2. Authentication Hijacking

The permanent authentication such as automatic login is usually enabled for the convenience use after once successful authentication. Even if the authentication protocol do not support long term session, the developer tries to implement virtual session by background authentication. In order to do it, the keeping of the user authentication cookies such as user's identity and password in the internal cache are required. These important data in the cache can be easily exposed by the software attack. The attackers are able to hijack the user authentication through the cached data.

5.3. Stored Data Exfiltration

The C_i is stored in the non-volatile storage area under *MA* in the *MA-AR*. The attackers can extract the values through software attack. If the device is rooted, the attack is easier to occur. Because of that, it is recommended that the value is not stored permanently or data should be stored encrypted. Once the encryption key is exposed, the protection using cryptography is useless. After that, the attackers are able to generate the B_i without big difficulty. The data in local storage is very vulnerable. In rooted device, the malware get the read permission under other application's control. The attacker also can access the file system on the device directly or inspect the runtime memory.

5.4. Internal Algorithm Disclosure

The developers usually assume that the compiled source code will be safe from the logic disclosure. However, the powerful tools such as decompiler and logic analyzer are able to disclose the internal logic. The decompiler is able to recover the source codes using the binaries. The logic analyzer can draw the flow graph to help the understanding the logic. It can expose which cryptographic algorithms are used and how to assemble the parameters to construct credentials. In case of the use of the original designed two-dimensional barcode formats for the security purpose, the algorithms should not be disclosed and the software modules for the algorithms should not to be exposed and extracted for other application. If it is possible, the attackers can produce forgery algorithms to perform same work.

5.5. Network Message Eavesdropping

The attackers have various methods to intercept the messages on the mobile device. The general purpose smart devices can be installed any application without any restrictions. The A_{it} could install a network packet monitoring tool on their mobile device to eavesdrop the messages between *MA* and the *AC*. While the mobile platform operates under some policies that restrict the packet monitoring permissions, the A_{it} can attempt to bypass the restriction by using rootkit. In rooted device, all application could obtain root permission.

5.6. IPC Message Eavesdropping

In the *MA-AR*, the *QG* can be located in separated module from *MA*. This is the case in order to use the common module to generate barcode provided by the platform. In this case, the *IPC* (*Inter Process Communication*) communication is generally used. If the *IPC* channel is not secure, messages C_i through *IPC* can be exposed to malware. For example, the *Broadcast* is popularly used in *Android* platform to pass the values to other application. However the implicit *broadcast message* could be broadcasted to all the application in the device not excluding malwares.

5.7. Communication Protocol Vulnerabilities

The secure protocols such as *SSL/TLS* prevent the leakage of the network packet. In this case, the attacker can manipulate the victim's network packet to pass through the malicious *HTTP Proxy* to incapacitate the secure communication. The *HTTP Proxy* can perform the *man-in-the-middle attack* to obtain the secrete key of the channel. *Callegati et al.* presented the *man-in-the middle attack* to the *HTTPS* through *ARP Spoofing* [10]. A_{ex} also can install rogue access point in the near place that the authentication communication usually happened. Because the authentication procedure has occurred in fixed places such as the gateway of the buildings or in front of the unmanned machines, A_{ex} could intercept the communications to proceed authentication through the rogue access point. *Marlinspike et al.* showed *SSL Strip Attack* that do a *man-in-the-middle attack* on *SSL* connections [11].

The Internal logic for communication in the software should be hidden as much as possible. The address of the destination server, *Restful API* names, the configuration of the parameters and the keyword for the message format are usually hard-coded inside mobile applications. The skilful attacker can reconstruct the protocol completely using a piece of this information.

6. FORGERY ATTACK RESULTS

We performed a forgery attack based on the discussed vulnerability in the previous section against the real world *BAS-MA*. The used attacker model was the internal attacker A_{it} and the architecture of the targeted system was *MA-AR*. The A_{it} registered his device to the *AC* through user authentication phase. Then the *MA* downloaded the C_i and generated B_i for displaying. The B_i had expired in 5 minutes to prevent replay attack. The *MA* was able to create a new B_i every 5 minutes without additional network communication with *AC*. This means that the targeted *MA* had stored not only user's credential but also the refresh logic in itself. The targeted *BAS-MA* also used its original two-dimensional barcode format (*Inter-Code*). The attack procedure is described as follows. The A_{it} intercepted the packets of the user authentication phase. The packets were not encrypted and the network channel also wasn't secure. A_{it} could find a plain XML document containing C_i in the intercepted packets. The C_i had the member identification number and organization code. The *AC* received the user's mobile phone number as parameter then returned the C_i. Anyone with a member's phone number can obtain member's credential data. The *QG* was included in the install package of the *MA*. The binary package of *MA* could be extracted from the targeted device without root permissions. The A_{it} could obtain the *QG* by uncompressing the package. The *QG* was a kind of shared object for the targeted device platform. A_{it} tried to decompile the binaries to understand the interface and the parameters of the *QG*. Thus the interface module was implemented in *Java*, the analysis didn't need high difficulty. The free decompile and analysis tools were available without commercial license. A_{it} could find the logic that how to assemble the U_i and timestamps to make the parameters for the *QG*. The shared object could be linked and executed with the forgery application on the same platform. Finally, A_{it} could produce the forgery application that received the victim's mobile number as parameter, then displayed the cloned B_i as *Inter-Code* format. The B_i could be accepted by the verification on the real world *BAS-MA* without any restrictions.

7. Recommendations For Secure Implementation

In Section 5, the various software vulnerabilities on the *BAS-MA* caused by immaturity implemented were discussed. The case study of the forgery attack through the software vulnerabilities on real system was reported in Section 6. Based on this, we suggest the countermeasures against the premature software implementations.

7.1. Interfered Screen Capturing

The screen capture function is usually enabled on common smart devices. It is recommended to block the function while the *MA* shows the B_i. The watermarking on the captured image is also good to give a warning to the A_{it}. It might be able to stop the harmful behaviour of the A_{it}.

7.2. Expiring QR-Code Available Period

The expiration of already issued B_i could cancel the replay attack. Some of salt value has to be mixed with the C_i before *QG* generates the B_i. *Lee et al.* [7] proposed a timestamps based authentication scheme suited for mobile device environment, in which users can be authenticated using a QR-Code. A secure authentication system proposed by *Shamal et al.* [12] that uses a two factor authentication by combining a password and a camera equipped mobile phone, where mobile phone is acting as a authentication token. If the nonce for expiring is hard-coded in the *MA*, it should be obfuscated against to be disclosed by malicious inspection.

7.3. Maintain Local Storage Cleanliness

The candidate data willing to be stored in local storage should be reviewed carefully. The local storage is always in danger to be examined by attackers. If it is possible, the empty of the local area is recommended. The session between *MA* and *AC* should be implemented based on secure protocol. The developer should not arbitrarily implement features beyond the protocol.

7.4. Deliberate Data Storing

The cryptography does not guarantee complete data protection. However, storing data without encryption is more dangerous. The encryption key supplied from trusted party over secure channel is better than included in the binary package. The white-box cryptography helps to hide the key inside the application [13]. If the decryption of the stored value is not required, one-way hash function could be a good choice. In other hand, hardware-backed approach such *as ARM TrustZone* [14] offers a high level security assurance. It provides a completely isolated memory area and data storages physically or logically. However, the application requires a close collaboration with hardware manufacturers to use this approach, and there are many restrictions on the implementation.

7.5. Obfuscation

The software obfuscation is a process to transform the source code into obfuscated code. The variable names are replaced by meaningless name and the execution flows become skewed without any logical error. It is a time consuming and laborious process to analyze the obfuscated binary. The obfuscation is the most effective and comprehensive defence against reversing attack. The obfuscation tool produces a obfuscated result with the source code or compiled binary. This can be fully integrated and automated build process. In fact, the developers do not need to worry to apply the obfuscation. It can be easily accomplished using the 3rd-party tools. In *Android*, the *ProGuard* is included as default in build system. Ensure good performance, it is good to use a commercial tool. The high performance obfuscation tools replace the strings, the constants and the hard-coded encryption keys by encrypted themselves. The obfuscation of control-flow had been proposed by *Chow et al.* [15]. In case that the *MA* includes the encryption key, the obfuscated use of key storage is the essential of the successful encryption.

7.6. Securing Messaging over IPC

When the mobile application structure uses the external *QG* through *IPC*, the *IPC message* carrying C_i to *QG* has to be attention not to be eavesdropped. According to the type platforms, the secure communication between processes might not be supported. The C_i should not be sent through a non-secure IPC. *Chin et al.* examine *Android* application interaction and identify security risks in application components. They provide a tool, *ComDroid*, that detects application communication vulnerabilities [16]. The *TaintDroid* is an extension to the *Android* mobile-phone platform that tracks the flow of privacy-sensitive data through third-party applications [17].

7.7. Detecting Rootkit and Malware

The attack on the rooted device is always more critical. The restricted policies are able to be cancelled easily. The applications are free to invade the protected area each other. It is recommended that let *MA* not launch on the rooted device. A rootkit is a set of malicious tools, in order to achieve the top level privilege of the system. *Kruegel et al*. presents a technique that exploits binary analysis to ascertain, at load time [18]. In case of the difficulty to implement to detect rootkit or malware itself, the use of 3rd-party solution could be alternative.

7.8. Detecting Rogue Access Point

The rogue AP is a wireless access point installed on a wired enterprise network without authorization from the network administrator. They are installed by a legitimate user who is unaware of its security implications or easily smuggled onto enterprise premises by an outsider. It also allows the attacker to conduct a *man-in-the-middle attack*. The white-list based detection approach [19] is appropriate to *BAS-MA*. The valid access point nearby *SP* could be listed up. It also can be an alternative to use the mobile network only such as *3G* or *LTE* instead of *WIFI*.

7.9. Restricting Allowed Device

The number of authentication enabled devices should be limited against the hijacking. Thus the B_i could be downloaded in only predefined mobile device. The device unique key is useful to implement the restrictions. The *IMEI* (*International Mobile Station Equipment Identity*) is a unique number to identity *3GPP* (*GSM, LTE*) mobile device. The *IMEI* could be a seed of the device unique key. The key is transmitted to the server along with the request for B_i. The server verifies the equality of the unique key from mobile device and the registered key in the server. Therefore, even if an attacker acquires a login cookie, it cannot be used in other devices. Of course the device unique key should not be stored in the local cookies.

7.10. Forged Verification

The executable is recommended to be signed by trusted certificate authority. This allows that the platform is able to verify whether the binary is forged. The application can verify the certification itself using platform API. The signature of the binary package could be accessed using the *PackageManager* in *Android*.

7.11. Limiting Distribution Path

The public on-line application store such as *Google Play Store* can be accessed without any limitation to download some mobile applications registered in the store. If the *MA* is registered on the public store, A_{ex} can download and install on their own devices in order to analyze the internal structures. The applications for internal users, such as *MA* are recommended to distribute on a private path. The system operators can send a URL containing the application downloadable link to the valid user via *SMS* or *E-mail*.

8. CONCLUSION

We discussed the vulnerabilities on the QR-Code authentication system caused by an immature software implementation. The main issue of the software security on the QR-Code authentication system is the protection of the logic and data around the QR-Code generator. These vulnerabilities can be overcome by a mature implementation. The improvisation implementation beyond the specification of the protocol is to be eradicated. The locally stored data should be cherry-picked and encrypted. The obfuscation can complete the robust software implementation. With these recommendations, the QR-Code authentication system would be a good solution for physical access system.

REFERENCES

[1] Starnberger, G., Froihofer, L., & Göschka, K. M., (2009) "QR-TAN: Secure mobile transaction authentication", Availability, Reliability and Security, pp578-583.

[2] Liao, K. C., & Lee, W. H., (2010). "A novel user authentication scheme based on QR-code" Journal of Networks, No.5(8), pp937-941.

[3] Lee, Y. S., Kim, N. H., Lim, H., Jo, H., & Lee, H. J., (2010) "Online banking authentication system using mobile-OTP with QR-code", Computer Sciences and Convergence Information Technology, pp644-648.

[4] Oh, D. S., Kim, B. H., & Lee, J. K., (2011) "A study on authentication system using QR code for mobile cloud computing environment", Future Information Technology, pp500-507

[5] Kao, Y. W., Luo, G. H., Lin, H. T., Huang, Y. K., & Yuan, S. M., (2011) "Physical access control based on QR code" Cyber-enabled distributed computing and knowledge discovery pp285-288

[6] Harini, N., & Padmanabhan, T. R., (2013) “2CAuth: A new two factor authentication scheme using QR-code”, International Journal of Engineering and Technology, No.5, pp1087-1094.

[7] Lee, Y., Kim, J., Jeon, W., & Won, D., (2012) “Design of a simple user authentication scheme using QR-code for mobile device”, Information Technology Convergence, Secure and Trust Computing, and Data Management, pp241-247.

[8] Kale, V., Nakat, Y., Bhosale, S., Bandal, A., & Patole, R. G., (2015) “A Mobile Based Authentication Scheme Using QR Code for Bank Security”

[9] Malik, J., Girdhar, D., Dahiya, R., & Sainarayanan, G., (2014) “Multifactor Authentication Using a QR Code and a One-Time Password”, Journal of Information Processing Systems, No.10.

[10] Callegati, F., Cerroni, W., & Ramilli, M., (2009) “Man-in-the-Middle Attack to the HTTPS Protocol”, IEEE Security & Privacy, pp78-81.

[11] Marlinspike, Moxie, (2009) “New tricks for defeating SSL in practice”, BlackHat DC.

[12] Shamal, S., Monika, K., & Neha, N., (2014) “Secure Authentication for Online Banking Using QR Code”, IJETAE–International Journal for Emerging Technology and Advance Engineering

[13] Chow, S., Eisen, P., Johnson, H., & Van Oorschot, P. C., (2003) “White-box cryptography and an AES implementation”, Selected Areas in Cryptography, pp250-270.

[14] Winter, J., (2008) “Trusted computing building blocks for embedded linux-based ARM trustzone platforms”, Proceedings of the 3rd ACM workshop on Scalable trusted computing, pp21-30.

[15] Chow, S., Gu, Y., Johnson, H., & Zakharov, V. A., (2001) “An approach to the obfuscation of control-flow of sequential computer programs”, Information Security, pp144-155.

[16] Chin, E., Felt, A. P., Greenwood, K., & Wagner, D., (2011) “Analyzing inter-application communication in Android”, Proceedings of the 9th international conference on Mobile systems, applications, and services, pp239-252.

[17] Enck, W., Gilbert, P., Han, S., Tendulkar, V., Chun, B. G., (2014) “TaintDroid: an information-flow tracking system for realtime privacy monitoring on smartphones”, ACM Transactions on Computer Systems.

[18] Kruegel, C., Robertson, W., & Vigna, G., (2004), “Detecting kernel-level rootkits through binary analysis”, Computer Security Applications Conference, pp91-100.

[19] Park, J., Park, M., & Jung, S., (2013). “A whitelist-based scheme for detecting and preventing unauthorized AP access using mobile device”, The Journal of Korean Institute of Communications and Information Sciences, No.38, pp632-640.

APPLICATION-BASED QOS EVALUATION OF HETEROGENEOUS NETWORKS

Farnaz Farid, Seyed Shahrestani and Chun Ruan

School of Computing, Engineering and Mathematics,
Western Sydney University, Sydney, Australia
farnaz.farid@westernsydney.edu.au,
s.shahrestani@westernsydney.edu.au,
c.ruan@westernsydney.edu.au

ABSTRACT

Heterogeneous wireless networks expand the network capacity and coverage by leveraging the network architecture and resources in a dynamic fashion. However, the presence of different communication technologies makes the Quality of Service (QoS) evaluation, management, and monitoring of these networks very challenging. Each communication technology has its own characteristics while the applications that utilize them have their specific QoS requirements. Although, the communication technologies have different performance assessment parameters, the applications using these radio access networks have the same QoS requirements. As a result, it would be easier to evaluate the QoS of the access networks and the overall network configuration depending on the performance of applications running on them. Using such application-based QoS evaluation approach, the heterogeneous nature of the underlying networks and the diversity of their traffic can be adequately taken into account. In this paper, we propose an application-based QoS evaluation approach for heterogeneous networks. Through simulation studies, we show that this assessment approach facilitates better QoS management and monitoring of heterogeneous network configurations.

KEYWORDS

QoS; QoS metric; Dynamic weight; Unified QoS Metric; application weight; weight

1. INTRODUCTION

The advancement and proliferation of modern wireless and cellular technologies have changed the way people work and communicate. By 2018, the data traffic over mobile networks is expected to reach 15.9 exabytes per month, with 69 percent of that consisting of video. There will be over 10 billion mobile-connected devices by 2018, which will exceed the world's expected population at that time [1]. To deal with this growing number of devices and the massive increases in traffic, the networks are moving towards an all-heterogeneous architecture. Any heterogeneous network constitutes of different communication technologies. These technologies have distinct bandwidths, coverage area, and operating frequencies. Their QoS characteristics, such as delay, throughput, and packet loss, as well as usage and implementation costs also differ from each other. As a result, the adaptation of heterogeneous network-based architecture for the provision of different applications especially multimedia applications faces significant

David C. Wyld et al. (Eds) : NETCOM, NCS, WiMoNe, CSEIT, SPM - 2015
pp. 161–172, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51613

challenges. Among these challenges, QoS-related issues such as the effective QoS evaluation, management, and monitoring still top the list [2].

Managing QoS for video or voice applications over heterogeneous networks is a challenging task. A research from Nemertes shows that the companies invest a significant amount of their budget to manage VoIP applications over these network architectures. For small enterprises, the annual costs range from $25,000, and for global enterprises this cost is around $2 million [3]. Therefore, the enterprises need to dedicate a lot of their effort to ensure service quality at every level of the network. System downtime is another challenge for businesses, which could often happen due to poor network management and monitoring. According to Gartner research, the hourly cost of system downtime for large enterprises was $42,000, with a typical business on average, experiencing 87 hours of downtime per year [4]. As a result, the QoS of any service-based network should be monitored, managed, and evaluated on an ongoing basis.

In this paper, we introduce some unified metric measurement functions that can help with assessing the application-based performance of heterogeneous networks. By taking the relevant performance-related parameters into account, these functions quantify the underlying network and the application-related QoS with a numerical value. The proposed approach considers the effects of the QoS-related parameters, the available network-based applications, and the available Radio Access Networks (RANs) to characterize the network performance with a set of three integrated QoS metrics. The first metric denotes the performance of each possible application in the network. The second one is related to the performance of each of the radio access networks present in the network. The third one characterize the QoS level of the entire network configuration. The core of this method is considering the effects of different application and radio access networks on the QoS of heterogeneous networks.

The rest of the paper is organized as follows: Section 2 discusses the background and motivations of this work, Section 3 illustrates the concept of unified QoS metric. Section 4 presents the application weight calculations in detail. The impact and the significance of the applications for QoS analysis are then discussed in Section 5. The last section gives the conclusions and proposes the future works.

2. RELATED WORK AND MOTIVATIONS

QoS evaluation in heterogeneous networks has been an active area of research [5, 6]. Most of the existing research focuses on the partial QoS evaluation of a heterogeneous network by deriving the performance level of a single access network and a single application present within that environment. Also, different studies have come up with various performance metrics for QoS evaluation of these networks. The conventional methods do not consider the performance of all the applications running on a network. For example, if there are voice and video conferencing running over a UMTS network, these methods do not include the performance analysis of these applications to quantify the overall network QoS. Furthermore, there is no unified metric to quantify the QoS of a network, which considers the performance of all the access networks present in it. For example, in a heterogeneous environment, there are three access networks, such as UMTS, WiMAX, and LTE. At present, no unified metric can represent the performance of this network configuration using the QoS-related parameters of these access networks.

Multi-criteria decision-making (MCDM) or Multi-Attribute Decision Making (MADM) algorithms have been widely employed in the area of the heterogeneous networks from vertical handover perspectives [7, 8]. The most common criteria, which are considered during this ranking process, are service, network, and user related [9]. These can include received signal strength, type of the service, minimum bandwidth, delay, throughput, packet loss, bit error rate, cost, transmit power, traffic load, battery status of the mobile unit, and the user's preferences. To facilitate the combining of these attributes into a single value, based on their relative importance, a weight is assigned to each attribute.

The weights for QoS-related parameters have both subjective and objective elements in it [10]. The network attributes, for example, the importance of received signal strength and bandwidth are objective in nature. Application related attributes such as delay, packet loss, and jitter show some objectivity. However, some studies have already revealed their potential subjective natures. For example, a study conducted in Tanzania shows that the users give moderate importance to end-to-end delay over packet loss [11]. The study by ETSI reveals that the users give strong importance to end-to-end delay over packet loss [12]. Therefore, the importance of application-related performance parameters can vary based on changing contexts, for example, between home and industrial environments or urban and rural areas. The significance of applications can vary depending on the context as well. For example, an application related to the education services can have higher importance compared to one that provides some entertainment services. Moreover, the absence or presence of an application will affect the weights of others in the network.

For weight assignment, the available literature on QoS evaluation in network selection has mostly used the Analytic Hierarchy Process (AHP) method, which is primarily developed by Saaty [7, 8, 13]. Some studies have also assigned fixed weights to these parameters based on their importance to service performance [14]. Both AHP and fixed weight methods are unable to handle the subjective and ambiguous factors related to weight determination such as context-based significance. In this study, Fuzzy Analytical Hierarchy Process (FAHP) with the extent analysis method is applied to bring the context-based information into the picture. This method is capable of handling ambiguity in any particular subject. It is also possible to assign the weights dynamically to the relevant parameters by using this method.

3. THE UNIFIED QOS METRIC

The quality of service on any network or application is usually evaluated through a set of specific metrics. For example, to assess the performance of any voice application, the delay, jitter and packet loss are measured and compared with the acceptable values of these parameters. Similarly, the QoS of any network is evaluated through parameters such as delay, packet loss, throughput, and available bandwidth. The presence of different types of communication technologies and applications in a heterogeneous network makes its QoS assessment method a challenging task. To deal with such challenges, this paper introduces unified metric measurement functions.

The QoS evaluation approach proposed in this work considers any heterogeneous network as a set of three layers; these are the application layer, the radio access network layer and the network configuration layer. Each of these layers uses a function to quantify a unified QoS metric, which flows to the next layer and derive the combined metric of that layer. Figure 1 shows the flowchart of the proposed approach. In the application layer, a function is defined to derive the QoS of each application through a unified metric. This function combines the values of several application-related performance metrics. As such, the QoS of a network-based application is treated as a function of QoS-related parameters. This can be expressed as:

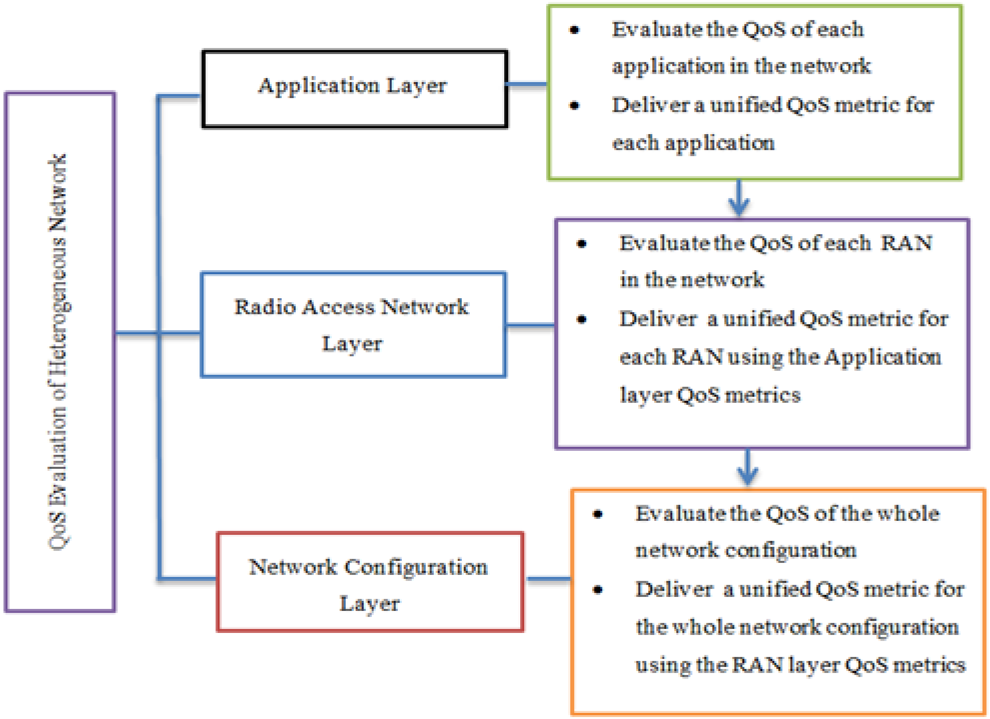

QoS Evaluation of Heterogeneous Network
Application Layer
Evaluate the QoS of each application in the network
Deliver a unified QoS metric for each application
Radio Access Network Layer
Evaluate the QoS of each RAN in the network
Deliver a unified QoS metric for each RAN using the Application layer QoS metrics
Network Configuration Layer
Evaluate the QoS of the whole network configuration
Deliver a unified QoS metric for the whole network configuration using the RAN layer QoS metrics

Figure 1. The concept of Unified QoS Metric

$$QoSAM_A = f\left(QP_1, QP_2, \ldots, QP_p\right) \tag{1}$$

where *A* denotes a network-based application, and *QP* refers to the QoS-related parameters. Then in the radio access network layer or RAN layer, the QoS of each access network, which are present in the network, is evaluated. This evaluation is conducted based on the performances of the active applications in those access networks. Hence, the QoS of an access network is viewed as a function of the application QoS metrics. It can be expressed as:

$$QoSRM_R = f\left(QoSAM_{A_{i=(1,2,\ldots\ldots,m)}}\right) \tag{2}$$

where *R* denotes any radio access network, and *i* refers to the number of active applications present on a network. Finally, to evaluate the QoS of the overall network configuration another function is defined, which uses the radio access network metrics as its input. This can be expressed as:

$$QoSCM = f\left(QoSRM_{R_{j=(1,2,\ldots\ldots,n)}}\right) \tag{3}$$

where *j* refers to the number of radio access networks present on a network.

4. The Application Weight Calculation

The weights of applications are considered during the QoS evaluation of radio access networks. The radio access network metric can be used to label a network as a particular service-oriented network such as education or health by integrating the application weights. For instance, if the QoS metric of a radio access network that is mainly used for health services, is always good, then that network can be taken as a suitable health service-oriented one for future use.

Table 1. Example of Application Weight Calculation

Networks	**Considered Parameters**			**Weights**	
	Application	**Service**	**Number of Users**	A_1	A_2
N_1	A_1	Education	20	$w_{A_1}^{N_1}$	$w_{B_1}^{N_1}$
	A_2	Entertainment	18		
N_2	A_1	Health	10	$w_{A_1}^{N_2}$	$w_{B_2}^{N_2}$
	A_2	Education	6		

The importance of the applications is subject to change depending on the requirements of particular networks. The service operators can update these criteria according to their particular circumstances. The criteria for this study have been formulated using the studies relevant to distance education-based service models [15]. For instance, in these service models, Videoconferencing (VC) bears more significance compared to voice-based applications. On the other hand, in a more general sense, VC may be less significant than voice-based applications as the latter is more easily amiable to the users. Therefore, the criteria to decide the applications weights in this regard can be the number of users using the application and the purpose and the context of that application usage.

Table 1 shows the example of two networks, N_1 and N_2, which have applications A_1 and A_2 with different number of users. The application weights are expressed as $w_{A_1}^{N_1}$, $w_{A_1}^{N_2}$, $w_{B_1}^{N_1}$, and $w_{B_2}^{N_2}$. In this paper, the weights of these applications are defined based on two criteria, the importance of the service, to which it belongs, and the number of users using that application. Other evaluation rules can be integrated based on individual needs. In the network N_1, it is considered that $w_{A_1}^{N_1}$ > $w_{B_1}^{N_1}$. This is determined depending on the fact that in the network N_1, the application A_1 is used by more users than the application A_2 as the application A_1 is used for educational services, whereas, the application A_2 is used for entertainment services. Therefore, when the QoS metric in the network N_1 is calculated considering these application weights, the QoS metric value reflects the significance of the service the application provides. As a result, if the QoS value is good for the application A_2 and poor for the application A_1, the outcome of the QoS level of the network will be poor. That is because, the majority of the users in that network experience poor performance for a valuable service. If, in any case, the entertainment service application has more users, the result will also be same as the education service is set to have higher significance than the entertainment service. These findings can change based on specific network requirements.

On the other hand, the Network N_2 supports both education and health services. As more users are using the health services compared to the education services, the application weight of A_1 is greater than A_2 , $w_{A_1}^{N_2} > w_{B_2}^{N_2}$. If the QoS value of the network N_2 is good, then it can be categorized as a health service-oriented network. Therefore, the configurations of N_2 can be recommended for any network that aims to deploy network-based health services in the future. Service operators can input these criteria to change the weights dynamically for any network.

The weight calculation involves two steps. At first, the alternatives, criteria, and the fuzzy judgement matrix are defined. Then in the second step, the actual weight is calculated based on those measures. FAHP-based calculations include: establishing a set of alternatives $X = \{x_1, x_2, ..., x_m\}$, a set of goal or evaluation criteria $G = \{g_1, g_2, ..., g_n\}$, a fuzzy judgement matrix (FJM), with elements $\widetilde{r_{ij}}$ that represents the relative importance of each pair of criteria *i* and *j*, and a weighting vector $w = (w_1, w_2, ..., w_n)$. Both steps involve the concept of Triangular Fuzzy Number (TFN) and fuzzy addition and multiplication operations. To derive the FJM for the first step the importance scale presented in Table 2 is used. It shows the TFN $K_t = (l_t, m_t, u_t)$ where t=1, 2,…, 9, and l_t, u_t and m_t are the lower, upper and the middle value of the fuzzy number K_t respectively. Table 3 shows the pair-wise comparison matrix for VC, voice, and VS applications formed based on the cited studies. The importance scale of Table 1 is used for the comparisons. If one of the applications is absent from the network, these pair-wise comparison matrices are subject to change.

For the second step of FAHP, different methods are proposed. The most prominent one is Chang's extent analysis method [16]. This method is chosen as it provides easy and flexible options for the weight calculation. The steps of the extent analysis method are as follows:

At first, the sums of the each row of the defined fuzzy comparison matrix are calculated. Then the normalization of the row sums is conducted using fuzzy multiplication to obtain fuzzy synthetic analysis. Therefore, in the fuzzy comparison matrix, the fuzzy synthetic analysis of criteria G_i of alternative X_m is calculated as:

$$D_{G_i}^{X_m} = \sum_{j=1}^{n} \widetilde{r_{ij}} \otimes \left[\sum_{i=1}^{n} \sum_{j=1}^{n} \widetilde{r_{ij}}\right]^{-1} = \left(\frac{\sum_{j=1}^{n} l_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{n} u_{ij}}, \frac{\sum_{j=1}^{n} m_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{n} m_{ij}}, \frac{\sum_{j=1}^{n} u_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{n} l_{ij}}\right) \tag{4}$$

where $i, j = \{1, 2, 3.............n\}$ and *n* is the number of criteria. In step 2, in order to rank the criteria against each alternative, the degree of possibility of two fuzzy numbers is applied. Therefore, $D_{G_2}^{X_m}(l_2, m_2, u_2) \geq D_{G_1}^{X_m}(l_1, m_1, u_1)$ is computed by the following equation:

Table 2. A FAHP-based Pair-wise Comparison Importance Scale

Fuzzy Numbers	**Definition**	**Triangular Fuzzy Number**
$k_1(l_1,m_1,u_1)$	Equal importance	(1,1,1)
$k_2(l_2,m_2,u_2)$	Intermediate values	(1/2,3/4,1)
$k_3(l_3,m_3,u_3)$	Moderate importance	(2/3,1,3/2)
$k_4\left(l_4,m_4,u_4\right)$	Intermediate values	(1,3/2,2)
$k_5\left(l_5,m_5,u_5\right)$	Strong importance	(3/2,2,5/2)
$k_6\left(l_6,m_6,u_6\right)$	Intermediate values	(2,5/2,3)
$k_7\left(l_7,m_7,u_7\right)$	Very strong importance	(5/2,3,7/2)
$k_8\left(l_8,m_8,u_8\right)$	Intermediate values	(3,7/2,4)
$k_9\left(l_9,m_9,u_9\right)$	Extreme importance	(7/2,4,9/2)

$$V\left(D_{G_2}^{X_m} \geq D_{G_1}^{X_m}\right) = \sup\left[\min\left(\mu_{D_{G_1}^{X_m}}(x), \mu_{D_{G_2}^{X_m}}(y)\right)\right] \tag{5}$$

It can be also expressed as:

$$V\left(D_{G_2}^{X_m} \geq D_{G_1}^{X_m}\right) = hgt\left(D_{G_2}^{X_m} \cap D_{G_1}^{X_m}\right)$$

$$= \mu_{D_{G_2}^{X_m}}\left(d\right) = \begin{cases} 1 & if \quad m_1 \geq m_2 \\ 0 & if \quad l_1 \geq l_2 \\ \dfrac{l_2-u_2}{(m_2-u_2)-(m_1-l_1)} & otherwise \end{cases} \tag{6}$$

and

$$V\left(D_{G_1}^{X_m} \geq D_{G_2}^{X_m}\right) = hgt\left(D_{G_1}^{X_m} \cap D_{G_2}^{X_m}\right)$$

$$= \mu_{D_{G_1}^{X_m}}\left(d\right) = \begin{cases} 1 & if \quad m_1 \geq m_2 \\ 0 & if \quad l_2 \geq u_1 \\ \dfrac{l_2-u_1}{(m_1-u_1)-(m_2-l_2)} & otherwise \end{cases} \tag{7}$$

where d is the ordinate to validate if the highest intersection point D is between $\mu_{D_{G_2}^{X_m}}$ and $\mu_{D_{G_1}^{X_m}}$. Both the values of $V\left(D_{G_2}^{X_m} \geq D_{G_1}^{X_m}\right)$ and $V\left(D_{G_1}^{X_m} \geq D_{G_2}^{X_m}\right)$ are required to compare $\mu_{D_{G_2}^{X_m}}$ and $\mu_{D_{G_1}^{X_m}}$. For large numbers of criteria, the degree of possibility is applied as:

$$V\left(D_{G_1}^{X_m} \geq D_{G_2}^{X_m}, D_{G_3}^{X_m}, \ldots\ldots\ldots\ldots, D_{G_n}^{X_m}\right) = V\left[\left(D_{G_1}^{X_m} \geq D_{G_2}^{X_m}\right) \text{ and } \left(D_{G_1}^{X_m} \geq D_{G_3}^{X_m}\right) \text{ and.....} \left(D_{G_1}^{X_m} \geq D_{G_n}^{X_m}\right)\right]$$

$$= \min V\left(d_{G_1}^{X_m} \geq d_{G_n}^{X_m}\right) \tag{8}$$

Table 3. Pair-wise Comparison Matrix for different Applications

Applica-tions	Criteria	VC		Voice		VS	
VC	Purpose of Usage	(1, 1, 1)		(3/2,2,5/2)	(1.09, 1.5, 2)	(2/3,1, 3/2)	(0.84, 1.25, 0.75)
	Number of Users	(1, 1, 1)		(2/3,1,3/2)		(1,3/2,2)	
Voice	Purpose of Usage	(2/5,1/2,2/3)	(0.54, 0.75, 1.09)	(1, 1, 1)		(2/3,1, 3/2)	(1.59, 2, 2.5)
	Number of Users	(2/3,1,3/2)		(1, 1, 1)		(5/2,3, 7/2)	
VS	Purpose of Usage	(2/3, 1,3/2)	(0.59, 0.84, 1.25)	(2/3,1,3/2)	(0.48, 0.67, 0.95)	(1, 1, 1)	
	Number of Users	(1/2,2/3,1)		(2/7,1/3,2/5)		(1, 1, 1)	

Assume that $d'\left(C_{G_n}^{X_m}\right) = \min V\left(d_{G_1}^{X_m} \geq d_{G_n}^{X_m}\right)$

In step 3, the weight vector $\mathbf{w}$ for each alternative is calculated. This is obtained as:

$$\mathbf{w}_m^{'} = \left(d'\left(C_{G_1}^{X_m}\right), d'\left(C_{G_2}^{X_m}\right), \ldots\ldots\ldots\ldots, d'\left(C_{G_n}^{X_m}\right)\right)^T \tag{9}$$

In step 4, the normalized weight vector is calculated for each alternative as:

$$\mathbf{w}_m = \left(d\left(C_{G_1}^{X_m}\right), d\left(C_{G_2}^{X_m}\right), \ldots\ldots\ldots\ldots\ldots\ldots\ldots, d\left(C_{G_n}^{X_m}\right)\right)^T$$

$$= \left(\frac{d\left(C_{G_1}^{X_m}\right)}{\sum_{j=1}^{n} C_{G_n}^{X_m}}, \frac{d\left(C_{G_2}^{X_m}\right)}{\sum_{j=1}^{n} C_{G_n}^{X_m}}, \ldots\ldots\ldots, \frac{d\left(C_{G_n}^{X_m}\right)}{\sum_{j=1}^{n} C_{G_n}^{X_m}}\right) \tag{10}$$

The weight vector of the considered applications is calculated as:

$$\mathbf{w}_A^{'}(\text{VC}, \text{Voice}, \text{VS}) = (1, \quad 0.94, \quad 0.56)$$

The normalization weight vector is as follows:

$$\mathbf{w}_A = \left(w_{\text{VC}}, w_{\text{Voice}}, w_{\text{VS}}\right) = (0.4, 0.38, 0.224)$$

5. IMPACT OF APPLICATION IMPORTANCE

This section evaluates the impacts of dynamic application weights on the unified QoS metrics. The performance of a few network-based service models have been analyzed using the fixed weight-based method in our previous work [17]. In this work, a detailed analysis is conducted to present the effects of dynamic weights on the performance of the same service models. The weights are calculated for each application according to the changing circumstances of the network. These weights are entered as inputs to derive the network QoS metric.

In the previous work, the voice application is set to have a higher importance than the VS application and the weights have been fixed as 0.6 and 0.4. In this work, those weights are set to change based on the pair-wise comparison matrices presented in Table 3. Figure 2 shows the QoS analysis of the scenario with twelve voice calls and one VS session on the network. The figure clearly indicates that when the voice and VS applications have equal importance, the access network has a good QoS level (e.g. 0.81). It shows an average QoS level (e.g. 0.62) for voice and a good QoS level (e.g. 1) for VS. When the importance level of voice application has been changed from having equal to extreme importance over VS application, the access network QoS comes down to an average value of 0.63. Although, the performance of the VS application is good, because of having a lower importance, it has less effect on the network QoS level. On the other hand, the voice application, being extremely important, has a greater impact on the network QoS level.

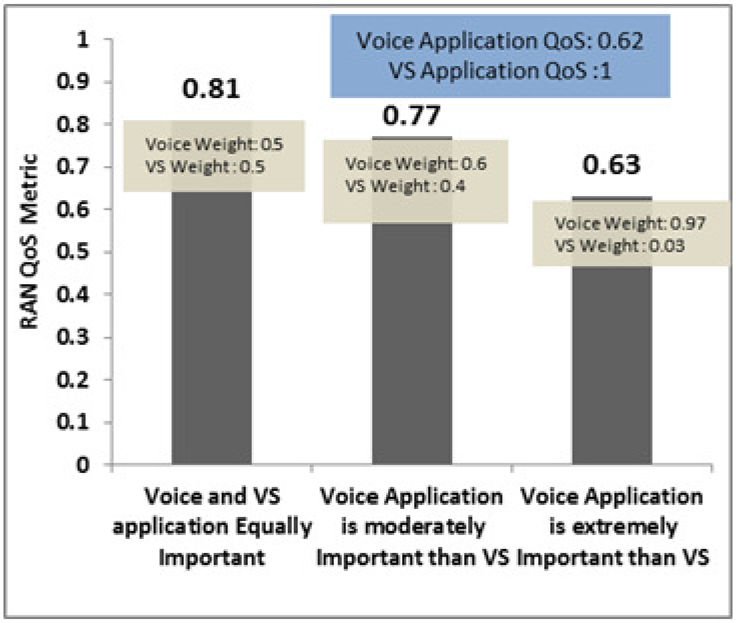


Figure 2. The Effect of Application Importance on Network QoS

Figure 3 shows a similar type of analysis with the altered importance of voice and VS applications. When the VS application has extreme importance, the network QoS improves due to the impact of application weights on the network performance.

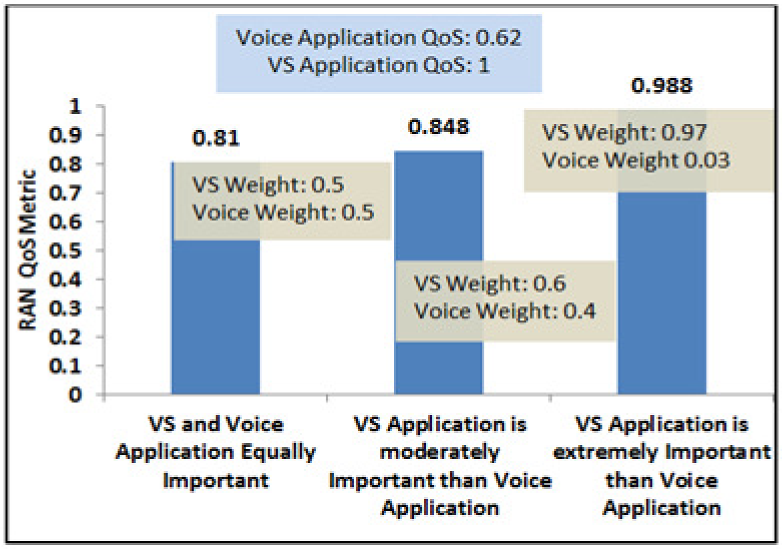


Figure 3. The Effect of Application Importance on the Network Performance

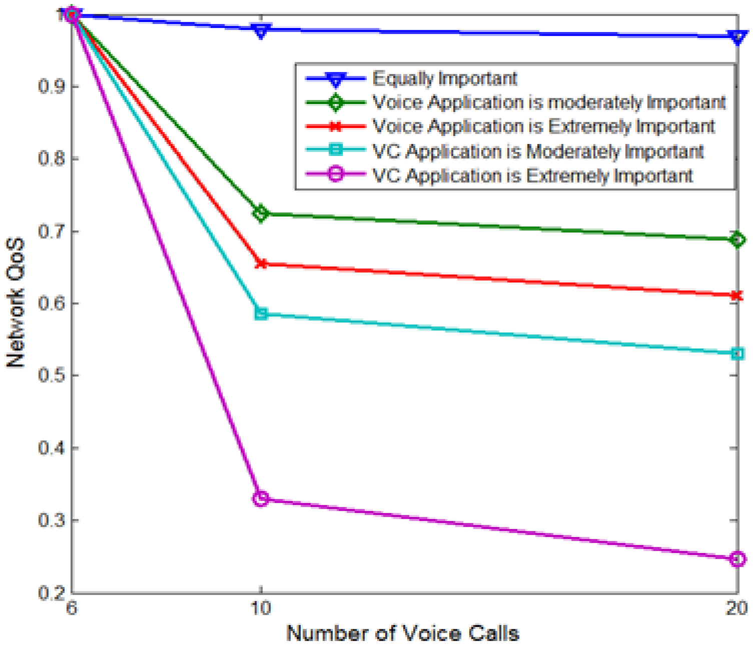


Figure 4. Network Performance Analysis with changing Application Importance

Figure 4 illustrates the QoS in a voice-based network for different number of calls when the importance of the application changes. When the VC application has extreme and moderate importance over voice application respectively, the network shows a poor QoS level. The reason is that the VC application with ten and twenty voice calls on the network experience a poor quality. On the other hand, the network takes an average QoS level with ten and twenty voice calls when the voice application has moderate and extreme importance over VC application respectively.

6. CONCLUSIONS

In this paper, an application-based QoS analysis method has been proposed and evaluated. In assessing the overall QoS level of a heterogeneous network, the levels of importance of applications are included as weights. The key contributions of this work include the proposing of a methodical approach for calculating and applying these weights. Extensive simulation studies, utilizing these weights for QoS assessment of various heterogeneous configurations supporting a variety of applications, have also been carried out. These studies demonstrate how the inclusion of the application importance weights for QoS evaluations, can assist in a systemic choosing of a fitting network configuration. In our future works, we intend to include several other factors that can influence the QoS provisions of a heterogeneous network supporting real-time applications.

REFERENCES

[1] Cisco, Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012–2017,Cisco, 2013.

[2] M. C. Lucas-Estañ and J. Gozalvez, “On the Real-Time Hardware Implementation Feasibility of Joint Radio Resource Management Policies for Heterogeneous Wireless Networks,” Mobile Computing, IEEE Transactions on, vol. 12, no. 2, 2013, pp. 193-205; DOI 10.1109/TMC.2011.256.

[3] I. Lazar and M. Jude, “Network design and management for video and multimedia applications,”2008; http://searchenterprisewan.techtarget.com/feature/Network-design-and-management-for-video-and-multimedia-applications.

[4] M. Perlin, “Downtime, Outages and Failures - Understanding Their True Costs,” 2013; http://www.evolven.com/blog/downtime-outages-and-failures-understanding-their-true-costs.html.

[5] S. M. Kantubukta Vasu, Sudipta Mahapatra, Cheruvu S Kumar, "QoS-aware fuzzy rule-based vertical handoff decision algorithm incorporating a new evaluation model for wireless heterogeneous networks," EURASIP Journal on Wireless Communications and Networking, vol. 2012, 2012.

[6] A. Alshamrani, S. Xuemin, and X. Liang-Liang, "QoS Provisioning for Heterogeneous Services in Cooperative Cognitive Radio Networks," Selected Areas in Communications, IEEE Journal on, vol.29, pp. 819-830, 2011.

[7] A. Sgora, P. Chatzimisios, and D. Vergados, "Access Network Selection in a Heterogeneous Environment Using the AHP and Fuzzy TOPSIS Methods," in Mobile Lightweight Wireless Systems.vol. 45, P. Chatzimisios, C. Verikoukis, I. Santamaría, M. Laddomada, and O. Hoffmann, Eds., ed: Springer Berlin Heidelberg, 2010, pp. 88-98.

[8] S. Qingyang and A. Jamalipour, "Network selection in an integrated wireless LAN and UMTS environment using mathematical modeling and computing techniques," Wireless Communications, IEEE, vol. 12, pp. 42-48, 2005.

[9] W. Lusheng and D. Binet, "MADM-based network selection in heterogeneous wireless networks: A simulation study," in Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology, 2009. Wireless VITAE 2009. 1st International Conference on, 2009, pp. 559-564.

[10] Z. Wenhui, “Handover decision using fuzzy MADM in heterogeneous networks,” Proc. Wireless Communications and Networking Conference, 2004. WCNC. 2004 IEEE, 2004, pp. 653-658 Vol.652.

[11] E. Sedoyeka, Z. Hunaiti, and D. Tairo, “Analysis of QoS Requirements in Developing Countries,” International Journal of Computing and ICT Research, vol. 3, no. 1, 2009, pp. 18-31.

[12] ETSI, Review of available material on QoS requirements of Multimedia Services, ETSI 2006.

[13] E. Stevens-Navarro, L. Yuxia, and V. W. S. Wong, “An MDP-Based Vertical Handoff Decision Algorithm for Heterogeneous Wireless Networks,” Vehicular Technology, IEEE Transactions on, vol. 57, no. 2, 2008, pp. 1243-1254; DOI 10.1109/tvt.2007.907072.

[14] Wen-Tsuen and S. Yen-Yuan, “Active application oriented vertical handoff in next-generation wireless networks,” Proc. Wireless Communications and Networking Conference, 2005 IEEE, 2005, pp. 1383-1388.

[15] O. I. Hillestad, A. Perkis, V. Genc, S. Murphy, and J. Murphy, “Delivery of on-demand video services in rural areas via IEEE 802.16 broadband wireless access networks,” Proc. Proceedings of the 2nd ACM international workshop on Wireless multimedia networking and performance modeling, ACM, 2006, pp. 43-52.

[16] D.-Y. Chang, “Applications of the extent analysis method on fuzzy AHP,” European Journal of Operational Research, vol. 95, no. 3, 1996, pp. 649-655; DOI http://dx.doi.org/10.1016/0377-2217(95)00300-2.

[17] F. Farid, S. Shahrestani, and C. Ruan, “QoS analysis and evaluations: Improving cellular-based distance education,” Proc. Local Computer Networks Workshops (LCN Workshops), 2013 IEEE 38th Conference on, 2013, pp. 17-23.

AUTHORS

Farnaz Farid is pursuing her PhD degree in Information Technology and Communications at the Western Sydney University. Prior to that she has worked in China as a web application developer and web business SME at IBM. Her research interests include wireless and cellular networking, web engineering, and technology for development.

Seyed Shahrestani completed his PhD degree in Electrical and Information Engineering at the University of Sydney. He joined Western Sydney University in 1999, where he is currently a Senior Lecturer. He is also the head of the Networking, Security and Cloud Research (NSCR) group at Western Sydney University.

Chun Ruan received her PhD degree in Computer Science in 2003 from the University of Western Sydney. Currently she is a lecturer in the School of Computing, Engineering and Mathematics at Western Sydney University. Prior to that, she worked as an associate professor, lecturer and associate lecturer at the Department of Computer Science, Wuhan University, China.

BLE-BASED ACCURATE INDOOR LOCATION TRACKING FOR HOME AND OFFICE

Joonghong Park[1], Jaehoon Kim[2] and Sungwon Kang[3]

[1]Samsung Electronics, Suwon, Republic of Korea
joonghong.park@samsung.com
[2]Samsung Electronics, Suwon, Republic of Korea
jaehoonk@samsung.com
[3]Korea Advanced of Science and Technology (KAIST),
Daejeon, Republic of Korea
sungwon.kang@kaist.ac.kr

ABSTRACT

Nowadays the use of smart mobile devices and the accompanying needs for emerging services relying on indoor location-based services (LBS) for mobile devices are rapidly increasing. For more accurate location tracking using Bluetooth Low Energy (BLE), this paper proposes a novel trilateration-based algorithm and presents experimental results that demonstrate its effectiveness.

KEYWORDS

Indoor location tracking, Indoor positioning, Distance-based filtering algorithm, Trilateration algorithm, Location-based services (LBS)

1. INTRODUCTION

Nowadays the use of smart mobile devices, such as smartphones, has increased explosively and the needs for emerging services relying on indoor location-based services (LBS) for mobile devices are also rapidly increasing.

For indoor location tracking, WiFi has been used most widely. Recently, the Bluetooth technology has introduced an energy efficient Bluetooth Low Energy (BLE) version [1]. Because of its lower transmitter power usage and its simpler receiver architecture without any sophisticated techniques for dealing with multipath such as MIMO or RAKE, BLE has been found to be a more suitable technology for indoor location tracking with respect to both accuracy and energy efficiency [2].

For computing distances, the BLE AP just broadcasts a short packet periodically with an advertising interval. After receiving this short packet, the BLE receiver can compute an approximate distance by comparing the Received Signal Strength Indication (RSSI) and the

David C. Wyld et al. (Eds) : NETCOM, NCS, WiMoNe, CSEIT, SPM - 2015
pp. 173–181, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51614

broadcasting power of the BLE AP [3, 4]. If there are 3 or more distance information from the installed BLE APs in the given indoor area, the current location can be calculated through the Trilateration algorithm [5, 6].

Not only WiFi but also BLE uses the same 2.4 GHz Industrial Scientific Medical (ISM) radio band, and because of the license free characteristics of the ISM band, they can easily have interferences of signals, which cause signal strength fluctuations [7]. Due to this interference, RSSI value is not always accurate as the result of calculating distance values changes in every advertising interval. To make location tracking more accurate, many studies were conducted such as using numerous BLE APs or using the fingerprint pattern matching instead of the Trilateration algorithm [7, 8]. Beyond a threshold, increasing the number of BLE APs does not improve and may even deteriorate the location tracking accuracy [7]. In the case of the method using the fingerprint pattern matching, a pre-processing is required that creates a database of signal strength values for all sampling locations [8].

In this paper, we propose a novel Trilateration-based location tracking solution that uses the algorithms that are designed based on the result of analyzing BLE signal characteristics. Then we verify the effectiveness of the proposed solution through several experimental results in a realistic environment.

2. THE PROPOSED ALGORITHM

Figure 1 is the flow chart for our proposed algorithm. Once the receiver receives signals from BLE APs, the receiver extracts RSSI values and calculates distances through the RSSI values. In the following, we explain our proposed algorithm in detail by presenting each of its core component algorithms, which are shown in blue in Figure 1, in a sub-section. The proposed method of reducing distance errors is presented in Section 2.1 (Applying the Kalman filter for the first time). Section 2.2 explains the steps from removing error distances to removing error locations in Figure 1. The proposed distance-based filtering algorithm, which dramatically minimizes location tracking error, is presented in Section 2.3. And the method for getting a more accurate location point is presented in Section 2.4 (Applying the Kalman filter for the second time).

2.1. Applying the Kalman filter for the first time

A distance value of a BLE AP is calculated from RSSI in every advertising interval and the calculated distance values can fluctuate due to other signals in the ISM band [7].

There are many ways for reducing fluctuation. We considered both the moving average method and the Kalman filter [9] that is an effective and the most common noise filtering algorithm. We tested them to compare their performances. Testing checked location fluctuations in a small area (3.9m * 3.9m). Testing results from our testing application are shown in Figure 2, in which dots are calculated locations and lines are differences between the current locations and the previous locations. To compare performance the following methods are used:

(F1) Raw - without any filter

(F2) 3 moving average method – calculates the average result of the recent 3 calculated locations.

(F3) 5 moving average method – calculates the average result of the recent 5 calculated locations.

(F4) Kalman filter – calculates the result of the Kalman filter.

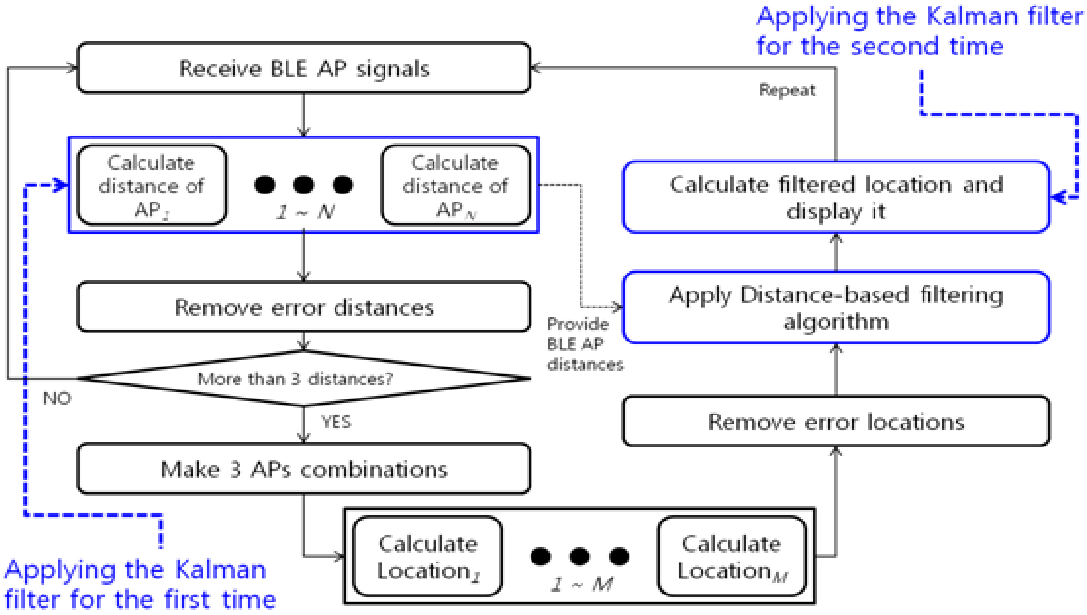

Applying the Kalman filter for the second time
Receive BLE AP signals
Repeat
Calculate distance of AP_1
$1 \sim N$
Calculate distance of AP_N
Calculate filtered location and display it
Remove error distances
Provide BLE AP distances
Apply Distance-based filtering algorithm
More than 3 distances?
NO
YES
Remove error locations
Make 3 APs combinations
Calculate $Location_1$
$1 \sim M$
Calculate $Location_M$
Applying the Kalman filter for the first time

Figure1. Flow chart for the proposed algorithm

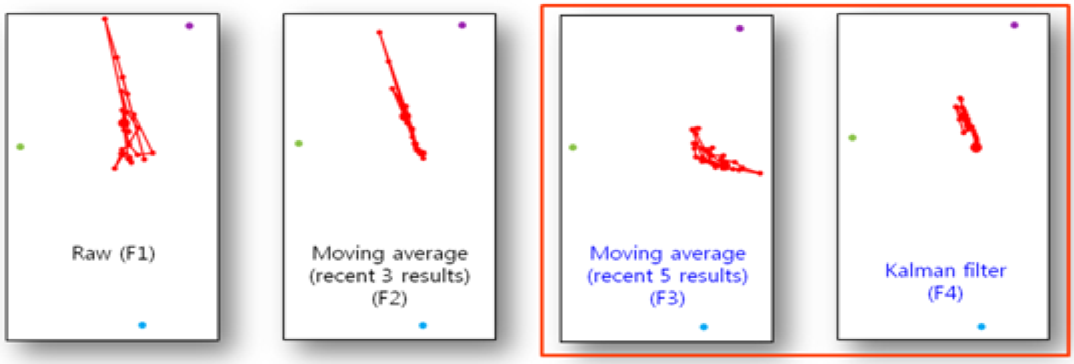


Figure 2. Filter comparison results

Figure 2 shows that both the 5 moving average method (F3) and the Kalman filter (F4) are superior to the others. To get the first location result in the case of the 5 moving average method (F3), however, waiting a minimum of 5 times of the advertising interval time is required. For instance, 5 seconds is required when the advertising interval is 1 second. This is a critical problem when the user moves repeatedly in a short time interval. So we decided to use the Kalman filter to reduce fluctuations in both distance and location.

Based on the above filter comparison results, we apply the Kalman filter to each distance value to reduce fluctuation effectively.

2.2. Removing error distances and locations

The most important thing for calculating accurate locations is the accurate distance information and, even though we apply the powerful Kalman filter, it may not be enough to get the accurate distance because of the inherent limitation of the BLE signal characteristics. Due to the fluctuation of the results, some distances should be discarded if they are out of the actual maximum distance.

If there are only three BLE APs, we can calculate only one location point through the Trilateration algorithm [5, 6]. If we use more than three BLE APs, however, any combinations of the three BLE APs can be selected for calculating location points. For example, if we use 6 BLE APs, a total of 20 combinations for calculating location points can be selected. Then each combination can be used to calculate a location point. However, because signals propagate making a circle without direction, calculated location points should be discarded if they are out of the boundary of the actual space. To get one accurate location, we can consider different choices of algorithms that calculate it by using the location points that remain after discarding the out-of-boundary location points.

2.3. The distance-based filtering algorithm

We developed the distance-based filtering algorithm by analyzing the characteristics of RSSI-distance variations. According to the path loss model [10], attenuation of RSSI becomes larger as the distance becomes longer. We picked the BLE AP closest from the current location as the reference point. Compared to the distances of the other BLE APs, the distance of this reference point is probabilistically more accurate.

As explained in Section 2.2, there are remaining location points calculated from a three BLE AP's combination. We can calculate distances between the reference point and the remaining location points and then choose N location points by sorting the distances where N is the value pre-defined by the environment. Now we can calculate an average location point through N location points as a candidate for the current location.

2.4. Applying the Kalman filter for the second time

Candidates of the current location are calculated in every advertising interval of BLE AP, for example in every 250ms. Still there may be some location tracking error in each candidate. So we apply the Kalman filter for the second time to each candidate location point to minimize location tracking error.

3. IMPLEMENTATION

To develop the solution, we implemented the receiver and configured the parameters of the BLE AP. The following sub-sections present configuring the BLE AP in Section 3.1, and implementation of the receiver in Section 3.2.

3.1. The BLE AP

The BLE AP periodically broadcasts a short packet to all receivers in every pre-defined advertising interval. Then the receiver approximates the proximity by converting the RSSI of this packet to distance depending on the broadcasting power of the BLE AP. When calculating a location using BLE, the role of the BLE AP is just advertising packets periodically. For the BLE AP we just need to adjust two parameters “the advertising interval” and “the broadcasting power”, and most of the BLE products support these two parameters. So we decided to use Estimote products as the BLE APs.

3.2. The receiver

We used Android phone to implement the receiver, and Figure 3 shows the software architecture of the receiver. The implemented Android application includes the following modules:

(D1) The BLE AP manager

(D2) The movement detector

(D3) The accuracy algorithm module

(D4) UI

We used the following public libraries:

(D5) The Kalman filter

(D6) The Trilateration algorithm

Because we used Estimote product, we also used the following library:

(D7) Estimote SDK/ service

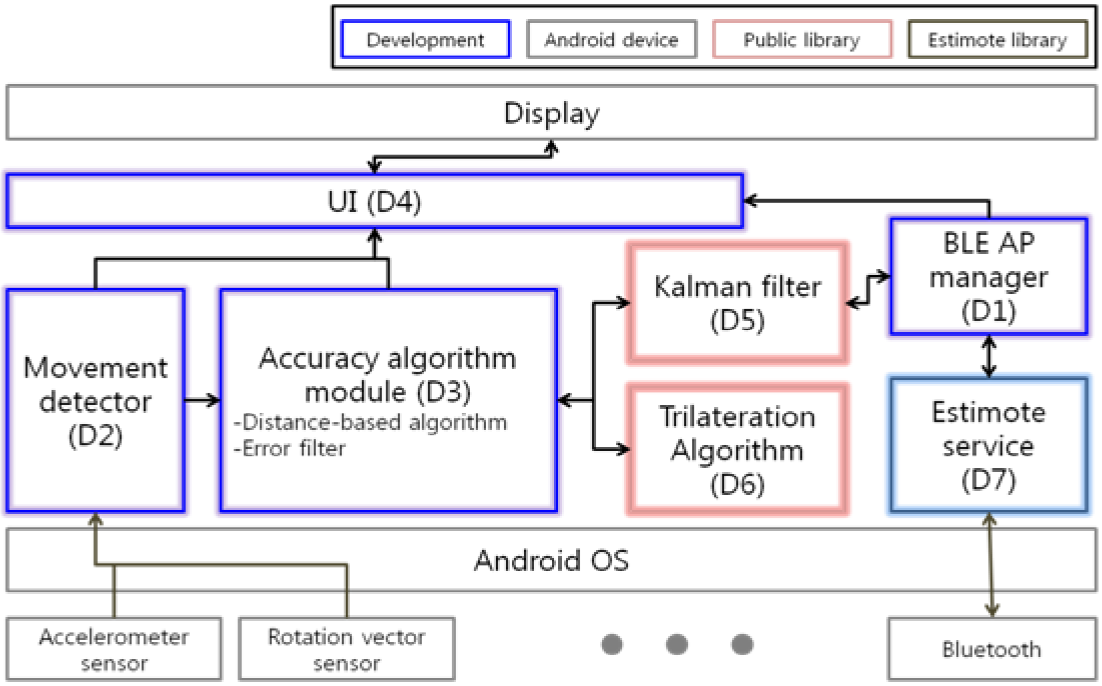


Figure 3. The software architecture of the receiver

The BLE AP manager (D1) maintains information of each BLE AP including its MAC address, location, and distance. The Estimote service (D7) notifies distance through event listener in the BLE AP manager in every advertising interval. Some distances are discarded because they are out of the boundary of actual space, and the Kalman filter (D5) is applied to get more accurate distances as explained in Section 2.1.

The movement detector (D2) is developed using the Android accelerometer sensor for detecting user movement. When moving such as walking or running occurs, the acceleration of gravity in the Android device changes. Android accelerometer sensor provides X, Y, Z values [11]. X value indicates left or right movement, Y value indicates up or down movement, and Z value indicates forward or backward movement. Through these values we can get the amount of change of the device acceleration of gravity. This information is passed to the accuracy algorithm module (D3) to make it detect user location changes.

The accuracy algorithm module (D3) was developed to get an accurate user location based on the Trilateration algorithm (D6) and includes the distance-based filtering algorithm explained in Section 2.3 and applies the Kalman filter (D5) for the second time as explained in Section 2.4.

4. Evaluation

The testbed consists of 6 Estimote BLE APs in a polygonal living room (8.5m × 13m) of a real house and 1 Android phone (Samsung Galaxy Note 3 with Android 5.0) as the receiver. Figure 4 shows the installed BLE APs in the living room.

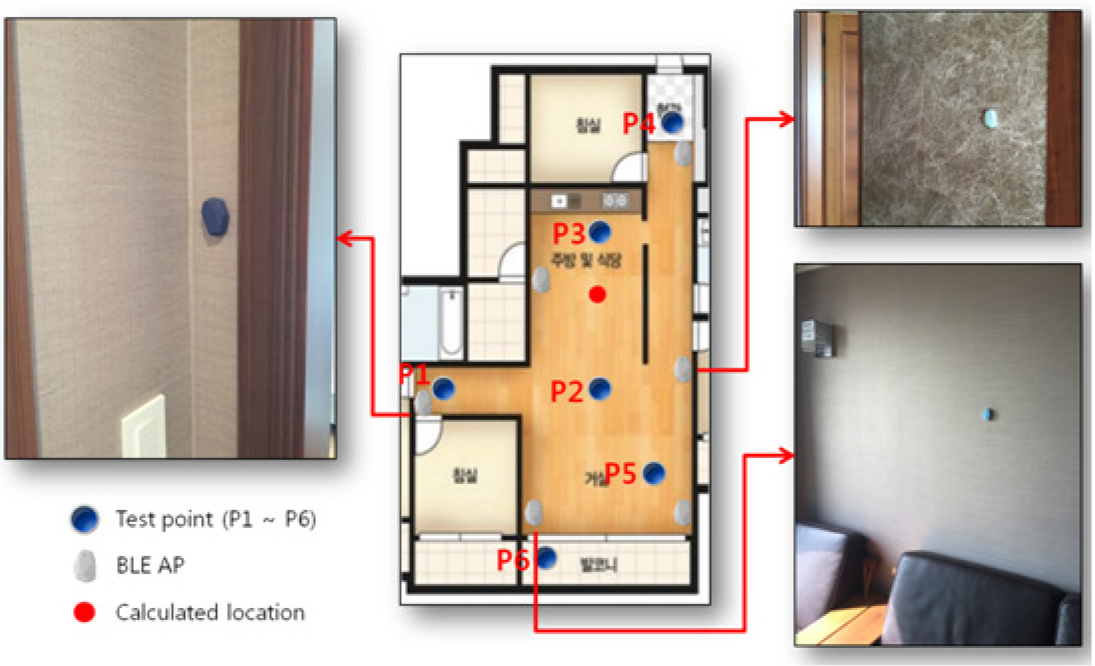


Figure 4. BLE APs and test points in the testbed

4.1. Evaluation scenario and method

We set 6 test points as in Figure 4 to measure accuracy by checking distance differences, called *location tracking errors,* between test points and calculated locations. For measuring accuracy the measurer moves from test point 1 (P1) to test point 6 (P6) sequentially.

When approaching each test point the measurer touches the test point button in the testing application, which was developed especially for evaluation, the testing application automatically calculates the location tracking error and displays its value next to the test point button. To evaluate performance, the following algorithms are considered:

(A1) The raw algorithm – calculates the average result from all calculated locations without the Kalman filter

(A2) The average algorithm – calculates the average result from all calculated locations using the Kalman filter

(A3) The proposed algorithm – calculates the average result from distance-based filtered locations using the Kalman filter

In the case of the raw algorithm (A1), only the minimal filter that excludes values outside the boundaries of the actual distance and the actual location is used. The result is the average value of all calculated locations from the BLE AP combinations.

In the case of the average algorithm (A2), the minimal filter and the Kalman filter are used. The result is the average value of all calculated locations from the BLE AP combinations.

In the case of the proposed algorithm (A3), in addition to the minimal filter and the Kalman filter, the distance-based filtering algorithm explained in Section 2.3 is used. The result is the average value of locations selected by the distance-based filtering algorithm.

4.2. Experimental results

We performed testing 10 times for each of the algorithms A1, A2 and A3. The advertising interval of the BLE AP is 250ms, and its broadcasting power is 4dBm.

Table 1 shows the average result from each test point, and Figure 5 shows the result graph when using the raw algorithm (A1). The average location tracking error was about 3.01m.

Table 2 includes average result from each test point and Figure 6 shows the result graph when using the average algorithm (A2). The average location tracking error was about 2.93m. The location tracking error was slightly decreased due to the use of the Kalman filter.

Table 1. Result of the raw algorithm

	Point1	Point2	Point3	Point4	Point5	Point6	Average
Average of 10 tests	1.94	1.51	2.61	5.57	2.63	3.79	3.01

Table 2. Result of the average algorithm

	Point1	Point2	Point3	Point4	Point5	Point6	Average
Average of 10 tests	2.37	1.05	2.62	5.43	2.46	3.65	2.93

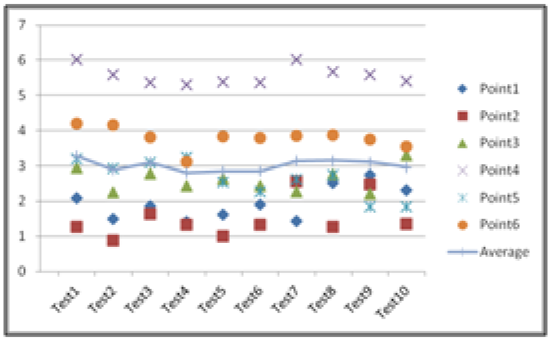


Figure 5. Result of the raw algorithm

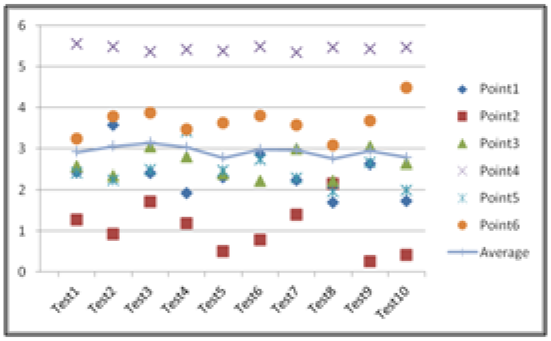


Figure 6. Result of the average algorithm

As Table 3 and Figure 7 show, the location tracking error dramatically decreased when using the proposed algorithm. The average location tracking error was about 1.77m and the reduction ratios are about 41% from the raw algorithm and about 39% from the average algorithm.

Table 3. Result of the proposed algorithm

	Point1	Point2	Point3	Point4	Point5	Point6	Average
Average of 10 tests	1.21	1.70	2.33	2.46	1.17	1.75	1.77

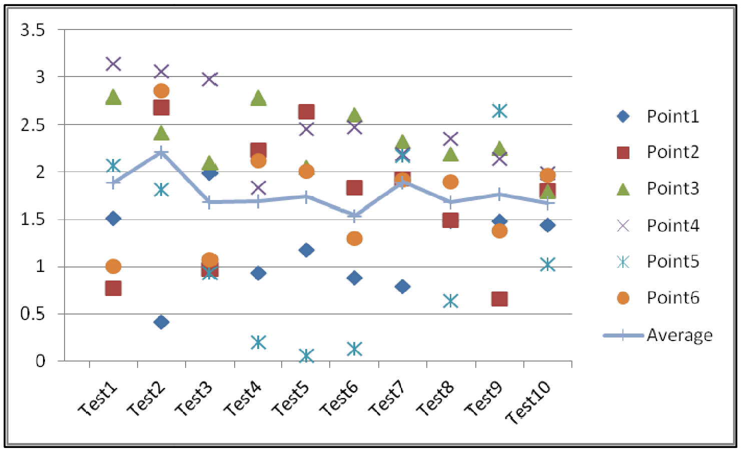


Figure 7. Result of the proposed algorithm

5. CONCLUSION

An indoor location tracking technology is necessary for emerging services or applications. In this paper we proposed a novel distance-based filtering algorithm and a solution, which are based on the Trilateration method that does not require any pre-processing such as creating a database of signal strength values at all sampling locations.

Through several experimental results, we demonstrated the effectiveness of the solution in significantly reducing location tracking errors. We are convinced that our indoor location tracking technology using the proposed algorithm and solution can further contribute to making more convenient the emerging services such as the Online to Offline (O2O) service and to creating new services in the IoT environment.

REFERENCES

[1] The Bluetooth Special Interest Group, Specification of the Bluetooth System, Covered Core Package, Version:4.0, Kirkland WA, USA, 2010.

[2] X. Zhao, Z. Xiao, A. Markham, N. Trigoni, Y. Ren, “Does BTLE measure up against WiFi? A comparison of indoor location performance,” The European Wireless(EW) Conference 2014, Berlin, Germany, 2014.

[3] Subhan, F., Hasbullah, H., Rozyyev, A., & Bakhsh, S. T., "Analysis of Bluetooth signal parameters for indoor positioning systems," The IEEE International Conference on Computer & Information Science (ICCIS), Vol. 2, pp. 151–156, USA, IEEE Computer Society, 2012.

[4] M. Hossain and S. Wee-Seng, "A comprehensive study of Bluetooth signal parameters for localization," The IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, pp. 187-192, 2007.

[5] L. Peneda, A. Azenha, and A. Carvalho, "Trilateration for indoors positioning within the framework of wireless communications," The 35th Annual Conference of IEEE Industrial Electronics IECON, pp. 2732 - 2737, 2009.

[6] Jeffrey Hightower, and Gaetano Borriello, “Location Sensing Techniques,” Technical report, University of Washington, Seattle, Washington, USA, 2001.

[7] R. Faragher, "An Analysis of the Accuracy of Bluetooth Low Energy for Indoor Positioning Applications," The 27th International Technical Meeting of The Satellite Division of the Institute of Navigation, pp. 201-210, Tampa, Florida, September 2014.

[8] Perez Iglesias, H. J., Valentın Barral, and Carlos J. Escudero, "Indoor person localization system through RSSI Bluetooth fingerprinting," The 19th IEEE International Conference on Systems, Signals and Image Processing (IWSSIP), 2012.

[9] E. Brookner, Tracking and Kalman Filtering Made Easy. New York: Wiley, 1998.

[10] T. Rappaport, Wireless Communications: Principles and Practice, Prentice-Hall, Englewood Cliffs, NJ, 1996.

[11] http://developer.android.com/reference/android/hardware/SensorEvent.html

INTENTIONAL BLANK

AN ADAPTIVE REMOTE DISPLAY FRAMEWORK TO IMPROVE POWER EFFICIENCY

Dong Hyun, Jo[1] and Dae Young, Kim[2]

[1]Department of Mobile Communication,
Samsung Electronics, Suwon, South Korea
donghyun.jo@samsung.com
[2]Department of Computer Science, KAIST, Daejeon, South Korea
kimd@kaist.ac.kr

ABSTRACT

As computing performance and network technology have evolved, mobile device users can enjoy high quality multimedia more easily. Remote Display - the technology which mirrors the screen of one device to another device - allows handheld mobile devices to share their screen contents with larger-sized display devices such as TVs. However, there is general concern about high power consumption caused by complex computation for encoding and continuous data transmission in the mobile devices.

In this paper, we present an adaptive remote display framework considering and utilizing the processing capability of display device. By supporting the Content Mirroring Mode, we can skip unnecessary steps and perform core activities to improve power efficiency and extend overall processing capability.

KEYWORDS

Remote display, Content Mirroring

1. INTRODUCTION

Recently, mobile devices such as smartphones and tablet PCs have become a part of everyday life. And advanced computing performance, high display resolution, and fast connectivity allow mobile device users to access and enjoy various multimedia services anytime, anywhere. Moreover, the cloud-based services provide appropriate content according to the type of connected device. Searching and downloading multimedia content that is compatible with the mobile device is no longer necessary.

However, in contrast with TV which has increasing screen size according to growing display resolution, the mobile devices which emphasize mobility and portability have restrictions in terms of screen size, so users are not able to maximize the experience of multimedia services.

David C. Wyld et al. (Eds) : NETCOM, NCS, WiMoNe, CSEIT, SPM - 2015
pp. 183–193, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51615

The remote display technology helps users to overcome the limit of screen size by providing a chance to mirror mobile device screen to another device with a large screen such as a TV. Using the remote display technology, users can also share the multimedia experience with family members or friends.

The visual quality and the end-to-end latency have been the technical challenges of the remote display. To mirror a high-resolution screen in real time, the screen has been captured and compressed using an encoding scheme to reduce the amount of data transfer. The efficient encoding scheme which provides better display quality and consumes lower network bandwidth has been researched [1].

Wi-Fi Miracast, a representative remote display solution, transmits screens encoded by H.264 video codec via peer-to-peer networking using a Wi-Fi direct (IEEE 802.11) connection. It provides good quality by supporting high resolutions of up to 1920 x 1080 pixels and low end-to-end latency. However, whereas the high compression ratio of H.264 provides savings in network bandwidth, the high computational complexity of H.264 causes high power consumption and it still remains a problem on power-limited mobile devices.

Previous researches have mainly focused on an encoding scheme to improve the problem [2] and there has been a lack of interest in overall framework to solve the problem.

2. RELATED WORKS

Bo-yun Eom et al [1] have proposed a power-aware remote display framework which uses a hybrid encoding scheme in VNC protocol. It aims to improve power efficiency by switching encoding modes adaptively to the battery level of client devices.

Ji-su Ha et al [2] have proposed a scheme to implicitly analyse the dynamics of a video file and uses the screen dynamics score to compute an ideal frame rate in run-time with respect to the multimedia content context. The proposed work estimates the screen dynamics by calculating the I-type macroblocks in a target interval which can be configured and skips frames based on the normalized I-type macroblock count, the screen dynamics score. The video with low dynamics, like video lectures, have shown a lower screen dynamics score than a dance-genre music video. Using the screen dynamics score, which is related to the actual dynamics, the frame rate can be controlled to minimize the transmission and power consumption without visible quality loss.

There are several remote display technologies. Chih-Fan Hsu et al [3] have measured and compared the performance of those various technologies. The work provides the result of the performance evaluations in various aspects: frame rate, resolution, bitrate, packet loss and so forth. Most of them are proprietary solutions designed by manufacturers, which means it may not be compatible with some devices.

3. BACKGROUND AND MOTIVATION

A Remote display framework is comprised of three major parts: The source device which captures, encodes, and transmits the screen, the display device which decodes the received data and displays it, the data transmission protocol which defines the format of the data that is to be transmitted from a source device to a display device.

The remote display is advantageous in that it can support any type of content if the source device can handle and display it. Conversely, the display device is responsible only for displaying a mirrored screen and the processing capability of display devices is not taken into account. However, various consumer electronics, such as TVs and refrigerators, have evolved into smart devices which have processing capability for various content and network connectivity [4]. We can utilize the display device as a content processing unit to extend the overall capability of the entire system.

The source device performs the intrinsic function which is content processing for local display. At the same time, it executes data processing and transmissions for remote display.

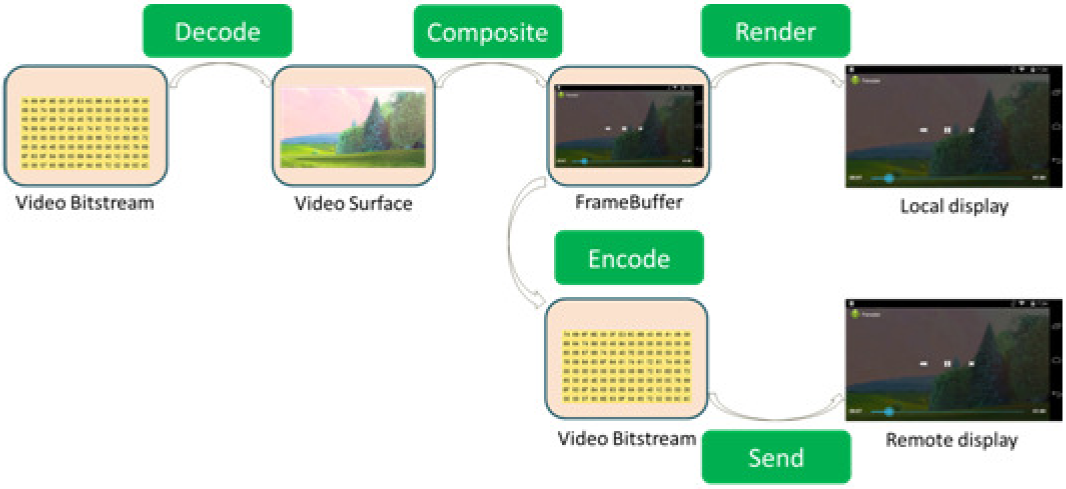


Figure 1. Process for the local display and the remote display

When playing a video, the following procedures are performed.

First, the media framework extracts the video bitstream from the media source and decodes it using the video decoder. Then, it passes the resulting video surface to the graphics engine. The video surface is resized to fit the render area. If necessary, the color format conversion is also carried out. The post-processed video surface is composited with a UI controller to make the final image for local display.

In addition to the above steps, the composited frame buffer is resized to fit the remote display and encoded to reduce network bandwidth. And then, it is sent to the remote display.

This intensive real-time processing results in lots of power consumption. Furthermore, in order to improve the visual quality of the remote display, more computation is required and more power consumption is generated. So, it is necessary to reduce and optimize the processes related to local and remote display.

In this paper, we present an approach considering an overall framework, including source device, data transmission protocol, and display device, from a broader perspective. As a result, we suggest a novel framework to improve power efficiency and extend overall processing capability for multimedia content.

4. ADAPTIVE REMOTE DISPLAY FRAMEWORK

In Figure 1, the transmitted video bitstream is almost same as the original video bitstream except the composited UI controller, resolution and bitrate. To make the similar video bitstream for the remote display, power-consuming processes such as resizing and encoding are carried out. If it does not have to be exact same screen, it is possible to reduce the steps by sending the original video bitstream.

In certain scenarios, such as a media file playback and slide show, the source device works mainly as a controller, and the local display in the source device are not essential and generate unnecessary power consumption. We can consider it as an optional process and reduce power consumption of the source device by skipping these processes.

Adaptive remote display is based on the extended processing capability: The content can be processed in any device that is capable of handling it. In case of remote processing in which the content is processed in display device, the source device transmits the original content such as media file, streaming URI instead of encoded video stream. The display device processes the received content using its own framework and displays it. In addition, the source device is able to skip local processing and display, if necessary.

According to the transmitted data type, two modes are defined for data transmission protocol in adaptive remote display framework: the Screen Mirroring Mode (SMM) and the Content Mirroring Mode (CMM). The Screen Mirroring Mode is the same one used in typical remote display solution, which the source device process content and transfer the encoded screen to the display device.

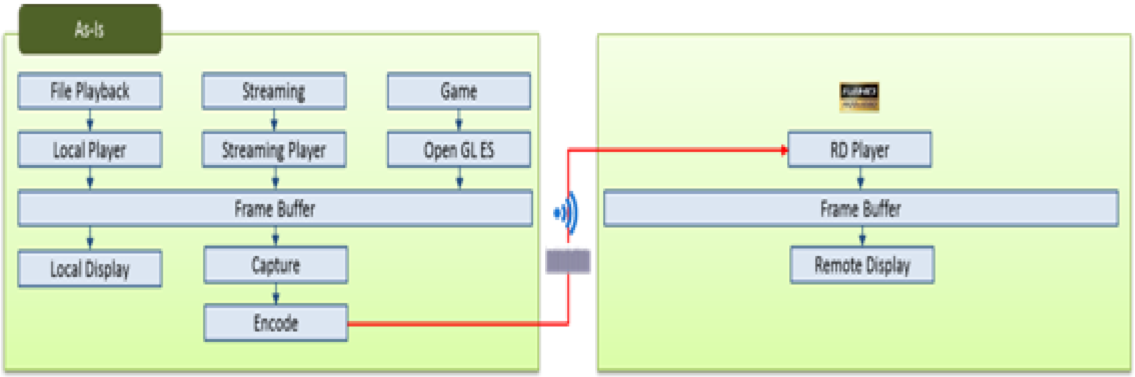


Figure 2. Block diagram for the Screen Mirroring Mode

In the Content Mirroring Mode, the content is transmitted instead of screen to the display device. In this mode, the frameworks are considered in an integrated way to support extended capability.

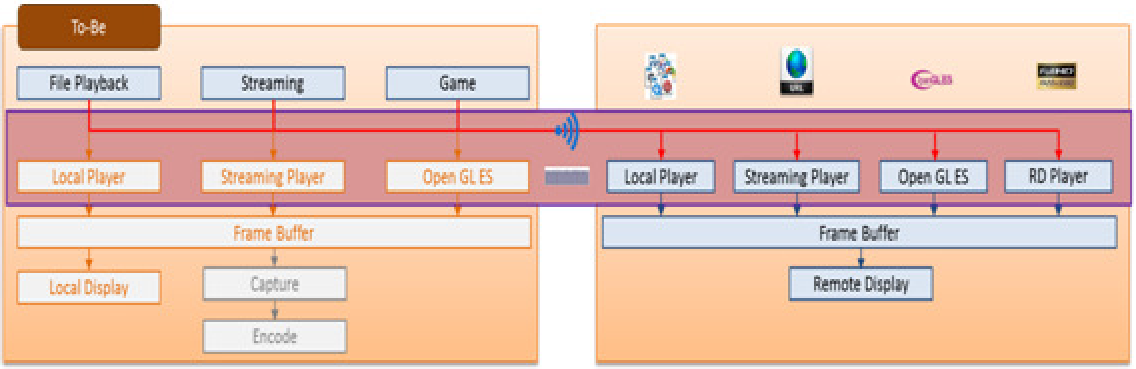


Figure 3. Block diagram for the Content Mirroring Mode

There are 3 main phases for the overall session of suggested solution.

4.1. Processing Capability Negotiation

Once a network connectivity completes successfully, the source and display device negotiate capability which both devices can process. The source device sends a request message to query the processing capability of the display device for contents such as media format/codec, streaming method, graphic library supported by the source device. The display device sends a response message listing contents that it is capable of processing, and then both devices finish the processing capability negotiation.

Table 1. Sample content for capability negotiation

Capability	Example
Text format	TXT
Media format	3GP, MP4, ASF, AVI, MKV
Audio format	3GA, M4A, WMA, MP3
Image format	GIF, PNG, BMP
Streaming protocol	RTSP, HTTP, HTTP live streaming,
Video codec	H.264, HEVC, MPEG4, VP8/VP9,
Audio codec	LPCM, AAC, WMA, MP3, AMR, FLAC, DTS, AC3, RA
Network access	Wi-Fi, LTE
OpenGL version	3.0

If the source device is not capable of playing the AC3 audio codec and the display device has a AC3 decoder, the audio or video file can be played in the display device. In this way, the overall capability can be extended.

4.2. Real-time Traffic Optimizer

User can access contents in the source device while the display device is connected. First, based on the result of processing capability negotiation, adaptive remote display framework determines if currently accessed content can be processed by the display device. If the content is supported by the display device, one of the mirroring mode is selected based on the comparison result of the expected network bandwidth. If the content is not supported by the display device, this step is ignored and the Screen Mirroring Mode is kept.

Table 2. An example of the network bandwidth comparison

Content	Screen Mirroring Mode	Content Mirroring Mode
HD 1.9 Mbps	5 Mbps * 2h 18m = 5.05 GB	1.9 Mbps * 2h 18m = 1.90 GB
HD 5.4 Mbps	5 Mbps * 2h 18m = 5.05 GB	5.4 Mbps * 2h 18m = 5.42 GB

When playing the HD 1.9 Mbps video, the Content Mirroring Mode sends 1.90 GB and it saves 62% network bandwidth compared to the Screen Mirroring Mode. However, when playing the HD 5.4 Mbps video, the Screen Mirroring Mode shows advantage in network bandwidth. In this case, the overall power consumption including encoding and networking should be considered.

4.3. Switching Mirroring Mode

Once a mirroring mode is determined by the remote display framework, the source device informs the display device of new mirroring mode by sending the mode switching message. The message includes the type and detail information about the transmission data. For instance, when switching to video streaming, the message includes data type for video streaming and streaming URI. Once the display device receives the message, it unloads previous processing engine and loads appropriate processing engine.

5. EVALUATIONS

To evaluate the power consumption, Galaxy Note 5 (CPU: Quad-core 1.5 GHz Cortex-A53 & Quad-core 2.1 GHz Cortex-A57, GPU: Mali-T760MP8, Resolution: 1440 x 2560, Wi-Fi: 802.11 a / b / g / n, Android 5.1.1 Lollipop) and the Power Monitor (Monsoon Solutions Inc.) has been used. All the conditions including brightness, network connectivity have been controlled.

For the resolution and bitrate of the remote display, FHD (Full High Definition, 1920x1080) 10 Mbps and UHD (Ultra High Definition, 3840x2160) 10 Mbps have been used.

MX Player and YouTube have been used for local video playback and video streaming, respectively.

The various video clips have been selected for the evaluations.

Table 3. The test video clips

Content	Properties of texture	Properties of movement
Nature Food Landscape	Complexity: low Texture change: mid	Movement: low Background moves slowly
Sport Sea	Complexity: mid Texture change: high	Movement: high
Music Video Movie	Complexity: mid Texture change: high	Movement: mid

5.1. Screen Mirroring Mode vs. Content Mirroring Mode w/ local display

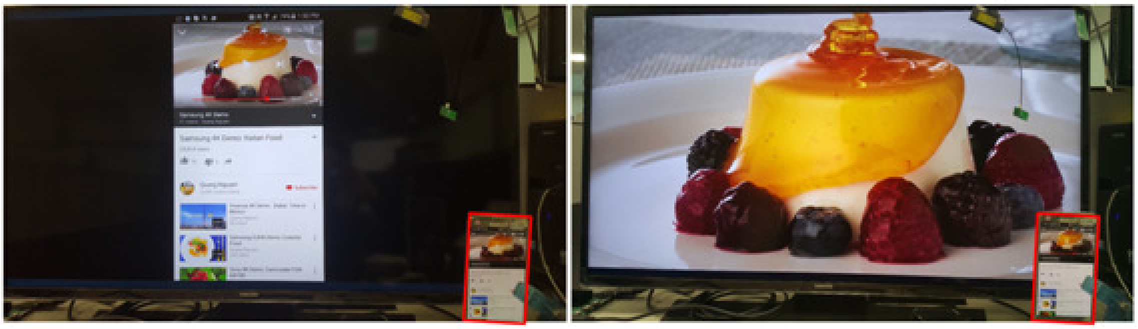


Figure 4. Screen Mirroring Mode (left), Content Mirroring Mode w/ local display (right)

As shown in Figure 4, the Screen Mirroring Mode mirrors the current screen to the remote display. The transmission data type is the encoded screen. So, it shows the same screen on the remote display.

The Content Mirroring Mode transmits raw data such as streaming URI, media file, and audio/video bitstream. The transmission data type is determined according to the capabilities of the display device.

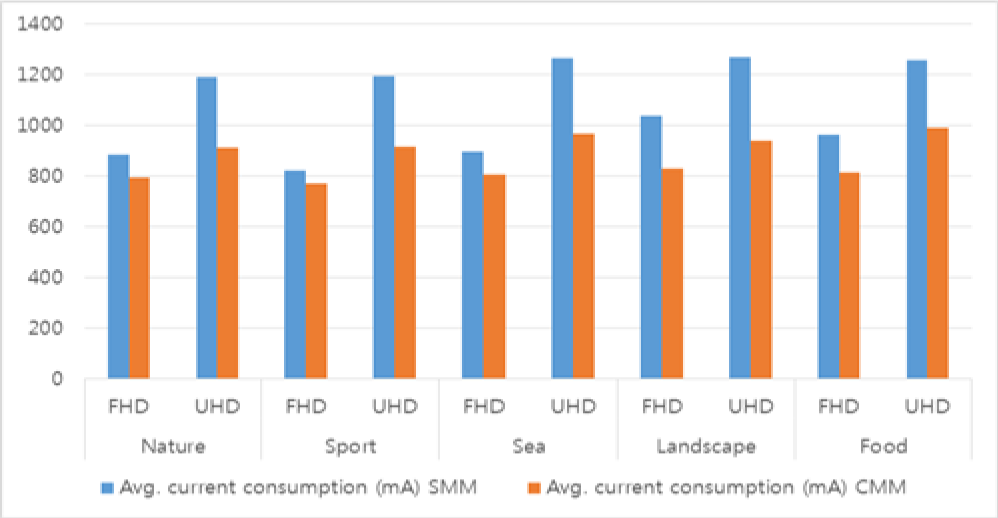


Figure 5. Comparison of the power (current) consumption:
Screen Mirroring Mode vs. Content Mirroring Mode (w/ local display, type: a/v bitstream)

Figure 5 shows the comparison result of the Screen Mirroring Mode and the Content Mirroring Mode which sends the a/v bitstream. In this case, the Content Mirroring Mode skips the resizing and encoding steps. The local display is also performed the same as the Screen Mirroring Mode.

The result shows a 12.4% (FHD) and 23.4% (UHD) improvement in power consumption.

5.2. Screen Mirroring Mode vs. Content Mirroring Mode w/o local display

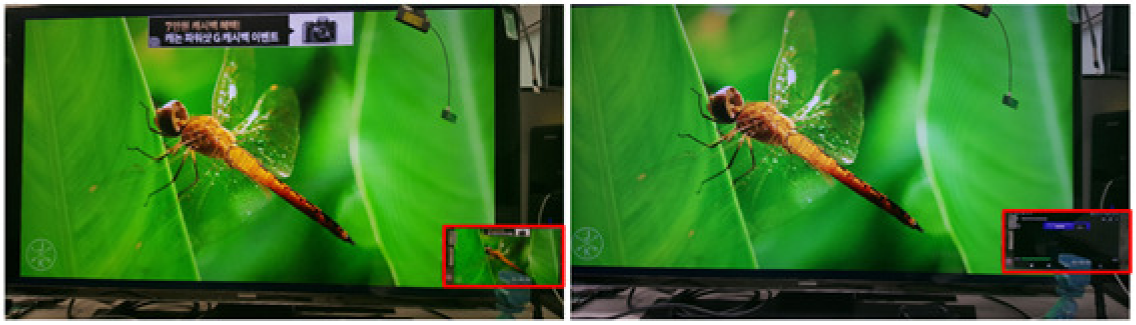


Figure 6. Screen Mirroring Mode (left), Content Mirroring Mode w/o local display (right)

When the user watches a video using the remote display technology, the local display may not be necessary because the video is played in the larger display. So the local display is a possible option we can skip. The Figure 6 shows the screen of the Screen Mirroring Mode which enables the local display and the screen of the Content Mirroring Mode which disables the local display. In this case, the power consumption caused by the local display can also be reduced in the Content Mirroring Mode.

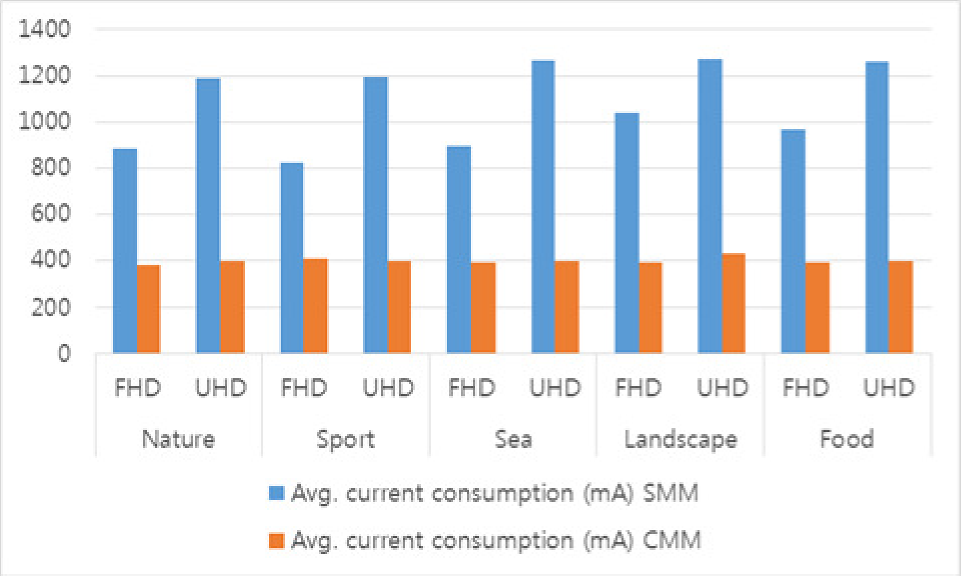


Figure 7. Comparison of the power (current) consumption:
Screen Mirroring Mode vs. Content Mirroring Mode (w/o local display, type: streaming URI)

When the display device is capable of processing streaming URI, which means it has a network connection such as wi-fi and it has an http streaming engine, the source device can transmit streaming URI. In this case, the source device skips the real-time processing for streaming video. As a result, the Content Mirroring Mode shows a 57.1% (FHD) and a 67.4% (UHD) improvement in power consumption.

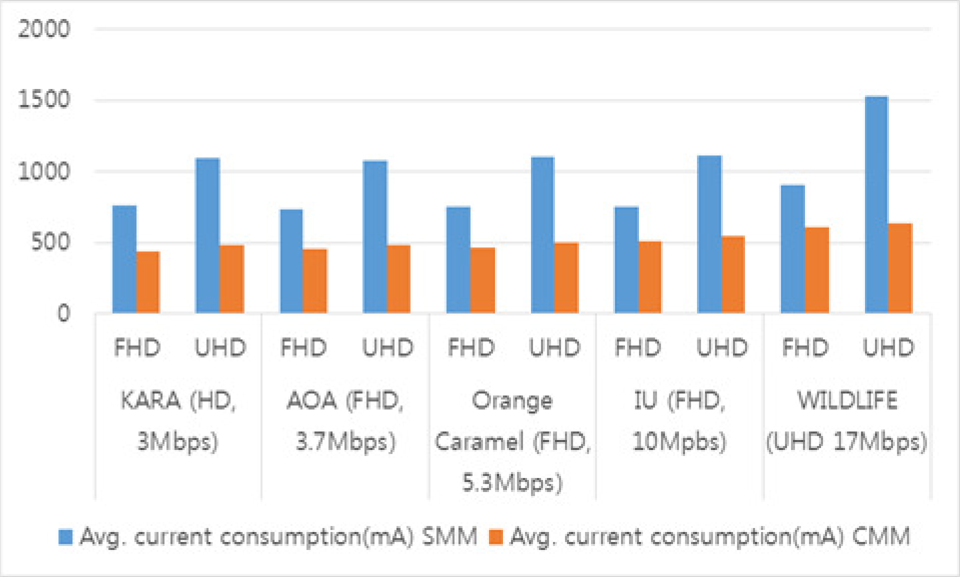


Figure 8. Comparison of the power (current) consumption:
Screen Mirroring Mode vs. Content Mirroring Mode (w/o local display, type: a/v bitstream)

The result shows a 36.5% (FHD) and a 55.0% (UHD) improvement in power consumption.

5.3. Content Mirroring Mode w/ LCD on vs. Content Mirroring Mode w/ LCD off

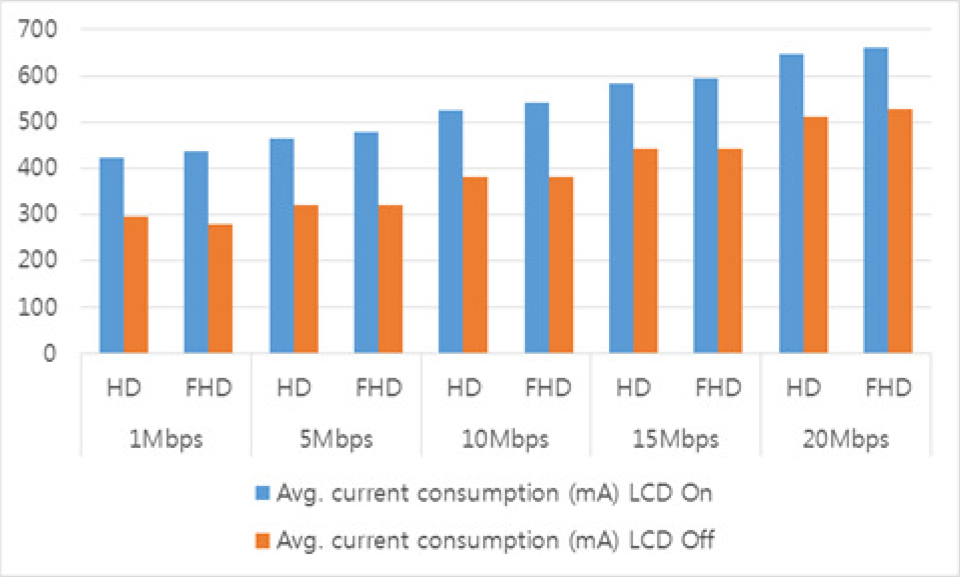


Figure 9. Comparison of the power (current) consumption:
Content Mirroring Mode (with LCD on) vs. Content Mirroring Mode (with LCD off)
(w/o local display, type: a/v bitstream)

If the local display is not necessary, there is another advantage in the Content Mirroring Mode. It can turn the LCD of the source device off. When the user watches a movie which has long running time, this option saves the battery of the source device. When the LCD is off, the result shows an additional improvement of 27.8% in the power consumption is obtained.

5.4. Content Mirroring Mode w/ bitrate change

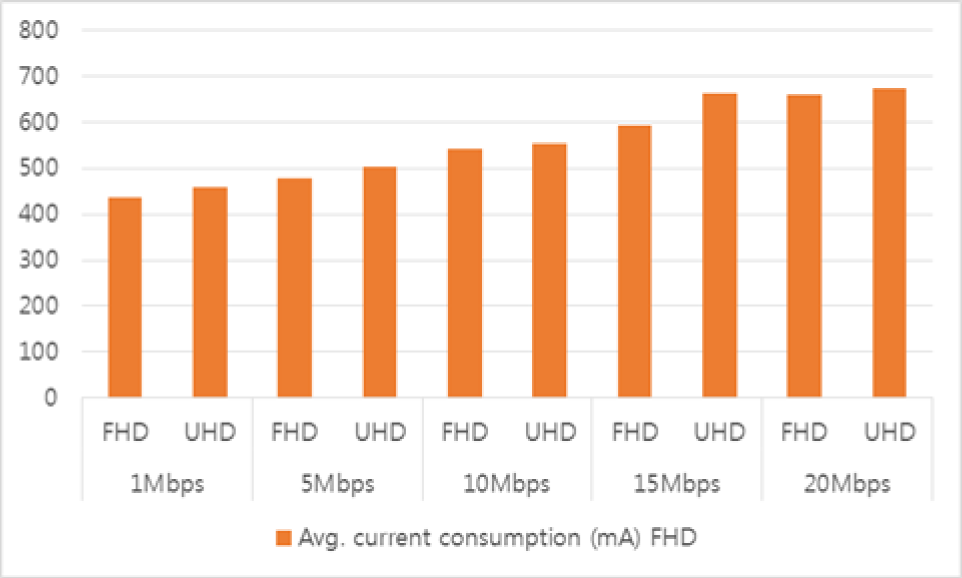


Figure 10. Comparison of the power (current) consumption:
Content Mirroring Mode (w/o local display, type: a/v bitstream)

As the bitrate of a video file increases, the power consumption also increases for data transmission. The result shows that power consumption of 10.5% is added when the bitrate increases by 5Mbps.

6. CONCLUSIONS

The remote display enables users to enjoy multimedia contents on a large screen. An adaptive remote display framework keeps the advantage and extends the overall capability. By reducing the content processing, the source device consumes lower power. Besides, in the Content Mirroring Mode, the amount of data transmitted can also be reduced. As a result, the power efficiency of mobile device is improved.

According to the evaluations, the Content Mirroring Mode which transmits a/v bitstream has shown a 12.4% (FHD 10 Mbps, w/ local display) to a 55% (UHD 10 Mbps, w/o local display) improvement in power consumption compared to the Screen Mirroring Mode. In the case of transmitting the streaming URI, the Content Mirroring Mode shows a 57.1% (FHD 10Mbps) and a 67.4% (UHD 10Mbps) improvement in power consumption. Furthermore, there is additional improvement of 27% in the power consumption if LCD is turned off, which is possible in the Content Mirroring Mode.

REFERENCES

[1] Boyun Eom, Choonhwa Lee, “An adaptive remote display scheme to deliver mobile cloud services”, IEEE Transaction Consumer Electronics, Vol. 60, Aug. 2014

[2] Jisu Ha, Puleum Bae, Keun-Woo Lim, JeongGil Ko, Young-Bae Ko, " Mobile contents on the big screen: adaptive frame filtering for mobile device screen sharing”, Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems, pp. 360-361, Nov. 2014

[3] Screencast in the Wild: Performance and Limitations, Chih-Fan Hsu and De-Yu Chen, ACM MM ’14

[4] http://www.samsung.com/sec/SmartHome/sHome.html

[5] How is Energy Consumed in Smartphone Display Applications?, Xiang Chen, Yiran Chen, ACM HotMobile'13

[6] Profiling Power Consumption on Mobile Devices, Luca Ardito, Energy 2013

[7] Casual Mobile Screen Sharing, Proceedings of the Fifteenth Australasian User Interface Conference (AUIC2014)

[8] Cloud Mobile Media: Reflections and Outlook, IEEE Transactions on Multimedia, vol. 16, no. 4, June, 2014

[9] Mirror, Mirror, On The Wall: Collaborative Screen-Mirroring for Small Groups, TVX 2014, June, 2014

[10] Wi-Fi CERTIFIED Miracast™: Extending the Wi-Fi experience to seamless video display, Wi-Fi Alliance, September, 2012

INTENTIONAL BLANK

HIGHER EDUCATIONAL EMPOWERMENT, ITS IDEAL PARAMETERS AND THEIR CO-RELATION WITH ECONOMY

Mohit Satoskar[1] and Anuprita Satoskar[2]

[1]Department of Computer Engineering, Ramrao Adik Institute Of Technology,
Navi Mumbai, India
mohitsatoskar@gmail.com
[2]Department of Commerce, Mumbai University, Mumbai, India
satoskar.anuprita@gmail.com

ABSTRACT

Co-relation between economy and post-graduation studies are highlighted. Bridging the gap between unemployment and post- graduates are essential in this competitive world. Detailed assessment shows human transition phases. The ideal factors to choose precise post-graduate career is analyzed. Problem related to unemployment is discussed briefly. It is shown how STEM related streams requires concealed professional mid -term work-experience in course structure.

KEYWORDS

Economy, Post-Graduation, Assessment, Unemployment, STEM, Work-Experience

1. INTRODUCTION

To foster and permeate better improved quality life, it is essential to alter change the higher educational structure which will produce successful professional and strengthen economy for betterment of society. In 2013-2014, in India, total vacant seats in Engineering/Technology were 4,62,580 [6]. Speaking about American Education, more than 36 million Americans- a fifth of the working age population, have gone off to college and left without a degree*. [4] However, there are several crucial factors such as tuition hike, lack of skills and opportunities. Non updated curriculum or course structure straight away impact on skilled jobs which further impacts on economy.

Higher educational institutions should deliver generic professionals with the excellence in education which will become pivotal 'ThinkTank' to industry. However, professional work experience in industry during studies will strengthen students ability and will try to boast their confidence level.

However, varsities should contribute industry oriented professional work experience as a top-up in course structure. Ultimately, the process will start from developing and highlighting

David C. Wyld et al. (Eds) : NETCOM, NCS, WiMoNe, CSEIT, SPM - 2015
pp. 195–201, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51616

entrepreneurship skills. By 2025, two thirds of all jobs in United States will require education beyond high school*. [4]

2. AIMS OF RESEARCH

a) To design ideal course structure for STEM courses which will contain professional work-experience as a core module to sharpen practical abilities of students.

b) To analyze inter-dependency between economy and post graduation studies with human transition phases.

c) This study will provide generic solution which will act as 'ThinkTank' for students to choose specialization stream in higher education.

d) This should reduce the rate of unemployment in fresher's.

3. OBJECTIVES

i. This study can built up holistic growth and good employability under competitive market conditions.

ii. The main objective of this research is to improve economy constraints and produce successful professional.

4. CO-RELATION BETWEEN ECONOMY AND POST-GRADUATION STUDIES

The Average life span lived by an individual is grouped according to the decisions the choices that he has to make during that time span and are analyzed accordingly. The choices that the individual makes affects the rest of his life in various aspects. The following groups will clearly state about the importance of the various phases and the importance of the choices.

- Education (15-18): The decisions in this phases are concerned with the potential quality of the rest of the life of the concerned individual. The decisions are regarding the educational directions which later shape their futures.

- Professional Transition (25-30): This is the practical phase of the individuals life span where the person has to choose job. The working environment has to be in such a way where the best of the knowledge acquired in the previous stages can be implemented in the real life.

- The Thirties Assessment (30-35): This is the phase where the person has to think about various other factors as the social factor, or the financial factor, or the decisions regarding family and settlement are taken. The knowledge which he has acquired has to be given back to the society.

- The Forties Transition (35-45): During this time the person analyses the way of life he has lived. The disparity between the dreams and aims he had set for his life the position where he stands is realized by him.

- Mid-age Assessment (45-55): The importance of life of meaning is realized by the individual. The consequences of the decisions and the choices made are to be faced.

- Pre-Retirement Transition (55-65): If the decisions taken in the earlier stages are of the right path they lead to the destination of satisfaction of achieving of the preset goals and aims. And if the decisions are in the any contradiction they lead to aimlessness and the search of the connecting a bridge between the ideal situation and the actual situation.

- The Seventies Transition (65-75): The phase of passing the knowledge to the next generations. The experiences and the consequences are passed to the next generation for their good.

- Seniority Transition (75-85): here the shifting process starts. the individual gradually shifts towards the phase of dependency from the various phases of independency.

5. THE IDEAL AND REALISTIC FACTORS TO CHOOSE POST-GRADUATION AREAS

Factors to be taken into consideration while choosing the path towards the profession of information technology stream are the initiative taken towards being a professional. Presently the factors that has to be considered and the factors which are considered in the reality deviate drastically. And the gap between the ideal factorial model and the present conditions has widened to bridge which concludes in the global problem of unemployment. Currently while choosing following factors are considered by the individual while opting for higher education in the field of information technology-

5.1. Priorities

The individual rather than thinking and considering the most vital factors while opting for this particular profession thinks of the priorities. The priorities such as the luxuries offered by this profession or even the illusion created by the industries about the stability of the profession and the life provided by it.

5.2. Capacity

Capacity of any person is of three types namely physical, emotional, and intellectual. If the person does not think of pushing his capacity as a professional he will not succeed. The capacity is the polished constraint. The capacity of the student is often misunderstood with the natural capabilities that e possess. And such misunderstanding affect the direction chosen by the student while choosing a post graduation program.

5.3. Generalized Theory

While choosing the direction towards being an professional in the field of information technology the students takes into consideration the most generalized factors which are into the minds of the society traditionally over a long period of time. While taking such things into consideration the student forgets the difference between the general constraints and his individual capabilities.

5.4. Financial Constraints

In this fast paced modern era students while choosing the course for their post graduation level make this factor their priority. Rather than analyzing their own abilities and skills they possess naturally and those which they can develop over a period of time by practicing they are more concerned about the financial future that the field of the information technology has to offer and also the financial needs for opting that course.

5.5. Market Analysis

Market conditions is the most crucial external factor affecting the student's decisions. The job opportunities which are actually present for the post graduates and the illusion which depicts the opportunities affect the decision. If the illusion does not provide for a stable future and financial stability the student is hesitant about the career and if the illusion created by the industries suggest anything otherwise there arises the problem of overcrowding which ultimately leads to the problem of unemployment.

5.6. Secondary Experiences

Each graduate taking into consideration the future courses to pursue has to think about the risk factor that comes along. In this unstable and ever-changing era of modernization while turning towards any direction they take into account the experiences of the professionals already from that particular area. Such kind of information is the second hand or the secondary information. The main constraint of individuality is ignored when the decisions taken are based on such secondary information.

6. IDEAL SITUATION CONSIDERATION

The factors which are considered in reality and the factors that actually have to be considered deviate from each other. The factors of the deviation and the constraints that affect the deviation along with the consequences are discussed below:

6.1. Capabilities

Capabilities are nothing but the natural abilities and the gifts that the student possesses. A graduate opting for his post graduation in the profession of information technology has to have certain natural abilities. These natural abilities are the natural gifts and not the ones which the person can develop over the period of time. The most crucial ability that a information technology professional has to possess is the logicality. He has to be very strong in logics and solving the problems with these logics.

6.2. Skills

Skill is the ability which is polished over a period of time with rigorous and continuous practice. The professional from the field of technology has to polish various skills as his communication, the skills of analyzing a particular problem, the skill of providing the most optimal solution, skills of taking the right decisions for the satisfaction of the particular problem in the optimum time period.

6.3. Interests

Interest can be aptly stated as the creative energy that motivates the professional to do the particular thing and take certain decisions. The technology professional must be passionate about the technology, the advancements, and the various other factors that constitutes the technology. He can develop his particular interest areas by opting for the corresponding course areas while choosing a course structure while pursuing his post graduation education.

6.4. Goals

Goals are nothing but the ambitions of the person. In the stream of technology the ambitions is the most primary factor which is supported along with the other factors for being successful in that particular field. Pre determined goals only help one to decide the path to achieve those. Accomplishment of such goals provides the industries with the accomplished information technology professionals.

6.5. Principals and Personality

While taking an decision of striving towards the profession of the information technology the factors of the personality and the principles of the individual even has to be considered. The student has to be constantly aware of the advancement of this fast paced ever-changing world of the technology. To keep the pace along with the quick and ever changing market conditions he has to be a decision taker and a risk bearer. Such qualities will help in accomplishing his preset goals and being a successful individual.

6.6. Combination of Goals and Reality

The preset goals and the ways to achieve these goals along with the real time constraints affect the quality of the professional that is produced by the industry of technology. Not only the goals but the complimentary course structure, the exposure to the post graduate student will develop his knowledge making him ready to face the real time challenges in the working environment and providing the most correct logical solution in the optimum time and resources.

7. PROBLEM OF UNEMPLOYMENT

Because of the various factors discussed above the problem of unemployment arises which can be stated and explained as follows* [2]:

$$(\mathrm{ISC} - RSC) + (ICS - RCS) + Other = \mathrm{Unemployment} \qquad 1$$

ISC= Ideal situation considerations; RSC= Real situation considerations; ICS= Ideal course structure; RCS= Real existing course structure; Other= Other variables; Unemployment= Problem of unemployment;

In the above stated mathematical representation it is stated that the deviation between the ideal situation considerations and the real situation considerations, along with the difference between the ideal course structure and the real existent course structure if combined with various other variables namely the market expectations and conditions or the economies constitute a crucial global problem of unemployment. These all deviations created a huge gap between the industry expectations and the requirements and the quality of the knowledge that the student actually possess which is hard to bridge.

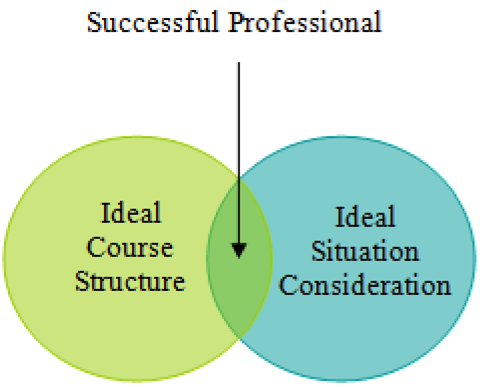


Figure 1 Ideal Strategy for Successful Professional

when the individual is furnished with an ideal course structure where he has considered ideal situation considerations, along with many other dependant variables males a successful professional. Ideal course structure will enhance his abilities as a professional the abilities which he had taken into account while choosing that respective field. And when a successful professional will work in the real time environment his efficiency will not only benefit the organization he is working for but also the economy of the particular nation. This model which proposes the ideal strategy will help us solve many economical issues such as unemployment, poverty, overcrowding and many more.

7. CONCLUSION

Providing a solution to the problem of unemployment through providing a better course structure to the pupils which will enhance their skills and strengthen their abilities to work in the competitive environment efficiently. Also it takes into consideration various ideal factors while opting for that particular field and striving towards success.

However, the outcome will be more realistic for the employer as well which will help to reduce the unemployment rate in developing countries.

ACKNOWLEDGEMENTS

Authors would like to convey sincere gratitude to Hon'ble Anant Tibile (Eminent Indian Novel Writer).

REFERENCES

[1] “General Economics", The Institute of Chartered Accounts of India. New Delhi. ISBN: 978-81-8441-035-8

[2] Satoskar, Mohit, Satoskar, Anuprita. " Strategic Model of Creative Higher Education to Employment for Information Technology". Proceedings of The World Congress on Engineering and Computer Science 2015. San Francisco. U.S.A. pp 320-324. Vol 1. ISBN: 978-988-19253-6-7.

[3] Mohit Satoskar, "Cognitive Prior-Knowledge Testing Method for Core Development of Higher Education of Computing in Academia", FECS'15- The 2015 Int'l Conference on Frontiers in Education: Computer Science and Computer Engineering, Las Vegas, July 2015. Paper ID: FEC6191.

[4] http://www.gatesnotes.com/Education/11-Million-College-Gradss (accessed: 15 November 2015)

[5] "World Education Report", 2000, United Nations Educational, Scientific and Cultural Organization.

[6] http://www.aicte-india.org/downloads/parliment_questions/PQ_Supplementary_362.pdf (accessed: 20 November 2015)

[7] http://www.uis.unesco.org/FactSheets/Pages/Literacy.aspx

[8] Report on Adult and Youth Literacy, 1990-2015, UNESCO. Canada. ISBN: 978-92-9189-117-7

[9] Satoskar¹, Mohit and S. Mali. "Comprehensive Curriculum of Programming for Engineers, its Teaching Models and in-Lab Monitoring Technique", Proceedings of the International Multi-Conference of Engineers and Computer Scientists. Vol. 1. 2015. ISBN: 978-988-192530209

AUTHORS

Mohit Satoskar received Bachelors of Engineering (Hons) degree in Computer Systems Engineering from London Metropolitan University, United Kingdom in 2012. Currently, Mohit is a Research Associate with Department of Computer Engineering in Ramrao Adik Institute of Technology, Navi Mumbai. India. Mohit has published several research papers, see the latest http:// independent.academia.edu/ MohitSatoskar/Activity. He did consulting for ECHS- Indian Armed Forces. He is a Certified Ethical Hacker. Mohit's current research interests are: natural multi-language processing, educational technology, e-governance, management information systems, curriculum design and implementation.

Anuprita Satoskar is a student of Institute of Chartered Accountants of India. She is also with University of Mumbai pursuing Bachelor of Commerce. Anuprita is a National Award Winner of TATA Building India competition.

INTENTIONAL BLANK

REAL HUMAN FACE DETECTION FOR SURVEILLANCE SYSTEM USING HETEROGENEOUS SENSORS

Yoon-Ki Kim[1], Doo-Hyun Hwang[2] and Chang-Sung Jeong[3]

[1,2,3]Department of Electrical Engineering, Korea University, Seoul, South Korea
[1]vardin@korea.ac.kr
[2]doohh88@korea.ac.kr
[3]csjeong@korea.ac.kr

ABSTRACT

Face detection algorithms are used to detect the human in various industry fields. A typical face detection algorithm such as Haar Feature-based Cascade Classifier gives us an easier way to detect human face. It consists of several classifiers which contain complicated arithmetic operations. Several classifiers constitute the cascade which can detect each element of human face. The more cascades are contained in the algorithm to detect elements of human face, the more it takes a time to detect human face. The previous cascade hardly recognize real human, since previous cascade processes only one source from image source. In this paper, we present a new cascade method for human face detection which exploits several classifiers for data not only from image source but also various heterogeneous sensors. Cascades consist of various sensors based on tuple data type could be operated quickly. It provides more accuracy of real human face detection, reduces the number of classifier for high speed processing in real-time.

KEYWORDS

Face Detection, Heterogeneous Sensor, Real-Time Processing, Haar-Like Feature

1. INTRODUCTION

In Internet of Things environment with wired/wireless sensor networks, efficient sensor data process are very significant for various useful data analysis [1]. Various sensors such as CCTVs, thermo-graphic camera and temperature sensors can be processed at the same time for more accuracy analysis. Those sensors notice different signal respectively. For example, CCTV notices image signal to detect face shape, thermo-graphic camera notices image signal to detect face temperature and gas sensor notices amount of gas in air. This heterogeneous sensors detect not only one sense but also various senses. Various sensors can enhance the accuracy of real human face detection in real-time environment.

Haar Feature-base Cascade is a useful algorithm in wide range of object detection application [2]. Cascading is a particular case of ensemble learning based on the concatenation of several classifiers, using all information collected from the output from a given classifier as additional information for the next classifier in the cascade [3]. In face detection field, classifier processes

David C. Wyld et al. (Eds) : NETCOM, NCS, WiMoNe, CSEIT, SPM - 2015
pp. 203–210, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51617

multimedia data from one source. A typical method is that first classifier detects the face shape, and then next classifiers can detect other shape such as eyes, mouth and nose in face shape. However, the more cascades are contained in that, the more it takes a time to finish. Consequentially, there is a trade-off between accuracy of result and processing speed. In this paper, we present a new cascade method for human face detection which exploits several cascades for data not only from image source but also various heterogeneous sensors. It provides more accuracy of real human face detection and reduces a number of classifier to high speed processing in real-time detecting. For this approach, we need to synchronize between each sensor, so that sensors data can be processed at the same time. Using this method, we can improve the accuracy of face detection.

The outline of our paper is as follows: In Section 2, we describe related works for introducing Haar Feature-base Cascade. And time synchronization method for various sensor. Then, in Section 3, we explain a new method using classifiers which process data from various sensors. Section 4 explains implementation of proposed method and shows its experimental results. Lastly Section 5 summarizes the conclusion of our research.

2. RELATED WORKS

Haar Feature-based Cascade is fast object detection algorithm [2] using Haar-like features and a cascade of classifiers. It has good detection rate depends on training data. And it calculates 2 frames data per second so that it can process in real-time. This algorithm consists of 4 stages. First stage is haar feature selection. Haar-like features can be made by calculating difference of the sum of pixels of areas inside rectangle. There are many haar features in a frame. This feature has too many operation to service in real-time. For this, it use second stage which has integral method to calculate quickly. And third stage is Adaboost training. Adaboost selects useful haar feature in total haar features using weight of each haar filter [4]. Each selected features can be trained data to classify true positive. Then, Adaboost can compose strong classifier which is consists of weak classifier. Last stage is to make cascading classifiers. These cascading classifiers is step by step method which is made by several weak classifiers. Firstly, top simple classifier judge the features whether it is true or false. If first classifier classify features as a true positive, it could be passed the next stage which consists of another weak classifiers. This method can reduce a lots of operation by using classifier cascading. Once, a classifier reject the features, It is regarded as false so that it cannot be passed next classifiers. All features pass the classifier cascading, it is targeted as an object.

There are various researches for processing sensor data from heterogeneous sensors[5-7]. Those sensors data are detected different elements respectively so that enhance the accuracy of detection result. This various elements can be used source of classifier cascade. For example, temperature sensor and weight sensor are a great help to detect real human.

Previous face detection approaches consider only multimedia source from a camera. Those methods have so many classifiers for high detection rate that it takes a great time. Our research goal is to enhance the true positive rate of detection using various sensors to reduce a number of classifiers.

3. FACE DETECTING USING HETEROGENEOUS SENSORS

In this section, we present a new architecture of face detection system using heterogeneous sensors for detecting real-human. Unlike typical face detecting systems, our system has additional classifiers to process various sensor data

3.1 Key features

This system has several key features as follows:

(1) It offers minimum number of cascades so that it reduce volume of operation. Typically, to enhance the true positive rate of detection, it would be a lot of cascades such as face cascade, eye cascade, nose cascade or mouth cascade. However, there is a trade-off between accuracy of result and processing speed. The more cascade are contained, the more it takes a time. our system select minimum number of cascades for high speed processing

(2) It offers time-stamp for processing the various sensors data at the same time. And those data come separately. Thus, it needs to synchronization for various sensors data. Our system set the time-stamp for synchronization.

(3) It offers real-face detecting except picture, doll using feature of human. A surveillance system has to detect real human, this system can extract feature of human using sensor such as temperature sensor. Moreover, multiple source enable system to detect various sense not only vision but also touch sense, weight sense, heat sense and so on. It is helpful to detect objects exactly which we want. In a cascade step, it judges sensor data whether it is necessary or unnecessary by using several classifier. If it is considered as true positive, it is passed next phase of cascade from other sensors.

3.2 System Model

The overall operation of our system model as shown fig 1. Our system model consists of train phase, synchronization phase and cascading phase. This operations shall be explained bellows.

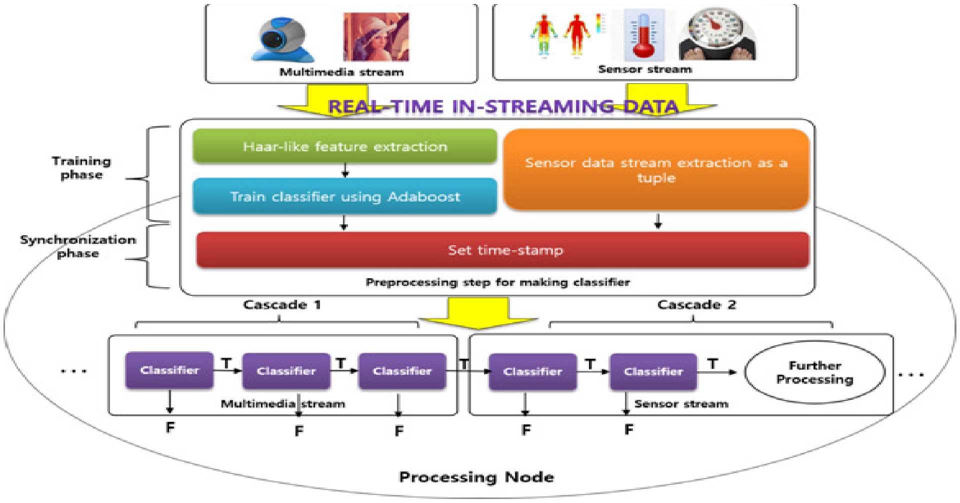


Figure 1. The overall model of face detecting system using heterogeneous sensors

3.2.1 Training Phase

In training phase, there are two section to make classifiers. First section is for training multimedia data from camera sensor in real time. Haar-like feature extraction and Adaboost make the classifier based on multimedia data. Second section is for extraction tuple from various sensors data except camera. Its section collects the sensors data and extracts tuple in real time. However, it is different time between each section yet. So it needs synchronization of time in next phase.

3.2.2 Synchronization Phase

In synchronization phase, it synchronizes the time between haar-like features and sensor tuples. Those sensors data are detected different elements respectively. Thus, it has different time stamp. To synchronize their time, sensor data set their time every frame-rate cycle so that it reduces volume of calculation. If its frame rate is 12 fps, other sensors data set their time every 12 frame. The time of Multimedia is standard-time. Figure 2 shows an example of this method.

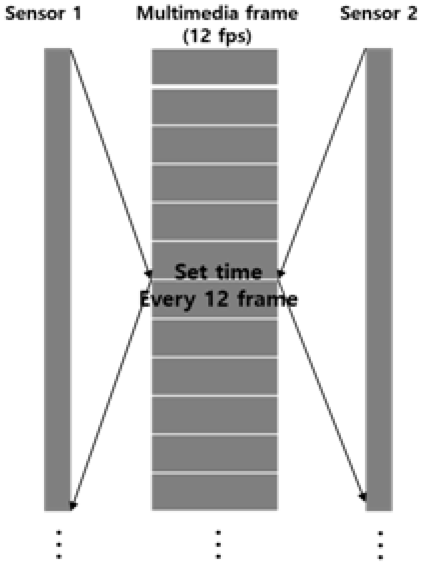


Figure 2. An example of synchronization method

3.2.3 Cascading Phase

In cascade phase, it makes robust classifier which consists of week classifier. Its cascades is connected to one another. To detect face, face cascade is set on head stage. Then, other cascades are set on next stage. It is mandatory that prior cascade judges true feature before posterior cascades. If prior cascade judges false, it doesn’t pass the opportunity to next cascade. There are cascades made by Adaboost algorithm for multimedia sensor process. The rest of cascades are made by range detector for heterogeneous sensors. The cascade consists of various sensors data classifier as shown figure 3.

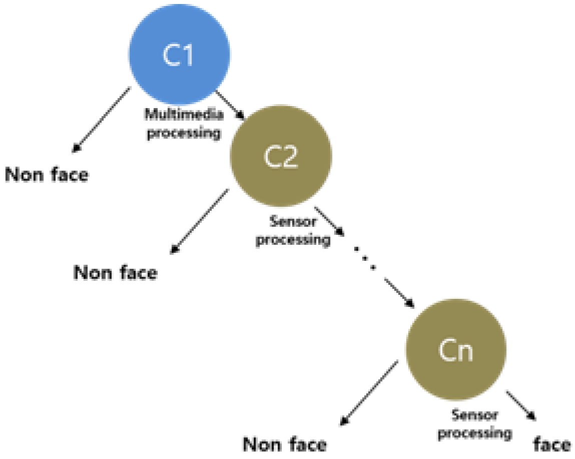


Figure 3. The model of cascade including multiple source processing

4. IMPLEMENTS

In this section, we shell show the implementation of our new system. We implement a face detection with a temperature data. There are six cases for this implementation. First is a real-human detecting with face cascade and no temperature cascade. Second case is the picture of human with face cascade and no temperature cascade. Third case is the picture of human with face cascade and temperature cascade. Fourth case is the real-human with face cascade, eyes cascade and no temperature cascade. Fifth case is a picture of human with face cascade, eyes cascade and no temperature cascade. The last case is a picture of human with face cascade, eyes cascade and temperature cascade. The implement cases of implementation as shown blows.

Table 1. Various cases of implement

Case Number	Object Type	The number of Cascades	Additional Sensor
Case 1	Real face	1 (face)	No sensor
Case 2	Picture	1 (face)	No sensor
Case 3	Picture	1 (face)	Temperature sensor
Case 4	Real face	2 (face, eyes)	No sensor
Case 5	Picture	2 (face, eyes)	No sensor
Case 6	Picture	2 (face, eyes)	Temperature sensor

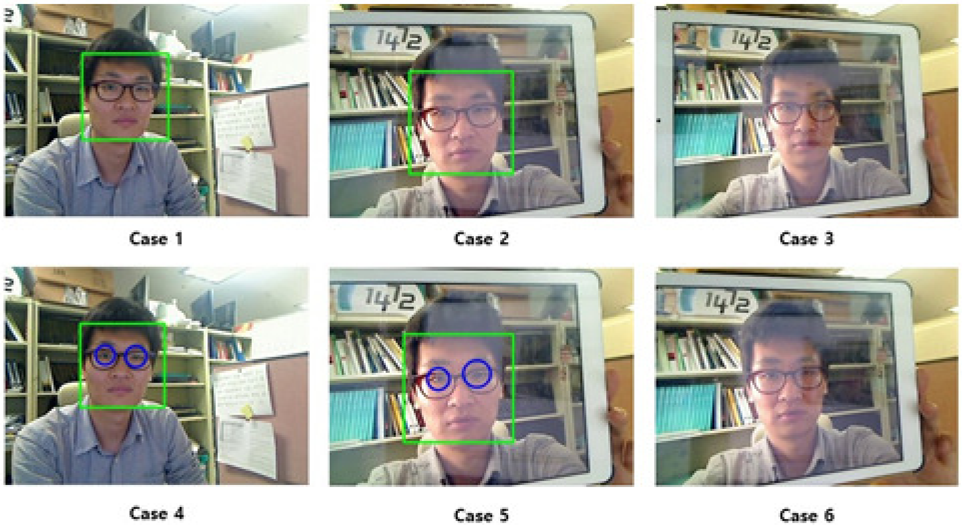
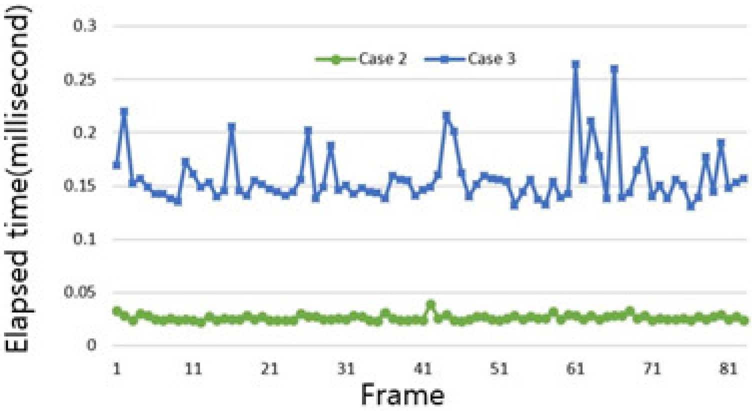


Figure 4. Result of implementation

5. EXPERIMENTAL RESULTS

We implement this system on 1 node which has Intel® core™ quad CPU Q6600 2.40 GHz processors and 8GB memory. The experimental results show that real human face-detecting system which has cascades from various sensors enhance the accuracy of detecting real human. Case 2 without temperature cascade detects the face. It is false positive. However, case 3 has no detection of face. It is true negative. Figure 5 shows the relation between the numbers of cascades and elapsed time. Case 4, 5 has eyes cascade additionally. Those case take a lot time to calculate than case 1, 2, 3. Because it contains many operation to extract haar feature since it contains eyes cascade. Case 2 is faster than case 3. It means that the more cascades are contained in that, the more it takes a time to detect. Because cascade based on multimedia data which contains many 'for statement', it can reduce the elapsed time by reducing cascade or using cascade form sensor data instead of multimedia data. Although case 6 contains eyes cascade, it processes the detection step faster than case 4 and 5. Because Temperature cascade does not pass the opportunity to eyes cascade. As a result, the composition consists of various cascades from heterogeneous sensors are the helpful to detect of real human-face.



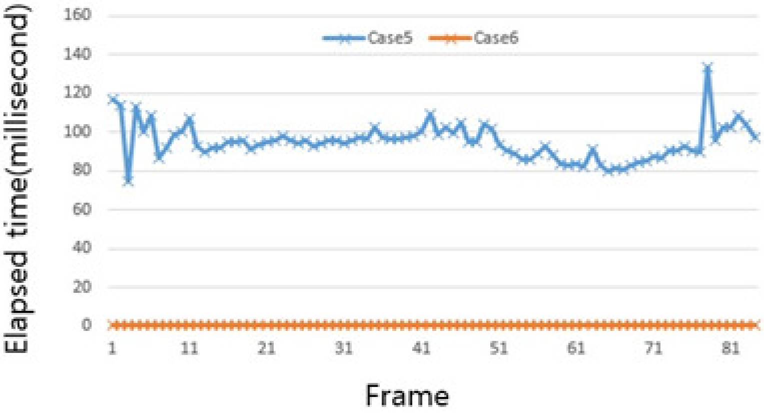


Figure 5. Result of experiments

6. CONCLUSIONS

In this paper, we have presented a cascade method for human face detection in surveillance system which exploits several classifiers for data not only from image source but also various heterogeneous sensors. A typical face detection algorithm such as Haar Feature-based Cascade Classifier gives us an easier way to detect the face. However, it consists of several classifiers which contain complicated arithmetic operations so that it takes a great deal of time to achieve a result. And it is hard to recognize real human face in short time, since previous cascade processes only one image source. Our method can enhance an accuracy of face detection using heterogeneous sensors. It uses a cascade which consists of classifiers. Each classifier processes data from not only image source but also various sensors data. It provides more accuracy of real human face detection and reduces the number of classifiers for high speed processing in real-time detecting.

ACKNOWLEDGMENTS

This research was supported by Korea university and MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support pro-gram (IITP-2015-H8501-15-1004) supervised by the IITP(Institute for Information & communications Technology Promotion)

REFERENCE

[1] Yu, Byunggu, Ranjan Sen, and Dong H. Jeong. "An integrated framework for managing sensor data uncertainty using cloud computing." Information Systems 38.8 (2013): 1252-1268.

[2] P. Viola and M. J. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," In Proc. of CVPR 2001.

[3] Gama, João, and Pavel Brazdil. "Cascade generalization." Machine Learning41.3 (2000): 315-343.

[4] VIOLA, Paul; JONES, Michael. Fast and robust classification using asymmetric adaboost and a detector cascade. Advances in Neural Information Processing System, 2001, 14.

[5] Yu, Byunggu, et al. "On managing very large sensor-network data using bigtable." Cluster, Cloud and Grid Computing (CCGrid), 2012 12th IEEE/ACM International Symposium on. IEEE, 2012.

[6] Jung, I. Y., Kim, K. H., Han, B. J., & Jeong, C. S. (2014). Hadoop-Based Distributed Sensor Node Management System. International Journal of Distributed Sensor Networks, 2014.

[7] Kui, X., Sheng, Y., Du, H., & Liang, J. (2013). Constructing a CDS-based network backbone for data collection in wireless sensor networks. International Journal of Distributed Sensor Networks, 2013.

AUTHORS

Yoon-Ki Kim is currently working toward the ph.D degree in Electronic and Computer Engineering at the Korea University. His research interests include real-time distributed and parallel data processing, IoT, Sensor processing and computer vision.

Du-Hyun Hwang is currently working towards a master’s degree at Department of Electrical Engineering, Korea University. His current research interests are distributed parallel computing, computer vision and GPU processing

Chang-Sung Jeong is a professor at the department of EE/CE at Korea University. He received his MS.(1985) and Ph.D.(1987) from Northwestern University, and B.S.(1981) from Seoul National University. Before joining Korea University, he was a professor at POSTECH during 1982-1992. He also worked as an associate researcher at UCSC during 1998-1999.

REAL-TIME PEDESTRIAN DETECTION USING APACHE STORM IN A DISTRIBUTED ENVIRONMENT

Du-Hyun Hwang[1], Yoon-Ki Kim[2] and Chang-Sung Jeong[3]

[1,2,3]Department of Electrical Engineering, Korea University, Seoul, Republic of Korea
[1]doohh88@korea.ac.kr
[2]vardin@korea.ac.kr
[3]csjeong@korea.ac.kr

ABSTRACT

In general, a distributed processing is not suitable for dealing with image data stream due to the network load problem caused by communications of frames. For this reason, image data stream processing has operated in just one node commonly. However, we need to process image data stream in a distributed environment in a big data era due to increase in quantity and quality of multimedia data. In this paper, we shall present a real-time pedestrian detection methodology in a distributed environment which processes image data stream in real-time on Apache Storm framework. It achieves sharp speed up by distributing frames onto several nodes called bolts, each of which processes different regions of image frames. Moreover, it can reduce the overhead caused by synchronization by computation bolts which returns only the processing results to the merging bolts.

KEYWORDS

Distributed stream processing, Apache storm, Image Processing, Pedestrian Detection.

1. INTRODUCTION

Recently, the data in IT industry has been dramatically increasing. Besides, the volume of the data is also increasing continuously. Especially, size of digital pictures and resolution of video is bigger than before. In addition to this, the sharp data increase for services in an era of Internet of Things(IoT) makes it difficult to process image data stream in real-time in just one node. Therefore it is essential to process large-scaled stream image data in a distributed environment.

In this paper, we propose a pedestrian detection methodology which is an efficient model for dealing with image data stream in the distributed environment. We use Apache Storm[1] for the implementation which is a distributed stream processing framework in real-time. Apache Storm runs topologies on a Storm cluster which consists of spouts and bolts. The spout is a streamer task which makes sequence of tuples that is a data model of Storm and bolts are tasks for processing jobs. A target of our distributed image stream processing is a pedestrian detection. The pedestrian detection is an important technique which can help to prevent many accidents in

David C. Wyld et al. (Eds) : NETCOM, NCS, WiMoNe, CSEIT, SPM - 2015
pp. 211–218, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51618

an automobile field and creates profit by analysing the number of customers on shops. Therefore we consider for pedestrian detection and propose an efficient way for detecting pedestrians on the distribute environment.

Our methodology is has several advantages as follows: Firstly, it speeds up the processing by being operated in parallel on the each node by distributing frames onto several computation. Secondly, it can reduce the overall computation load by dividing the frame into several region to detect on the each frame. Finally, the overhead caused by synchronization can be reduced by returning only the processing result.

The outline of our paper is as follows: In section 2, we describe related works about Apache Storm framework and pedestrian detection algorithm which we selected. In section 3, we present a model for processing pedestrian detection efficiently on the distributed environment and explain a topology for a workflow running on the Storm cluster. In section 4, we explain our results of experiment.

2. RELATED WORKS

2.1. Distributed Stream Processing

Recently, big data systems like Hadoop [2] have been used on various fields. Hadoop is populist platform on the big data systems. However, applications become more various, and users have needed to get process results more quickly. Therefore many stream process platforms appear to provide services which can process the big data in real-time.

Big data needs to process data in distributed and parallel environment because it is not structured like data saved in database. For this solution, Hadoop appear which can process big data using MapReduce [3] that is parallel processing framework. However, according to increase of data like RFID, twit and CCTV, batch systems like Hadoop have reached limit to process these large-scaled data. For resolving this problem, various techniques have been suggested to process it.

2.2. Apache Storm

Apache Storm is a distributed stream data processing system, which has been used in Twitter for various critical computations. Apache Hadoop is an essential framework for distributed processing large-scaled data and already has been used in Hadoop ecosystem in many ways these days. However, it does not cover real-time stream processing. For this reason, Apache Storm appeared and has been a solution for real-time processing. It is possible to use on the Hadoop or alone with Zookeeper [4] that is nodes manager.

Apache Storm has various features. Firstly, it guarantees fault-tolerant and high availability. If errors happen on the process works, Storm reassign it immediately. Therefore, the works can be operated continuously. Secondly, its latency of process is short, because it does not save data and processes in real-time. Third, it is scalable. It can add additional nodes on the Storm cluster easily. Finally, it guarantees reliable. Every data can be processed without any loss.

Storm architecture is similar with Hadoop. It consists of one master node which is called Nimbus and one or a couple of worker nodes which are called Supervisor. In addition, Storm relies on Zookeeper for managing nodes in Storm cluster. Zookeeper gives information of supervisor’s state to Nimbus. Then, Nimbus assign works to supervisors and each supervisor processes assigned tasks.

Storm's data model is unbounded sequence of tuples which is consist of a field and a value. The stream of tuples flows through topologies, which are directed graphs. The topology's vertices represent computations and the edges represent the data flow. The vertices divided into two type, Spout and Bolt. Spout is a supplier of stream, which reads tuples form sources and provides it to topologies. And bolt is processing unit. Every works of Storm is on the Bolt, and it emits the results of processing to other bolts.

2.3. Pedestrian Detection

Pedestrian detection techniques are for finding people on the load. Mostly research in this filed has focused in finding people standing than sitting or lying, although it could be useful for saving a life in a disastrous situation. Anyway, this technique is useful in many ways like counting people shopping in shops and warning to people about dangers ahead. The applications of pedestrian detection technique are striking.

Detectors in OpenCV[5], open vision library, are representative among published detectors.

1. The histogram of Oriented gradients (HOG) [6] (INRIA) – HOG detector of Dalal2005 [6], which is learned through INRIA Person Database. It is difficult to detect object of a template which is smaller size than 64(w)x128(h).

2. HOG(Daimler) – HOG detect which is learned through Daimler Pedestrian Dataset. It could detect objects of a small template.

3. Hogcascades – Detector which is applied cascade technique in HOG feature.

4. Haarcascades – Detector of Viola2001 [7], whose speed to detect objects is fast that others.

Haarcascades was used as our detector for finding pedestrians, since our purpose of this paper is just detecting in distributed environment rather than focusing in accuracy of detection.

3. Distributed Pedestrian Detection Model

To detect pedestrians in the distributed environment, we propose a methodology and a workflow for the detections.

3.1. Methodology of Detection

The frame rate of video is higher than before. The 30fps's CCTV translates 30 frames during 1 secondly. In this situation, the location of objects have changed little between a few continuous frames since the frame rate is too fast for pedestrians to move to other locations. Figure 1 shows that there are little changes of pedestrian's location between continuous frames of the video. We suggest the efficient methodology for detecting pedestrians in the distributed environment in real-time using the characteristic explained above.

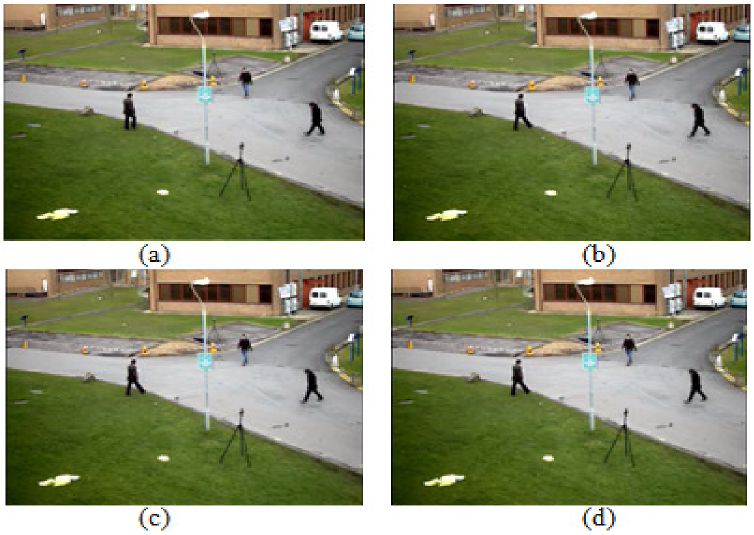


Figure 1. The continuous frames of ftp 15 video (a) Frame 1, (b) Frame 2, (c) Frame 3 and (d)

3.1.1. Distributing Frames Phase

In distributing frames phase, the frames of the stream source are distributed onto several nodes for processing to detection in parallel. If the fps of image data stream is very high, sampling frames is needed for reducing workload. Since the locations of the pedestrians has little changed between continuous frames, sampling can reduce execute time.

3.1.2. Detecting Pedestrian Phase

In detecting pedestrian phase, the pedestrian detection is operated in each node in parallel. As noted above, the detection does not need to be operated in every part of the frame since the locations of pedestrians have not changed between a few frames in case of image data stream of high fps. Therefore, each node operates to detect pedestrians on the different regions of the different frames to reduce the overall computation and processing time.

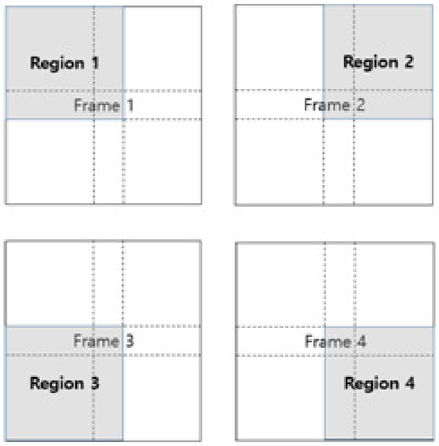


Figure 2. Detect Regions of each Frames

3.1.3. Processing Results Transmission phase

Each node which is operated for detection does not transmit the frames as the results, but only the information about the locations of pedestrians. For this method, a node for merging this results receives the original frames from the source separately. Transmitting only the information about the location of pedestrians can reduce network load caused by communications, since it can reduce the size of total traffic. Moreover, the total execute time can be reduced by removing a function of synchronization, since the original frames are transmitted separately. If the frames are transmitted as the results from the each node which process to detect pedestrians, the synchronization is needed for arranging the frames in order.

3.2. Workflow

In this section, we explain the topology which is run on the Storm cluster. For reducing the network load and the works, features are delivered instead of frames. If the frames as the result of the detection are delivered, synchronization between the frames will be needed and happen network load for delivering every frame again.

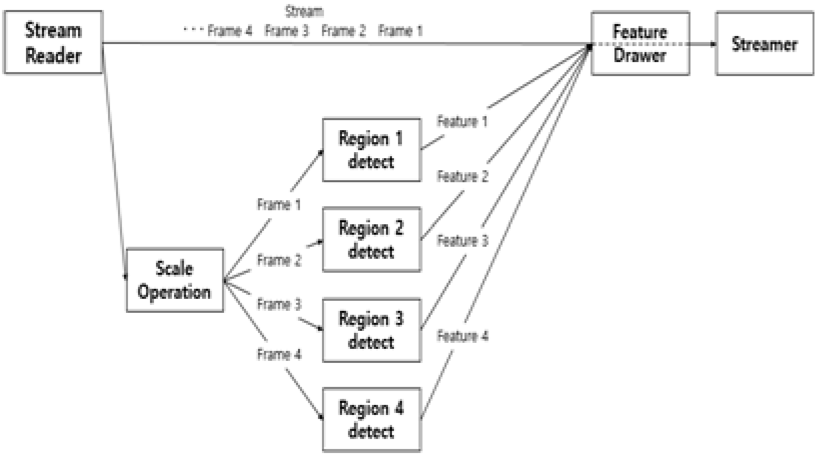


Fig 3. The topology of the detection

The workflows are as follows.

(1) Spout: The spout generates stream of sequential tuples from video sources. Video frames are serialized for delivering as tuples and emitted to Bolt for scale operation and feature drawer. Received tuples are deserialized to frames for processing image.

(2) Scale Bolt: It controls the volume level of the frame properly to reduce the load of each operations. However, if frames' volume level become too small, it comes difficult to operate the detection.

(3) Detect Bolt: Its operation is to detect pedestrians in parallel. Each bolt receives the frames in sequence and detects the pedestrians in particular area. For example, detect bolt 1 receives the frame 1 and finds pedestrians in some part of the frame. Detect bolt 2 receives the frame 2 and find pedestrians as well but in another part of the frame 2. If pedestrians are detected, then the bolt emits the features to a next bolts, which are information of the detection result.

(4) Feature Drawer: feature drawer bolt receives features from the detect bolts, which are information about detected pedestrians. At the same time, it receives frames directly from the spout simultaneously. Then, it draws the features on the received frame.

(5) Streamer: Streamer bolt emits the tuple stream of the result as mjpeg format to web service to watch the result of processing.

4. IMPLEMENTATION

4.1. Experiment Environment

Our system described above is implemented on a laptop with an Intel® Celeron® CPU B800 @ 1.50GHz processor running Window8. Program is developed by Java using eclipse luna using Storm framework. The library is used the OpenCV. Distributed environment is composed by Oracle VM ViktualBox. Three virtual machine are generated and run Linux Ubuntu 14.04 64bit. Apache-Storm 0.9.5 and Zookeeper 3.4.6 is installed for running Storm.

4.2. Detection Result

Figure 4 shows the result of pedestrian detection. Since the location of pedestrians is little changed between frames, it is possible to detect every pedestrians separately.

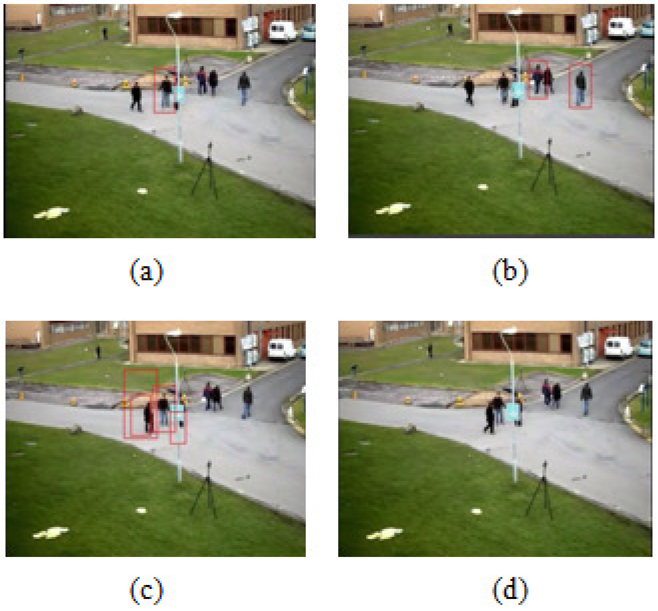


Figure 4. Detection pedestrians on each part of sampled frames. (a) upper left, (b) upper right, (c) bottom left, (d) bottom right

4.3. Performance Evaluation

In this section, we explain about the evaluation of our experiment.

Figure 5 shows overall information of Topology which operates pedestrian detection and the data flows. Table 1 and Table 2 show detail information of spout and bolts. Executors are threads in a worker process. Pedestrian bolt which detects pedestrians in a frame has 4 executors for reducing execute latencies.

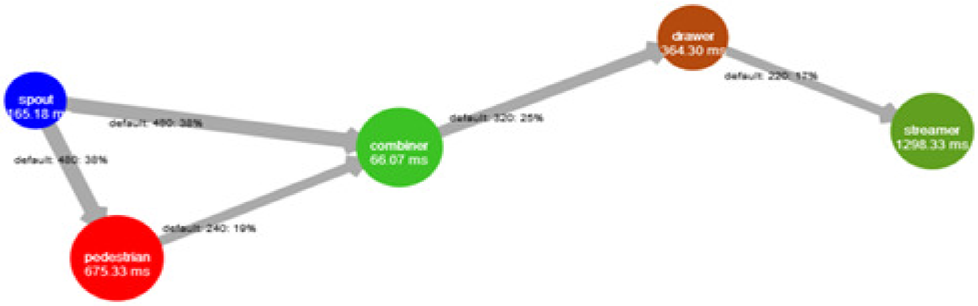
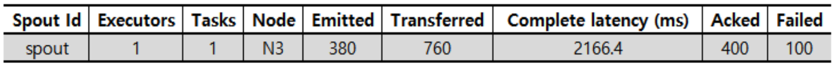


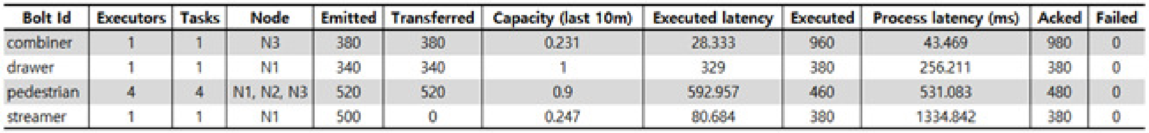
Figure 5. Topology stats visualized

Table 1. Spouts stats



Spout Id	Executors	Tasks	Node	Emitted	Transferred	Complete latency (ms)	Acked	Failed
spout	1	1	N3	380	760	2166.4	400	100

Table 2. Bolts stats



Bolt Id	Executors	Tasks	Node	Emitted	Transferred	Capacity (last 10m)	Executed latency	Executed	Process latency (ms)	Acked	Failed
combiner	1	1	N3	380	380	0.231	28.333	960	43.469	980	0
drawer	1	1	N1	340	340	1	329	380	256.211	380	0
pedestrian	4	4	N1, N2, N3	520	520	0.9	592.957	460	531.083	480	0
streamer	1	1	N1	500	0	0.247	80.684	380	1334.842	380	0

5. Conclusions and Future Works

In this paper, we have presented the pedestrian detection methodology in the distributed environment using Apache Storm framework. Actually, the image processing in the distributed environment is not appropriate due to network load problem caused by communications between nodes for translating frames. However, we cannot but use distributed processing since the only one node cannot cope with processing high resolution and high fps image data stream in real-time. To conclude, we have suggested the methodology for detecting pedestrian in real-time in the distributed environment to reduce works as distributing the frames and dividing the region for detection into several parts. Besides, it reduces the overhead cause by synchronization by returning the only processing result. In the future, stream image processing in the distributed environment will be an essential technique and need research to reduce network load problem between nodes.

Acknowledgements

This research was supported by Korea university and MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2015-H8501-15-1004) supervised by the IITP(Institute for Information & communications Technology Promotion)

References

[1] Toshniwal, Ankit, et al. "Storm@ twitter." Proceedings of the 2014 ACM SIGMOD international conference on Management of data. ACM, 2014.

[2] Shvachko, Konstantin, et al. "The hadoop distributed file system." Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on. IEEE, 2010

[3] Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." Communications of the ACM 51.1 (2008): 107-113

.

[4] Hunt, Patrick, et al. "ZooKeeper: Wait-free Coordination for Internet-scale Systems." USENIX Annual Technical Conference. Vol. 8. 2010.

[5] Bradski, Gary. "OpenCV." Dr. Dobb's Journal of Software Tools (2000).

[6] Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 1. IEEE, 2005.

[7] Viola, Paul, and Michael Jones. "Rapid object detection using a boosted cascade of simple features." Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on. Vol. 1. IEEE, 2001.

[8] Kothiya, Shraddha V., and Kinjal B. Mistree. "A review on real time object tracking in video sequences." Electrical, Electronics, Signals, Communication and Optimization (EESCO), 2015 International Conference on. IEEE, 2015.

[9] Im, Dong-Hyuck, Cheol-Hye Cho, and IlGu Jung. "Detecting a large number of objects in real-time using apache storm." Information and Communication Technology Convergence (ICTC), 2014 International Conference on. IEEE, 2014.

[10] Benenson, R., Mathias, M., Timofte, R., & Van Gool, L. (2012, June). Pedestrian detection at 100 frames per second. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on (pp. 2903-2910). IEEE.

AUTHORS

Du-Hyun Hwang is currently working towards a master's degree at Department of Electrical Engineering, Korea University. His current research interests are distributed parallel computing, computer vision and GPU processing

Yoon-Ki Kim is currently working toward the PhD degree in Electronic and Computer Engineering at the Korea University. His research interests include real-time distributed and parallel data processing, IoT, Sensor processing and computer vision.

Chang-Sung Jeong is a professor at the department of EE/CE at Korea University. He received his MS.(1985) and Ph.D.(1987) from Northwestern University, and B.S.(1981) from Seoul National University. Before joining Korea University, he was a professor at POSTECH during 1982-1992. He also worked as an associate researcher at UCSC during 1998-1999.

ROBUST HUMAN TRACKING METHOD BASED ON APPEARANCE AND GEOMETRICAL FEATURES IN NON-OVERLAPPING VIEWS

Binbin Liang[1], Songchen Han[1], Yan Zhu[2], Liping Di[1]

[1]School of Aeronautics and Astronautics, Sichuan University, Chengdu, China
liangbinbin110@126.com
[2]College of Civil Aviation, Nanjing University of Aeronautics and Astronautics,
Nanjing ,China
735506286@qq.com

ABSTRACT

This paper proposes a robust tracking method which concatenates appearance and geometrical features to re-identify human in non-overlapping views. A uniformly-partitioning method is proposed to extract local HSV(Hue, Saturation, Value) color features in upper and lower portion of clothing. Then adaptive principal view selecting algorithm is presented to locate principal view which contains maximum appearance feature dimensions captured from different visual angles. For each appearance feature dimension in principal view, all its inner frames get involved in training a support vector machine (SVM). In matching process, human candidate filtering is first operated with an integrated geometrical feature which connects height estimate with gait feature. The appearance features of the remaining human candidates are later tested by SVMs to determine the object's existence in new cameras. Experimental results show the feasibility and effectiveness of this proposal and demonstrate the real-time in appearance feature extraction and robustness to illumination and visual angle change.

KEYWORDS

Human Tracking, Non-Overlapping Views, HSV Appearance, Geometrical Features, SVM

1. INTRODUCTION

Video-surveillance is an increasingly developed technology in various domains such as on-site security surveillance, shoplifting evidencing and intelligent recognition. In many circumstances, video-surveillance system consists of multiple cameras without overlapping views, thus making spatial and temporal information of moving objects unavailable for parts of the path. So it's significantly desirable to explore a robust and real-time human tracking method in non-overlapping views.

The human tracking task is usually composed of three main aspects, namely, foreground detection, feature extraction and object matching. Foreground detection is a fundamental

David C. Wyld et al. (Eds) : NETCOM, NCS, WiMoNe, CSEIT, SPM - 2015
pp. 219–235, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51619

component in visual tracking system. It extracts the people of interests and separates it with backgrounds. The performance of foreground detection directly determines the accuracy of human tracking. Feature extraction is a crucial component in visual tracking system. Human feature in computer vision consists of many descriptors which will help machine to understand the real world. Object matching is another crucial component in visual tracking system. When human features are extracted in independent surveillance area, they need to be matched from one area to another, making decision on “what is what” and “who is who”. The human tracking task in non-overlapping views depends upon these three aspects, especially the latter two aspects.

1.1. Related Work

While numerous acceptable techniques have been proposed so far for foreground detection, there is still a need to produce more efficient algorithms in term of adaptability to multiple environments, noise resilience, and computation efficiency. Related literatures suggest that Visual Background extraction (ViBe) outperforms recent and proven state-of-the-art methods in terms of both computation speed and detection rate[1].

As for feature extraction, previous literatures have concentrated on appearance features to track human objects across disjoint cameras[2-6]. Gianfranco et al.[4] reviewed the development of a set of models that capture the overall appearance of an individual. Some of the models provide a holistic description of a person, and some others require an intermediate step where specific body parts need to be identified. Hyun-Uk et al.[5] segmented each person by a criterion with appearance and estimated the segmented regions as Gaussian mixture model (GMM) for correspondence. Both experimental results demonstrated that appearance feature has superior performance in identifying individuals, but illumination changes pose a major problem to the invariance of appearance features. In order to solve this problem, Javed et al. [6] proposed to compensate illumination variations using brightness transfer functions but his approach is limited to compensate for different illumination in different regions of the images. It still needs more researches on the solution of illumination variations in disjoint human tracking task.

Individuals can always be discriminated based on their appearances, except the notable situation where exists people in uniforms. In addition, appearance features will probably alter along with the change of visual angle which occurs frequently when people walk across cameras. Related researches combined geometrical features with appearance features and managed to overcome such problems to some extent [7-11]. Madden et al. [7] focused on a framework based on robust shape and appearance features extraction and performed well in some specific scenarios. However, his proposal only employed height as robust shape feature without considering the limitation of height in discriminating human beings due to the close resemblance of human stature. The popular remedy of this limitation is to combine gait feature with height to strengthen the discrimination of human beings. Takayuki et al. [9] proposed a method that tracks a walking human using the features of gait to calculate a robust motion signature and showed the potential validity of the proposed method in a typical surveillance system. Despite of the well-done performance of gait feature in human tracking task, it has deficiencies that a high identification rate and low computational cost are still far from being achieved[10]. Specifically, the typical gait feature is tough to adjust to the visual angle change in disjoint tracks.

In general, a discriminative appearance feature which is robust to illumination change and uniform disturbance needs to be extracted. Moreover, a discriminative geometrical feature which can adjust to the visual angle change expects more explorations.

1.2. Contributions of This Paper

This paper proposes a new robust human tracking method based on appearance and geometrical features in non-overlapping views. ViBe is initially adopted to detect and segment foregrounds of each single individual. Then a uniformly-partitioning method is presented to obtain the local low-bins HSV histograms. The uniformly-partitioning method is intended to eliminate the disturbance of illumination change. Thereafter, a new geometrical feature is created which integrates height estimate, height estimate difference parameters and the sinusoidal periodicity of walking. These three features can reflect the height and gait movement simultaneously. To our knowledge, it's original to integrate such three intrinsic features in human tracking which avoids the complexity of feature extraction and can robustly counter the visual angle change even across disjoint cameras. For a high hit rate, an adaptive principal view selecting algorithm (APVSA) and appearance feature dimension determining algorithm (AFDDA) are innovatively presented and SVM is also adopted for the same purpose. APVSA facilitates the robustness to appearance change and SVM strengthens the discrimination of human individual.

The proposed method has a configuration as shown in Fig.1. In this method, the foregoing cameras train and update SVMs. When new camera comes, all its inner frames are preprocessed and the foregrounds are detected. Then the appearance and geometrical features are extracted. The geometrical features filter the object candidates and the appearance features determine the tracked object continuously on the basis of principal view selecting mechanism. Section 2 introduces ViBe-based motion detection briefly. Section 3 and Sectionn4 describe appearance and geometrical features. Section 5 constructs the matching mechanism based on human features, APVSA, AFDDA and SVM. Section 6 discusses the experimental results in simple and complex scenario. Finally, Section 7 concludes the highlights of this proposal.

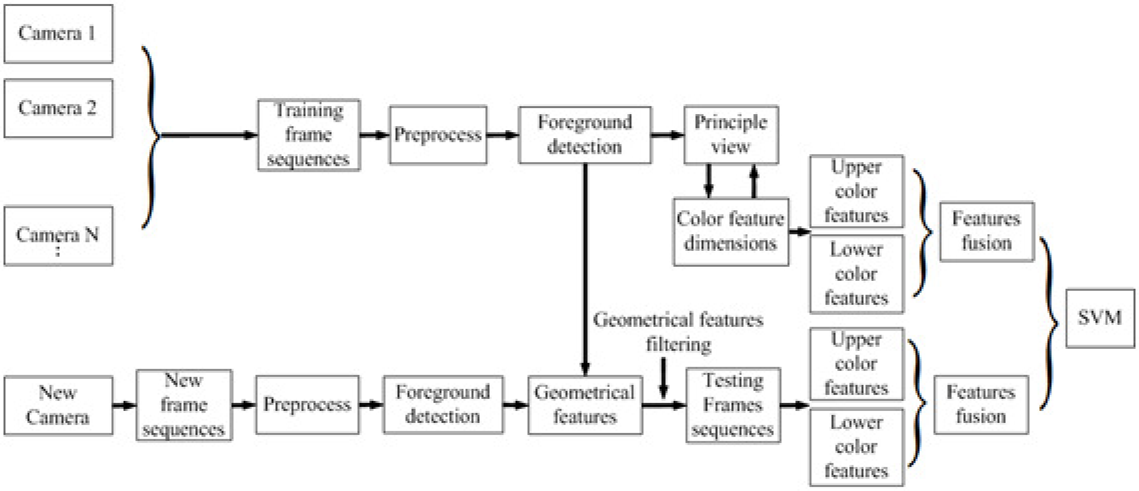


Fig.1. The process diagram for the proposed method

2. MOTION DETECTION

Motion detection is the key low-level fundamental work in intelligent video tracking task which requests accurate segmentation of foreground from background. The emerging ViBe algorithm is proven to perform satisfactorily in terms of both computation speed and detection rate. It provides an excellent robustness to scene changes and extracts foreground rapidly from early frames. Thus, this paper employs ViBe to detect and segment moving objects in video scenes in the way of [12]. After an individual is segmented, an external bounding box is used to contain it.

3. APPEARANCE FEATURES EXTRACTION

This proposal is inspired by [7] to extract local appearance of upper and lower portion in clothing which can almost discriminate individuals instead of a global appearance. These features allow a more sensitive reaction to appearance change and fasten the extraction speed as well. In this paper, the upper 30-50 percent from the top of the external bounding box is chosen to be the upper portion, while the lower 65-85 percent from the top is chosen to be the lower portion.

Appearance features often vary with the illumination change. In order to deal with this, HSV color space is employed to model appearance color in this paper. HSV is a color space in which the effect from color brightness can be suppressed by decreasing value characteristic (V). In light of this, it can reduce the appearance dissimilarity caused by illumination change. Appearance feature is partitioned into upper and lower portion, and each of them will generate HSV color features. Moreover, each HSV component will generate a histogram and thus make massive bins which will cost a lot of time consumption. To cope with this, a uniformly-partitioning method of HSV space is presented on the basis of [13].In this method, Hue component is divided uniformly into Q_H intervals, Saturation is divided uniformly into Q_S intervals and Value is divided uniformly into Q_V intervals. Consequently, H, S, V component is converted to quantization level as H_C, S_C, V_C, such that

$$H_C = \left\lfloor \frac{H * Q_H}{360} \right\rfloor, S_C = \lfloor S * Q_S \rfloor, V_C = \lfloor V * Q_V \rfloor, \tag{1}$$

Then H_C, S_C and V_C are integrated as a vector γ_{HSV} with different weight coefficients in the following formula

$$\gamma_{HSV} = H_C * Q_S * Q_V + S_C * Q_V + V_C, \tag{2}$$

This method reduces the amount of total histogram bins and improves the efficiency of appearance feature extraction. Moreover, since Q_S and Q_V are greater than 1, the weight of Value characteristic is lessened. This is conducive to the robustness to brightness change.

4. GEOMETRICAL FEATURES EXTRACTION

Height estimate is a commonly utilized geometrical feature because it keeps almost invariable across camera views. But height estimate is quite limited to identify individual because of the significant resemblance of people’s height. The popular remedy tends to combine gait feature with height estimate to enhance individual’s uniqueness. Nevertheless, the complication of gait features and poor robustness to visual angle change unveil its flaws in disjoint tracking tasks. In this paper, a new geometrical feature which is enlightened by gait movement is innovated.

4.1. Geometrical Features

This paper assumes people in scenes nearly walk in a constant speed and stay in a same behavior model. In each frame, the apparent height of individual is defined as the vertical length from the top of head to the bottom position of feet. Assuming Section 1 achieves accurate foreground segmentation and human being stays upright in most cases, the apparent height in this paper is computed as the length from the middle of the top pixel row to the middle of the bottom of bounding box instead.

The point $p(x, y)$ on the image plane can be described with a homogeneous coordinate $p = [\mathrm{x}, \mathrm{y}, 1]^{\mathrm{T}}$ and its corresponding point in the world coordinate system is $p_{real} = [\mathrm{x}, \mathrm{y}, \mathrm{z}, 1]^{\mathrm{T}}$. The projection from the real world point p_{real} to the image point p is given by

$$\delta p = K[Rt]p_{real}\ , \tag{3}$$

where δ is a nonzero scale factor and R, t denote rotation matrix and translation vector respectively. K is the intrinsic parameter matrix such that

$$K = \begin{bmatrix} f_u & s & u_0 \\ & f_v & v_0 \\ & & 1 \end{bmatrix}, \tag{4}$$

where f_u and f_v are the camera focal length expressed in pixel units along two axes, s denotes a skew coefficient. u_0, v_0 represent the principal axe coordinate respectively.

Morphological researches indicate that human walking involves rhythmic up-and-down displacement of the upper body, leading to the apparent bobbing of head. Furthermore, these vertical movements must occur in a smooth sinusoidal manner for the conservation of energy[14]. The apparent height is exhibited as a sinusoidal curve $h(t) = \mu_h + \sigma_h sin(\omega t + \phi)$. When the legs combine closest together, the maximum apparent height h_{max} occurs. When the legs separate furthest, the minimum apparent height h_{min} occurs.

Setting the maximum apparent height as the apparent stature, so the true height $h_{stature}$ can be projected as below

$$h_{stature} = \delta \left[Rt\right]^{-1} K^{-1} h_{max}\ , \tag{5}$$

Let $\Delta h_{i,i+1}$ denotes the apparent height difference between adjacent frames i and $i+1$. The mean μ_Δ and standard variance σ_Δ of $\Delta h_{i,i+1}$ in N_{biom} frames could be calculated statistically as

$$\mu_{\mathrm{V}} = \frac{\sum_{i=1}^{N_{biom}} \Delta h_{i,i+1}}{N_{biom}}\ , \tag{6}$$

$$\sigma_\Delta = \sqrt{\frac{1}{N_{biom}} (\Delta h_{i,i+1} - \mu_\Delta)^2}\ , \tag{7}$$

Another intrinsic feature, the walking periodicity $F_{periodic}$, equals to the frames number $N_{periodic}$ counting from the previous h_{max} frame to the next h_{max} frame, that is

$$F_{periodic} = N_{periodic}\ , \tag{8}$$

Thus, three intrinsic features: height estimate $h_{stature}$, height estimate difference parameters $(\mu_{\Delta}, \mathrm{v}_{\Delta})$ and walking periodicity $F_{periodic}$ are ascertained to be geometrical features. Notably, μ_{Δ} and v_{Δ} depend on the vertical movements and $F_{periodic}$ relies upon time interval, all of which avoid the complexity of computing in previous literatures and can keep unchanged in various visual angles.

4.2. Geometrical Features Based Matching

Whether two geometries match or not is determined by the similarity coefficient ρ_{biom} which is computed as below

$$\rho_{biom} = \alpha_{stature} norm_{stature} + \alpha_{\Delta} norm_{\Delta} + \alpha_F norm_{periodic} , \tag{9}$$

Here, $\alpha_{stature}, \alpha_{\Delta}, \alpha_F$ represent the weighting coefficient of stature(height estimate), height estimate difference parameters and walking periodicity respectively such that $\alpha_{stature} + \alpha_{\Delta} + \alpha_F = 1$. $norm_{stature}, norm_{\Delta}$, $norm_{periodic}$ are normalized functions in formulation (10), (11) and (12)

$$norm_{stature} = c_{stature} abs(h_{stature} - h'_{stature}) , \tag{10}$$

$$norm_{\Delta} = c_u \lambda_{\mu} abs(\mu_{\Delta} - \mu'_{\Delta}) + c_v \lambda_v abs(v_{\Delta} - v'_{\Delta}) , \tag{11}$$

$$norm_{periodic} = c_{periodic} abs(F_{periodic} - F'_{periodic}) , \tag{12}$$

where $c_{stature}, c_{\mu}, c_v, c_{periodic}$ denote normalization coefficients and they are valued with relevance to real need. λ_{μ}, λ_v denote the weighting coefficient of $\mu_{\Delta}, \mathrm{v}_{\Delta}$ such that $\lambda_{\mu} + \lambda_v = 1$. $abs(*)$ represents an absolution function.

5. OBJECT MATCHING

In this section, an adaptive principal view selecting algorithm (APVSA) will be introduced and the appearance feature dimensions in principal view are to be determined. For each appearance feature dimension, SVM is trained and is expected to fuse the upper and lower appearance feature together to generate a classifier. In matching process, below-threshold candidates will be filtered by geometrical features. Then the remaining candidates will be tested by the trained SVMs.

5.1. Principal View

Principal View (PV) refers to the view field which contains maximum appearance features captured from different angles, such as front, side and back. In non-overlapping wide area, principal view $V_{principle}$ is selected manually and empirically. But the known image sequences are usually captured from non-principal views. In this case, an interim principal view (IPV) $V_{interim}$ would be determined from the known views. This paper establishes an adaptive principal view selecting algorithm (APVSA) as presented in Algorithm 1.

Algorithm1 Adaptive principal view selecting algorithm

```
Initial view $V_{initial}$
if initial view is principal view $V_{principle}$ then
  principal view is selected, $V_{principle} = V_{initial}$
end if
else
  Set initial view as interim principal view , $V_{interim} = V_{initial}$
    for each subsequent view $V_i$ do
        if $V_i$ is principal view $V_{principle}$ then
      principal view is selected, $V_{principle} = V_i$ break;
        end if
        else if the tracked object is contained in $V_i$ then
      Count the appearance feature dimension number $N_{D_i}$ in $V_i$
        if $N_{D_i} > N_{D_{interim}}$ then
                Substitute $V_i$ for $V_{interim}$, $V_{interim} = V_i$
            end if
 end else if
      end for
end else
```

5.2. Appearance Feature Dimensions

In this paper, each over-threshold appearance feature is recognized as one appearance feature dimension (AFD). Different features in different angles have a unique AFD. Principal view is expected to have maximum AFDs.

Given there are $N_{training}$ frames captured from principal view, their HSV histograms are generated and analyzed. Two histograms will be gathered into the same group if their similarity is larger than a set threshold. Otherwise, they will be categorized into two different groups. If the frames in a group outnumber the set threshold, this very group will generate an appearance feature dimension.

The similarity of two appearance features F_1, F_2 is codetermined by the correlations of their upper appearances and lower appearances. The upper appearance generally contributes more to discriminate individuals, so it is weighted 65%. The lower appearance contributes less, so it is weighted 35%. Herewith, since the left half of HSV color histogram represents the upper appearance and the right half represents the lower appearance, the similarity of two histograms H_1, H_2 is computed from the correlations of their left halves and right halves. The correlation of two histograms H_1, H_2 is calculated statistically as

$$\rho\left(\mathrm{H}_1, \mathrm{H}_2\right) = \frac{\sum_{i=0}^{n} \mathrm{H'}_1(i)\mathrm{H'}_2(i)}{\sqrt{\sum_{i=1}^{n} \mathrm{H'}^2{}_1(i)\mathrm{H'}^2{}_2(i)}}, \tag{13}$$

where,

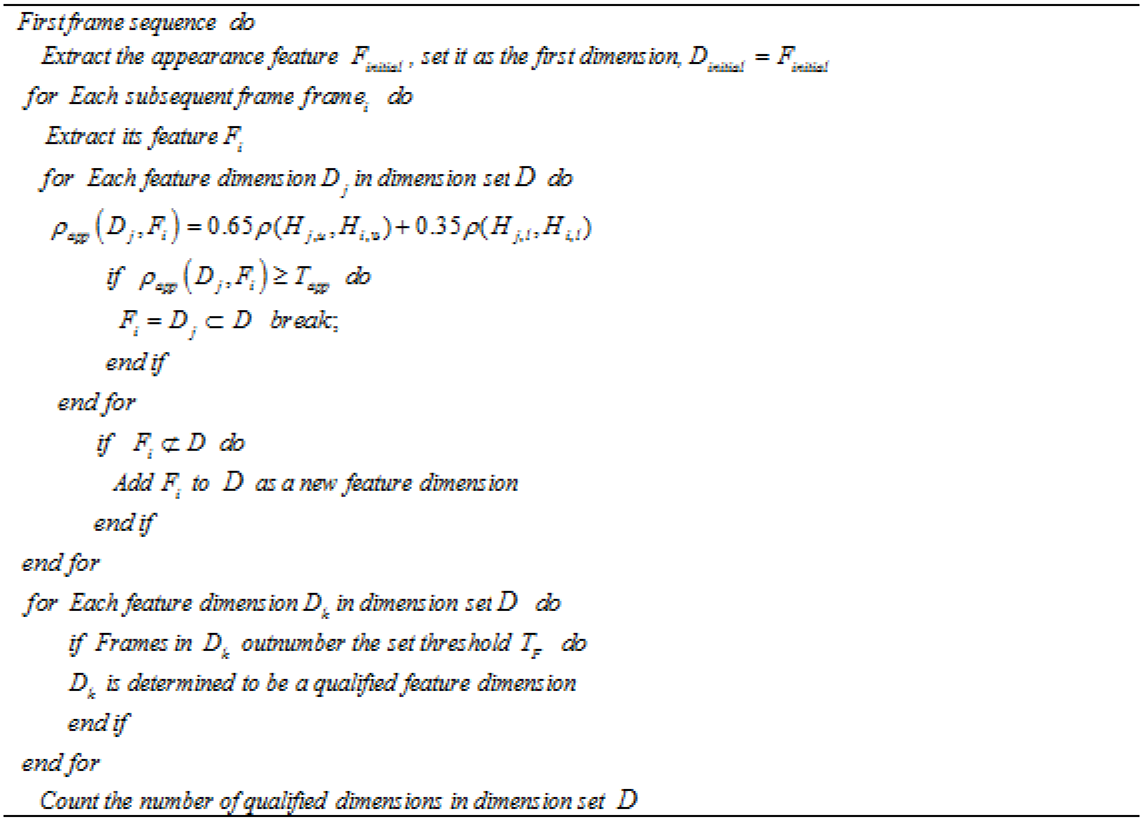
$$H'_k(i) = v_k(i) - \left(\sum_{j=1}^{n} v_k(j) \right) / \mathrm{n}\ , \tag{14}$$

where $v_k(i)$ denotes the value of dot i in histogram k. The similarity of two histograms is formulated as following

$$\rho_{app}(F_1, F_2) = 0.65\rho(H_{1,u}, H_{2,u}) + 0.35\rho(H_{1,l}, H_{2,l})\ , \tag{15}$$

Algorithm 2 shows the pseudo code of appearance feature dimension determining algorithm (AFDDA).

Algorithm 2 Appearance feature dimensions determining algorithm



```
First frame sequence  do
  Extract the appearance feature $F_{initial}$ , set it as the first dimension, $D_{initial} = F_{initial}$
for  Each subsequent frame $frame_i$  do
  Extract its feature $F_i$
  for  Each feature dimension $D_j$ in dimension set $D$  do
    $\rho_{app}(D_j, F_i) = 0.65\rho(H_{j,u}, H_{i,u}) + 0.35\rho(H_{j,l}, H_{i,l})$
      if  $\rho_{app}(D_j, F_i) \geq T_{app}$  do
        $F_i = D_j \subset D$  break;
      end if
  end for
    if  $F_i \not\subset D$  do
      Add $F_i$ to $D$ as a new feature dimension
    end if
end for
for  Each feature dimension $D_k$ in dimension set $D$  do
    if  Frames in $D_k$ outnumber the set threshold $T_F$  do
    $D_k$ is determined to be a qualified feature dimension
    end if
end for
  Count the number of qualified dimensions in dimension set $D$
```

5.3. SVM-Based Object Matching

5.3.1. Support Vector Machine

Support Vector Machine (SVM) is a powerful classifier which shows many special advantages in solving classification of nonlinear high-dimensional pattern recognition problems with small samples[15]. This paper adopts SVM to map from nonlinear space to a higher dimensional space and construct the optimal separation hyper-plane. More specifically, SVM is used to fuse upper and lower appearance features in training data.

Given a set of labeled samples $\{x_i, y_i\}, i = 1,2,...,n$, x_i is an m-dimensional vector, $y_i \in \{1,0\}$ is the topic label. The samples labeled with $y_i = 1$ belong to positive class while the samples labeled with $y_i = 0$ belong to negative class. After training the known sample data, a classifying function is generated as formula

$$f(x) = \text{sgn}\{W^T X + b\}, \tag{16}$$

The output of $f(x)$ determines the class of input vector X . After being mapped by nonlinear projection, the samples become linearly separable. According to [16], a proper kernel function $K(x_i, x_j)$ can help to classify the mapped data. Then there is an optimal separating function

$$max\, O(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j), \tag{17}$$

where, $\alpha_i \geq 0, i = 1,2,...,l$ is Lagrange multiplier. Consequently, the classifying function becomes

$$f(x) = sgn\left\{\sum_{i=1}^{n} \alpha_i^* y_i K(x_i, x) + b^*\right\}, \tag{18}$$

Here, α_i^*, b^* represent the optimal Lagrange multiplier and the corresponding classifying threshold. This paper employs Gaussian radial function as a kernel function as following

$$K(x_i, x_j) = exp\left\{-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right\}, \tag{19}$$

5.3.2. SVM Training and Testing

The SVM is trained manually by the source frames of tracked object collected before-hand at the beginning of tracking process. When the first camera comes, all its inner frames are tested by the trained SVMs to position the tracked object. The frames of the tracked object captured in the first camera will train new SVMs and replace the old SVMs to test the new coming frames in its next camera. The training stage will continue until the tracked object appears in principal view. The SVMs trained in principal view are the final SVMs and all the frames in later cameras will be tested by the final SVMs thereafter. Each of appearance feature dimensions in principal view(or in interim principal view) corresponds to a $SVM_j \ (j = 1,2,...,N_{hist})$.

For an object candidate set $set\{O_i\}$, the similarity of each object candidate O_i with the tracked object O_0 is computed. If $\rho_{biom}(O_i, O_0) \leq T_{biom}$, then O_i will be canceled from $set\{O_i\}$. The remaining m_{obj} candidates render test frames to SVM_j . For each candidate, if at least one of SVM_j outputs 1, the candidate would be recognized as the tracked object and will be contained in a red bounding box with the tracked label. If all the outputs are zero, the candidate would not be the tracked object.

6. EXPERIMENTAL RESULTS AND DISCUSSION

6.1. Experimental Setup

Numerous experiments are carried out to verify the feasibility and effectiveness of this proposal. The shooting facilities include two types of camera （Canon Power Shot A3200 IS, Canon PC1356） and they are installed in fixed height in 7 non-overlapping views. The testing platform is under VC++ and OpenCV 2.3.1 with Intel Core i7 870(2.93 GHz) CPU, 8 GB RAM and Windows 7 64bits OS.

The experiments are conducted in simple scenario and complex scenarios. In simple scenario, there exists only one or two moving objects and their moving trajectories are relatively regular. The simple scenario satisfies the single variable requirement of experiment and facilitates a detailed analysis of the advantages in this proposal. The complex scenario is composed of 7 non-overlapping views and there exists several individuals in field of views where may witness many mutual occlusions.. The experimental complex scenarios reflect closer to real-life situations and reveal the toughness of video surveillance. Before the tracking process starts, the SVM is trained in advance by the manually selected source frames of tracked person.

6.2 Experimental Results in Simple Scenario

An experiment in simple scenario where exists a short individual (object 1) and a tall individual (object 2) in same uniforms is initially conducted (see Fig.2). Each individual's geometrical parameters are computed statistically through the first 30-50 frames. Fig.2 a) lists the heights of Object 1 and Object 2 in frames within two rows respectively. Table 1 indicates that Object 2 will cease to be analyzed because its geometrical similarity with the tracked object is 0.68, less than threshold 0.8. Object 1 will continue being analyzed since its geometrical similarity with the tracked object is 0.92, more than 0.8. The sequences of Object 1 will be tested later by the SVMs which are generated through source sequences. Fig.2 b) exhibits the HSV histograms in frame sequences of Object 1 and the outputs of SVMs indicate the match result of each sequence.

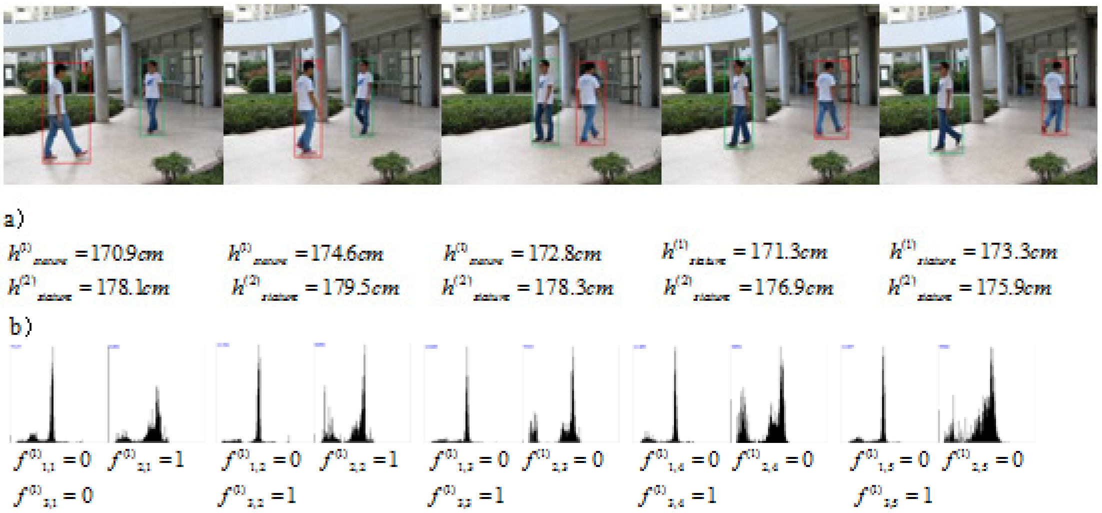

a)
$h^{(1)}_{stature}$ =170.9cm
$h^{(2)}_{stature}$ = 178.1cm
$h^{(1)}_{stature}$ =174.6cm
$h^{(2)}_{stature}$ = 179.5cm
$h^{(1)}_{stature}$ = 172.8cm
$h^{(2)}_{stature}$ = 178.3cm
$h^{(1)}_{stature}$ = 171.3cm
$h^{(2)}_{stature}$ =176.9cm
$h^{(1)}_{stature}$ = 173.3cm
$h^{(2)}_{stature}$ = 175.9cm
b)
$f^{(1)}_{1,1}=0$
$f^{(1)}_{2,1}=1$
$f^{(1)}_{3,1}=0$
$f^{(1)}_{1,2}=0$
$f^{(1)}_{2,2}=1$
$f^{(1)}_{3,2}=1$
$f^{(1)}_{1,3}=0$
$f^{(1)}_{2,3}=0$
$f^{(1)}_{3,3}=1$
$f^{(1)}_{1,4}=0$
$f^{(1)}_{2,4}=0$
$f^{(1)}_{3,4}=1$
$f^{(1)}_{1,5}=0$
$f^{(1)}_{2,5}=0$
$f^{(1)}_{3,5}=1$

Fig.2 Object tracking of individuals with different geometrical features. $h^{(1)}_{stature}$ in a) gives stature of object 1, $h^{(2)}_{stature}$ gives stature of object 2. The left half of HSV histogram in b) stands for the upper appearance feature and the right half of HSV histogram stands for the lower appearance feature. $f^{(k)}_{i,j}$ is the testing output of the k_{th} object from i_{th} SVM in frame j.

Table 1 Geometrical parameters of moving objects

	Stature/cm	μ_Δ /cm	v_Δ /cm	$F_{periodic}$ /frames	Similarity
Tracked object	174.2	0.320	0.021	13	---
Object 1	174.6	0.351	0.019	13	$0.92 > 0.8$
Object 2	179.5	0.392	0.020	15	$0.68 < 0.8$

Table 2 demonstrates the slump of erroneous matching rate with the increase of the geometrical components. Comparing 2nd row with 3rd and 4th row, the omit matching rate declines obviously while the matching rate has slight drops.

Table 2 Matching parameters of different geometrical feature groups

Group	Matching rate(%)	Erroneous matching rate(%)	Omit matching rate(%)
None	95.8	82	8.2
$h_{stature}$	93	14.2	18.1
($h_{stature}$, (μ_Δ, v_Δ))	92.4	4.7	12
($h_{stature}$, (μ_Δ, v_Δ) , $F_{periodic}$)	91.2	0.6	9.1

Note: Matching rate refers to the proportion of frames in which object matching occurs, it includes correct matching and incorrect matching; Erroneous matching rate refers to the proportion of frames in which the non-tracked object is falsely matched as the tracked object; Omit matching rate refers to the proportion of frames in which the tracked object is not matched with.

Another experiment in simple scenario is conducted where exists two geometrically similar individuals whose clothes are in huge contrast. Table 3 lists their geometrical parameters. The similarities of the two objects with the tracked object are both greater than the set threshold 0.8, and both objects would continue being analyzed. In the following process, appearance feature is analyzed to discriminate each individual and the tracking result is indicated by the outputs of SVMs, as shown in Fig.3.

Table 3 Geometrical parameters of moving objects

	Stature/cm	μ_Δ /cm	v_Δ /cm	$F_{periodic}$ /frames	Similarity
Tracked object	174.2	0.320	0.021	13	---
Object 1	174.3	0.341	0.020	13	$0.94 > 0.8$
Object 2	173.5	0.332	0.025	13	$0.87 > 0.8$

The footages in Fig.3 show the success of human tracking in this experimental situation even across two camera views where the shooting angles and illumination conditions change vastly.

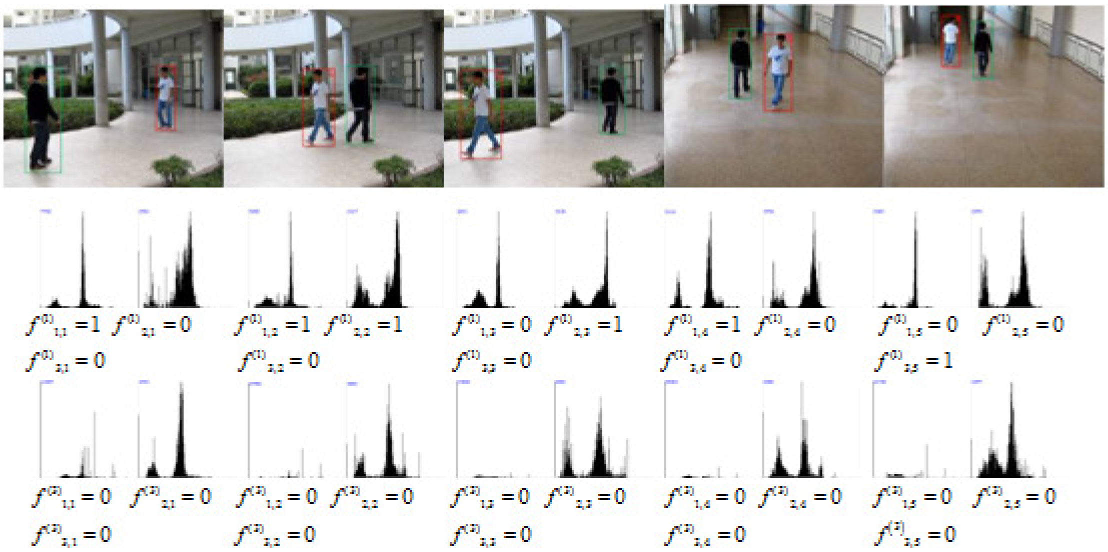

$f^{(1)}_{1,1}=1$ $f^{(1)}_{2,1}=0$
$f^{(1)}_{3,1}=0$
$f^{(1)}_{1,2}=1$ $f^{(1)}_{2,2}=1$
$f^{(1)}_{3,2}=0$
$f^{(1)}_{1,3}=0$ $f^{(1)}_{2,3}=1$
$f^{(1)}_{3,3}=0$
$f^{(1)}_{1,4}=1$ $f^{(1)}_{2,4}=0$
$f^{(1)}_{3,4}=0$
$f^{(1)}_{1,5}=0$ $f^{(1)}_{2,5}=0$
$f^{(1)}_{3,5}=1$
$f^{(2)}_{1,1}=0$ $f^{(2)}_{2,1}=0$
$f^{(2)}_{3,1}=0$
$f^{(2)}_{1,2}=0$ $f^{(2)}_{2,2}=0$
$f^{(2)}_{3,2}=0$
$f^{(2)}_{1,3}=0$ $f^{(2)}_{2,3}=0$
$f^{(2)}_{3,3}=0$
$f^{(2)}_{1,4}=0$ $f^{(2)}_{2,4}=0$
$f^{(2)}_{3,4}=0$
$f^{(2)}_{1,5}=0$ $f^{(2)}_{2,5}=0$
$f^{(2)}_{3,5}=0$

Fig.3 Human tracking of individuals with different appearance features. The HSV histograms of the individual in white are arranged in the second row, the HSV histograms of the individual in black are arranged in the third row.

Table 4 lists the comparative matching parameters of this proposed method with Hyun-Uk's proposal in [5]. The figures show the increase of accuracy rate and the decrease of omit matching rate.

Table 4 Matching parameters of different human tracking proposals in simple scenario

Tracking proposal	Accuracy rate(%)	Erroneous matching rate(%)	Omit matching rate(%)
Proposed Method	93.3	0.6	2.8
Hyun-Uk's in [5]	78.6	0.4	26.2

Note: Accuracy rate refers to the proportion of frames in which the tracked object is matched with accurately.

To verify the real-time of the appearance feature extraction in this paper, all the 898 frames captured from two views are preprocessed into 720×540 bmp images and their appearance features are extracted through the proposals of Hyun-Uk's [5] and Madden's [7] as well as this proposed method. The triple comparison diagram of computation time is illustrated in Fig.4.

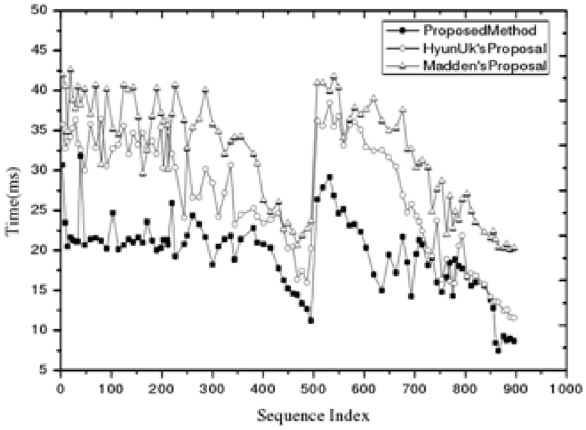
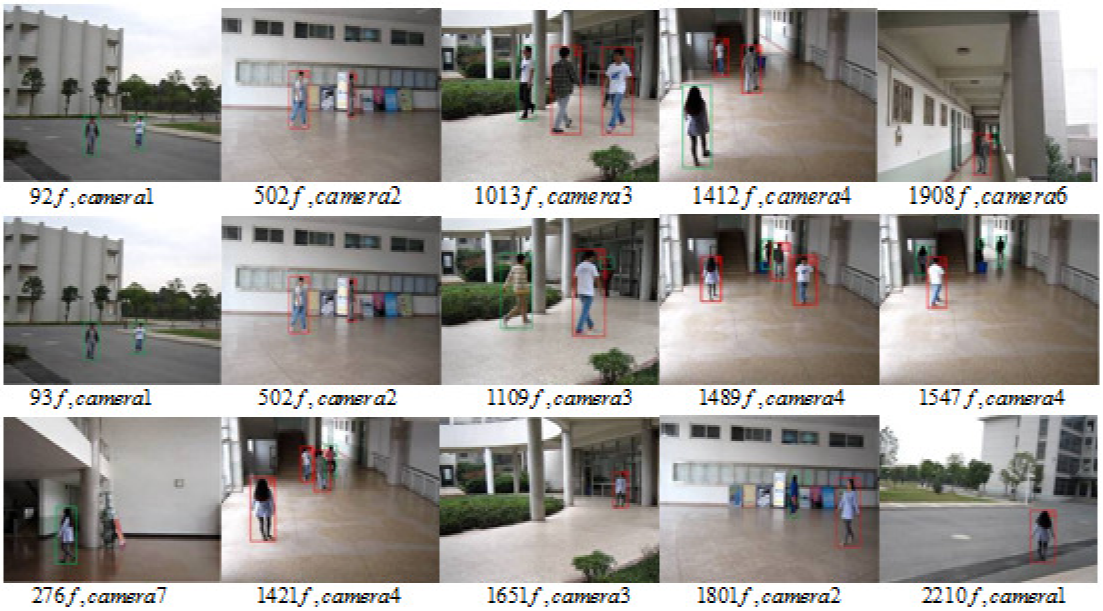


Fig.4 Comparison diagram of computation time of three different appearance feature extracting methods.

6.3 Experimental Results in Complex Scenarios

Another three experiments in complex scenarios are carried out to verify the feasibility of this proposal to track multi humans simultaneously. Seven cameras are installed in a wide area as shown in Fig.6 and none of them has pair-wise overlapping views. The view field of Camera 2 is manually selected as principal view.

When a moving object is first detected, its geometrical and appearance features will be stored in the system and it will be contained in a green bounding box. Fig.5 shows the tracks of object1, object2 and object5 in non-overlapping views.



92 f, camera1 | 502 f, camera2 | 1013 f, camera3 | 1412 f, camera4 | 1908 f, camera6

93 f, camera1 | 502 f, camera2 | 1109 f, camera3 | 1489 f, camera4 | 1547 f, camera4

276 f, camera7 | 1421 f, camera4 | 1651 f, camera3 | 1801 f, camera2 | 2210 f, camera1

Fig.5 The tracks of object 1,2,5 in five non-overlapping views exhibited in Row1, Row2 and Row3 respectively. The caption underneath indicates the image sequence index and camera belonging.

Note: The aspect ratio of the footages above may be reassigned for layout alignment.

The proposed method is further compared with other human tracking methods like Hyun-Uk's [5] Madden's [7] and Trevor's [11]. Table 5 lists the matching parameters of each method. The figures in first column not only reveal the decrease of accuracy rate when compare with the simple scenario but also show that this proposal significantly outperforms other three tracking methods in terms of accuracy rate and erroneous matching rate. However, this proposal has a higher omit matching rate than [5] and [7], but much lower than [11].

Table 5 Matching parameters of different human tracking proposals in complex scenarios

Proposal	Accuracy rate(%)	Erroneous matching rate(%)	Omit matching rate(%)
Proposed Method	84.3	5.6	10.8
Hyun-Uk's [5]	62.6	45.4	6.2
C. Madden's [7]	69	11	9.6
Trevor's [11]	77.6	8.1	23

According to temporal relation, it can retrieve the walking path of each tracked object. In this wide area, Object 1 walks along such a path: (North Gate, F1) to (North Hall, F1) to (Outdoor Corridor, F1) to (Left Stair, F2) to (Right Corridor,F2) as the blue route illustrated in Fig.6; Similarly, Object 2 walks along a path like the red route and Object 5 walks along a path like the yellow route.

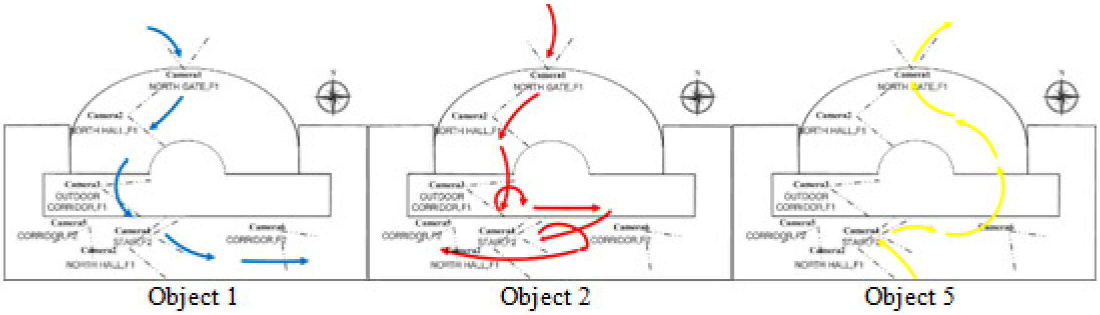


Fig.6 Walking path of each tracked object

6.4. Discussion

Further analysis demonstrates the effectiveness of combining stature (height estimate), $\mu_{\Delta}, \mathrm{v}_{\Delta}$ and $F_{periodic}$ together in decreasing erroneous matching rate and omit matching rate. The obvious decline in term of omit matching rate from 2nd row to 3rd and 4th row in Table 2 means that the over filtering problem resulting from foreground errors could be suppressed by adding μ_{Δ} and v_{Δ} into geometrical feature group. Over filtering problem is referred to the false filtering of capable object candidates in this paper.

The figures in Table 4 show the advantage of this proposed method in elevating the accuracy rate and reducing the omit matching rate. This attributes to the adoption of APVSA because it is designed to counter the appearance change in different visual angles. Whereas, since [5] extracts a global appearance feature rather than local major appearances, it outperforms this proposal slightly in term of erroneous matching rate. Fig.4 illustrates a triple comparison diagram of computation time. The contrast lines demonstrate the advantage of this proposal in computation time consumption. Comparing with [7], the proposed method benefits from the HSV uniformly - partitioning method as well as the choice of local appearance features. Furthermore, the proposed method is more applicable than [13] owing to the flexibility of the partitioning method.

The decrease of accuracy rate in complex scenario comparing with simple scenario is mainly due to the mutual occlusions of objects that occur frequently in complex scenario. But figures in Table 5 also show that this proposal significantly outperforms other tracking methods in terms of accuracy rate and erroneous matching rate. It attributes to the use of height estimate difference parameters $\mu_{\Delta}, \mathrm{v}_{\Delta}$ and walking periodicity $F_{periodic}$ in object filtering operation as well as the design of APVSA to maximize appearance feature dimensions of the training samples. The fusion of $\mu_{\Delta}, \mathrm{v}_{\Delta}$ and $F_{periodic}$ with height estimate strengthens the robustness to height disturbances. The adoption of APVSA strengthens the robustness to visual angle change. However, this proposal performs worse than [5] and [7] in term of omit matching rate, but much better than [11]. This is mainly due to the over filtering problem caused by foreground extraction errors.

The numeric results and further analyses in these experiments demonstrate the superior performance of this proposal in disjoint human tracking tasks. This proposal moves the tracking methods forward as it can counter the illumination, visual angle change simultaneously as well as cut the time consumption of appearance feature extraction.

7. CONCLUSION

This paper presents a robust method to re-identify human object across non-overlapping views based on appearance and geometrical features. All the features extracted in this paper are intended to keep unchanged in different views. This proposal mainly benefits the disjoint human tracking in three aspects:

1, The local uniformly-partitioned HSV color features are extracted in a real-time speed which manages to resist the illumination change as well as cut computation time consumption in appearance feature extraction phase.

2, A new geometrical feature which integrates height estimate, height estimate difference parameters and the sinusoidal periodicity of walking is created with avoidance of the complexity of extraction and can robustly counter the visual angle change even across disjoint cameras.

3, An innovative matching method is presented based on the designs of APVSA, AFDDA and the adoption of SVM. APVSA facilitates the robustness of people re-identification to appearance change in different visual angles and SVM strengthens the discrimination of human individual.

Experimental results in simple and complex scenario indicate the accuracy and efficiency of this proposal. Further results demonstrate the effectiveness of the combined geometrical features in reducing the erroneous matching rate and the effectiveness of APVSA in decreasing omit matching rate. In spite of the abovementioned advantages, this proposal cannot overcome the occlusion problem which often occurs in complex scenario. The future work will focus on this.

ACKNOWLEDGEMENTS

We thank the reviewers for helping us to improve this paper. This work is supported by Startup Project for Introduced Talents of Sichuan University.

REFERENCES

[1] Olivier Barnich, Marc Van Droogenbroeck, ViBe: A Universal Background Subtraction Algorithm for Video Sequences, transactions on image processing,.20(.6) (2011) 1709-1724.

[2] Yao Lu , Ling Li , Patrick Peursum ,Human Pose Tracking Based on Both Generic and Specific Appearance Models, 12th International Conference on Control, Automation, Robotics & Vision,2012,p.1071-1076

[3] T. D'Orazio, P.L.Mazzeo, P.Spagnolo, Color Brightness Transfer Function Evaluation for Non-overlapping Multi Camera Tracking, Third ACM/IEEE International Conference on Distributed Smart Cameras, 30 August-2 September,2009.

[4] Gianfranco Doretto , Thomas Sebastian, Peter Tu, Jens Rittscher, Appearance-based person re-identification in camera networks: problem overview and current approaches, Journal of Ambient Intelligence and Humanized Computing 2(2) (2011) 127-151.

[5] Hyun-Uk Chae, Kang-Hyun Jo, Appearance Feature Based Human Correspondence under Non-overlapping Views, Emerging Intelligent Computing Technology and Applications, 5754 (2009) 635-644.

[6] O. Javed and K. Shafique and M. Shah, “Appearance Modeling for Tracking in Multiple Non-Overlapping Cameras,” IEEE Conference on Computer Vision and Pattern Recognition, 2 (2005) 26-33.

[7] C. Madden , M. Piccardi. A Framework for Track Matching Across Disjoint Cameras using Robust Shape and Appearance Features, IEEE Conference on Advanced Video and Signal Based Surveillance, 5-7 September 2007.

[8] Yu-Chih Lin, Yu-Tzu Lin, Human recognition based on plantar pressure patterns during gait, Journal of Mechanics in Medicine and Biology,13(2),2013

[9] Takayuki Hori ; Jun Ohya ; Jun Kurumisawa, Identifying a walking human by a tensor decomposition based approach and tracking the human across discontinuous fields of views of multiple cameras, Computational Imaging VIII, 75330 (2010).

[10] YuChih Lin, BingShiang Yang, Yu Tzu Lin,Yi Ting Yang, Human Recognition Based on Kinematics and Kinetics of Gait, Journal of Medical and Biological Engineering, 31(4) (2010) 255-263.

[11] Trevor Montcalm, Bubaker Boufama, Object Inter-camera Tracking with non -overlapping views: A new dynamic approach, 2010 Canadian Conference Computer and Robot Vision,June,2010, 354-361.

[12] Chenhui Yang, Weixiang Kuang, Robust Foreground Detection Based on Improved ViBe in Dynamic Background. International Journal of Digital Content Technology and its Applications (JDCTA) 7(4) (2013) 754-763.

[13] Muhammad Riaz, Gwangwon Kang, Youngbae Kim, Sungbum Pan, and Jongan Park, Efficient Image Retrieval Using Adaptive Segmentation of HSV Color Space. International Conference on Computational Sciences and Its Applications ICCSA 2008, vol.55 (2008) 491 – 496.

[14] Walking. Williams and Wilkins , Human Walking. (V.Inman, H.J.Ralston ,and F.Todd. ,1981).

[15] Dawei Li, Lihong Xu, Erik D. Goodman, Yuan Xu and Yang Wu, Integrating a statistical background-foreground extraction algorithm and SVM classifier for pedestrian detection and tracking. Integrated Computer-Aided Engineering, 20 (2013) 201–216.

[16] John Wiley & Sons, New York, Statistical learning theory. (Vapnik, V.N.,1998).

AUTHORS

Binbin Liang was born in 1990. He got a Master Degree in Engineering from the College of Civil Aviation at Nanjing University of Aeronautics and Astronautics. His research interests include civil aviation emergency management and computer vision.

Dr. Songchen Han was born in 1964. He obtained his Ph.D. in Engineering at Harbin Institute of Technology. He is currently a professor in Sichuan University in China. His research interests include (1) next generation air traffic management system (2) air traffic planning and simulation.

Yan Zhu was born in 1991. She is studying for a Master Degree in Engineering from the College of Civil Aviation at Nanjing University of Aeronautics and Astronautics. Her research interests include civil aviation emergency management and general aviation emergency rescue.

Liping Di was born in 1964. She obtained his Bachelor Degree in Aircraft Control at Harbin Institute of Technology, China. She is currently a teacher in Sichuan University in China. Her research interest is aircraft control.

INTENTIONAL BLANK

A Generalized Sampling Theorem Over Galois Field Domains For Experimental Design

Yoshifumi Ukita

Department of Management and Information,
Yokohama College of Commerce, Yokohama, Japan
ukita@shodai.ac.jp

Abstract

In this paper, the sampling theorem for bandlimited functions over $GF(q)^n$ domains is generalized to one over $\prod_{i=1}^{n} GF(q_i)$ domains. The generalized theorem is applicable to the experimental design model in which each factor has a different number of levels and enables us to estimate the parameters in the model by using Fourier transforms. Moreover, the relationship between the proposed sampling theorem and orthogonal arrays is also provided.

Keywords

Digital Signal Processing, Sampling Theorem, Experimental Design, Orthogonal Arrays, Fourier Analysis

1. Introduction

In digital signal processing [3], the sampling theorem states that any real valued function f can be reconstructed from a sequence of values of f that are discretely sampled with a frequency at least twice as high as the maximum frequency of the spectrum of f. This theorem can also be applied to functions over finite domain [4] [8]. For example, Ukita et al. obtained a sampling theorem over $GF(q)^n$ domains [8], which is applicable to the experimental design model in which all factors have the same number of levels. However, this sampling theorem is not applicable to the model in which each factor has a different number of levels, even though they often do [2], [7]. Moreover, a sampling theorem for such a model has not been provided so far. In this paper, the sampling theorem for bandlimited functions over $GF(q)^n$ domains is generalized to one over $\prod_{i=1}^{n} GF(q_i)$ domains. The generalized theorem is applicable to the experimental design model in which each factor has a different number of levels and enables us to estimate the parameters in the model using Fourier transforms. In addition, recently, the volume of the data has grown up rapidly in the field of Big Data and Cloud Computing [11] [12], and the generalized theorem can also be used to estimate the parameters for Big Data efficiently. Moreover, the relationship between the proposed sampling theorem and orthogonal arrays [1] is provided.

David C. Wyld et al. (Eds) : NETCOM, NCS, WiMoNe, CSEIT, SPM - 2015
pp. 237–247, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51620

2. PRELIMINARIES

2.1 Fourier Analysis on Finite Abelian Groups

Here, a brief explanation of Fourier analysis on finite Abelian groups is provided. Characters are important in the context of finite Fourier series.

2.1.1 Characters [5]

Let G be a finite Abelian group (with the additive notation), and let S^1 be the unit circle in the complex plane. A character on G is a complex-valued function $\chi: G \to S^1$ that satisfies the condition

$$\chi(\boldsymbol{x} + \boldsymbol{x}') = \chi(\boldsymbol{x})\chi(\boldsymbol{x}') \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in G. \tag{1}$$

In other words, a character is a homomorphism from G to the circle group.

2.1.2 Fourier Transform [4]

Let $G_i, i = 1,2,\cdots,n$, be Abelian groups of respective orders $|G_i| = g_i, i = 1,2,\cdots n$, $g_1 \leq g_2 \leq \cdots \leq g_n$, and

$$G = \times_{i=1}^{n} G_i \quad and \quad g = \prod_{i=1}^{n} g_i. \tag{2}$$

Since the character group of G is isomorphic to G, we can index the characters by the elements of G, that is, $\{\chi_{\boldsymbol{w}}(\boldsymbol{x}) | \boldsymbol{w} \in G\}$ are the characters of G. Note that $\chi_{\boldsymbol{0}}(\boldsymbol{x})$ is the principal character, and it is identically equal to 1. The characters$\{\chi_{\boldsymbol{w}}(\boldsymbol{x}) | \boldsymbol{w} \in G\}$ form an orthonormal system:

$$\frac{1}{g}\sum_{x \in G} \chi_{\boldsymbol{w}}(\boldsymbol{x})\chi_{\boldsymbol{z}}^{*}(\boldsymbol{x}) = \begin{cases} 1, & \boldsymbol{w} = \boldsymbol{z}, \\ 0, & \boldsymbol{w} \neq \boldsymbol{z}, \end{cases} \tag{3}$$

where $\chi_{\boldsymbol{z}}^{*}(\boldsymbol{x})$ is the complex conjugate of $\chi_{\boldsymbol{z}}(\boldsymbol{x})$.

Any function $f: G \to \mathbb{C}$, where $\mathbb{C}$ is the field of complex numbers, can be uniquely expressed as a linear combination of the following characters:

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{w} \in G} f_{\boldsymbol{w}}\, \chi_{\boldsymbol{w}}(\boldsymbol{x}), \tag{4}$$

where the complex number

$$f_{\boldsymbol{w}} = \frac{1}{g}\sum_{x \in G} f(\boldsymbol{x})\, \chi_{\boldsymbol{w}}^{*}(\boldsymbol{x}), \tag{5}$$

is the $\boldsymbol{w}$-th *Fourier coefficient* of f.

2.2 Fourier Analysis on $\prod_{i=1}^{n} GF(q_i)$

Assume that $q_i, i = 1,2,\cdots n,$ are prime powers. Let $GF(q_i), i = 1,2,\cdots n,$ be a Galois fields of respective orders $q_i, i = 1,2,\cdots n,$ which contain finite numbers of elements. We also use $\prod_{i=1}^{n} GF(q_i)$ to denote the set of all n-tuples with entries from $GF(q_i), i = 1,2,\cdots n.$ The elements of $\prod_{i=1}^{n} GF(q_i)$ are expressed as vectors.

Example 1: Consider $GF(2) = \{0,1\}$ and $GF(3) = \{0,1,2\}$. Then, if $n = 3$ and $q_1 = 2, q_2 = 2, q_3 = 3,$

$$\prod_{i=1}^{3} GF(q_i) = \{000,001,002,010,011,012,100,101,102,110,111,112\}.$$

Specifying the group G in Sect. 2.1.2 to be the group of $\prod_{i=1}^{n} GF(q_i)$ and $g = \prod_{i=1}^{n} q_i$, the relations (3),(4) and (5) also hold over the $\prod_{i=1}^{n} GF(q_i)$ domain.

Then, the characters $\{\, \chi_{\boldsymbol{w}}(\boldsymbol{x}) | \boldsymbol{w} \in \prod_{i=1}^{n} GF(q_i)\}$ form an orthonormal system:

$$\frac{1}{\prod_{i=1}^{n} q_i} \sum_{\boldsymbol{x}\in\prod_{i=1}^{n} GF(q_i)} \chi_{\boldsymbol{w}}(\boldsymbol{x})\chi_{\boldsymbol{z}}^{*}(\boldsymbol{x}) = \begin{cases} 1, & \boldsymbol{w} = \boldsymbol{z}, \\ 0, & \boldsymbol{w} \neq \boldsymbol{z}, \end{cases} \tag{6}$$

Any function $f: \prod_{i=1}^{n} GF(q_i) \to \mathbb{C}$, can be uniquely expressed as a linear combination of the following characters:

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{w}\,\in\prod_{i=1}^{n} GF(q_i)} f_{\boldsymbol{w}}\, \chi_{\boldsymbol{w}}(\boldsymbol{x}), \tag{7}$$

where the complex number

$$f_{\boldsymbol{w}} = \frac{1}{\prod_{i=1}^{n} q_i} \sum_{\boldsymbol{x}\,\in\prod_{i=1}^{n} GF(q_i)} f(\boldsymbol{x})\, \chi_{\boldsymbol{w}}^{*}(\boldsymbol{x}), \tag{8}$$

is the $\boldsymbol{w}$-th *Fourier coefficient* of f.

3. EXPERIMENTAL DESIGN

In this section, a short introduction to experimental design [2], [7] is provided.

3.1 Experimental Design Model

Let $F_1, F_2, \cdots, F_n$ denote the n factors to be included in an experiment. The levels of factor F_i can be represented by $GF(q_i)$, and the level combinations can be represented by the n-tuples $\boldsymbol{x} = (x_1, x_2, \cdots, x_n) \in \prod_{i=1}^{n} GF(q_i)$.

Example 2:

Let Machine (F_1) and Worker (F_2) be factors that might influence the quantity of a product.
F_1 : new machine (level 0), old machine (level 1),

F_2: highly skilled worker (level 0), average skilled worker (level 1), unskilled worker (level 2).

For example, $\boldsymbol{x} = 01$ represents a combination of new machine and average skilled worker. Then, the effect of the machine, averaged over all workers, is referred to as the effect of main factor F_1. Similarly, the effect of the worker, averaged over both machines, is referred to as the effect of main factor F_2. The contrast between the effect of the machine for a highly skilled worker, the effect of the machine for an average skilled worker, and the effect of the machine for an unskilled worker is referred to as the effect of the interaction of F_1 and F_2.

Next, an explanation of the model in the context of experimental design is given. In previous works [8], [9], [10], all factors were restricted to have the same number of levels. In this paper, I give the definition of the generalized model in which each factor has a different number of levels as follows.

Definition 1: Generalized Model

y(x) is used to denote the response of the experiment with level combination x and assume the model

$$y(\boldsymbol{x}) = \sum_{\boldsymbol{w} \in I_A} f_{\boldsymbol{w}} \chi_{\boldsymbol{w}}(\boldsymbol{x}) + \epsilon_{\boldsymbol{x}}, \tag{9}$$

where

$$I_A = \{\, (\mathrm{b_1 a_1}, \mathrm{b_2 a_2}, \dots, \mathrm{b_n a_n}) | \boldsymbol{a} \in A, \mathrm{b_i} \in \ GF(q_i)\}. \tag{10}$$

The set $A \subseteq \{\ 0,1\ \}^{\mathrm{n}}$ represents the general mean, main factors, and interactive factors included in the model.

(For example, consider $A \subseteq \{\ 000,100,010,001,110\ \}$.Then, $000,100,010,001,110$ indicate the general mean, main factor of F_1, main factor of F_2, main factor of F_3, and interactive factor of F_1 and F_2, respectively.) The model includes a random error

$\epsilon_{\boldsymbol{x}}$ satisfying the expected value $E(\epsilon_{\boldsymbol{x}}) = 0$ and constant variance σ^2.

In addition, it is usually assumed that the set A satisfies the following monotonicity condition [2].

Definition 2: Monotonicity

$$\boldsymbol{a} \in A \rightarrow \boldsymbol{a}' \in A \quad \forall \boldsymbol{a}' \ (\boldsymbol{a}' \sqsubseteq \boldsymbol{a}), \tag{11}$$

where $\boldsymbol{a} = (a_1, a_2, \cdots, a_n), \boldsymbol{a}' = (a'_1, a'_2, \cdots, a'_n)$ and $\boldsymbol{a}' \sqsubseteq \boldsymbol{a}$ means that if $a_i = 0$ then $a'_i = 0, i = 1,2, \cdots, n$.

Example 3:

Consider $A = \{\,00000,10000,01000,00100,00010,00001,11000,10100,10010\,\}$.
Since the set A satisfies (11), A is monotonic.

Next, let $\boldsymbol{w} = (w_1, w_2, \cdots, w_n)$. The main effect of F_i is represented by $\{ f_{\boldsymbol{w}} | \, w_i \neq 0 \; and \; w_k = 0 \; for \; k \neq i\}$. The interaction of F_i and F_j is represented by $\{ f_{\boldsymbol{w}} | \, w_i \neq 0 \; and \; w_j \neq 0 \; and \; w_k = 0 \; for \; k \neq i, j\}$

Example 4:

Consider A given in Example 3 and $q_1 = 2, q_i = 3, i = 2, \dots, 5$. Then, I_A is given by

$$I_A = \{00000, 10000, 01000, 02000, 00100, 00200, 00010, 00020, \\ 00001, 00002, 11000, 12000\}.$$

For example, the main effect of F_1 is represented by f_{10000}, and the interaction of F_1 and F_2 is represented by f_{11000} and f_{12000}.

In experimental design, we are given a model of the experiment. In other words, we are given a set $A \subseteq \{0,1\}^n$. Then, we determine a set of level combinations $x \in X$, $X \subseteq \prod_{i=1}^{n} GF(q_i)$. The set X is called a design. Next, we perform a set of experiments according to the design X and estimate the effects from the result, $\{ (\boldsymbol{x}, y(\boldsymbol{x})) | \boldsymbol{x} \in X \}$.

An important standard for evaluating designs is the maximum of the variances of the unbiased estimators of effects calculated from the result of the experiments. It is known that, for a given number of experiments, this criterion is minimized in an orthogonal design [6].

3.2 Orthogonal Designs

In this subsection, a definition of Orthogonal Designs for the generalized model is provided.

Definition 3: Orthogonal Designs

At first, define $v(\boldsymbol{a}) = \{i \, | a_i \neq 0, 1 \leq i \leq n\}$.

For $\boldsymbol{a_1} = (a_{11}, a_{12}, \dots, a_{1n})$, $\boldsymbol{a_2} = (a_{21}, a_{22}, \dots, a_{2n}) \in \{0,1\}^n$, the addition of vectors $\boldsymbol{a_1}$ and $\boldsymbol{a_2}$ is defined by $\boldsymbol{a_1} + \boldsymbol{a_2} = (a_{11} \oplus a_{21}, a_{12} \oplus a_{22}, \dots, a_{1n} \oplus a_{2n})$, where $\oplus$ is the *exclusive or* operation.

An *orthogonal design* $C^{\perp}$ for $A \subseteq \{0,1\}^n$ is satisfies the condition that for any $\boldsymbol{a}, \boldsymbol{a'} \in A$,

$$\left| C^{\perp}_{i_1,\dots,i_m}(\varphi_1, \dots, \varphi_m) \right| = \frac{|C^{\perp}|}{q_{i_1} q_{i_2} \cdots q_{i_m}}, \\ \varphi_1 \in GF(q_{i_1}), \dots, \varphi_m \in GF(q_{i_m}) \qquad (12)$$

where $i_1, \dots, i_m$ are defined by $v(\boldsymbol{a} + \boldsymbol{a'}) = \{i_1, \dots, i_m\}$, and $C^{\perp}_{i_1,\dots,i_m}(\varphi_1, \dots, \varphi_m) = \{\boldsymbol{x} | x_{i_1} = \varphi_1, \dots, x_{i_m} = \varphi_m, \boldsymbol{x} \in C^{\perp}\}$.

Example 5:

Consider A given in Example 3 and $q_1 = 2, q_i = 3, i = 2, \dots, 5$.

Then, an orthogonal design $C^{\perp}$ for A is given as follows.

Table 1. Example of orthogonal design $C^{\perp}$.

	x_1	x_2	x_3	x_4	x_5
1	0	0	0	0	0
2	0	0	1	1	1
3	0	0	2	2	2
4	0	1	0	0	1
5	0	1	1	1	2
6	0	1	2	2	0
7	0	2	0	1	0
8	0	2	1	2	1
9	0	2	2	0	2
10	1	0	0	2	2
11	1	0	1	0	0
12	1	0	2	1	1
13	1	1	0	1	2
14	1	1	1	2	0
15	1	1	2	0	1
16	1	2	0	2	1
17	1	2	1	0	2
18	1	2	2	1	0

The *Hamming weight* $Hw(\boldsymbol{a})$ of a vector $\boldsymbol{a} = (a_1, a_2, \cdots, a_n)$ is defined as the number of nonzero components. As a special case, if $A = \{\boldsymbol{a} | Hw(\boldsymbol{a}) \leq t, \boldsymbol{a} \in \{0,1\}^n\}$**.** $C^{\perp}$ corresponds to the set of rows of a subarray in a mixed level orthogonal array of strength $2t$ [1]. Hence, $C^{\perp}$ can be easily obtained by using the results of orthogonal arrays.

However, because it is generally not easy to construct an orthogonal design for A, it is important to consider efficiency in making the algorithm to produce the design. However, because the main purpose of this paper is not to construct the orthogonal design, the algorithm is not included in this paper.

4. Sampling Theorem For Functions Over Galois Field Domains For Experimental Design

In this section, I provide a sampling theorem for bandlimited functions over Galois field domains, which is applicable to the experimental design model in which each factor has a different number of levels.

4.1 Bandlimited Functions

The range of frequencies of f is defined by a bounded set $I \subset \prod_{i=1}^{n} GF(q_i)$. Then, $f_{\boldsymbol{w}} = 0$ for all $\boldsymbol{w} \in \prod_{i=1}^{n} GF(q_i) \setminus I$. Any function whose range of frequencies is confined to a bounded set I is referred to as bandlimited to I.

4.2 A Sampling Theorem for Bandlimited Functions over $\boldsymbol{\prod_{i=1}^{n} GF(q_i)}$ Domains

Theorem 1:

Suppose a set A is monotonic and $f(\boldsymbol{x})$ is expressed as

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{w} \in I_A} f_{\boldsymbol{w}} \, \chi_{\boldsymbol{w}}(\boldsymbol{x}), \tag{13}$$

where $I_A = \{ (\mathrm{b}_1\mathrm{a}_1, \mathrm{b}_2\mathrm{a}_2, \dots, \mathrm{b}_\mathrm{n}\mathrm{a}_\mathrm{n}) | \boldsymbol{a} \in A, \mathrm{b}_\mathrm{i} \in \ GF(q_i)\}$. Then, the Fourier coefficients can be computed by

$$f_{\boldsymbol{w}} = \frac{1}{|C^{\perp}|} \sum_{\boldsymbol{x} \in C^{\perp}} f(\boldsymbol{x}) \, \chi_{\boldsymbol{w}}^{*}(\boldsymbol{x}), \tag{14}$$

where $C^{\perp}$ is an orthogonal design for A.

The proof of Theorem 1 requires the following three lemmas.

Lemma 1:

For any non principal character χ of H,

$$\sum_{\boldsymbol{h} \in H} \chi(\boldsymbol{h}) = 0, \tag{15}$$

Proof: This follows immediately [5, Lemma 2.4].

Lemma 2:

Suppose a set A is monotonic, and $C^{\perp}$ is an orthogonal design for A.

Then, for $\boldsymbol{w}, \boldsymbol{z} \in I_A$,

$$\begin{gathered} \left| C^{\perp}_{i_1,\dots,i_m}(\varphi_1, \dots, \varphi_m) \right| = \frac{|C^{\perp}|}{q_{i_1} q_{i_2} \cdots q_{i_m}}, \\ \varphi_1 \in GF\left(q_{i_1}\right), \dots, \varphi_m \in GF\left(q_{i_m}\right) \end{gathered} \tag{16}$$

where $i_1, \dots, i_m$are defined by $v(\boldsymbol{z} - w) = \{i_1, \dots, i_m\}$.

Proof: Let $S_A = \ \{ v(\boldsymbol{a} + \boldsymbol{a}') | \boldsymbol{a}, \boldsymbol{a}' \in A \}$. Because a set A is monotonic, $v(\boldsymbol{z} - w) \in S_A$ holds for $\boldsymbol{w}, \boldsymbol{z} \in I_A$. Hence, by the definition of $C^{\perp}$, equation (16) holds.

Lemma 3:

Suppose a set A is monotonic, and $C^{\perp}$ is an orthogonal design for A. Then,

$$\sum_{\boldsymbol{x}\in C^{\perp}} \chi_{\boldsymbol{z}}(\boldsymbol{x})\chi_{\boldsymbol{w}}^{*}(\boldsymbol{x}) = \begin{cases} |C^{\perp}|, & \boldsymbol{w}=\boldsymbol{z}; \\ 0, & otherwise, \end{cases} \tag{17}$$

for all $\boldsymbol{z} \in I_A$.

Proof: If $\boldsymbol{w} = \boldsymbol{z}$, then $\chi_{\boldsymbol{z}}(\boldsymbol{x})\chi_{\boldsymbol{w}}^{*}(\boldsymbol{x}) = \chi_{\boldsymbol{0}}(\boldsymbol{x})$=1 for any $\boldsymbol{x}$. Hence, $\sum_{\boldsymbol{x}\in C^{\perp}} \chi_{\boldsymbol{z}}(\boldsymbol{x})\chi_{\boldsymbol{w}}^{*}(\boldsymbol{x}) = |C^{\perp}|$.

Next, consider the case that $\boldsymbol{w} \neq \boldsymbol{z}$. Define $\boldsymbol{u} = \boldsymbol{z} - \boldsymbol{w}$ and let $v(\boldsymbol{u}) = \{i_1, \ldots, i_m\}$. Then,

$$\sum_{\boldsymbol{x}\in C^{\perp}} \chi_{\boldsymbol{z}}(\boldsymbol{x})\chi_{\boldsymbol{w}}^{*}(\boldsymbol{x}) = \sum_{\boldsymbol{x}\in C^{\perp}} \chi_{\boldsymbol{u}}(\boldsymbol{x}) \tag{18}$$

$$= \sum_{\boldsymbol{x}\in C^{\perp}} \chi_{u_{i_1},\ldots,u_{i_m}}\left(x_{i_1}, \ldots, x_{i_m}\right) \tag{19}$$

$$= \frac{|C^{\perp}|}{q_{i_1}q_{i_2}\cdots q_{i_m}} \left(\sum_{\boldsymbol{h}\in\prod_{j=1}^{m} GF\left(q_{i_j}\right)} \chi_{u_{i_1},\ldots,u_{i_m}}(\boldsymbol{h}) \right) \tag{20}$$

where $\chi_{u_{i_j}}\left(x_{i_j}\right) = 1$ for $u_{i_j} = 0$, was used for the transformation from (18) to (19), and Lemma 2 was used for the transformation from (19) to (20). Then, by (20) and Lemma 1, $\sum_{\boldsymbol{x}\in C^{\perp}} \chi_{\boldsymbol{z}}(\boldsymbol{x})\chi_{\boldsymbol{w}}^{*}(\boldsymbol{x}) = 0$ is obtained.

Proof of Theorem 1: The right hand side of Equation (14) is given by

$$\frac{1}{|C^{\perp}|}\sum_{\boldsymbol{x}\in C^{\perp}} f(\boldsymbol{x})\,\chi_{\boldsymbol{w}}^{*}(\boldsymbol{x}) = \frac{1}{|C^{\perp}|}\sum_{\boldsymbol{x}\in C^{\perp}} \left(\sum_{\boldsymbol{z}\in I_A} f_{\boldsymbol{z}}\,\chi_{\boldsymbol{z}}(\boldsymbol{x}) \right)\chi_{\boldsymbol{w}}^{*}(\boldsymbol{x})$$

$$= \frac{1}{|C^{\perp}|}\sum_{\boldsymbol{z}\in I_A} f_{\boldsymbol{z}} \left(\sum_{\boldsymbol{x}\in C^{\perp}} \chi_{\boldsymbol{z}}(\boldsymbol{x})\,\chi_{\boldsymbol{w}}^{*}(\boldsymbol{x}) \right) \tag{21}$$

$$= f_{\boldsymbol{w}} \tag{22}$$

where Lemma 3 was used for the transformation from (21) to (22). Hence, Theorem 1 is obtained.

Theorem 1 is applicable to the generalized model given in Definition 1. When we experiment according to an orthogonal design $C^{\perp}$, we can obtain unbiased estimators of the $f_{\boldsymbol{w}}$ in (9) using Theorem 1 and the assumption that $\epsilon_{\boldsymbol{x}}$ is a random error with zero mean,

$$\hat{f}_{\boldsymbol{w}} = \frac{1}{|C^{\perp}|}\sum_{\boldsymbol{x}\in C^{\perp}} f(\boldsymbol{x})\,\chi_{\boldsymbol{w}}^{*}(\boldsymbol{x}), \tag{23}$$

Hence, the parameters can be estimated by using Fourier transforms.

5. Relationship Between The Sampling Theorem and Orthogonal Arrays

Experiments are frequently conducted according to an orthogonal array. Here, the relationship between the proposed sampling theorem and orthogonal arrays will be provided.
At first, mixed level orthogonal arrays of strength t are defined as follows.

Definition 4: Orthogonal Arrays of strength t [1]

An Orthogonal Array of strength t is $N \times n$ matrix whose i-th column contains q_i different factor-levels in such a way that, for any t columns, every t-tuple of levels appears equally often in the matrix.

The N rows specify the different experiments to be performed.

Next, the definition of orthogonal arrays of strength t can be generalized by using a bounded set A instead of the strength t. The definition of the generalized mixed level orthogonal arrays is provided as follows.

Definition 5: Orthogonal Arrays for A

An orthogonal array for A is an $N \times n$ matrix whose i-th column contains q_i different factor-levels in such a way that, for any $\boldsymbol{a}, \boldsymbol{a}' \in A$, and for any m columns which are i_1-th column, ..., i_m-th column, where $i_1, \dots, i_m$ are defined by $v(\boldsymbol{a}+\boldsymbol{a}') = \{i_1, \dots, i_m\}$, every m-tuple of levels appears equally often in the matrix.

If $A = \{\boldsymbol{a} | Hw(\boldsymbol{a}) \leq t, \boldsymbol{a} \in \{0,1\}^n\}$, an orthogonal array for A is identical to a mixed level orthogonal array of strength $2t$. In other words, Definition 4 is a special case of Definition 5.

Moreover, by Definition 3 and Definition 5, it is clear that the set of rows of an orthogonal array for A is an orthogonal design $C^{\perp}$ for A. Hence the following Corollary is obtained from Theorem 1 immediately.

Corollary 1:

Suppose a set A is monotonic and $f(\boldsymbol{x})$ is expressed as

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{w} \in I_A} f_{\boldsymbol{w}} \chi_{\boldsymbol{w}}(\boldsymbol{x}), \tag{24}$$

where $I_A = \{ (\mathrm{b_1 a_1}, \mathrm{b_2 a_2}, \dots, \mathrm{b_n a_n}) | \boldsymbol{a} \in A, \mathrm{b_i} \in GF(q_i)\}$. Then, the Fourier coefficients can be computed by

$$f_{\boldsymbol{w}} = \frac{1}{|C^{\perp}|} \sum_{\boldsymbol{x} \in C^{\perp}} f(\boldsymbol{x}) \chi_{\boldsymbol{w}}^{*}(\boldsymbol{x}), \tag{25}$$

where $C^{\perp}$ is the set of rows of an orthogonal array for A defined in Definition 5 and $|C^{\perp}| = N$.

This corollary shows the relationship between the proposed sampling theorem and orthogonal arrays.

6. CONCLUSIONS

In this paper, I have generalized the sampling theorem for bandlimited functions over $GF(q)^n$ domains to one over $\prod_{i=1}^{n} GF(q_i)$ domains. The generalized theorem is applicable to the experimental design model in which each factor has a different number of levels and enables us to estimate the parameters in the model by using Fourier transforms. I have also provided the relationship between the proposed sampling theorem and orthogonal arrays.

ACKNOWLEDGEMENTS

This paper was partially supported by the Academic Committee of Yokohama College of Commerce.

REFERENCES

[1] A.S. Hedayat, N.J.A. Sloane and J. Stufken, (1999) Orthogonal Arrays: Theory and Applications, Springer.

[2] T. Okuno and T. Haga, (1969) Experimental Designs, Baifukan, Tokyo.

[3] A.V. Oppenheim and R.W. Schafer, (1975) Digital Signal Processing, Prentice-Hall.

[4] R.S. Stankovic and J. Astola, (2007) “Reading the Sampling Theorem in Multiple-Valued Logic: A journey from the (Shannon) sampling theorem to the Shannon decomposition rule,” in Proc. 37th Int. Symp. on Multiple-Valued Logic, Oslo, Norway.

[5] E.M. Stein and R. Shakarchi, (2003) Fourier Analysis: An Introduction, Princeton University Press.

[6] I. Takahashi, (1979) Combinatorial Theory and its Application, Iwanami Syoten, Tokyo.

[7] H. Toutenburg and Shalabh, (2009) Statistical Analysis of Designed Experiments (Third Edition), Springer.

[8] Y. Ukita, T. Saito, T. Matsushima and S. Hirasawa, (2010) “A Note on a Sampling Theorem for Functions over GF(q)^n Domain,” IEICE Trans. Fundamentals, Vol.E93-A, no.6, pp.1024-1031.

[9] Y. Ukita and T. Matsushima, (2011) “A Note on Relation between the Fourier Coefficients and the Effects in the Experimental Design,” in Proc. 8th Int. Conf. on Inf., Comm. and Signal Processing, pp.1-5.

[10] Y. Ukita, T. Matsushima and S. Hirasawa, (2012) “A Note on Relation Between the Fourier Coefficients and the Interaction Effects in the Experimental Design,” in Proc. 4th Int. Conf. on Intelligent and Advanced Systems, Kuala Lumpur, Malaysia, pp.604-609.

[11] Sharma, S., Tim, U. S., Wong, J., Gadia, S., and Sharma, S. (2014) “A Brief Review on Leading Big Data Models,” Data Science Journal, 13(0), pp.138-157.

[12] Sharma, S., Shandilya, R., Patnaik, S., and Mahapatra, A. (2015) Leading NoSQL models for handling Big Data: a brief review, International Journal of Business Information Systems, Inderscience.

AUTHORS

Yoshifumi Ukita has been a professor of the Department of Management Information at Yokohama College of Commerce, Kanagawa, Japan since 2011. His research interests are artificial intelligence, signal processing and experimental designs. He is a member of the Information Processing Society of Japan, the Japan Society for Artificial Intelligence and IEEE.

AUTHOR INDEX

Ahmed Zouinkhi 69
Ali Gholami 131
Amel Ltifi 69
Anuprita Satoskar 195
Binbin Liang 21, 219
Chang-Sung Jeong 203, 211
Chun Ruan 161
Dae Young, Kim 183
Dong Hyun, Jo 183
Dongho Won 151
Doo-Hyun Hwang 203
Du-Hyun Hwang 211
Erwin Laure 131
Farnaz Farid 161
Haixia Xu 115
Hon Cheung 09
Hussein Suleman 83
Hyeong-Kyu Lee 51
Jaehoon Kim 173
Ji Hyoung Ahn 59
Jinmok Kim 151
Jongho Mun 151
Joo Rak Kang 39
Joonghong Park 173
Joonghwan Lee 151
Josiah Chavula 83
Liping Di 21, 219
Melissa Densmore 83
MinGyu Lee 51
Mohamed Salim Bouhlel 69
Mohit Satoskar 195
Nabil Giweli 09
Patrick Kreidl O 99
Peili Li 115
Ridha Bouallègue 01
Seyed Shahrestani 09 , 161
Shangfeng Gao 21
Shigang Chen 99
Siwon Sung 151
Songchen Han 21, 219
Sonia Zaibi Ammar 01
Sungwon Kang 173
Tae-Jin Lee 39, 51, 59
Xinmin Tang 21
Yan Zhu 219
Yian Zhou 99
Yoon-Ki Kim 203, 211
Yoshifumi Ukita 237
You Zhou 99
Youssef Zid 01
Yuanyuan Ji 115
Yunmin Kim 59
Zhiyuan Shen 21