Lecture Notes in Electrical Engineering 362

Hamzah Asyrani Sulaiman
Mohd Azlishah Othman
Mohd Fairuz Iskandar Othman
Yahaya Abd Rahim
Naim Che Pee
Editors

Advanced Computer and Communication Engineering Technology

Proceedings of ICOCOE 2015

Springer

Lecture Notes in Electrical Engineering

Volume 362

About this Series

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

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

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

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

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

Hamzah Asyrani Sulaiman
Mohd Azlishah Othman
Mohd Fairuz Iskandar Othman
Yahaya Abd Rahim · Naim Che Pee
Editors

Advanced Computer and Communication Engineering Technology

Proceedings of ICOCOE 2015

Springer

Editors
Hamzah Asyrani Sulaiman
Universiti Teknikal Malaysia Melaka
Durian Tunggal, Melaka
Malaysia

Mohd Azlishah Othman
Universiti Teknikal Malaysia Melaka
Durian Tunggal, Melaka
Malaysia

Mohd Fairuz Iskandar Othman
Universiti Teknikal Malaysia Melaka
Durian Tunggal, Melaka
Malaysia

Yahaya Abd Rahim
Universiti Teknikal Malaysia Melaka
Durian Tunggal, Melaka
Malaysia

Naim Che Pee
Universiti Teknikal Malaysia Melaka
Durian Tunggal, Melaka
Malaysia

ISSN 1876-1100 ISSN 1876-1119 (electronic)
Lecture Notes in Electrical Engineering
ISBN 978-3-319-24582-9 ISBN 978-3-319-24584-3 (eBook)
DOI 10.1007/978-3-319-24584-3

Library of Congress Control Number: 2015953809

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Contents

RWA: Novel Heuristic Algorithm for Optical Networks with Dynamic Traffic

Arturo Rodriguez, Washington Fernández and Leonardo Ramírez

Abstract This article proposes the snake-one heuristic for the solution of the problem of routing and wavelength assignment in WDM optical networks with dynamic traffic. This heuristic is simulated in the NSFNET network with 3 other heuristics such as simulated annealing, genetic algorithms and tabu search. The results show that this new heuristic improves the blocking probability indicator, until 140 Erlangs. However, the network utilization indicator aggravated by any load. This heuristic allows the emergence of a new generation of algorithms where the objective is focused on improving network utilization indicator. The research team is already working in two variations of this heuristic, called snake-two and snake-three. It is expected that both algorithms, improve indicators of way simultaneously.

Keywords Tabu search · Simulated annealing · Genetic algorithm · Wavelength

1 Introduction

Since data transmission through IPv4, the transport of information has grown at a substantial rate, mainly due to service generation, increasing data demand. The network was not ready for these services, so it had to evolve to IPv6, and so did the

A. Rodriguez (✉)
Department of Industrial Technologies, Research Group New Technologies (GINT), Universidad Santiago de Chile, Santiago, Chile
e-mail: arturo.rodriguez@usach.cl

W. Fernández
Department of Electrical Engineering, Research Group New Technologies (GINT), Universidad Del Bio Bio, Concepción, Chile
e-mail: wfernand@ubiobio.cl

L. Ramírez
Division of Technology Development and Innovation Research Group on Telemedicine (TIGUM), Universidad Militar de Nueva Granada, Bogotá, Colombia
e-mail: leonardo.ramirez@unimilitar.edu.co

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_1

number and quality of services. On the other hand, wireless networks grew to the same or greater extent due to the generation of services in mobile telephony, which evolved from 1G to 4G, with a very large amount of services. Meanwhile, the transport networks supported the increased load and routing generated much delay compared to switching, and this made the transport networks evolve until the appearance of fiber optics, and this technology in turn to WDM (Wavelength Division Multiplexing) networks. AON (All Optical Network) requires OXC (Optical Cross Connect) commuters and it can commute different wavelengths, and each wavelength develops up to 2.5 Gbps. These networks require establishing ligthpaths, which are sets of links in the network where the data travel along different wavelengths [1, 2].

When the networks use different wavelengths along the route, they are called Wavelength Conversion Networks, which have the advantage of developing a good use of the network, but with the disadvantage of increased delay in establishing the connection and higher network implementation cost.

Constraint Continuity Wavelength (CCW) is established. On the other hand, finding a route and its wavelength with CCW to solve this problem is called RWA (Routing Wavelength Assignement), and the state of the art shows different optimization strategies, algorithms and criteria. In general, first we find the route and then we assign a wavelength, generating two algorithmic processes. There are other proposals, called integral solution strategies, that allow finding the route and the wavelength in the same algorithmic process. Moreover, the characteristics of the demand required from these networks varied from static to dynamic, where the former is a demand scenario that can be optimized, while the latter is a scenario where the problem becomes NP-Complete. Therefore, different algorithmic processes were established, with heuristic algorithms prevailing over conventional ones, basically because the conventional ones only find a route, while the heuristic are capable of finding several routes more readily and they only look for good routes, while the conventional ones look for the optimum. The most extensively reviewed algorithms in the state of the art are Genetic [3–7], Particle Swarm Optimization [7], Artificial Bee Colony [7], Ant Colony [8], Simulated Annealing [5, 9, 10, 11, 12, 13], Tabu Search [5, 8, 9, 11, 14, 15] and very innovative combinations [9, 10, 11, 15, 16, 17, 18] using different criteria and optimization methods [19–23]. The TIGUM and GINT research groups in Colombia and Chile propose new algorithmic processes called Snake One and Snake Two which will be tested in dynamic demand scenarios with an integral strategy, simulating them in the NSFNET Network (National Science Foundation NETwork) [7]. The present paper will show the results of the Snake One algorithm.

2 Simulation Scenario

Optical networks have evolved to what is called WDM (Wavelength Division Multiplexing) networks, which allow different wavelength traffic on the same fiber, which is commonly of the monomode low dispersion type so that it allows multiplexing without the appearance of nonlinear phenomena. Some of the most

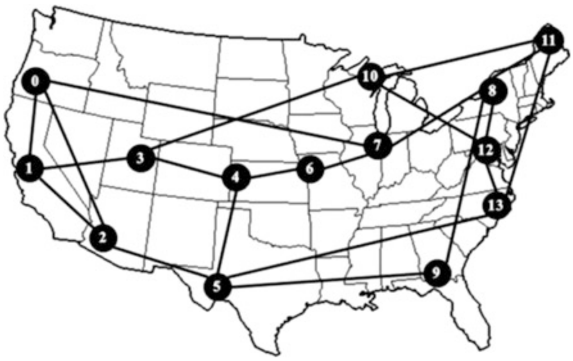


Fig. 1 National Science Foundation NETwork (NSFNET)

important WDM networks are NSFNET (National Science Foundation NETwork) [24], ARPANET (Advanced Research Project Agency NETwork) [2], EON (European Optical Network), UKNET (United Kingdom NETwork), and others. WDM networks can be with or without wavelength conversion. Networks with wavelength conversion allow more efficient use of the network's resources than networks without wavelength conversion, but the former have greater delay than the latter. Figure 1 shows the network used in this research (NSFNET), with 14 nodes and 8 wavelengths, without wavelength conversion [16, 17].

The parameters used were similar to those presented in [24], and comparisons were made of the relevant indicators, such as the probability of blocking and using the network, varying the load in the [0,180] interval with increments of 5 erlangs.

The number of connections made during the simulation in each and every one of the scenarios, either Simulated Annealing (SA), Genetic Algorithms (AG), Tabu Search, or Snake One (SNK1), was 10^8 connection requests. The nonheuristic algorithm used as reference is the one shown in [24], which made 10^4 requests and loads up to 160 erlangs.

3 New SNAKE ONE Algorithm

The heuristic that will be shown simulates the movement of a snake in the cost matrix of the network, with horizontal and vertical movements within a matrix, building the path until it reaches its destination. For that purpose it is established that requests arrive (with Poisson distribution) to the border nodes (Edge OXC) bringing with them three parameters that must be satisfied, otherwise the request must be blocked (Eq. 1):

$$d_i^S = (r_O, r_D, n_C, t_C) \tag{1}$$

where d_i^S is the vector that represents the ith request that gets to the sth border node, r_O is the identification number of the node of origin of the input demand, r_D is the identification number of the node of destination of the input demand, n_C is the number of requested connections, and t_C is the requested connection time for the (r_O, r_D) pair. If N is the number of nodes in the network, n_{CX} is the maximum number of connections at each link, and n_W is the number of wavelengths at each link (assumed to be equal in all the network's links).

Five matrices have been established: the links matrix and the cost matrix (E, C, Eq. 1), the Lambda matrix and the time matrix (λ, T, Eq. 2), and the Snake matrix (S, Eq. 3). Matrix E will retain the network topology, in this case the NSFNET, however, any network that it is desired to simulate can be configured. The definitions of the matrices are given below. Two more columns were introduced in matrix S with the purpose of saving partial calculations like total cost of the rows and the node's label.

$$E = \left\{ \begin{array}{l} e_{ij}/e_{ij} = 0\,\forall\, i = j, \wedge\, e_{ij} = G \wedge i \neq j\, NEE \\ \wedge e_{ij} = 1 \wedge i \neq j\, SEE \wedge G \gg n_C \\ \wedge i, j \in [0, N-1] \end{array} \right\} \tag{2}$$

$$C = \left\{ \begin{array}{l} c_{ijk}/c_{ijk} = n_{CX} * e_{ij} \wedge i, j \in [0, N-1] \\ \wedge k \in [0, n_W - 1] \end{array} \right\} \tag{3}$$

$$\text{SEE} = \text{There is a Link} \quad \text{NEE} = \text{There is no Link}$$

$$\lambda = \left\{ \begin{array}{l} \lambda_{ijk}/\big(\lambda_{ijk} = 0\, LA \vee NEE\big) \vee \\ \big(\lambda_{ijk} = k + 1 \wedge LD\big) \forall i, j \in [0, N-1] \\ \wedge k \in [0, n_w - 1] \end{array} \right\} \tag{4}$$

$$T = \left\{ \begin{array}{l} t_{ijk}/\big(t_{ijk} = -t_c\, LA \wedge t_{ijk} = 0\, NEE\big) \vee \\ \big(t_{ijk} = 1\, LD\big) \forall i, j \in [0, N-1] \wedge k \in [0, n_w - 1] \end{array} \right\} \tag{5}$$

$$\text{LA} = \text{Lambda Activated} \quad \text{LD} = \text{Lambda Deactivated}$$

Figure 2, shows the procedures that are carried out to get the LP in a single algorithmic process. For example, let the seventh request of the demand at node 1 be denoted by $d_7^1 = (1, 4, 1, 500)$ in a network of 8 nodes (N = 8) and 3 wavelengths, i.e., it has node 6 as destination and it requests one connection for a time of 500 ms. Then the Snake matrix is established as a function of the cost matrix, adding two columns, the first one for the labels of the nodes and the other one for the sum of the cost (Eq. 6), where the required LP will be resolved.

$$S = \left\{ \begin{array}{ll} S_{i0k} = i & \forall\, i \in [0, N-1] \wedge\, k \in [0, n_w - 1] \\ S_{ijk}/S_{ijk} = c_{ijk} * t_{ijk} & \forall\, i \in [0, N-1] \wedge j \in [1, N] \wedge\, k \in [0, n_w - 1] \\ S_{i(N+1)k} = \sum_{j=1}^{N} S_{ijk} & \forall\, i \in [0, N-1] \wedge j \in [1, N] \wedge\, k \in [0, n_w - 1] \end{array} \right\} \tag{6}$$

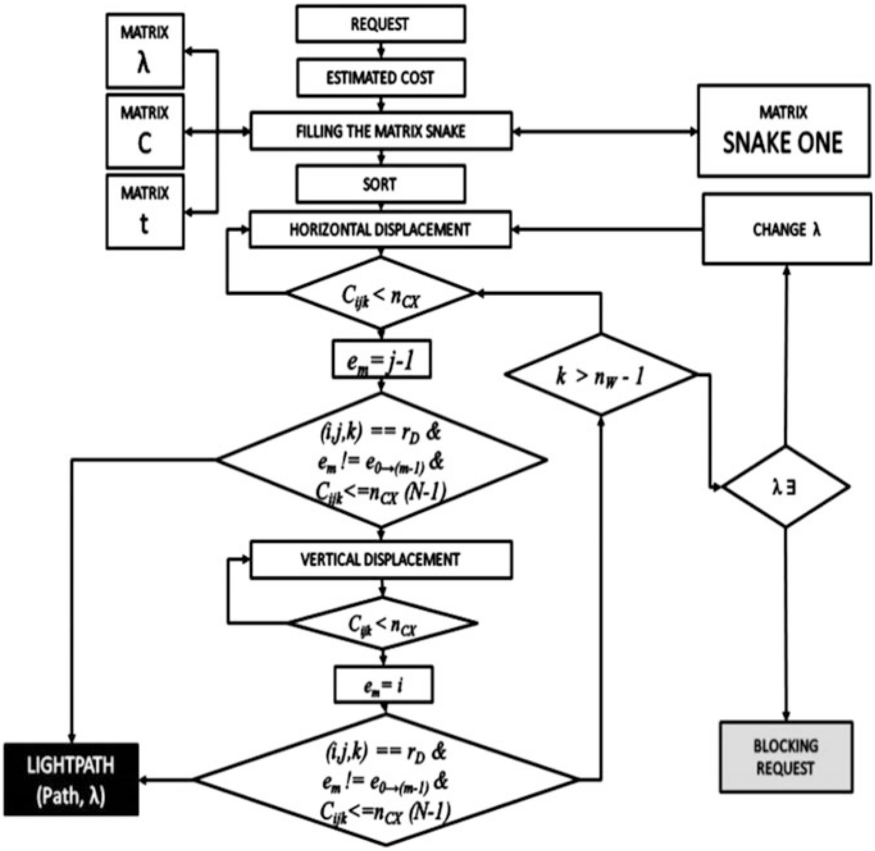


Fig. 2 Flow chart of the SNAKE-One algorithm

$$S_{ij} = \begin{vmatrix} 0 & 0 & 16 & 1000 & 1000 & 1000 & 1000 & 1000 & 16 & 5032 \\ 1 & 16 & 0 & 16 & 16 & 1000 & 1000 & 1000 & 1000 & 4048 \\ 2 & 1000 & 16 & 0 & 1000 & 1000 & 16 & 1000 & 1000 & 5032 \\ 3 & 1000 & 16 & 1000 & 0 & 16 & 1000 & 1000 & 16 & 4048 \\ 4 & 1000 & 1000 & 1000 & 16 & 0 & 16 & 16 & 1000 & 4048 \\ 5 & 1000 & 1000 & 16 & 1000 & 16 & 0 & 16 & 1000 & 4048 \\ 6 & 1000 & 1000 & 1000 & 1000 & 16 & 16 & 0 & 16 & 4048 \\ 7 & 16 & 1000 & 1000 & 16 & 1000 & 1000 & 16 & 0 & 4048 \end{vmatrix}$$

Fig. 3 Snake matrix for wavelength 0

The Snake-One algorithm will be applied to the matrix of Fig. 3 and a route will be sought; if it is not found, a route will be sought in the next wavelength (FF), and if a route is not found in the available wavelengths, the request is blocked. It is supposed that the matrix of times determines the total availability, that is $t_{ijk} = 1$.

$S_{ij}=$

Node	0	1	2	3	4	5	6	7	Cost
0	1	2	3	4	5	6	7	8	9
1	16	0	16	16	1000	1000	1000	1000	4048
3	1000	16	1000	0	16	1000	1000	16	4048
4	1000	1000	1000	16	0	16	16	1000	4048
5	1000	1000	16	1000	16	0	16	1000	4048
6	1000	1000	1000	1000	16	16	0	16	4048
7	16	~~1000~~	~~1000~~	16	1000	1000	16	0	4048
0	0	16	1000	1000	1000	1000	1000	16	5032
2	1000	16	0	1000	1000	16	1000	1000	5032

Fig. 4 Vertical displacement for wavelength 0

Let the request be $d_7^1 = (1, 4, 1, 500)$, so Matrix S is arranged in decreasing order according to the cost column. Looking at column 0 of matrix S (Nodes), we locate the row corresponding to the Node of Origin 1 (*0,0*), as seen in Fig. 4, then we move along horizontally until we locate the first cost greater than n_C and less than 1000, (Value that allow differentiate a link does not existing) which is located at node 0 (*0,1*).

Since node 0 does not correspond to the destination, we move vertically until we find a cost greater than 0 and less than 1000, which is located at node 7 (*5,1*). Since node 7 does not correspond to the destination, we move horizontally again, locating node 3 (*5,4*), and then we move vertically locating destination node 4 (2,4), ending the algorithm (Fig. 4), because it is the destination proposed in the input request. Therefore, the displacement in matrix S is:

$$(0,0) \rightarrow (0,1) \rightarrow (5,1) \rightarrow (5,4) \rightarrow (2,4)$$

which corresponds to the lightpath 1-0-7-3-4, with wavelength 0, as seen in Fig. 5.

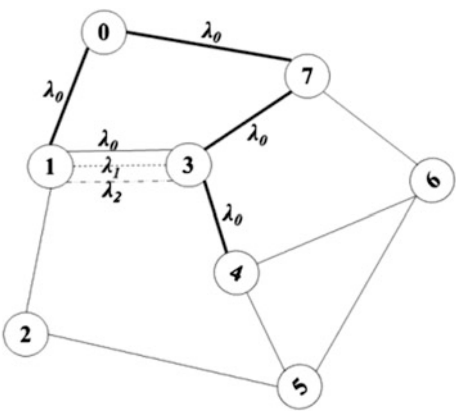


Fig. 5 Solution for the example network of 3 wavelengths λ_0, λ_1, λ_2

4 Comparison of Results

The simulation was made under dynamic traffic conditions, and the heuristic algorithms were subjected to a load variation of 0–180 erlangs. It should be stressed that the referential work found in [25] only made the simulation up to 160 erlangs, but this does not prevent the comparison. The performance of the proposed Snake One algorithm is important from the standpoint of the blocking probability.

Table presents the heuristics and the mean values of the blocking probability using as nonheuristic referential (RF) element the work shown in [24]. The mean value of the Snake One algorithm is 0.32, which is the smallest of all. However, when looking at the use of the network, the algorithm cannot decrease it, and it increases to a mean value of 68.22 %, which is the largest of all the heuristics, as seen in Figs. 6 and 7 (Table 1).

When the performance of the blocking probability in the load distribution is examined (Fig. 8), the Snake algorithm achieves better performance up to 130 erlangs with respect to the reference and up to 140 erlangs with respect to the other heuristics. However, when we look at the network use (Fig. 9), it shows no improvement at any load interval, and it appears as the heuristic that makes the largest use of the network's resources.

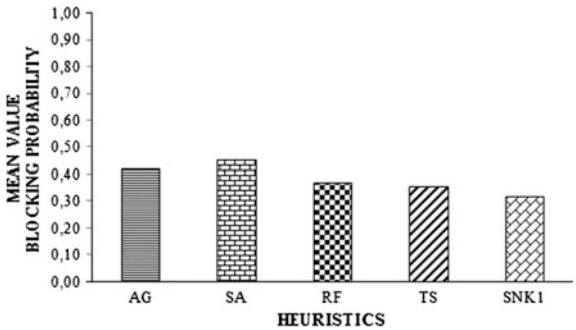


Fig. 6 Comparison of the mean value blocking probability

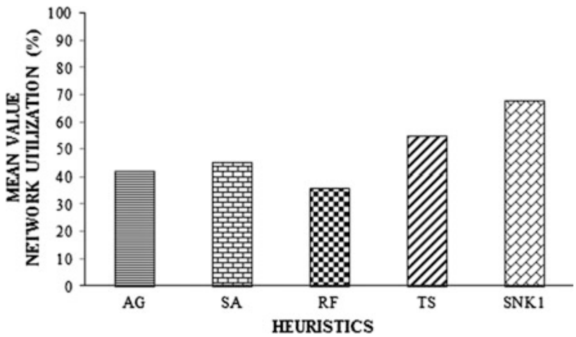


Fig. 7 Comparison of the mean value network utilization

Table 1 Comparison of mean values

	AG	SA	RF	TS	SNK1
Blocking probability	0.42	0.45	0.32	0.36	0.32
Use of the network (%)	42.17	45.30	32.17	55.17	68.22

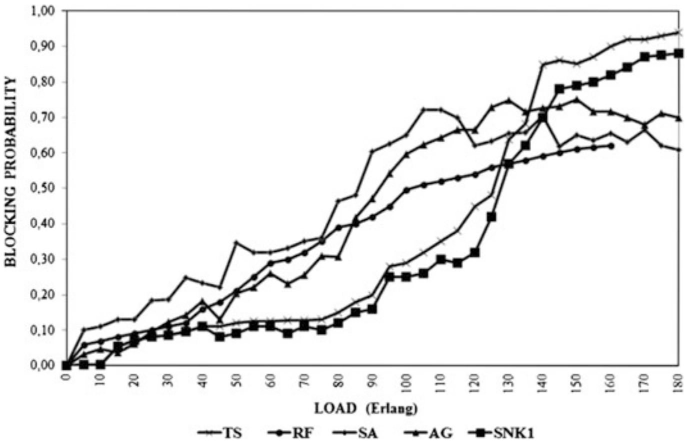


Fig. 8 The blocking probability

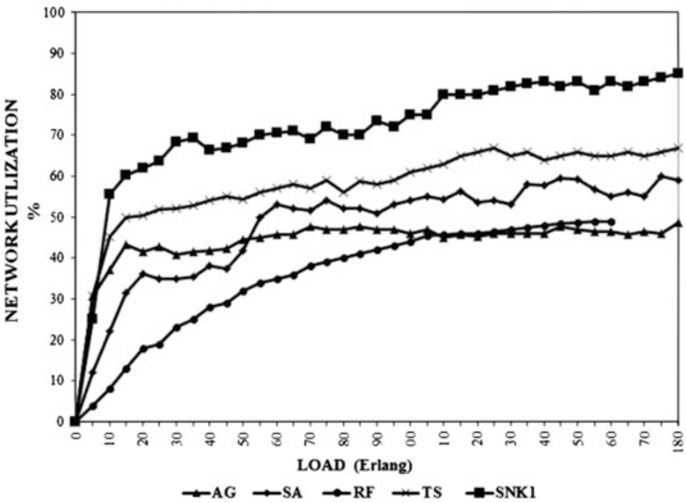


Fig. 9 The network utilization

5 Conclusions

The SNK1 algorithm achieves good performance in satisfying the demand, but its performance in network use is not acceptable, mainly because its search is made only from the standpoint of the cost matrix, without taking into account the traffic, and the next step will be to improve the performance by modifying the algorithm so that it will concentrate the traffic in specific sectors (SNK2). On the other hand, the algorithm responds well up to 130 erlangs, allowing it to be recommended in cases of dynamic high load traffic. It is important to keep on doing research on new algorithms and strategies that allow better performance of the indicators.

References

1. Zang, H., Jue, J., Mukherjee, B.: A review of routing and wavelength assignment approaches for wavelength–routed optical WDM networks. Opt. Netw. Mag. **1**, 47–60 (2000)
2. Khan, A., Thompson. D.: Solving the WDM network operation problem using dynamic synchronous parallel simulated annealing. In: Proceedings of IEEE SoutheastCon, 8–10 Apr 2005, pp. 296–301. IEEE, USA (2005)
3. Rodríguez, A., Saavedra. F., Ramírez. L.: Solución Simultánea del Enrutamiento y asignación de longitud de onda en redes WDM con algoritmos genéticos. In: IEEE COLCOM 2008, 24–27 Sept 2008. IEEE, Popayan (2008)
4. Rodríguez, A., Saavedra, F.: Optimización del Algoritmo Genético para la Solución Integral de Enrutamiento en Redes Fotónicas. Inf. Tecnol. **21**, 125–133 (2010)
5. Barpanda, R., Turuk, A., Sahoo, B., Majhi, B.: Genetic algorithm techniques to solve routing and wavelength assignment problem in wavelength division multiplexing all-optical networks. In: Communication Systems and Networks (COMSNETS), 4–8 Jan 2011, pp. 1–8. IEEE, Bangalore (2011)
6. Monoyios, D., Vlachos, K.: Multiobjective genetic algorithms for solving the impairment-aware routing and wavelength assignment problem. Opt. Commun. Networking **3**, 40–47 (2011)
7. Rashedi, A., Kavian, Y., Mahani, A., Strobel, O.: Evolutionary algorithms for solving routing and wavelength assignment problem in optical networks: a comparative study. Transparent Opt. Netw. **1**, 1–4 (2012)
8. Zhang-Liang, W., Yue-guang, L.: An ant colony algorithm with tabu search and its application. In: Intelligent Computation Technology and Automation (ICICTA), 28–29 March 2011, pp. 412–416. IEEE, Shenzhen (2011)
9. Assis, K., Ferreira, D., Giozza, W.: Hybrid algorithms for routing and assignment wavelengths in optical networks. Lat. Am. Trans. **8**, 214–220 (2010)
10. Hredzak, B., Diessel. O.: Optimization of placement of dynamic network-on-chip cores using simulated annealing. In: IECON 2011 IEEE, 7–10 November 2011, pp. 2400–2405. IEEE, Melbourne, Canada (2011)
11. Na, L., Haixing, L., Luo, Y.: A new MPLS fault restoration algorithm based on simulated annealing and tabu search. In: Wireless Communications, Networking and Mobile Computing (WiCOM), 23–25 Sept 2011, pp. 1–4. IEEE, Wuhan (2011)
12. Rodriguez, A., Saavedra, F., Ramirez, L.: Simulated annealing a proposed solution to the RWA problem in photonic networks. In: IEEE INTERCON2011 UNI, 8–13 Aug 2011. IEEE, Lima (2011)

13. Sakamoto, S., Oda, T., Kulla, E., Ikeda, M., Barolli, L., Xhafa, F.: Performance analysis of WMNs using simulated annealing algorithm for different temperature values. In: Complex, Intelligent, and Software Intensive Systems (CISIS), 3–5 July 2013, pp. 164–168. IEEE, Taichung (2013)
14. Charbonneau, N., Vokkarane, V.: Tabu search meta-heuristic for static manycast routing and wavelength assignment over wavelength-routed optical WDM networks. In: Communications (ICC), 23–27 May 2010, pp. 1–5. IEEE, Cape Town (2010)
15. Boonyopakorn, P., Meesad, P.: A hybrid GA and tabu search approach to find optimal node placement in IP networks. In: Wireless Communications, Networking and Mobile Computing (WiCOM), 23–25 Sept 2011, pp. 1–4. Wuhan, IEEE (2011)
16. Rodriguez, A., Gutierrez, A., Rivera, L., Ramirez, L.: Ruteo y Asignación de Longitud de Onda: Comparación de Algoritmos Genéticos y Simulated Annealing. Inf. Tecnol. **4**, 13–18 (2014)
17. Rodriguez, A., Ramirez, L., Rivera, L., Gutierrez, A.: Routing wavelength assignement: a solution with tabu search in dynamic traffic. Ingeniare. Rev. chil. ing. **4**, 495–503 (2014)
18. Rodriguez, A., Ramirez, L., Rivera, L., Gutierrez, A.: Comparing genetic algorithms and simulated annealing for dynamic traffic routing. Lect. Notes Electr. Eng. (Springer) **315**, 3–14 (2014)
19. Manichavelu, D., Vaidyanathan, R.U.: Particle swarm optimization (PSO)-based mode and link lifetime prediction algorithm for route recovery in MANET. EURASIP J. Wirel. Commun. Networking **2014** (2014) (article no. 107)
20. Shang, Y., Bouffanais, R.: Consensus reaching in swarms ruled by a hybrid metric-topological distance. Eur. Phys. J. B **87**(12) (2014) (article no. 294)
21. Barpanda, R., Sahoo, B., Turuk, A., Majhi, B.: Solving large problem instances of the RWA problem using genetic algorithms. In: Industrial and Information Systems (ICIIS), 29 July–1 Aug 2010, pp. 41–46. IEEE, Mangalore (2010)
22. Khelifi, L., Zidi, I., Zidi, K., Ghedira, K.: A hybrid approach based on multi-objective simulated annealing and tabu search to solve the dynamic dial a ride problem. In: Advanced Logistics and Transport (ICALT), 29–31 May 2013, pp. 227–232. IEEE, Sousse (2013)
23. Charbonneau, N., Vokkarane, V.: Static routing and wavelength assignment for multicast advance reservation in all-optical wavelength-routed WDM networks. IEEE/ACM Trans. Networking **1**, 1–14 (2012)
24. Zang, H., Jue, J., Sahasrabuddhe, L., Ramamurthy, R., Mukherjee, B.: Lightpath establishment in wavelength-routed WDM networks. IEEE Commun. Mag. **39**, 100–108 (2001)
25. Hui, Z., Jue, J.P., Sahasrabuddhe, L., Ramamurthy, R., Mukherjee, B.: Dynamic lightpath establishment in wavelength-routed WDM networks. IEEE Commun. Mag. 100–108 (2001)

A Multiple Path-Metrics Based Multi-path Routing Protocol for Providing Differentiated QoS to Various Data Traffic

Md Shohel Ahmed, Duc Van Le and Seokhoon Yoon

Abstract In this paper, we propose a novel multiple path-metrics based multi-path (MPMP) routing protocol, which can find the most suitable route that satisfies the requirements of various data traffic in mobile ad hoc networks (MANETs). In order to provide differentiated QoS to various traffic, we propose a weighted cost function based on multiple path metrics, and D-ETX (Data-driven ETX) and available bandwidth estimation methods. To our best knowledge, this work is the first attempt to consider the multiple metrics-based multi-path selections for providing differentiated QoS to various traffic in MANETs. Simulation results demonstrate that the proposed protocol can achieve a significant improvement in terms of packet delivery ratio and end-to-end delay over existing protocols.

1 Introduction

In this paper, we consider an MANET that carries multimedia traffic including streamed video and audio, which require differentiated QoS (Quality of Service). Multimedia streaming in MANET brings a lot of challenges [1, 2]. For example, multipath fading and shadowing effect may increase the variability of the link capacity and the transmission error rate.

In such a network, it is highly desired to find an appropriate route for ensuring the quality of multimedia streaming. However, most existing mobile ad hoc routing protocols focus on finding the minimum hop-count path [3]. Fewer works tried to

M.S. Ahmed · D. Van Le · S. Yoon (✉)
Department of Electrical and Computer Engineering, University of Ulsan, Ulsan, South Korea
e-mail: seokhoonyoon@ulsan.ac.kr

M.S. Ahmed
e-mail: sohelahmed86@gmail.com

D. Van Le
e-mail: anhduc.mta@gmail.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_2

find the best quality path in the presence of lossy links or congested paths. Even such studies have not paid attention for providing sufficient and differentiated QoS.

For example, single metric-based protocols treat all data traffic equally and focus on providing QoS with respect to throughput and delay [4, 5]. Therefore, single metric-based protocols may not satisfy the various requirements of different multimedia traffic.

A limited number of approaches considered multiple metrics [6, 7] as well as multiple paths [8, 9]. However, they do not offer various quality routes for differentiated services for different applications. In addition, most of those protocols considered route expiration time and energy consumption as key metrics for route selection. Those metrics may not reflect the QoS requirements of multimedia traffic.

In order to address the limitations of existing routing protocols and provide differentiated services for various applications, we propose a multiple path-metrics based multi-path (MPMP) routing protocol that uses a weighted cost function to select the most appropriate route according to the need of applications. We also propose D-ETX and available bandwidth estimation methods, which are incorporated with the weighted cost function to provide differentiated QoS.

In MPMP, the routing decision is made based on the traffic requirements since different applications require various types of QoS. For example, the data generation rate of video streams may be higher than best effort and voice traffic. Also, video transmissions need the minimum bandwidth to achieve a desired satisfactory level. On the other hand, voice streaming requires a more stringent deadline for delivery. Therefore, voice traffic demands the shortest reliable path. In addition, critical data needs higher reliability.

Our proposed protocol may choose the largest available bandwidth path for video traffic. For voice traffic, it may select the shortest reliable path. Highest reliable path is chosen for critical data by selecting a path with the lowest D-ETX. In other words, multiple paths based on different demands can provide an appropriate route to each of the data traffic i.e., the same source and destination pair may have multiple different paths for each data traffic.

Extensive simulations have been performed, and the results show that our proposed protocol can outperform existing ad hoc routing protocols in terms of packet delivery ratio and average packet delay.

The rest of this paper is organized as follows. In Sect. 2, we present MPMP routing protocol design in detail. Section 3 presents the performance study. Section 4 concludes the paper.

2 MPMP Routing Protocol

In our protocol architecture, the cost of a path is estimated by considering multiple path-metrics. A weighted cost function is used to determine the multiple routes from the source to the destination node according to data traffic requirements.

2.1 Cost Function Design Based on Multiple Path-Metrics

This sub-section describes the design of the weighted cost function. The objective of the function is to estimate the cost for each path using various path-metrics. In this paper, we consider three important metrics which are bottleneck bandwidth, the data-driven expected transmission count (D-ETX) and the end-to-end path length.

First, the cost for each metric is estimated independently. Then, the overall cost of a path is obtained by using those individual costs for different metrics. Consideration of multiple metrics for a single route gives the opportunity of multiple quality paths. Since different data traffic may demand various kinds of service, multiple quality paths may provide an appropriate route to each of the data traffic. The overall cost of a path combines the different cost functions (or metrics) as a weighted sum. The overall cost of a route is calculated as follows:

$$C_i = \alpha f(x_i) + \beta g(y_i) + \gamma h(z_i) \tag{1}$$

where C_i represents the overall cost of path *i,* and $\alpha + \beta + \gamma = 1, 0 \leq \alpha, \beta, \gamma \leq 1$. Also, N is the set of available paths between the same source and destination pair.

Functions $f(x)$, $g(y)$ and $h(z)$ are the individual cost function for different metrics. Also, α, β and γ are weights of those functions, respectively. The cost functions are defined as follows:

$$f(x) = \frac{x_{\max} - x}{x_{\max}}, \quad 0 \leq x \leq x_{\max} \tag{2}$$

$$g(y) = \frac{y - y_{\min}}{y_{\max} - y_{\min}}, \quad y_{\min} \leq y \leq y_{\max} \tag{3}$$

$$h(z) = \frac{z - z_{\min}}{z_{\max} - z_{\min}}, \quad z_{\min} \leq z \leq z_{\max} \tag{4}$$

where x, y and z represent the bottleneck bandwidth, hop count and D-ETX of end-to-end path, respectively.

The values of α, β and γ are selected based on the data traffic requirements. Therefore, the same source and destination pair may have multiple different paths for each data traffic i.e., multiple paths can be used to carry data from the source node to the destination node.

The function $f(x)$ in (2) computes the cost of the path in terms of available bandwidth, whereas x_{max} represents the maximum channel bandwidth. Note in (2) that, as the available bandwidth, x, increases, the bandwidth cost decreases i.e., higher the available bandwidth, lower the cost of the path.

Similarly, $g(y)$ in (3) estimates the cost using the end-to-end path length, y. The maximum length of a path, y_{max}, is estimated based on our analytical observation, which is presented in Sect. 3. Note that (3) indicates that a shorter path length leads to a lower cost i.e., a better path. Similarly, function $h(z)$ calculates the cost of the path using D-ETX. In this case, a lower D-ETX represents a lower cost of the path. Also note that all cost functions have the same scalar domain, which is [0, 1].

We introduce a weight for each function in order to put a priority on particular metrics. The weight values are varied according to the traffic requirements. For example, the video data requires a high bandwidth. Therefore, it is desired that a path with a high bandwidth is chosen for video traffic. In other words, the weight value of α needs to be higher than β and γ when calculating the path cost using (1).

Similarly, the weight of β may need to be higher than α and γ for voice data. This is because the voice traffic needs a short and reliable path. Likewise, for critical data, γ may need to be higher than α and β. A high weight value of γ reduces the path cost in terms of D-ETX and increases the possibility of selecting a more reliable path.

2.2 D-ETX (Data-Driven ETX) Design

First, in order to observe the distribution of link loss ratio of an ad hoc network, we perform experiments to estimate the link loss ratio in terms of pair-wise packet delivery ratio. We use 802.11b in physical layer and CBR as data traffic in application layer.

Figure 1 shows the pair-wise packet delivery ratio of each link which has participated in data transmission. The distribution of link loss ratio shown in Fig. 1 gives us two important observations. First, a large number of links have a high loss ratio. These links might not be able to deliver data packet. Another observation is *Hello* packet and *ACK* packet can obtain a higher delivery ratio than data packets.

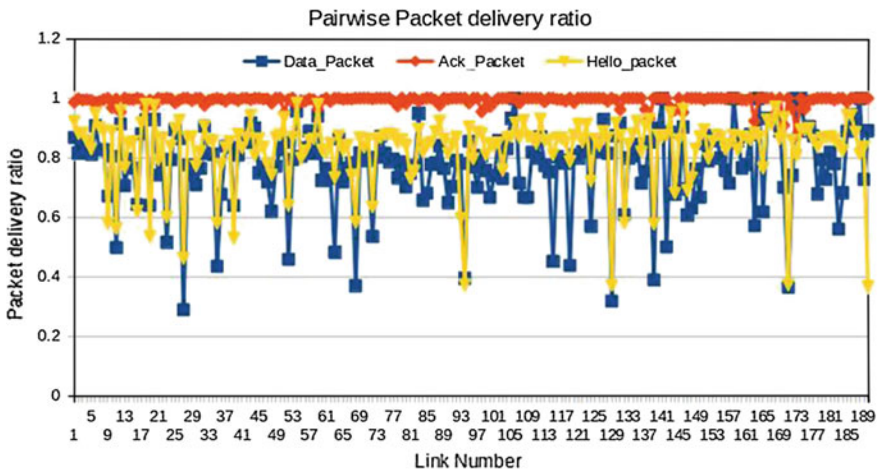


Fig. 1 Packet delivery ratio of Data, Hello and ACK packet

This indicates that *Hello* and *ACK* packet may not accurately reflect the real status of the link.

ETX represents the number of transmissions including retransmissions that is expected to deliver a packet. The original ETX [10] uses the forward and reverse delivery ratio which are calculated using the dedicated broadcast probe packets for each link. ETX of a routing path is the summation of ETX of each link of the route.

However, according to our observations discussed, the delivery ratio of probe packets may not correctly reflect the expected number of transmission or retransmission of a link since the broadcast probe packets are small in size comparing to multimedia data packet. Packet dropping probability of probe packet is very low. As a result, delivery ratio of probe packet can be high, even though the link is lossy.

Therefore, in our protocol, we propose the D-ETX calculation process for more accurate estimation. We consider the delivery ratio of actual data packet to estimate the D-ETX. We use the probe packets only when there is no data traffic over the link. Every node monitors the data traffic in the network and calculates the packet delivery ratio (d_{data}) of each link periodically.

$$d_{data}(t) = \frac{N_{ack}(t-T,t)}{N_{data}(t-T,t)} \tag{5}$$

where $N_{data}(t - T, t)$ and $N_{ack}(t - T, t)$ are the number of transmitted data packets and received acknowledgement packets at the node for last time window T, respectively.

Only when there is no data traffic of a link for a particular time, D-ETX is estimated based on the probe packet. In our protocol, we use *Hello* message as probe packet. Each host broadcasts a series of *Hello* packets during a fixed time window T and counted the number of packets sent and received. Suppose that every node broadcasts *Hello* packets of fixed size at every τ seconds. Then, the delivery ratio (d_{hello}) at time t becomes

$$d_{hello}(t) = \frac{N_{hello}(t-T,t)}{\frac{T}{\tau}} \tag{6}$$

where $N_{hello}(t - T, t)$ is the number of *Hello* packets received for last T seconds and T/τ is the number of probe packet it should receive.

Then, the data-driven expected number of transmission is calculated as

$$D - ETX = \frac{1}{d} \tag{7}$$

where d is the forward delivery ratio of the link based on type of packets i.e., d is set to $d_{data}(t)$ and $d_{hello}(t)$ of the forward link in cases of using data packets and *Hello* packets, respectively.

2.3 Available Bandwidth Estimation

In order to provide sufficient QoS for different traffic, a routing protocol needs to estimate the available bandwidth accurately. However, estimating the available bandwidth precisely using IEEE 802.11 MAC is a challenging task, since individual node has no knowledge about the status of neighboring nodes on carrying data traffic. Because the bandwidth is shared among the neighboring nodes under IEEE 802.11 DCF mode, estimating the consumed bandwidth of all nodes within the carrier sense range is necessary.

Therefore, we estimate the consumed bandwidth within the carrier sense range. Every node broadcasts its consumed bandwidth periodically using Hello messages i.e., the bandwidth usage of a host is piggybacked onto the Hello messages. We modify the AODV Hello message structure to include the consumed bandwidth information. Each node observes the total amount of data it fed into the network within a specific time window (e.g., 5 s).

The node can obtain the neighboring consumed bandwidth information through $Hello$ messages. However, node also needs to estimate the consumed bandwidth in its carrier sense range. Note that the interference range can be twice of the transmission range. Therefore, in this paper, we enhance the two-hop bandwidth estimation process [4] by taking into consideration the physical layer preamble and backoff time.

First, let BW_{av}, BW_{ch} and BW_{cn} denote the available bandwidth, channel bandwidth, and total consumed bandwidth, respectively. Then,

$$BW_{av} = \frac{BW_{ch} - BW_{cn}}{w} \tag{8}$$

where w is the weight factor and $w = (T_{overhead} + p)/p$ where p is packet size and $T_{overhead}$ is the overhead taken by routing protocol and MAC operations.

$$T_{overhead} = T_{PHY} + T_{MAC} + T_{UDP} + T_{IP} + DIFS + SIFS + T_{ACK} + T_{BCK} \tag{9}$$

where T_{PHY}, T_{MAC}, T_{UDP}, T_{IP} and T_{BCK} are physical layer preamble, MAC header, UDP header, IP header and average backoff time, respectively.

In order to more closely estimate the consumed bandwidth, the physical layer preamble and backoff time are also added for the weight factor. In this work, T_{BCK} is approximated by

$$T_{BCK} = \left(\frac{CW_{\min}}{2}\right) T_{slot} \tag{10}$$

where CW_{min} represents the minimum contention window and T_{slot} is time slot. The available bandwidths of all hosts are estimated.

2.4 Best Route Selection Procedure

In MPMP, a modified route discovery mechanism of AODV is used to discover multiple paths between source and destination node. More specifically, when the source node has a packet to transmit and has no route to the destination, it initiates the route discovery by flooding route request (RREQ) messages. When an intermediate node receives an RREQ, it first updates the available bandwidth, D-ETX and the hop count in the RREQ header. When the destination node receives an RREQ, it first checks whether it is QoS traffic (e.g., video, voice and critical data). If the traffic requires QoS, the destination node stores the route in a temporary list and waits for the waiting period, which is the half of the route request time-out period. If the traffic does not require QoS, it just sends an RREP immediately to the source as in AODV.

After the waiting period, the destination node estimates the cost of each path using (1) and selects the lowest cost path according to the type of QoS traffic. Finally, destination node sends RREP to the source.

3 Performance Study

The proposed protocol (MPMP) is evaluated through simulation using NS2. We evaluate the performance of the protocol for video, voice and data traffic that require QoS. We assume a multi-hop wireless network of 50 nodes randomly distributed in 1000 × 1000 m area. IEEE 802.11 MAC protocol with RTS/CTS disabled is used. Note that RTS and CTS are used to alleviate the hidden terminal problem. However, for multimedia data transmissions it is recommended to disable RTS/CTS [11].

In our simulation, we consider 802.11b for physical layer which has the maximum channel data rate x_{max} of 11 Mbps. In this work, the H.264/MPEG-4 is used for video encoding, which supports a high video quality at a relatively low data rate due to a high compression ratio. We consider 25 frames per second where frame size of 480 × 360 with average background motion. By encoding the raw video data, the node load of video traffic becomes about 0.6048 Mbps.

For the best quality voice traffic, we use the G.711 encoding, which requires the data rate of 64 Kbps. From our observation, we notice that in a network of size 1000 × 1000 m with transmission range 250 m, the maximum number of hops (y_{max}) is about 5 on average. The maximum D-ETX of a link is about 2. Therefore, considering the worst case, the maximum D-ETX (z_{max}) of a path is set to $5 \times 2 = 10$ in this work.

In order to emulate the transmission of video, voice, and data, constant bit rate (CBR) traffic with data packets of 1500, 160, and 460 bytes is used on top of UDP, respectively.

Performance of the protocol is measured for different scenarios and topologies. In this paper, we present two sets (i.e., static and mobile network) of simulation results.

We use various number of QoS traffic connections over the static network and use various movement speeds over the mobile network. We compare MPMP with AODV, DSR, AODV with ETX, and Bandwidth Estimation-based QoS Routing Protocol (BEQRP) [5]. In the simulation scenario, we use video, voice, and critical data connections with node loads of 0.6048, 0.064 and 0.032 Mbps, respectively.

For simulations, different weight values for different traffic are considered. More specifically, for video, values of (α, β, γ) are set to (0.5, 0.3, 0.2), respectively. Also, value sets of (0.2, 0.5, 0.3) and (0.2, 0.3, 0.5) are used for voice and data, respectively.

3.1 Network with Stationary Nodes

We notice for video traffic that, initially the packet delivery ratio (PDR) of MPMP and AODV is close as shown in Fig. 2a. The reason is that initially network is not very loaded, and, as a result, all protocols show sufficient performance. As the number of connections increases, network load increases and PDR decreases for all protocols. MPMP shows a higher PDR in all the cases than AODV, DSR, ETX-AODV and BEQRP. This is because MPMP selects the highest bandwidth path with reliability which allows for avoiding the poor quality links and congested paths i.e., the highest bandwidth path can provide the required bandwidth and reduce the packet dropping probability.

MPMP also achieves PDR improvement for voice and data traffic comparing to other protocols as shown Fig. 2b, c. Note that, real-time characteristics of voice traffic make it delay sensitive, which requires to meet the deadline. By considering the requirements of voice traffic, MPMP selects the shortest path with reliability and bandwidth. Since MPMP can avoid the lossy and congested path by using D-ETX metric for data traffic, a low number of route breaks occurred compared to other protocols.

In terms of delay, as shown in Fig. 3a, MPMP shows a better performance than AODV, ETX-AODV and BEQRP which is essential for real-time video traffic. Note that AODV, DSR, and ETX-AODV determine a route without considering the

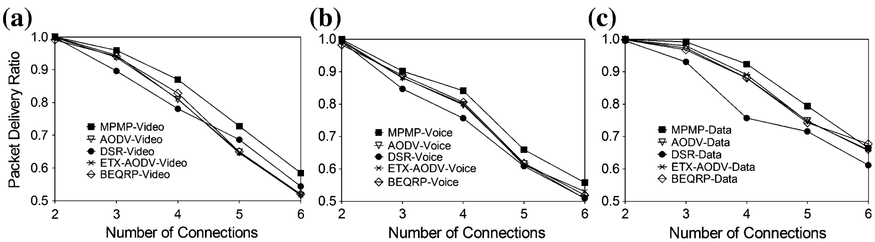


Fig. 2 Effects of variable traffic connection on packet delivery ratio. **a** Video traffic. **b** Voice traffic. **c** Data traffic

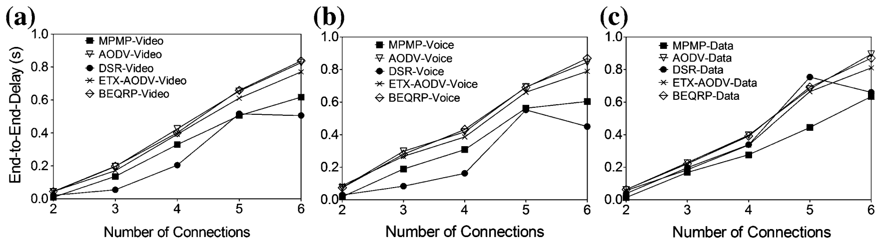


Fig. 3 Effects of variable traffic connection on end-to-end delay. **a**. Video traffic. **b**. Voice traffic. **c**. Data traffic

bottleneck bandwidth and congestion in the network. When the route is highly congested, packet needs to wait for a long time in the queue in AODV, and ETX-AODV. Note that DSR achieves the lowest delay for video and voice traffic, as shown in Fig. 3a, b at a cost of a very low throughput shown in Fig. 2a, b, which may make it impractical.

As shown in Fig. 3c, MPMP shows the lowest delay for data traffic, even though the protocol puts a priority on reliability. This is because a reliable path can also have a lower delay due to less packet retransmissions and re-discovery of routes.

3.2 Network with Mobile Nodes

We use the random waypoint model (RWP) to simulate the movements of nodes. The node speed is randomly chosen in (0, v_{max}), where v_{max} is maximum node speed. We collect performance results of protocols over variation of the node speed while keeping the same number of data connections. Specifically, we use 2, 5, and 5 connections for video, voice and data traffic, respectively.

As shown in Fig. 4, MPMP gains a high PDR compared to other protocols. Figure 4b shows the PDR of voice traffic. In this case, MPMP obtains the highest performance in all cases. Note that MPMP selects a path that has the shortest path

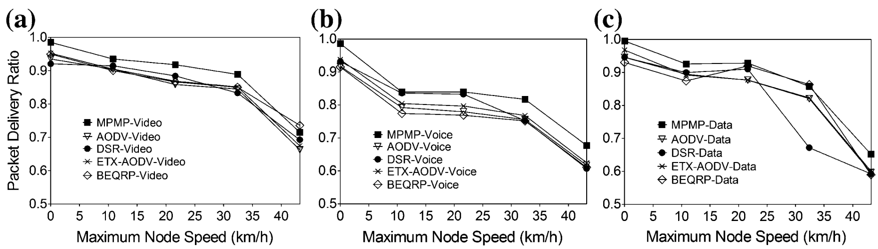


Fig. 4 Effects of variable node speed on packet delivery ratio. **a** Video traffic. **b** Voice traffic. **c.** Data traffic

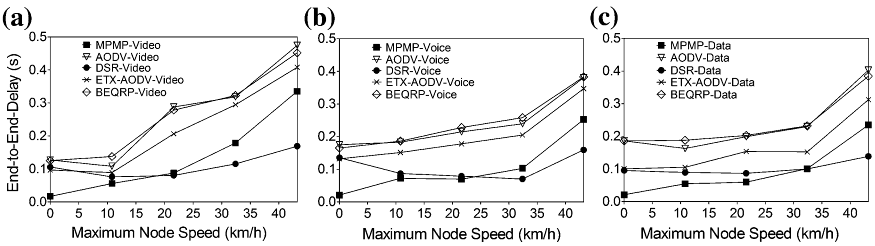


Fig. 5 Effects of variable node speed on end-to-end delay. **a** Video traffic. **b** Voice traffic. **c.** Data traffic

length and reliability. AODV, ETX-AODV and BEQRP show similar trends in all cases. The PDR values of data are shown in Fig. 4c. The PDR decreases while the node speed increases. However, MPMP protocol shows a better performance. This is because a lower number of route break occurred in MPMP.

In terms of delay, as shown in Fig. 5a, the frequent route break due to node mobility causes extra delay for each protocol. MPMP gains a low delay comparing to AODV, ETX-AODV and BEQRP. As shown in Fig. 5b, MPMP can achieve a low delay comparing to others protocols. AODV and BEQRP show the highest delay. For data traffic, as shown in Fig. 5c, the delay of AODV, ETX-AODV and BEQRP is high comparing to MPMP. This is because route break in AODV, BEQRP and also in ETX-AODV is more frequent.

4 Concluding Remarks

In this paper, we have proposed a novel multiple path-metrics based multi-path (MPMP) routing protocol for providing the best suitable route according to QoS requirements of applications. A new weighted cost function has been designed to determine the cost of a routing path based on enhanced bandwidth estimation, D-ETX and end-to-end length of the path. Simulation results show that the proposed protocol can outperform existing routing protocols in terms of packet delivery ratio and end-to-end delay.

Acknowledgement This work was supported by the 2014 Research Fund of University of Ulsan.

References

1. Chlamtac, I., Conti, M., Liu, J.J.N.: Mobile ad hoc networking: imperatives and challenges. Ad Hoc Netw. **1**(1), 13–64 (2003)
2. Antonio, P., Grimaccia, F., Mussetta, M.: Architecture and methods for innovative heterogeneous wireless sensor network applications. Remote Sens. **4**(5), 1146–1161 (2012)

3. Hanzo, L.I., Tafazolli, R.: A survey of QoS routing solutions for mobile ad hoc networks. IEEE Commun. Surv. Tutorials **9**(2) 50–70 (2007)
4. Chen, L., Heinzelman, W.: QoS-aware routing based on bandwidth estimation for mobile ad hoc networks. IEEE J. Sel. Areas Commun. **23**(3), 561–572 (2005)
5. Liu, C., Liu, K., Li, L.: Research of QoS-aware routing protocol with load balancing for mobile ad hoc networks. In: Proceedings of 4th International Conference on Wireless Communications, Networking and Mobile Computing, pp. 1–5 (2008)
6. Ahn, H., Seo, Y., Kim J., Ko, Y-B., Lee C., Kim, K.-J.: Multi-metric geo-routing protocol for tactical ad hoc networks. In: Proceedings of 2013 International Conference on Information Networking (ICOIN), pp. 89–94 (2013)
7. Cao, L., Sharif, K., Wang, Y., Dahlberg, T.: Adaptive multiple metrics routing protocols for heterogeneous multi-hop wireless networks. In: Proceedings of 5th IEEE Consumer Communications and Networking Conference (CCNC), pp. 13–17 (2008)
8. Balachandra, M., Prema, K.V., Makkithaya, K.: Multiconstrained and multipath QoS aware routing protocol for MANETs. Wirel. Netw. **20**(8), 2395–2408 (2014)
9. Hwang, Y., Varshney, P.: An adaptive QoS routing protocol with dispersity for ad- hoc networks. In: Proceedings of the 36th Annual Hawaii International Conference on System Sciences, pp. 302–311 (2003)
10. Couto, D.S.J.D., Aguayo, D., Bicket, J., Morris, R.: A high-throughput path metric for multi-hop wireless routing. In: MobiCom ‘03 Proceedings of the 9th Annual International Conference on Mobile Computing and Networking, pp. 134–146 (2003)
11. Sun, Y., Sheriff, I., Belding-Royer, E.M., Almeroth, K.C.: An experimental study of multimedia traffic performance in mesh networks. In: Proceedings of International Workshop on Wireless Traffic Measurements and Modeling, pp. 25–30 (2005)

Real-Valued Negative Selection Algorithms: Ensuring Data Integrity Through Anomaly Detection

Rihab Salah Khairy, Rozaida Ghazali and Ayodele Lasisi

Abstract The Real-Valued Negative Selection algorithms which are the focal point of this work generate their detector set based on the points of self data. Self data is regarded as the normal behavioural pattern of the monitored system. An anomaly in data alters the confidentiality and integrity of its content thereby causing a defect for making useful and accurate decisions. Therefore, to correctly detect such an anomaly, this study applies the real-valued negative selection with; fixed-sized detectors (RNSA) and variable-sized detectors (V-Detector) for classification and detection of anomalies. Classifier algorithms of Support Vector Machine (SVM) and K-Nearest Neighbour (KNN) are used for benchmarking the performances of the real-valued negative selection algorithms. Experimental results illustrate that RNSA and V-Detector algorithms are suitable for the detection of anomalies, with the SVM and KNN producing significant efficiency rates. It was also gathered that V-Detector yielded superior performances with relation to the other algorithms.

Keywords Artificial immune system · Real-valued negative selection algorithm · Variable detector · Anomaly detection

R.S. Khairy · R. Ghazali · A. Lasisi (✉)
Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja,
Batu Pahat, Johor, Malaysia
e-mail: lasisiayodele@yahoo.com

R.S. Khairy
e-mail: rehabsalah08@gmail.com

R. Ghazali
e-mail: rozaida@uthm.edu.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_3

1 Introduction

In modern day life, anomaly is one of the major cause of great losses. A number of anomaly detection techniques are proposed in handling issues related with protecting the integrity of data. These techniques are adequately applied to fault tolerance, robotic control, network intrusion detection, bioinformatics. In general, the problem of anomaly detection can be seen as a two or more class classification problem. Given an element from a given problem space, the system should classify it as normal or abnormal [1, 2]. However, this is a very general characterization since it can correspond to different problems depending on the specific context where it is interpreted. Therefore, from a statistical point of view, the problem can be seen as that of outlier detection which is referred to as an observation deviating from other observations and triggering uncertainty as to how it was generated [3].

Many modern techniques exist in literature that are based on Artificial Intelligence, Neural Network, Bayesian Network, Fuzzy logic, K-nearest Neighbour algorithm, Support Vector Machine, Decision Tree, Fuzzy Logic Based System, Sequence Alignment, Genetic Programming etc., and has evolved in detecting various anomalies [4].

The field of Artificial Immune Systems which began in the early 1990s serve as alternative and efficient algorithms for detecting anomalies to the already existing methods. They were inspired by the Biological Immune System (BIS) which is robust, decentralized, error tolerant, and adaptive in nature. The immune system is highly complicated and appears to be precisely tuned to the problem of detecting and eliminating infections [5]. There are a number of AIS models used in pattern recognition, fault detection, computer security, and a variety of other applications in the field of science and engineering. These AIS models tend to mimic the biological processes of negative selection, immune network, clonal selection, and dendritic cell/danger theory. These models emphasize on designing and applying computational algorithms and techniques using simplified models of various immunological processes and functionalities [6]. The negative selection algorithms which utilizes real-valued representations namely RNSA and V-Detector, as surveyed in [7] shall be applied for anomaly detection.

Hence, in this study, a performance analysis resting on the proficiency of real-valued negative selections; RNSA and V-Detector algorithms for anomaly detection are explored and examined. The structure of the paper is highlighted as follows: Sect. 2 describes the artificial immune system, its inspiration and some of its models. Negative Selection Algorithm and insight into the real-valued negative selection algorithms are discussed in Sect. 3. Experimental simulations, results and analysis are reflected in Sect. 4. The contribution of the study concludes with Sect. 5.

2 Artificial Immune System

The concept and theory of Artificial Immune System will be incomplete without the mention of the source of inspiration in bringing its algorithms into being, referred to the Biological Immune System (BIS). The body has different mechanisms to protect itself (*self* cells) from harmful foreign materials. One of these mechanisms is the natural immune system and its main purpose is to detect and destroy any unwanted foreign cells (*non-self* cells) that could be harmful to the body. These *non-self* cells are known as *antigens* and the natural immune system produces *antibodies* to bind to these antigens.

The Biological Immune System mainly consists of lymphoid organs that create lymphocytes. The two most familiar lymphocytes are the T-cell and B-cell formed in the bone marrow. Both T-cell and B-cell have receptors on their surfaces to bind with the antigen [8]. The immune system is a natural resistance to diseases using sophisticated adaptive mechanisms intended either to destroy the invaders or to neutralize their effects. The BIS can be classified according to functionality into two different layers of defence which are innate and adaptive. The innate immunity is the first line of defense and its non-specific. It is categorized as non-specific because does not concentrate on a particular type of pathogen. When an invasion bypass the innate immunity, the adaptive immunity line of defense is called into action. Adaptive immunity is specific as it targets, matches a particular pathogen, and stores in memory the structure of that pathogen for faster detection and elimination if encountered again [9].

Meanwhile, the artificial immune systems, techniques new to the scene of biological inspired computation and artificial intelligence, are based on metaphor and abstraction from theoretical and empirical knowledge of the mammalian immune system. Brownlee [10] stated that "a robust biological process is critical to combating of disease in the body. Furthermore, the immune system is known to be distributed in terms of control, parallel in terms of operation, and adaptive in terms of function, all of which are features desirable for solving complex or intractable problems faced in the field of artificial intelligence".

There are a number of AIS models used in pattern recognition, fault detection, computer security, and a variety of other applications in the field of science and engineering [6]. Most of these models emphasize on designing and applying computational algorithms and techniques using simplified models of various immunological processes and functionalities [11, 12]. Also, AIS has gained increasing interest among researchers in the development of immune-based models and techniques to solve diverse complex computational or engineering problems [13]. Researchers have explored the main features of the AIS mechanisms and exploited them in many application areas. Based on their aspects, some AIS techniques have been found to be more suitable for certain application areas compared to other AIS approaches. It has been found that negative selection models and algorithms are widely used in fault detection and computer security applications utilizing the self/non-self-recognition aspect. Alternatively, the artificial immune network approaches are used in clustering, classification, data analysis and data mining applications. The clonal selection models are used mostly for optimization problems [14]. The Danger Theory Project/Dendritic Cell Algorithm concludes the major AIS approaches that exist in literature, and they are targeted at anomaly detection and computer security applications based on the identification of danger rather than differentiating between self/non-self as highlighted by negative selection algorithm [15].

3 Negative Selection Algorithm

To guard the *self* cells and also eradicate unknown antigens (*non-self* cells), the Negative Selection Algorithm (NSA) inspired by the biological negative selection is equipped with the *self-non-self* discrimination process [16]. The T-cells are involved in the negative selection process, and starts from within the thymus at an immature state. For the T-cells to acquire maturation, they undergo a pseudo-random genetic rearrangement and are exposed to the *self* cells in the host. The T-cells that react to the *self* cells are eliminated via a process called *apoptosis* while those without reaction are granted permission to leave the thymus and circulate all around the body to detect and destroy *non-self* antigens. The result of such a mechanism is that while on the one hand the (released) matured T-cells kill the *non-self* antigens; they are, on the other hand, non-reactive to the *self* (body) cells. Thus, Negative Selection Algorithm (NSA) may be viewed as a mechanism to discriminate the *self* from *non-self* [17]. There exist two types of NSA based on the data representation, which are the string (or binary) negative selection algorithm, and the real-valued negative selection algorithm.

In illumination of the concept of NSA, the research group lead by Stephanie Forrest proposed the immune negative selection algorithm [18]. This first implementation initially used a binary representation for the elements in the self/non-self space. The main idea of the algorithm is to generate a set of detectors which do not harm *self* and distinguish the *non-self* (unauthorized user, virus, etc.) from *self* (authorized users, protected data files, etc.). This algorithm consists of two processes: censoring and monitoring. The censoring phase caters for the generation of mature detectors. Subsequently, the system being protected is monitored for changes by the detectors generated in the censoring stage. The real-valued negative selection algorithms, the focus of this study, are discussed in the subordinate sections.

3.1 Real-Valued Negative Selection with Fixed Detectors

The Negative Selection Algorithm proposition [18] suffers greatly from time complexity as it is exponential to the size of the matching window (the number of bits used to compare two binary strings). In order to tackle these problems, Gonzalez et al. [19] proposed a negative selection algorithm that uses real-valued representation of the self/non-self space. This algorithm, called Real-Valued Negative Selection Algorithm (RNSA) tries to alleviate the scaling issues of binary negative selection algorithms, while it uses various schemes to speed up the detector generation process.

The real-valued negative selection algorithm using fixed sized detectors is based on a pre-specified number of detectors [19]. This is not the best approach, and obviously provides no guarantee that the non-self space is completely covered. However, by selecting a large enough value for the number of detectors, the algorithm is expected to provide adequate results. The input to the algorithm is a set of self samples represented by n-dimensional points (vectors). The algorithm tries to evolve a complement set of points called *antibodies* or *detectors* that cover the non-self space. This is accomplished by an iterative process that updates the position of the detector driven by two goals: (1) Move the detector away from the self points, and (2) Keep the detectors separated in order to maximize the covering of the non-self space.

3.2 V-Detector Negative Selection Algorithm

The first implementation of the real-valued negative selection algorithm [19] generated detectors in which the distance threshold (or radius) was constant throughout the entire detector set. However, the detector features can reasonably be extended to overcome this limitation. Ji and Dasgupta [20] proposed a new scheme of detector generation and matching mechanisms for negative selection algorithms which introduced detectors with variable properties. The algorithm includes a new variable parameter, which is the radius of each detector. The threshold used by the distance matching rule defines the radius of the detectors; the choice of variability becomes paramount as the non-self hemisphere to be covered by detectors exhibit an option to be changeable with respect to its size.

The V-Detector and the RNSA share similar characteristics when the detection phase is concerned. The significant difference is with the detector threshold utilized for the unknown data detection. This is made possible as each detector is now assigned radius which differs from the RNSA having a constant radius for all the detectors. An unknown data is classified as non-self if the minimum distance to any detector is less than the detector variable radius, and else, it is classified as self.

4 Experimental Results and Analysis

Experiments are performed to provide empirical evidence on the comparative study of the real-valued negative selection algorithms; RNSA and V-Detector for anomaly detection problems, with two different anomaly detection techniques;

SVM and KNN. The MATrix LABoratory (MATLAB) is used for the implementation of the algorithms. Datasets have been retrieved from the UCI Machine Learning Repository and Knowledge Extraction based on Evolutionary Learning (KEEL), and they are: Iris plant data, Balance-Scale data, Lenses data, and Hayes-Roth dataset. The data partition for RNSA and V-Detector is based on the class distributions in the datasets. In order to pass the datasets as input for execution in MATLAB, for a two class dataset, the normal class is employed as the training data (100 %) and considered as *self* while the other class as *non-self*. In the case of datasets with three classes as registered in Iris, Balance-Scale, Lenses, and Hayes-Roth datasets, one of the classes is selected as the *self* and the remaining classes as *non-self*. This procedure is repeated for all the classes, which simply means that each of the class is employed as self for training, with others as non-self. For testing, all the data elements are used in classifying either as self or non-self. In all the experiments, 100 % of the training data is used and have an execution of 20 runs each, with the average values recorded. The Euclidean distance in (1) is used to measure the affinities between the detectors and real-valued coordinates. The parameters for RNSA are: detection radius $r_d = 0.1$, adaptation rate $\eta_o = 0.005$, age of the detector $t = 15$, and decay rate $\tau = 15$. Also, for V-Detector, the parameters are: self radius $r_s = 0.1$, estimated coverage $c_0 = 99.98$ % and Maximum Number of Detectors $T_{max} = 1000$.

$$D = \sqrt{\sum_{i=1}^{n} (d_i - x_i)^2} \tag{1}$$

where $d = \{d_1, d_2, \ldots, d_n\}$ are the detectors, $x = \{x_1, x_2, \ldots, x_n\}$ are the real-valued coordinates, and D is the distance.

4.1 Performance Evaluation

The target outcome of the simulations is to know the ability of algorithms that can perform best with two evaluations performance in consideration. They are the Detection Rate (DR) and False Alarm Rate (FAR) depicted in (2) and (3). If the algorithms perform well and meet the targets, it can then be applied to new data to detect anomalies in the future.

$$DR = \frac{TP}{TP + FN} \tag{2}$$

$$FAR = \frac{FP}{FP + TN} \tag{3}$$

where *TP* represent the number of *non-self* elements identified as *non-self*; *TN* represent the number of *self* elements identified as *self*; *FP* translate to the number of *self* elements identified as *non-self*; *FN* translate to the number of *non-self* elements identified as *self*.

4.2 Simulation Results

The simulation experiments are carried out using MATLAB (R2011b) on Petium4 Core i5 CPU. Results after series of experiments are tabulated, graphed and discussed.

The results shown in Table 1 illustrate the effectiveness of the anomaly detection techniques on Iris and Balance-Scale datasets. The RNSA and V-Detector generated detection rates of 93.93 % and 98.73 % respectively for the Iris data, with 97.38 % and 100 % for the Balance-Scale data. Their false alarm rates are at lowest minimum with RNSA accounting for a higher rate at 2.09 %. The high accuracy rates reveal that the RNSA and V-Detector are equipped with the capabilities of detecting anomalies. The SVM and KNN generated good detection rates with both reaching their highest rates at 97.30 % and 96.70 % for Iris data, and for Balance-Scale data, rates of 91.70 % and 80 % are produced. Higher false alarm rates are attributed to SVM and KNN at 2.20 % and 11.60 % respectively. With respect to all the algorithms, V-Detector proved to be superior and the graph representation translated in Fig. 1.

The results obtained for the detection rate based on Hayes-Roth dataset varied proportionally to real-valued negative selection algorithms, SVM and KNN algorithms as shown in Table 2 and diagrammatically displayed in Fig. 2. The SVM gained superiority over the V-Detector with detection rate of 86.90 % as against 85.12 % for V-Detector. Same could not be reported for the false alarm rate as the table turned against SVM by yielding a higher positive rate at 8.90 %. A 4.11 % false positive rate is attributed to V-Detector, which coincidentally is the rate for RNSA. For RNSA and KNN, detection rates of 82.65 % and 81.30 % are generated respectively. KNN gave the highest positive rate of 11.90 %.

Table 1 Performances analysis for iris and balance-scale

Algorithms	Iris		Balance-scale	
	Detection rate (%)	False alarm rate (%)	Detection rate (%)	False alarm rate (%)
RNSA	93.93	0.90	97.38	2.09
V-detector	98.73	0.00	100	0.00
SVM	97.30	1.30	91.70	2.20
KNN	96.70	1.70	80.00	11.60

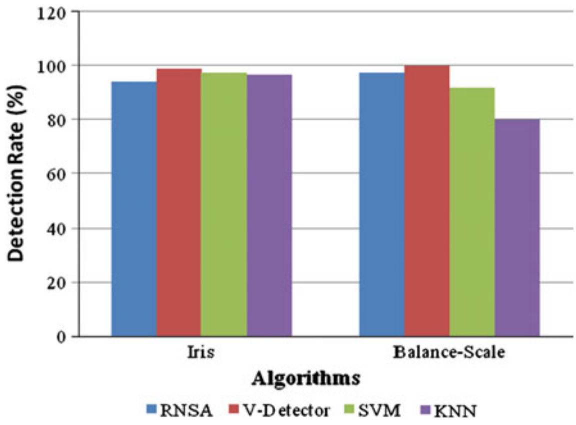


Fig. 1 The detection rates on iris and balance-scale

Table 2 Performances analysis for Lenses and Hayes-Roth

Algorithms	Lenses		Hayes-Roth	
	Detection rate (%)	False alarm rate (%)	Detection rate (%)	False alarm rate (%)
RNSA	99.92	0.00	82.65	4.11
V-detector	100	5.61	85.12	4.11
SVM	83.30	18.60	86.90	8.90
KNN	66.70	32.50	81.30	11.90

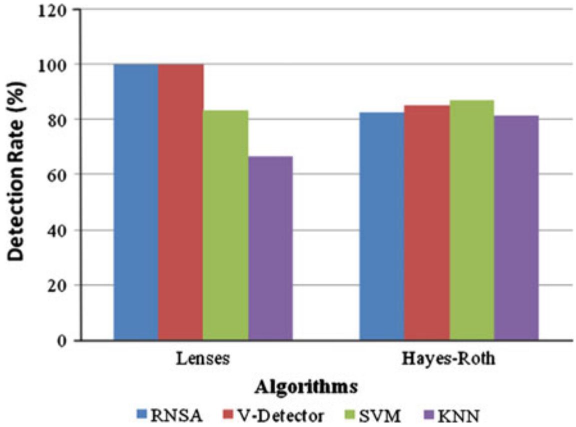


Fig. 2 The detection rates on lenses and Hayes-Roth

Consequently, the performance results for the Lenses data reached the climax of 100 % with V-Detector, followed by RNSA with 99.92 %, SVM with 83.30 %, and lastly KNN yielding 66.70 %. False alarm rates of 0.00 %, 5.61 %, 18.60 % and 32.50 % are recorded by RNSA, V-Detector, SVM and KNN respectively. Overall, the V-Detector surpassed all other algorithms performance wise.

5 Conclusion

The need for ensuring integrity and confidentiality in data has prompted computer scientists and researchers in proffering ways and avenues to adequately secure information. This stem from anomalies or abnormality, and therefore detection improvement requires continuous efforts in many fields, including Artificial Immune System (AIS). For several data, AIS classifiers have proven their ability in classifying successfully those data by revealing the abnormalities therein. As such, this research focuses on using Real-Valued Negative Selection algorithms with focus on fixed detector (RNSA) and variable detector (V-Detector) in classifying different datasets. Two benchmarked algorithms; SVM and KNN are used for comparison and simulated on datasets acquired from standard databases. Their performances are validated with two measuring criteria: detection rate, and false alarm rate, then a comparison carried out based on their performances on the datasets.

Overall, RNSA and V-Detector performed well on the datasets, and with compatible results from the benchmarking algorithms. Meanwhile, V-Detector was the best in terms of detection rate and false alarm rate. Consequently, it can be inferred that the V-Detector yielded more accurate results, provided that the choice of parameters are properly determined and thus affirm the real-valued negative selection algorithms suitability for detecting abnormalities.

Acknowledgements This work is supported by the Office for Research, Innovation, Commercialization, and Consultancy Management (ORICC), Universiti Tun Hussein Onn Malaysia (UTHM), and Ministry of Higher Education (MOHE) Malaysia under the Fundamental Research Grant Scheme (FRGS) Vote No. 1235.

References

1. Patcha, A., Park, J.-M.: An overview of anomaly detection techniques: existing solutions and latest technological trends. Comput. Networks. **51**, 3448–3470 (2007)
2. Lasisi, A., Ghazali, R., Herawan, T.: Comparative Performance Analysis of Negative Selection Algorithm with Immune and Classification Algorithms. Recent Advances on Soft Computing and Data Mining. pp. 441–452. Springer, Berlin (2014)
3. Gonzalez, F.: A study of artificial immune systems applied to anomaly detection (2003)
4. Tripathi, K.K., Ragha, L.: Hybrid approach for credit card fraud detection. Int. J. Soft Comput. Eng. **3**, 2231–2307 (2013)
5. Tuo, J., Ren, S., Liu, W., Li, X., Li, B., Lei, L.: Artificial immune system for fraud detection. Systems, Man and Cybernetics, 2004 IEEE International Conference on. pp. 1407–1411 (2004)
6. Aziz, A.S.A., Salama, M.A., Hassanien, A.E., Hanafi, S.E.-O.: Artificial Immune System Inspired Intrusion Detection System Using Genetic Algorithm. Inform. **36**, 347–357 (2012)
7. Dasgupta, D., Yu, S., Nino, F.: Recent advances in artificial immune systems: models and applications. Appl. Soft Comput. **11**, 1574–1587 (2011)

8. Andrews, P.S.: An investigation of a methodology for the development of artificial immune systems: a case-study in immune receptor degeneracy. University of York, Department of Computer Science, York (2008)
9. Lasisi, A., Ghazali, R., Herawan, T.: Negative selection algorithm: a survey on the epistemology of generating detectors. In: Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013), Lecture Notes in Electrical Engineering. pp. 167–176 (2014)
10. Brownlee, J.: Artificial immune recognition system (airs)-a review and analysis. Swinburne Univ. Technol. Melbourne, Aust. Tech. Rep. (2005)
11. De Castro, L.N., Timmis, J.: Artificial immune systems: a new computational intelligence approach. Springer, Berlin (2002)
12. Dasgupta, D.: Advances in artificial immune systems. Comput. Intell. Mag. IEEE. **1**, 40–49 (2006)
13. Al-Enezi, J.: Artificial immune systems based committee machine for classification application (2012)
14. Al-Enezi, J.R., Abbod, M.F., Al-Sharhan, S.: Advancement in artificial immune systems: a perspective of models, algorithms and applications. GCC Conference and Exhibition, 2009 5th IEEE. pp. 1–6 (2009)
15. Greensmith, J.: The dendritic cell algorithm. Doctoral dissertation, Nottingham Trent University (2007)
16. Chakraverty, S.: Mathematics of Uncertainty Modeling in the Analysis of Engineering and Science Problems. IGI Global (2014)
17. Dasgupta, D., Nino, F.: Immunological computation: theory and applications. CRC Press (2008)
18. Forrest, S., Perelson, A.S., Allen, L., Cherukuri, R.: Self-nonself discrimination in a computer. In: Proceedings of 1994 IEEE Computer Society Symposium on Research in Security and Privacy. pp. 202–212 (1994)
19. Gonzalez, F., Dasgupta, D., Kozma, R.: Combining negative selection and classification techniques for anomaly detection. In: Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02. pp. 705–710 (2002)
20. Ji, Z., Dasgupta, D.: Real-valued negative selection algorithm with variable-sized detectors. Genetic and Evolutionary Computation–GECCO 2004. pp. 287–298 (2004)

An Effective Distributed Service Model for Image Based Combustion Quality Monitoring and Estimation in Power Station Boilers

K. Sujatha, K. Senthil Kumar, T. Godhavari, R.S. Ponmagal and N.P.G. Bhavani

Abstract This research work deals with monitoring of combustion quality in power station Boilers using Service Oriented Architecture which is used to minimize the flue gas emissions at the exit. A model of distributed industrial boilers and integrate it with the Internet through Service Oriented Architectural paradigm is designed. This strategy can be applied to monitor and control the industrial boiler process parameters such as Flame Temperature and Flame intensity. It is proposed to consider the use of service oriented architecture to program and deploy the boiler process parameters. The cost effective technique to develop an intelligent combustion monitoring system is discussed in this paper. A combination of image processing algorithm with Bayesian Classifier is used. The feature extraction was done using Image J and feature reduction was done using Support Vector Machine (SVM). The classification of the flame images based on the features was done using the Bayesian approach whose results are also validated. The combination of the two techniques proved to be beneficial so as to monitor the combustion quality at the furnace level is made possible. Moreover the flue gas emissions are minimized which reduces air pollution. The Service Oriented Architecture is designed to access boiler combustion parameter such as flame intensity as a service and implemented in

---

K. Sujatha (✉) · K. Senthil Kumar · T. Godhavari
EEE/ECE/CSE Department, Center for Electronics Automation and Industrial Research (CEAIR), Dr. M.G.R. Educational and Research Institute University, Chennai, Tamil Nadu, India
e-mail: drksujatha23@gmail.com

K. Senthil Kumar
e-mail: ksenthilkumar@drmgrdu.ac.in

T. Godhavari
e-mail: t.godhavari@gmail.com

R.S. Ponmagal · N.P.G. Bhavani
EEE Department, Center for Electronics Automation and Industrial Research (CEAIR), Dr. M.G.R. Educational and Research, Chennai, Tamil Nadu, India
e-mail: rsponmagal@gmail.com

N.P.G. Bhavani
e-mail: sbreddy9999@gmail.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_4

such a way that for every specific requirement of the monitor/control center, the services of the boilers are invoked through a registry and the specific changes in the combustion parameters are also notified. Different boilers of a power plant can be networked together to monitor different combustion process parameters, and they have been integrated with Internet by registering them as services; hence a complete distributed integration environment is exploited. The aim of this paper is to construct a model of distributed industrial sensors and integrate it with the Internet through Service Oriented Architectural paradigm. This strategy can be applied to monitor and control the industrial process parameters such as Temperature, Pressure, the level of CO, CO_2, NO_x, and Combustion Quality. It is proposed to consider the use of service oriented architecture to program and deploy the sensed process and pollution parameters. The Service Oriented Architecture for sensor network has been extended to Cloud, to access sensor as a service and implemented in such a way that for every specific requirement of the monitor/control center, the assimilation regulator invoke the services of the sensors through a registry and the specific changes in the sensed parameters are also notified as auditable event using push interaction pattern of SOA. The sensed parameters and the combustion quality level can be viewed through mobile, using appropriate authentication.

Keywords Combustion quality · Support vector machine · Bayes net classifier · Naives bayes classifier · Image processing · Euclidean distance · Service oriented architecture · Sensor cloud

1 Introduction

This research work deals with monitoring of combustion quality so as to minimize the flue gas emissions at the exit. The boilers play a major role in the thermal power plants. For reheating furnaces, the steel industry is constantly pursuing more efficient and economical control systems to comply with increasingly stringent environmental regulations without compromising product quality. Degradation of the natural environment as a result of the emission of pollutants demands a better means of process control particularly in energy intensive industries. The reheating of steel and gas turbine power plants are such sectors that consume massive amounts of energy due to the combustion of gaseous fuels such as natural gas [1]. A multi-zone furnace is widely used in the steel reheating industry to ensure efficient and even heat transfer to the load. The furnace is normally monitored and optimised for the desirable overall excess air by monitoring the oxygen in the common stack and individual zone temperature profiles. However, this level of information is not sufficient to detect local imbalances in, for example, the air/fuel ratio of individual burners, which can result in poor combustion efficiency and elevated NO_x and CO emissions. NO_x is one of the pollutants which are emitted from the combustion units of industrial boilers. Strict environmental rules regarding air pollution are being implemented in different parts of the world [9]. These rules require combustion processes to limit NO_x emission to certain regulatory limits. Boiler operators and control systems require on-line NO_x measurements to operate the boilers at the best efficiency while maintaining the emission level within the regulatory limits [11]. The literature survey states that the CO emissions and the combustion quality are measured from the captured flame images using K-means clustering algorithm. The ON-OFF condition of the furnace flame was alone identified using Back Propagation Algorithm (BPA) [4, 5, 8]. Moreover the combustion quality was alone classified using Hidden Markov Model (HMM) [6, 10]. The 2D visualization technique is based on segmentation of the flame images with a threshold value so as to detect the edges of the flame [12]. But here, in this paper an attempt has been made to identify the flame temperature, emission of various flue gases using SVM based feature reduction and Bayes classifier.

The paper discusses about the Service Oriented Architecture for Combustion quality Estimation of Power station Boilers in Sect. 2. Section 3 explains about the Materials and Methods used for the combustion quality estimation. Section 4 discusses the results and Sect. 5 concludes the proposed paper. As a part of the future work incorporation of flue gas sensors can be carried out.

2 A Cloud Based Framework for Integration of Industrial Sensor

2.1 Service Oriented Architecture for Combustion Quality Estimation

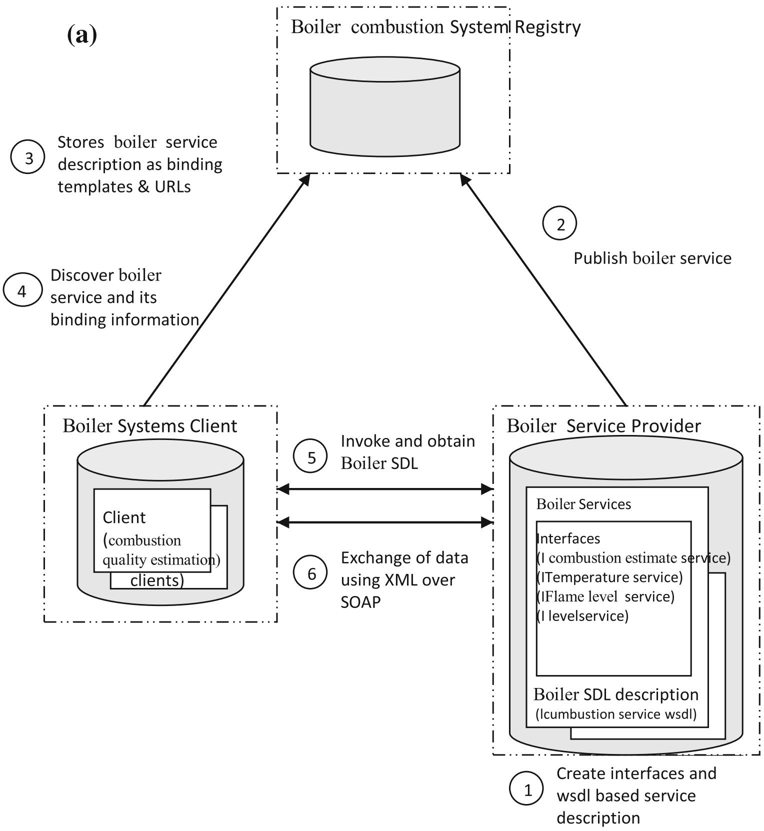
Large process industry systems are a complex (potentially very large) set of (frequently) multi-disciplinary, connected, heterogeneous systems that function as a complex system. Such systems should be based on process control algorithms, architectures and platforms that are scalable and modular (plug and play) and area applicable across several sectors, going far beyond what current Supervisory Data Acquisition and Control (SCADA), and Distributed Control Systems (DCS) and devices can deliver today. A first fast analysis of current implemented SCADA and DCS systems detects a set of major hinders for not completely fulfilling some of all those criteria: the big number of incompatibilities among the systems, hard code data, different view on how systems should be configured and used, co-existence of technologies from a very long period of time (more than 20 years), use of reactive process automation components and systems instead of having them working in a proactive manner. If we began hooking all these hinders, we would soon have an unmanageable mess of wiring, and custom software, and little or no optimal communication. To date, this has been the usual result, where “point solutions” have been implemented without an overall plan to integrate these devices into a meaningful “Information Architecture”.

To address the above issue, in this paper, an efficient Service Oriented Architectural model is proposed for a large process industry such as NLC which has many existing Thermal Power stations and many more to come. These TPS have many Boilers. This paper offers the solution of integrating these boilers to monitor the combustion quality at control centre. Here, it is described how the boiler combustion parameters are converted into web service and how they have been deployed. Web services are application components that are designed to support interoperable machine-to-machine interaction over a network. This interoperability is gained through a set of XML-based open standards, such as the Web Services Description Language (WSDL), the Simple Object Access Protocol (SOAP), and Universal Description, Discovery, and Integration (UDDI). These standards provide a common and interoperable approach for defining, publishing, and using web services. J2EE 1.4 SunAppserver is used as service provider. The J2EE 1.4 platform provides comprehensive support for web services through the JAX-RPC 1.1 API, which can be used to develop service endpoints based on SOAP as in Fig. 1a.

The interface and implementation files for the combustion parameters such as flame intensity and flame temperature, which are estimated as per the technology used in Sect. 3 are written. Configuration files are written to specify the XML namespace and target namespace. These file are compiled to generate WSDL (contains possible inputs and server's address) for client reference and mapping file (port number and service endpoint location) for server reference. With deployment tool war files are generated for the services written and deployed into the server. The lists of services are discovered and invoked by the boiler applications (client), i.e. monitor station, using SOAP messages. Through the above process many Boilers of a power plant can be efficiently integrated and monitored.

The architecture shown in Fig. 1b enables us to couple the services provided by remote industrial sensor networks over the Internet. Fig depicts the SOA Model for distributed networked industrial sensors. The sensor services are deployed into the application server as service description which has the location information and provides a service end point, a target namespace and a transport name. This component makes use of message exchanges in the xml data format. Since programming sensor networks remains too complex with existing programming languages and techniques, using XML messages in sensor networks will optimize the way of getting information from the network. To facilitate orchestration and aggregation of services into processes and applications, an eb-xml-registry-repository (register agent) is used. The registry-repository provides a single view of all services. The sensor services are published into the eb-xml registry using wsdl. The lists of services are discovered and invoked by the sensor applications (client), using XML-RPC and SOAP messages. The client communications are passed through the assimilation regulator (AR). The AR also takes care of the authentication of the users and delivering the required parameters using push interaction pattern. This pattern can be triggered by multitude of events, here an auditable event, trigger (when the process parameters exceeds some threshold) the message sent to the client.

(a)



(b)

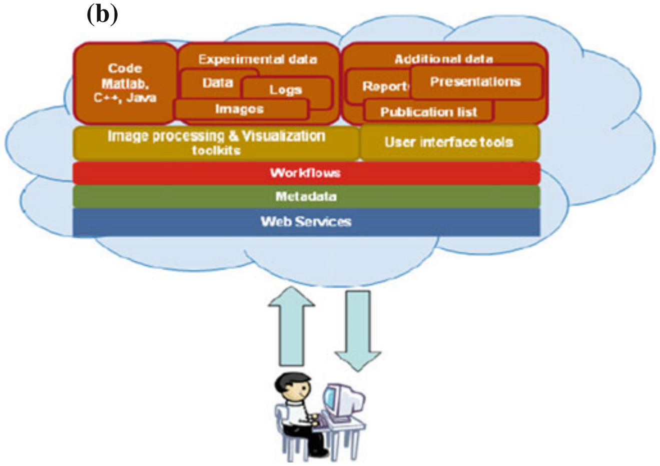


Fig. 1 **a** Service oriented architecture for combustion quality estimation of power station boilers. **b** Flame image analysis application on cloud

Data storage and sharing is difficult for these sensors due to the data inflation and the natural limitations, such as the limited storage space and the limited computing capability. Since the emerging cloud storage solutions can provide reliable and unlimited storage, they satisfy to the requirement of pervasive computing very well. The assimilation regulator is used to Integrate Sensor Networks with Cloud. The cloud provides scalable processing power and several kinds of connectable services. This distributed architecture has many similarities. With a typical wireless sensor network, where a lot of motes, which are responsible for sensing and local pre-processing, are interconnected with wireless connections. Since wireless sensor networks are limited in their processing power, battery life and communication speed, cloud computing usually offers the opposite, which makes it attractive for long term observations, analysis. The deployment of temperature sensor is done in IBM Bluemix cloud.

2.2 Methodology for Industrial Application on Cloud

Due to the massive number of programming languages and image processing toolkits, simple sharing of resources as plug and play functionality is difficult. While the proposed cloud computing framework in Fig. 1 allows a platform to have an independent access to remote computing services, its web services allow end-users to fully interact with data, information requests as well as applications with a low level of user interaction. In this proposal, I am going to study different components of this framework and develop them using existing software solutions, imaging toolkits and Microsoft technologies. For instance, Visual Studio can be utilized as the development environment, SQL server as the database manager, and Microsoft Workflow foundation as the workflow engine [12, 13]. A number of work packages have to be studied/developed for this framework:

- Evaluate existing software environment solutions.
- Design standard datasets, metadata and web services.
- Design workflow orchestration and enactment, for example by Windows Workflow Foundation.
- Incorporate existing imaging and visualization toolkits
- Design mechanisms and tools to enhance software usability for clinicians

A number of challenges have to be addressed. The adaptation of existing software will require an analysis and may require significant time. Also, merging toolkits with different languages (such as Matlab, Java and C++ as shown in Fig. 1b) and their communication will need a deep study. The design and development of suitable user interfaces for clinicians is another challenge. The idea is to use multi-touch technology to improve the interaction with image analysis applications as well as the learning experience. Finally, contributions not only should meet the quality but also facilitate the way researchers do it. The design of mechanisms to provide such facility must be analyzed and evaluated.

3 Materials and Methods

3.1 Existing Set up at Neyveli Lignite Corporation Ltd

The Neyveli Lignite Corporation (NLC), Neyveli, Tamil Nadu, India is one the major thermal power plants in India with highly modernized technology. The Thermal Power Station (TPS-Exp I) is the most recently renovated power station. The firing system is called as the tangential firing system as shown in Fig. 1. The flame monitoring system used here indicates only the presence or absence of the flame inside the furnace. The major idea to determine the ON-OFF status was to prevent the explosion of boiler. The repeated addition of coal without monitoring the status of the flame will collapse the entire set up [1]. It is also possible to estimate the temperature and flue gas emissions by monitoring the flame colour [7, 8]. To avoid this situation the above said flame monitoring system is used. As a part of the existing setup the temperature measurement is done using JK type of thermocouple and the measurement of flue gases done using gas analyzers (very expensive and prone to cold end corrosion) (Fig. 2).

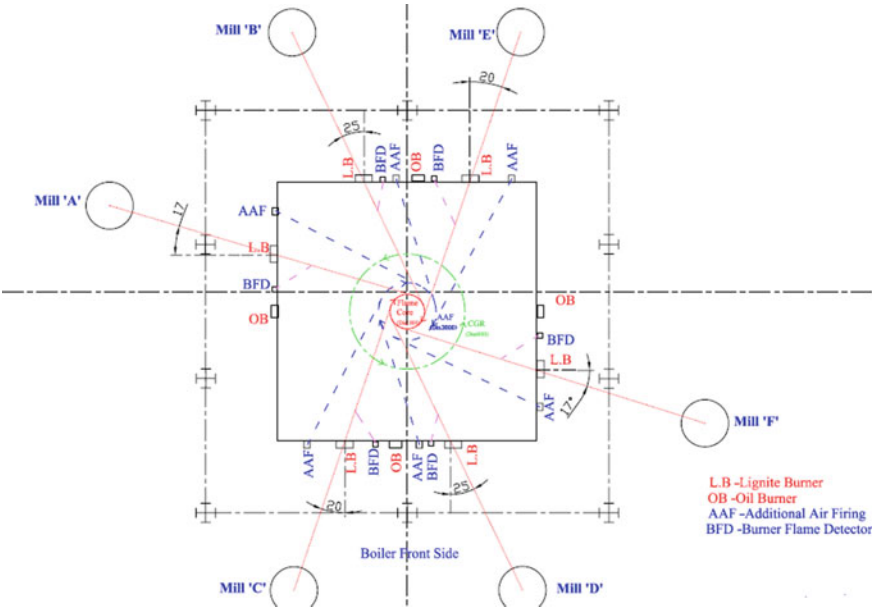


Fig. 2 Tangential firing system *(Courtesy Neyveli Lignite Corporation)*

Fig. 3 Proposed set up at NLC

3.2 Proposed Method for Flame Monitoring in Boilers Using Image Processing

The flame images are obtained from the flame video. The intensity of the flame varies with respect to temperature and flue gas emissions [1]. The features are extracted and then reduced using SVM. The reduced feature set is used as an input to the Bayesian classifier and finally the classification performance is validated with certain performance measures. Figure 3 shows the schematic representation of the flame monitoring system.

3.3 Procedure for Data Collection

The flame images corresponding to three different combustion conditions were collected by varying the air/fuel ratio and taking the corresponding readings of flame temperature, NO_x, SO_x, CO and CO_2 emissions [2, 3]. The sample images corresponding to the above said three categories with their respective parameters are tabulated in Table 1. The measurements for flue gas emissions are obtained from the gas analyzers placed at the exit of the chimney. The temperature measurement is made using thermocouple as mentioned earlier.

3.4 Preprocessing

Preprocessing is done to remove noise. The frame size acquired was 320 × 240. It is sufficient that we take 30 × 30 portion of the image for further analysis. The extracted flame images if corrupted with noise, then filtering need to be done. The various filters include Unsharp mask filter, Maximum filter, Minimum filter, Variance filter, Mean filter and Median filter. The filtered images using the above said filters are displayed in Fig. 4 shown below. It is inferred that the performance of median filter on noise removal was good when compared with the other filters. The mean filter also performs filtering to a certain extent but the clarity of the filtered image is slightly low when compared with the median filter. The other filter types are not suitable because the basic information present in the original image is lost. Hence it is better to use a median filter for noise removal pertaining to this

Table 1 Data for measurement of various flue gas emissions, combustion quality and temperature

Combustion category	Image	Flame temperature (deg Celsius)	CO emission in (ppm)	Combustion quality in (%)	Air/fuel ratio (no units)	Temperature of superheated steam in (°C)	CO_2 emission Nm^3/hr	NOx emissions mg/Nm^3
Class1 complete combustion		1250	100–120	80–100	4:1	530	400	70
Class 2 partial combustion		900	200–210	50–70	Slightly less than 4:1	240	700	120
Class 3 incomplete combustion		300	300	30–40	Very much less than 4:1	170	1000	200

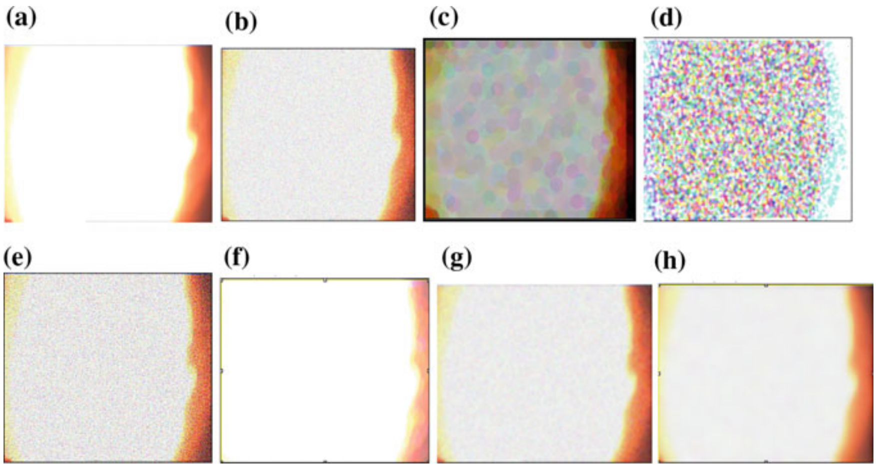


Fig. 4 Effect of different filtering techniques on the corrupted flame image. **a**.Original flame image. **b** Flame image corrupted with noise. **c** Effect of minimum filter. **d** Effect of variance filter. **e** Effect of unsharp mask filter. **f** Effect of maximum filter. **g** Effect of mean filter. **h** Effect of median filter

situation. The simulation for addition and removal of noise was done using Image J. The performance evaluation for noise removal is done using Mean Squared Error (MSE) and Peak Signal to Noise Ratio (PSNR) in decibel (dB) and the results are tabulated in the Table 2. The MSE is minimum and PSNR value is maximum in case of the median filtering. Hence it is inferred that median filter is effective in case of removing the salt and pepper noise. The $M \times N$ is the size of the image, O(m, n) denotes the original image and R(m, n) denotes the retrieved image. The expressions for MSE and PSNR are given by Eqs. 1 and 2 respectively.

$$\mathrm{PSNR} = 10 * \log_{10}\left[(255)^2/\mathrm{MSE}\right] \tag{1}$$

$$\mathrm{MSE} = \sum_{m,n} \sqrt{}[\mathrm{O}(m,n) - \mathrm{R}(m,n)]/(\mathrm{M} \times \mathrm{N}) \tag{2}$$

Table 2 Performance evaluation for noise removal by various filters

S.No	Name of the Filter	MSE	PSNR (dB)	S.No	Name of the Filter	MSE	PSNR (dB)
1.	Minimum filter	1.145	47.54	4.	Maximum filter	0.83	48.94
2.	Variance filter	2.228	44.65	5.	Mean filter	0.146	56.49
3.	Unsharp mask filter	4.327	41.77	6.	Median filter	0.0038	62.33

3.5 Feature Extraction

The features are the basic pattern present in the image that gets repeated in various directions. Features like area of the flame, mean intensity, maximum intensity, minimum intensity, kurtosis, centroid, mode, perimeter, standard deviation and median for all the groups of flame images are extracted using Image J.

3.6 SVM Based Feature Reduction

The features are reduced using SVM technique. Support vector machines (SVMs) are a set of related supervised learning methods used for classification. Intuitively, an SVM model is a representation of the points in space, mapped so that the features of the separate categories are divided by a clear gap that is as wide as possible. New features are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. Support vector machine constructs a hyper plane or set of hyper planes in a high or infinite dimensional space, which can be used for classification tasks. Here the attributes are ranked by the SVM technique [7]. The results obtained (screen shots) by SVM feature reduction with the variation of features in Fig. 5 along with the ranking of features by SVM in Table 3.

3.7 Training and Testing by Bayes Classification and Euclidean Distance Classifier

A Bayes classifier is a simple probabilistic classifier based on applying Bayes theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood [13].

The Euclidean distance classifier is a conventional classifier and it is based on the simple two point formula. The calculation of Euclidean distance is based on the L_2 norm.

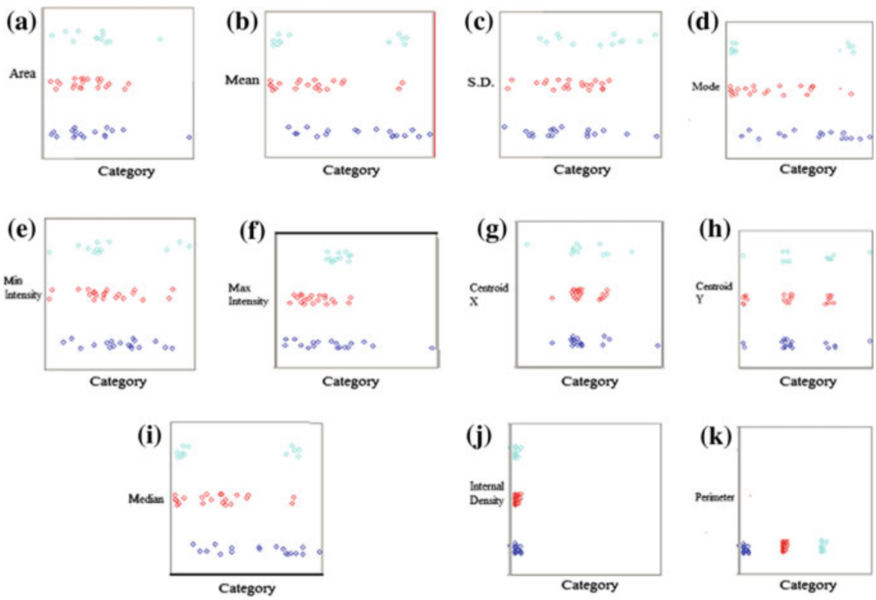


Fig. 5 Variation of features from SVM feature evaluator

Table 3 Ranking of Features by SVM

Ranking	Features	Ranking	Features	Ranking	Features
1	Perimeter	5	Centroid about Y	9	Mode
2	Area	6	Centroid about X	10	Standard deviation
3	Minimum Intensity	7	Maximum Intensity	11	Mean
4	Median	8	Kurtosis	–	–

4 Results and Discussion

4.1 *Output for Euclidean Distance Classifier*

The outputs for Euclidean Distance Classifier are shown in the Fig. 6. The number of images classified correctly for class 1, class 2 and class 3 are 11, 8 and 12 respectively.

4.2 *Training by Bayes Net Classifier*

Bayes net is a powerful tool for knowledge representation and inference under conditions of uncertainty [11].The classification is done with the feature set ranking

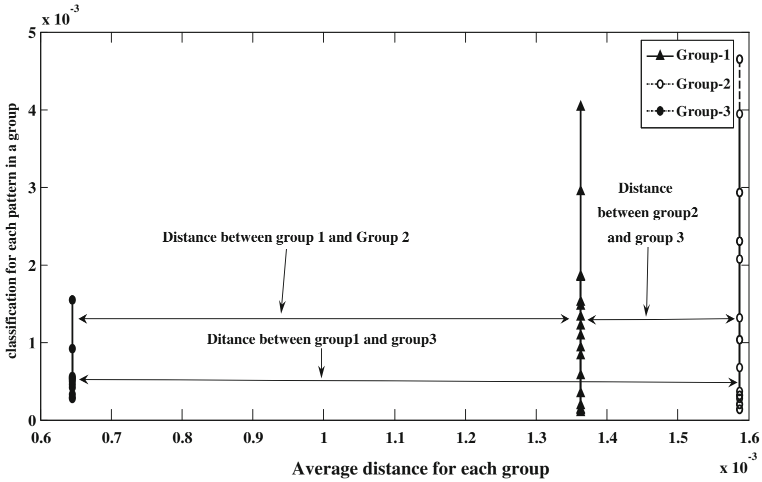


Fig. 6 Distribution of patterns for all the three classes using Euclidean classifier

Table 4 Estimation parameters with features ranking 1, 2, 3, 4 and 1, 2, 3, 4, 5, 6

Parameters	Value for Bayes net		Value for Naive Bayes	
Total no. of instances	51	51	51	46
Instances classified correctly (%)	100	100	90	90.16
Kappa statistic	1	1	0.8485	0.8496
Mean absolute error	0.0359	0.029	0.077	0.0785
Root mean squared error	0.0391	0.0377	0.1946	0.2183
Relative absolute error	0.0826	0.067	0.1737	0.1807
Root relative squared error	0.08233	0.0877	0.4177	0.4688

1, 2, 3, 4 and 1, 2, 3, 4, 5, 6. The classification was 18, 21 and 12 corresponding to the three categories. Since in both the cases all the 51 instances were classified it is chosen that the features 1, 3, 5, 7 can be considered for training. Tables 4 and 5 shows the various parameters during training of the classifier with two different feature set. The results for training by Bayes net classifier are shown in Table 5 with feature set 1, 2, 3, 4 and 1, 2, 3, 4, 5, 6 respectively.

Table 5 Evaluation parameters of training for Bayes net classifier

Feature set	TP rate		FP rate		Precision		Recall		F-measure	
Features ranking 1, 2, 3, 4 and 1, 2, 3, 4, 5, 6	1	0.778	0	0.03	1	0.993	1	0.778	1	0.646
	1	0.952	0	0.153	1	0.833	1	0.952	1	0.889
	1	1	0	0	1	1	1	1	1	1

Table 6 Evaluation parameters for training with Naives Bayes classifier

Feature set	TP rate		FP rate		Precision		Recall		F-measure	
Features ranking 1, 2, 3, 4	1	0.833	0	0.03	1	0.993	1	0.778	1	0.646
	1	0.905	0	0.1	1	0.864	1	0.906	1	0.889
	1	1	0	0.26	1	0.923	1	1	1	0.96

4.3 Training by Navies Bayes Classifier

A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes theorem with strong (naive) independence assumptions [10]. The classification was done with the feature set 1, 2, 3, 4 and 1, 2, 3, 4, 5, 6. The classification is 18, 21 and 12 corresponding to the three categories with the feature set ranked 1, 2, 3, and 4. Similarly the classification was 15, 19 and 12 corresponding to the three categories with the feature set 1, 2, 3, 4, 5, and 6. The number of instances classified correctly is less in number as compared to the Bayes net classifier. Hence the Bayes net and Naives Bayes classifier are used for testing. The results for training by Navies Bayes is shown in Table 6 with feature set 1, 2, 3, 4 and 1, 2, 3, 4, 5, 6 respectively.

4.4 Testing by Bayes Net and Naives Bayes Classifier

Testing the Bayes net classifier is done with 27 images (9 images from class 1, 9 images from class 2 and 9 images from class 3 respectively). The results for testing the Bayes net classifier is shown below with the feature set 1, 2, 3 and 4 and similarly the testing of Naives Bayes classifier is also done with the same set of 27 images. The results for testing the Bayes net and Naives Bayes classifier with the feature set 1, 2, 3 and 4 for the Bayes net and Naives Bayes classifier yields maximum number of flame images to be correctly classified. The testing was done by cross validation to evaluate the performance of the classifier. The performance metrics in Table 7 shows that the values of FP rate and various types of Bayes classifiers are within tolerance.

Table 7 Evaluation parameters during testing with features 1, 2, 3, 4 for Bayes net and Naives Bayes

Lassifier type	Category	TP rate		FP rate		Precision		Recall		F-measure	
Bayes net and Naives Bayes classifier	Class 1	0.889	1	0.167	0	0.727	1	0.889	1	0.8	1
	Class 2	0.778	1	0.056	0	0.889	1	0.889	1	0.889	1
	Class 3	1	1	0	0	1	1	0.778	1	0.875	1

4.5 Deployment of Industrial Sensor for Monitoring Process Parameters

Since the process parameters and combustion quality are monitored and controlled through cloud, the proposed system offers pollution free environment in industries, as there is no necessity of maintaining very large servers, which leads to Green Computing. Different types of sensors are networked together to monitor different parameters, and they have been integrated with Internet by registering them as services, hence a complete distributed, and integration environment is exploited. The results using Bluemix are shown in Figs. 7, 8 and 9 respectively.

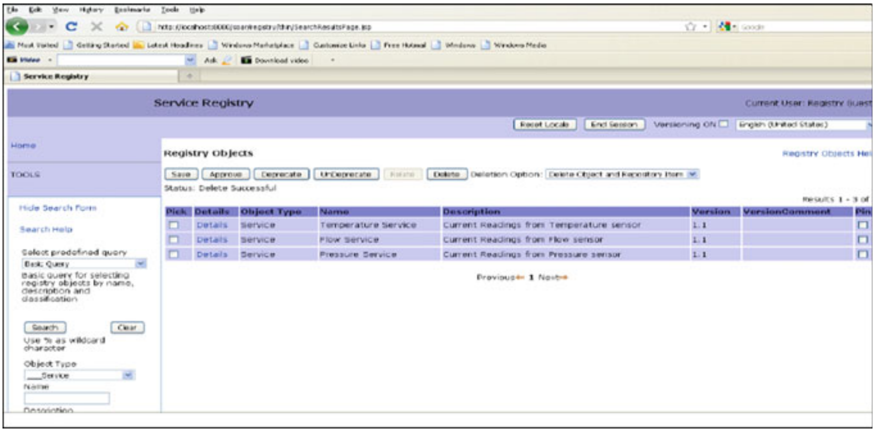


Fig. 7 Industrial sensor system registry deployed in the cloud

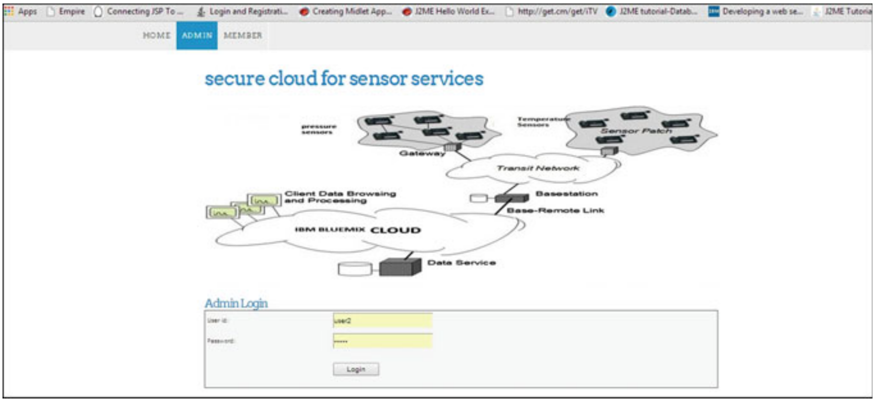


Fig. 8 Industrial sensor for temperature measurement deployed in the cloud

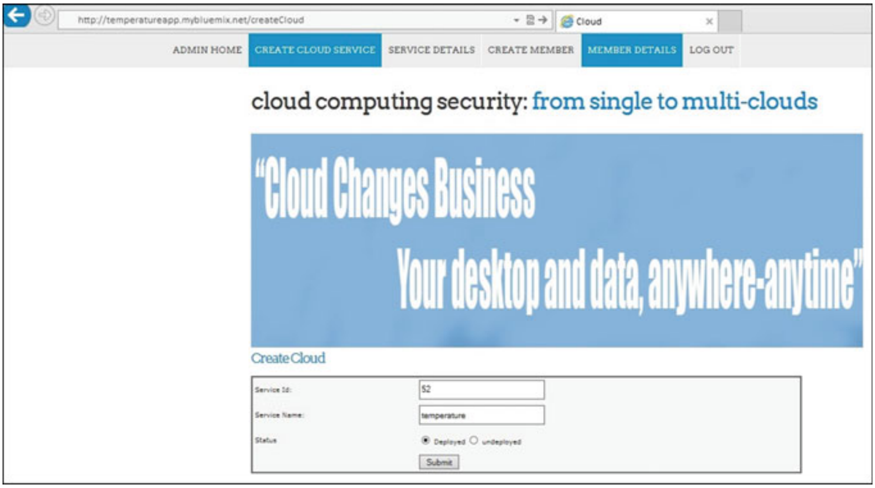


Fig. 9 Industrial sensor for temperature measurement deployed in the cloud

5 Conclusion

The flame images were collected from the control room through the Service Oriented Architecture for a boiler in the power station. Thirty nine correct images were identified. The images were preprocessed and features are extracted. The extracted features are reduced using SVM. Training is done with 51 images taken from class 1, class 2 and class 3 images. Testing results indicate that various types of Bayesian classifier give maximum classification performance. It is evident from the analysis that the SVM with various types of Bayesian classifier yields optimal values for recall and precision. Classification performance can be improved by further preprocessing the acquired images and also increasing the number of images in the database. The proposed solution of SOA for combustion quality estimation of boilers meets out the operational independence of the constituent systems, managerial independence of the constituent systems, geographical distribution of the constituent systems, evolutionary development and emergent behavior.

Thus an intelligent architectural system to analyze the process parameters with converging technologies such as Cloud computing is implemented and the result is investigated. The combination of wireless sensor networks, with their huge amount of gathered sensor data, with a cloud computing infrastructure makes it attractive in terms of integration of sensor network platforms for monitoring different process parameters, scalability of data storage, scalability of processing power for different kinds of analysis, worldwide access to the processing and storage, and be able to share the results more easily. Flawless incorporation of wireless sensors with web service architectures, support the ubiquitous provisioning of high quality industrial services in a distributed service provisioning environment, allow for real time implementation of the industrial conditions of remote parameters, independent of their location.

References

1. Bhavani N.P.G., Sujatha K., Kumaresan M., Pongamal R.S., Brahmananda reddy: Soft sensor for temperature measurement in gas turbine power plant. Int. J. Appl. Eng. Res. **9**(23), 21305 (2014). ISSN 0973-4562
2. Ponmagal R.S., Dinesh N., Rajaram U.: Design and development of secure cloud architecture for sensor services. Distributed Computing and Internet Technology Lecture Notes in Computer Science, vol. 8956, pp. 339–344. Springer, Berlin (2015)
3. Rajesh V., Gnanasekar J.M., Ponmagal R.S., Anbalagan P.: Integration of wireless sensor networks with cloud. In: IEEE International Conference on Recent Trends in Information, Telecommunication and Computing, pp. 321–323 (2010)
4. Sujatha, K., Pappa, N., Kumar, S.K., Nambi, U.S.: Monitoring Power Station Boilers Using ANN and Image Processing, vol. 631–632, p. 1154. Trans Tech Publications, Switzerland, Advanced Materials Research (2013)
5. Sujatha, K., Pappa, N.: Combustion quality monitoring in PS boilers using discriminant RBF. ISA Trans. **2**(7), 2623–2631 (2011)
6. Han, P., Zhang, X., Zhen, C., Wang, B.: Boiler Flame Image Classification Based on Hidden Markov Model. IEEE ISIE (2006)
7. Meyer, D., Leisch, F., Hornik, K.: The support vector machine under test. Neurocomputing **55** (1–2), 169–186 (2003)
8. Lu, G., Gilabert, G., Yan, Y.: Vision based monitoring and characterization of combustion flames. J. Phys. Conf. Ser. **15**, 194–200 (2005)
9. Wójcik, W: Application of fibre-optic flame monitoring systems to diagnostics of combustion process in power boilers. Bull Pol Acad Sci Technical Sci **56**(2) (2008)
10. Rish, I.: An empirical study of the naive Bayes classifier. IJCAI Workshop on Empirical Methods in Artificial Intelligence (2001)
11. Guo, Y., Greiner, R.: Discriminative model selection for belief net structures, AAAI. pp. 770–776 (2005)
12. Shahriyar, R., Bari, M.F., Kundu, G., Ahamed, S.I., Md. Akbar, M.: Intelligent mobile health monitoring system (IMHMS). Int. J. Control Autom **2**(3) (2009)
13. Doukas, C.: Mobile healthcare information management utilizing cloud computing and android OS. In: 32nd Annual International Conference of the IEEE EMBS, Buenos Aires, Argentina (2010)

Implementation of P2P File Sharing Using Bi-directional Chord Protocol Algorithm

Vasanthi Vatsavai, Sriram Suravarapu and Nader F. Mir

Abstract One of the key components of achieving efficiency when looking up on resources is the success of peer-to-peer (P2P) applications. As the structured P2P is an efficient way to locate resources, our study performs two improvements in a structured P2P resource lookup protocol of Chord-based algorithms which is widely accepted. The proposed improvements in detail are presented using two methods. In the first method, the routing information for accurate search of resource is low. In this method, only clockwise direction of resource lookup can be implemented and a unique algorithm is designed for enhancing the finger table in a Chord. A counter-clockwise finger table is also included for generating resource queries in two direction clock wise and anticlockwise increasing the density of neighboring fingers. The proposed model also employs a new functionality of eliminating excess fingers generated by the inclusion of fingers by the suggested model. The results of simulation on effectiveness of the search in the proposed model against the average lookup hops and average lookup delay are presented. The model of algorithm context is extended as finger table resulting in forwarding-storm of routing maintenance messages.

Keywords Peer-to-peer networks · P2P · Distributed content sharing · Distributed systems · Internet file sharing

1 Introduction

A *peer-to-peer* (P2P) network is utilized to form a procedure in between two peers to share or substitute data [1–4]. Usually in a peer-to-peer network, each and every peer can similarly engage in the network or action as client as well as server as

V. Vatsavai · S. Suravarapu · N.F. Mir (✉)
Department of Electrical Engineering, Charles W. Davidson College of Engineering, San Jose State University, San Jose, CA, USA
e-mail: Nader.Mir@sjsu.edu

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_5

equally source and a beneficiary [1, 5–7]. Presently, the peer-to-peer technology is extensively used in various programs using the Internet wherein users can instantly connect or use an average server [8–10]. A P2P network is an interconnected network of peers, where the information is stored in a allocated manner in all peers reducing the require for a focused server and a peer can consult data from any more peer more than a point of time [11]. There are two models of peer-to-peer networks:

1. *Structured peer-to-peer networks* that communicate models possess fixed associations overlay exactly where a distributed hash table dependent indexing is utilized commonly. Structured P2P uses DHT categorization similar as in the Chord system.
2. *Unstructured peer-to-peer networks* require association or search engine optimization of network associations lacking any algorithm [1]. Generally, there are three models of unstructured P2P what have become described as of currently:

   (a) Pure peer-to-peer techniques in which the complete network includes only peers with identical potential. So there is simply one routing layer here are no specialized or significant priority peers with any specialized infrastructure perform.
   (b) Hybrid peer-to-peer systems that include peers consisting of specialized or significant priority infrastructure peers with exclusive infrastructure function described as super peers.
   (c) Centralized peer-to-peer systems that utilize a central server for categorization applications and to load and determine the entire system of unstructured peers.

Napster frequently utilized for network file sharing is a focused system of P2P network example. Gnutella and Freenet incorporate the decentralized model while Kazaa accepts the hybrid model. In [12], structured P2P networks are recommended. A structured P2P model approach depends on the fact that a peer can browse all the remaining peers in search of a necessary file.

1.1 Distributed Hash Table (DHT)

A *distributed hash table* (DHT) is a "structured" P2P network and operates as a database that contains (*object name, address of object*) pairs. An *object name* can be a music title, video clip name, or a book title, and an *address of object* can be an address of a peer that holds the object such as an IP address of a peer. Clearly, a database held by a peer can store multiple (object name, address of object) pairs and an object may be stored by multiple peers.

The challenge in such a P2P networking is how to build a "distributed" database of (*object name*, *address of object*) pairs where each peer can store only a subset of all pairs. In such an environment, a peer must query a significantly large distributed database using a particular key. The DHT must then be able to locate those peers who have the (*object name*, *address of object*) pairs where the targeted key in this pair is contained, and then return the (key, value) pairs to the requesting peer. This type of distributed database is referred to as a *distributed hash table* (DHT).

In order to build such a distributed database of objects for P2P applications, a *hash function* is used. The details of hash functions are presented in [1]. Briefly, a hash function indicated by h is utilized to expose deliberate corruption of a message when any *message* is mapped to data of a fixed size called a *hash*, or a *message digest*. A hash function that operates in a computer network allows a host to easily verify that some input data is authenticated. Hash functions are typically used in *hash tables*, to agilely find a data record by mapping a search key to an index. The index in a hash table provides the location in the hash table where the corresponding record is stored.

1.2 Models for Chord Protocol

The chord protocol has been designed to manage a major problems associated with the P2P software i.e. inefficiency in distinguishing a peer holding a particular data item. Essentially an allotted lookup protocol generally maps a designated key to a peer. The Chord protocol offers a quick estimation of hash perform of keys mapped to peers utilizing regular hashing. Whenever a new peer is manufactured in the system its connected keys are consistently dispersed to all the peers. A chord peer obtains the hash advantages of its nearby peers and stores the data in a dispensed manner ensuing in effectively building a well stabilized load and subsequently it is a scalable protocol also. Frequent hashing is utilized to particularly name individual peer. It analyzes a peer along with its IP address and key hashed towards an m-bit identifier utilizing a hash function. The endeavor organizes all the identifiers in an identifier circle modulo. The peers are then designated keys evaluating the identifier of a key using the identifier of a peer as well as if identifier's value is additional than regarding of the key, the key gets allotted to a specific peer. The chosen peer is defined as the substitute of the appointed key denoted as successor (k), indicating successor of key k. The identifiers if illustrated in a circle of quantities starting from, then successor (k) is given the first peer when traversed in a clockwise location from k.

1.3 Chord Lookup Algorithm

Figure 1 demonstrates a Chord ring using $m = 6$ [13]. The Chord ring has 10 peers and stores five keys. The substitute of identifier 10 is peer 14, therefore key 10 might be situated at peer 14. Likewise, keys 24 and 30 might be situated at peer 32, key 38 at peer 38, and key 54 at peer 56″ [11]. In the procedure of determining the peer with the desired file and minimizing the time factor is a significant factor of the experiment. In the strategy of lookup, chord protocol affords each peer to search its

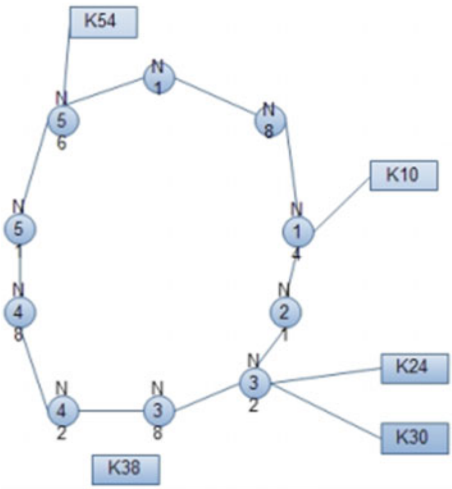


Fig. 1 An identifier circle (ring) consisting of 10 peers storing five keys [13]

substitute for the significant key. As each and every peer has an accessibility of its substitute the lookup traverses the chord ring additionally and again in lookup for the key which is a time using process. To manage this issue, the chord protocol maintains a finger table to reduce the time factor of the lookup for the key.

If the total number of bits in an identifier is supposed to be *m*, most peers managed by a routing table possess an optimum of *m* entries. This table is recognized as the finger table and is continued by every peer. The *i*th entry of the table maps the peer's first successor of *f* succeeding peers by at least $2^t - 1$ on the identifier circle i.e. $f = successor(n + 2^{t-1})$ and this peer *f* is called as peer n and is denoted as n · finger[i]. The finger table might be fully understood from Fig. 2.

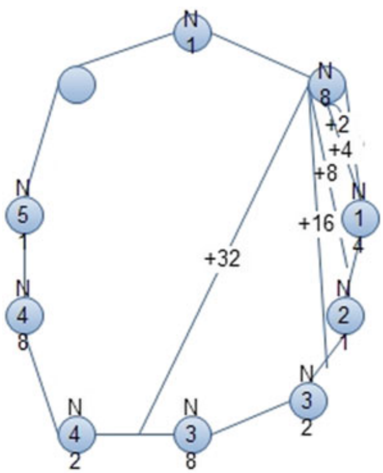


Fig. 2 Finger table entries for each peer in chord ring

The ring has two basic posts, one for the identifier and another for the IP Address of a specific peer.

Figure 2 demonstrates the finger table for peer N8. The first finger of peer 8 points to its descendant which is peer 14, because peer 14 is the first successor which succeeds $(8 + 2^0)$ mod $2^6 = 9$. Likewise, the last finger of peer 8 points to peer 42, as peer 42 is the first peer which is the descendant of $(8 + 2^5)$ mod $2^6 = 40''$. It is apparent from the figure that the first finger for every peer is its first successor.

In the Chord protocol model, a new peer getting into the system demands another peer to notice a successor for it and while the successor is discovered it is identified its successor. The successor is conscious that the new peer is its predecessor though the preceding predecessor of the new peer's successor is unmindful of the new peer launched in the system. To enhance the system of arrangement changes, every peer is rejuvenated in intervals of time and each and every peer requests its successor and predecessor for modifications. Thus if a peer is included or deleted from the present system, the predecessor peer of the peer to be deleted will consider all successor nodes of the deleted peer as it's successors. Each peer stores entries of its successive one and the past n peers in a table structure to maintain the order. Progressive messages are transmitted to some other peers in the table and responses are received from peer outcomes in the peer to store the obtained keys and label.

2 Proposed Model

Our study presents techniques with conquering the challenges affiliated with the primary Chord algorithm. We thrive to acquire a capable and fast lookup for peers incorporating keys by adjusting the algorithm. As the authentic Chord algorithm runs lookups for a particular key in clockwise position resulting in enhanced search time and quantity of routing messages in the P2P network, we recommend the bidirectional lookup. The performance of the bidirectional query in the ring is based on the fact that the search for the key is accomplished in the peers predecessor being different to the original version where the general Chord ring requires to be traversed. Thus, it necessitates fewer hops and is quicker.

The prevailing finger table stores the successors additionally their mapped keys for a particular peer even though the anti finger table retailers the list of predecessor peer. Opposite to a finger table that affiliate links to peers in clockwise location, the anti finger table links the peers in anti-clockwise location. If we choose two peers A and B with a prevailing link in finger table of a Chord from express peer A to peer B, an anti finger table links in opposite from peer B to peer A in anticlockwise direction.. Thus, a superior Chord protocol is made of three elements: Finger Table, Successor List, and Anti finger Table. In this customized protocol, the first two factors are similar as the unique Chord protocol.

2.1 Lookup Algorithm

The opposite finger tables require the revision of the unique algorithm for obtaining the keys. The authentic protocol key-data mappings are managed in the application of the customized algorithm. The increased algorithm executes the lookup portion in the anti finger table if the peers appreciate is greater compared to the key, and it retrieves the data rather of looking the completely ring. Also, if a key benefit is not found in the anti finger table subsequently, lookup is carried out in the finger table. The anti-clock is a practical search comparable to the clock wise search that benefits in the key's to be noticed quickly.

Also each program is built-in within the additional presentation evaluation features to calculate the routine algorithm. The guidelines process requires model of Chord algorithm utilizing Open Chord platform; for improvement we used My Eclipse 8.0; the source code produced to a.bat file; the control line is used to check the adjustments in the source code. The finger table continues the same thus is not necessitating any customizations to the table or the affirmation of the finger table description. The optimum number of items in the finger table is considered as m and the finger table size is $O(\log N)$ where N denotes the total number of peers in the network. Each and every peer also has an anti-finger table with m ingress and its size is $O(\log N)$. The BiChord lookup algorithm executes the lookup in iterations. The iterative look up of the required key is performed with a peer transmitting a search message and if there are no entries nearer than itself equally in finger and opposite finger table, the peer is identified as predecessor or substitute of the key. In other words, the algorithm will certainly search the fingers as well as anti-fingers of the routing table and ahead the lookup content to the next hop that is better than the current peer till it detects the peer past or making it the necessary key.

The result of the copied Chord algorithm finds exactly where the simulator is constructed on Java Virtual Machine. As the level of simulator peers is relatively small (generally less than 50 peers), it transfers up the lookup process by using the chord circle moderately. To examine the performance of lookup process, a timer demonstrates the better results that can be acquired is a millisecond. The timer though fails in giving appropriate results when the endeavor is executed numerous times in determining the cost. Consequently, we use the hop counter to compute the lookup procedure steps exactly where the hops counter is insensible by the hardware arrangement and gives appropriate results of the algorithm. The bi-directional lookup algorithms pseudo code is as below:

```
Peer findPredecessor(key, n){
Peer pred = n.getPredecessor();
if ( pred == null )
        return n; //n is the current peer
else if (key.isInInterval(pred.ID, n.ID) //check if thekey is between the pred and current peerreturn n;
 else {Peer n' = getClosestPrecedingPeer(key) //if not,
 track the closest preceding peer and lookup again
 return findPredecessor(key, n') // recursively
 find the predecessor of peer n'
 }
 }

    Peer findSuccessor(key, n){
    Peer succ = n.getSuccessor();
    if (succ == null)
    return n; //n is the current peer
    else if (key.isInInterval(n.ID, succ.ID) //check if the
key is between the current peer and successor's peer
    return n;
    else {
    Peer n'=getClosestPrecedingPeer(key) //if not,
track the closest preceding peer and lookup again
    return findSuccessor(key,n') // recursively find
the predecessor of peer n'
}
}

    Set < Serializable > retrieve_R(key){
    hops_R = 0; //initialized the hops counter in anti-finger table direction
    whild(!retrieved){
```

```
                Peer responsible_R=null;
                responsible_R= findPredecessor(id);
                hops_R+=1; //while not retrieve the desired key,
        add the hop counter by 1
                try{
        results_R= responsiblePeer_R.retrieveEntries(id);
// get the responsiblePeer to fetch the entry
                retrieved = true; //if successfully get the value,
        set retrieved state to true
                }catch(Exception e){}
                        continue;
                        }
                }
                if(results_R!=null) values1.add(entry.getValue());
//add the lookup result to the valueset
        final_hopsR = hops_R; //get the hop counter for the current lookup operation
                return values1;
        }
        Set<Serializable> retrieve(key){
                hops=0; //initialized the hops counter in finger table
                direction
                        whild(!retrieved){
                        Peer responsiblePeer=null;
                        responsiblePeer = findSuccessor(id);
                        hops+=1; //while not retrieve the desired key, add
the hop counter by 1
        try{
```

```
         result = responsiblePeer.retrieveEntries(id);
      //get the responsiblePeer to fetch the entry
            retrieved = true; //if successfully get the value,
      set retrieved state to true
            }catch(Exception e){}
            continue;
            }
      }
      if(result !=null) values.add(entry.getValue());
// add the lookup result to the valueset
       final_hops =hops; //get the hop counter for the current lookup operation
      return values;
}
```

In our Chord variant, the pseudo code sets a hop counter for lookup operation in both clockwise direction and inverse-clockwise direction. While retrieving a key the system returns a hop value in either search routine.

3 Simulation Results

Our P2P file sharing strategy was developed to conduct a thorough performance analysis. Consequently, an experimental model that consists of a desired number of peers and the connection between these peers established under the proposed Chord strategy. The given input file to a specific peer is partitioned into chunks. These chunks are stored in the connected eligible peers. In regards to the use of this file by the parent peer, one can track the peers and collects the chunks. This process of collecting chunks is done through the DDF-Chord. The screen casts of the devised simulation model are shown in Figs. 3 and 4.

Within the testing phase of the protocol, we simulated communities with 10, 20, 30,…80, 90 and 100 nodes respectively. In each community dimension, we report the number of nodes, a fair number of entries in hand table, a fair variety of records of reverse directional finger table, the lookup hops in every single direction as well as the total variety of entries. The performance of the DDF-Chord towards lookup ratio and time taken for lookups is explored in given graphs, which is compared with Chord protocol as shown in Figs. 5 and 6. In particular, Fig. 5 exhibits less look-up time is required for DDF-Chord compared with Chord.

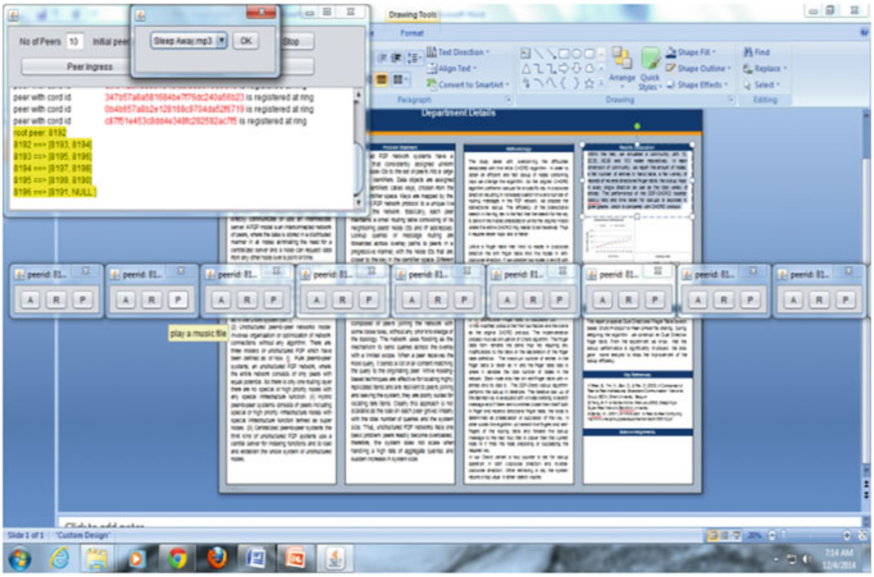


Fig. 3 The dialog that appears upon clicking the button with option ‘R’ that represents the removal of a file from source peer

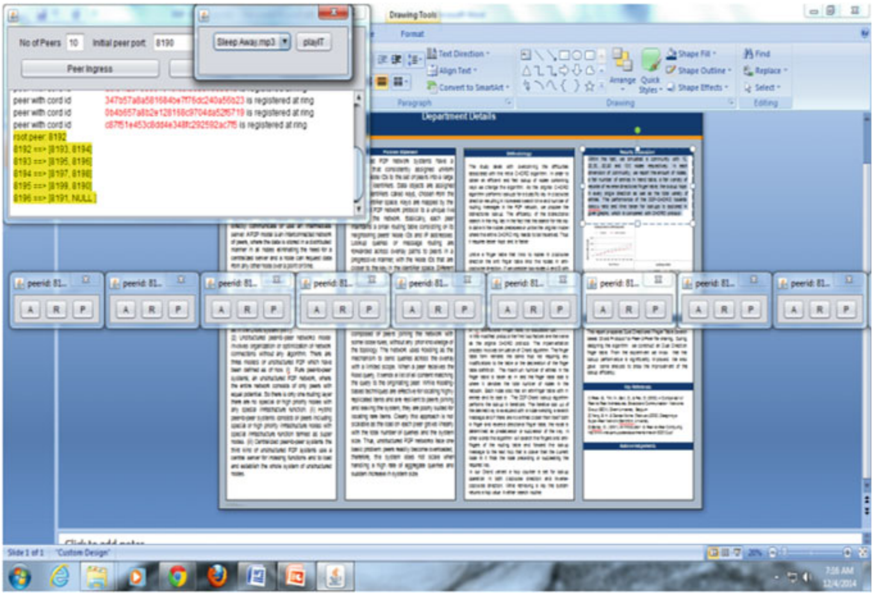


Fig. 4 The dialog opened against to the button ‘P’ clicked, which allows play the selected music file

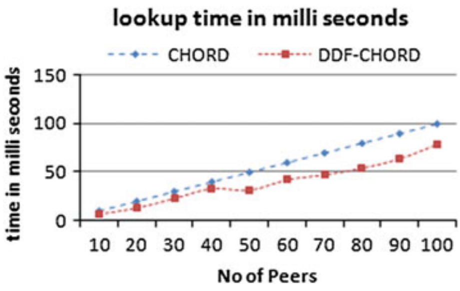


Fig. 5 The lookup time taken by CHORD and DDF-CHORD

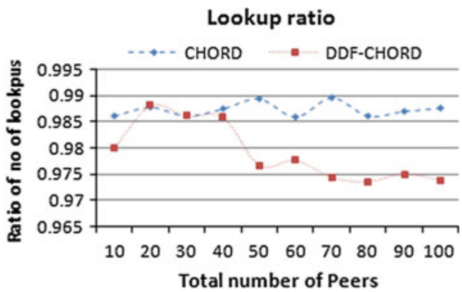


Fig. 6 The lookup ratio comparison between Chord and DDF-CHORD

4 Conclusion

This article proposed a Dual Directional Finger Table Search based Chord Protocol for Peer-2-Peer file sharing. In the design of the algorithm, we constructed a Dual Direction finger table. From the experimental results, we learned that the lookup performance is significantly improved. We also presented some analysis to show the improvement of the lookup efficiency.

References

1. Mir, N.F.: Computer and Communication Networks, 2nd edn. Pearson Prentice Hall, Upper Saddle River (2015). ISBN: 0133814742
2. Peter, B., Tim, W., Bart, D., Piet, D.: A Comparison of Peer-to-Peer Architectures, Broadband Communication Networks Group (IBCN). Department of Information Technology (INTEC), Ghent University, Belgium, pp. 1–2 (2012)
3. Minar, N.: Distributed Systems Topologies: Part 2, O'Reilly Network. http://www.openp2p.com/pub/a/p2p/2002/01/08/p2ptopologiespt2.html (2002)
4. Yang, B.H. & Garcia-Moline (February 2002), Designing a Super-Peer Network, Stanford University
5. Shirky, C.: In: Oram, A. (ed.) Listening to Napster, Peer-to-Peer: Harnessing the Power of Disruptive Technologies. O'Reilly Press, USA (2001)
6. Tyson, J.: Marshall Brain's HowStuffWorks. How Napster Worked. http://www.howstuffworks.com/napster1.Htm (2000)

7. Hebrank, M.: Gnutella & Napster. HUH? or What do I need to Know to Keep from Looking Like an Idiot. http://www.wacked.org/~heller/gnutella/
8. Turcan, E.: Peer-to-Peer: The Third Generation Internet. http://csdl.computer.org/comp/proceedings/p2p/2001/1503/00/15030095.pdf (2010)
9. LimeWire: The Gnutella Protocol Specification v0.4, Document Revision 1.2. http://www.clip2.com
10. Kan, G.: In: Oram, A. (ed.) Gnutella, Peer-to-Peer: Harnessing the Power of Disruptive Technologies. O'Reilly Press, USA (2008)
11. Dimitri, D., Antonio, G., Bill, K.: Analysis of Peer-to-Peer Network Security using Gnutella. http://www.cs.berkeley.edu/~daw/teaching/cs261-f02/reports/defig.pdf (2003)
12. Guilfoyle, J., Lempsink, E.: giFT's Interface Protocol, Source Forge. http://gift.sourceforge.net/docs.php?document=interface.html (2003)
13. Stoicay, I. et al.: Chord: A Scalable Peer-to-peer Lookup Protocol for Internet Applications. Transactions on Networking, vol. 11(1) (2003)

An Overview of Retinal Blood Vessels Segmentation

Fatimatufaridah Jusoh, Habibollah Haron, Roliana Ibrahim and Mohd Zulfaezal Che Azemin

Abstract In the recent past, the application of image processing in the fields of medicine and ophthalmology was widely used. Retina blood vessels are the only part of the human body that can be directly visualized non-invasively in vivo. Retina segmentation is important to help ophthalmologists detect various eyes diseases such as diabetic retinopathy, glaucoma, and age macular degeneration. Consequently, vessel segmentation is an important step in image analysis used to assess retinal abnormality. Vessel segmentation must be completed accurately to obtain good results for further image analysis. This paper reviews the algorithms used in previous studies on retinal vessel segmentation and discusses the problems associated with retina analysis.

1 Introduction

Medical imaging rose to prominence due to the advancement of computer and image technology [1, 2]. In early the 1960s, optical imaging played a significant role in clinical medicine. Developing technology in optical imaging was crucial to

F. Jusoh · H. Haron (✉) · R. Ibrahim
Faculty of Computing, University of Technology Malaysia,
81310 Skudai, Johor, Malaysia
e-mail: habib@utm.my

F. Jusoh
e-mail: efaridah88@gmail.com

R. Ibrahim
e-mail: roliana@utm.my

M.Z. Che Azemin
Kulliyyah of Allied Health Sciences, International Islamic University Malaysia,
Jalan Sultan Ahmad Shah, 25200 Kuantan, Pahang, Malaysia
e-mail: zulfaezal@iium.edu.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_6

advances in optics, data acquisition, and image processing [3]. In 1891, German ophthalmologist, Gerloff was the first to discover photographic retina images that showed blood vessels. The fundus camera was first developed by Gullstrand in 1910 [4].

Herbert and Michael [5] stated that the development of optical imaging and increasing research in retinal image analysis might be due to the needs of clinical practice to find better and cheaper ways of identifying, managing, and treating retinal disease. Interest and developments in this field may be the result of a desire on the part of clinical practitioners and the research community for a better understanding of the causes of retinal disease and disease progression that require detailed analysis.

Image segmentation of the morphological features of retinal blood vessel can be used in diagnosis, screening and for treatment purposes. The extracted features of blood vessel such as width, tortuosity, and bifurcation are used in image analysis to detect various eye diseases that may cause blindness [6–8]. Image analysis of retina vessels such as changes in diameter is used to diagnose hypertension, whereas, bifurcation angles, and tortuosity can assist in the detection of cardiovascular diseases and diabetic retinopathy [6, 7, 9]. Accurate detection of retinal vascular is crucial and can be a valuable aid for diagnosing ophthalmologic diseases [8, 10, 11]. In addition, retinal vascular has a unique structure for each person and blood vessel can be visualize non-invasive directly in vivo [6–8, 10, 12, 13]. Thus, the extracted retinal blood vessel is not only useful for diagnosis purposes, but also in the registration of retinal images and for biometrics applications [8, 9, 14].

Vascular networks are traditionally mapped by hand, which is a time-consuming process that requires both training and skill [13, 15]. The variety of vascular networks that have been recorded in retinal images and the irregularity of the image acquisition process makes retinal blood vessel segmentation a challenging task [16]. The segmentation of blood vessels is an important preprocessing step for the early detection of retinal diseases [13, 15]. According to [10], the measurements of blood vessel features are open to user-bias as they are obtained using semi-automated methods. Automated vessel segmentation algorithms have been introduced by researchers. These algorithms are categorized as being supervised or unsupervised methods [17]. Figure 1 provides an example of retinal images before and after the blood vessels was segmented.

This paper will use information from publicly available datasets for retina images (Sect. 2) and previous studies that examined blood vessel segmentation algorithms (Sect. 3). Further discussion is provided in Sect. 4 and the conclusion is covered in Sect. 5.

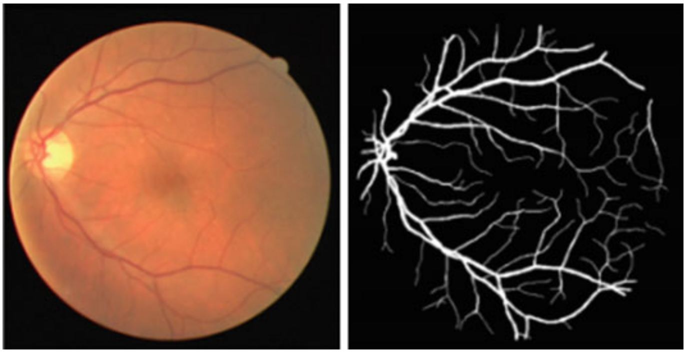


Fig. 1 Retinal image: before and after vessel segmentation

2 Retinal Image Dataset

Most of the retinal vessel segmentation methodologies were evaluated using two public databases known as DRIVE and STARE. Using the same retinal image dataset make it easier to compare results with the results of other studies. However, there are several others public database that contain retinal image such as the MESSIDOR database.

The photographs in the DRIVE database [18] were obtained from a diabetic retinopathy screening program in the Netherlands. Each image is a compressed JPEG. The DRIVE database was established to enable comparative studies on segmentation of blood vessels in retinal images. The retinal images were captured using 8 bits per color plane at 768 by 584 pixels and collected using a Canon CR5 non-mydriatic 3CCD camera with a 45 degree field of view (FOV). The FOV of each image is circular with a diameter of approximately 540 pixels. Retinal images in the database have been cropped around the FOV. For each image, a mask image was provided that delineated the FOV.

The STARE database [19] contains 20 images for blood vessel segmentation; ten of these images contain pathology. The images were acquired using a TopCon TRV-50 fundus camera with a 35 degree field of view and the images captured were 8 bits per color plane by 605 × 700 pixels. The images were manually segmented by two observers. The first observer segmented 10.4 % of the pixels as vessels and the second observer segmented 14.9 % of the pixels as vessels. This indicates that the second observer may have segmented more of the thinner vessels.

The MESSIDOR database [20] is the largest database and it contains 1200 retinal images acquired by three ophthalmologic departments using a color video 3CCD camera on a Topcon TRC NW6 non-mydriatic retinograph with a 45 degree field of view. The images were captured using 8 bits per color plane at 1440 × 960, 2240 × 1488, or 2304 × 1536 pixels. Eight hundred images were acquired with pupil dilation (one drop of Tropicamide at 0.5 %) and 400 without dilation.

3 Previous Studies on Retinal Segmentation Algorithm

Many different approaches for automated vessel segmentation have been proposed. Vessel tracking, matched filter responses, and morphological processing are categorized as unsupervised methods. Whereas, segmentation based on pixel classification is categorized as a supervised method. Supervised methods require a feature vector for each pixel and manually labelled images to train the algorithm [10, 14, 17, 21]. According to Patton et al. [7], four techniques that are commonly used in vascular segmentation are matched filter, vessel tracking, morphological processing, and neural networks.

Matched filter involves filtering the image using kernel and Gaussian functions to extract the linear segment of retinal blood vessels [7, 10, 13, 22]. Many improvements have been made on matched filtering approaches, including using global or local thresholding strategies [19, 23, 24]. The vessel tracking method is based on the continuity property of blood vessels that tracks each vessel from the initial point. According to [25], the vessel tracking approach can identify each vessel and provide more significant vessel structures than the pixel-based methods. Morphological methods have two main processes, namely dilation and erosion. In image processing, morphological methods are appropriate for analyzing shapes in images that use a priori feature of the vascular shape such as linear continuity. When segmenting blood vessels, the algorithms that extract linear shapes are useful [7]. Kee [26] stated that statistical parameters for each pattern class are estimated using a sample pattern in the neural network method. Neural networks consist of layers and units. Units are also known as neurons, which take an input and pass the output to the next layer. Neural networks are commonly designed to be fed-forward, in which a unit gives the output to all the units in the next layer, and the previous layer does not receive any feedback from next layer [27].

Based on previous studies on retinal blood vessel segmentation (Table 1), various techniques have been introduced for the efficient detection of retinal vessel that lead to highly accuracy results. These proposed algorithms were tested using the public retinal datasets DRIVE and STARE. Manoj et al. [27] also used the MESSIDOR dataset in addition to DRIVE and STARE.

Fraz et al. [28] proposed a supervised method for segmenting blood vessels using an ensemble classifier of bagged decision tree that used nine dimensional feature vectors based on gradient orientation analysis, morphological linear transformation, line strength measure, and Gabor filter response. The approach proposed by Fraz was fast and required fewer samples in the training phase.

A method proposed by Yin et al. [29] was an automatic tracking method based statistical Bayesian. The tracking process started with seed points and detected vessel edge points iteratively using the proposed Bayesian method.

Asad et al. [15] proposed an approach for retinal vessel segmentation using a stand-alone bio-inspired algorithm, Ant Colony System (ACS) without coalition with others method. Eight features were used in this study; however, the results of Asad's proposed method did not reveal the best segmentation approach. The performance of

Table 1 Previous studies on retinal blood vessel segmentation

Technique	Accuracy result		
Ensemble classifier (decision tree and bootstrap) [28]	DRIVE: 0.9480; STARE: 0.9534		
Tracking method based statistical Bayesian method [29]	STARE: 0.9290		
Ant colony system [15]	DRIVE: 0.9028		
Self-organizing map, K-means clustering [14]	DRIVE: 0.9459		
Neural network [30]	DRIVE: 0.9503		
Neural network [27]:	DRIVE:	STARE:	MESSIDOR:
(a) Feed forward back propagation neural network (FFBNN)	0.9623	0.9583	0.9541
(b) Radial basis function (RBF)	0.8995	0.8929	0.9061
(c) Multi-layer perceptron (MLP)	0.9007	0.8485	0.8276
Matched filtering [31]	DRIVE: 0.9340; STARE: 0.9341		
Matched filtering [32]	DRIVE: 0.9461; STARE: 0.9521		
Multi-scale line-tracking [33]	DRIVE: 0.9285		
2-D gabor wavelet and sharpening filters [34]	STARE: 0.9439		

the proposed approach was also considered as it depended on simple and fast computed features.

An automatic unsupervised method was introduced by Lupaşcu and Tegolo [14] and it was based on Self-Organizing Maps (SOM) and K-means clustering. SOM was trained using images that were divided equally between the training and testing phases. The map units were divided into two classes by K-means clustering. Similar to the most accurate supervised methods, the proposed unsupervised method did not have any a priori knowledge in pixel labels because it relied on knowledge of vessel network morphology. Compared to supervised methods, which are computationally more expensive, the proposed method is computationally fast.

Franklin and Rajan [30] used a back propagation algorithm in the neural network for vessel segmentation. Back propagation is a systematic method for training multi-layer artificial neural networks. This method could assure efficiency in the classification phase and resulted in excellent performance, even though it needed a training phase.

A supervised method proposed by Manoj et al. [27] employed nine dimensional feature vector based on the orientation analysis of a gradient vector field, morphological transformation, line strength measures, and Gabor filter responses. The pixels were classified using three neural network classifier techniques, which were a Feed Forward Back propagation Neural Network (FFBNN), Radial Basis Function (RBF) and Multi-Layer Perceptron (MLP). The effectiveness and its speed of classification made the method proposed by Manoj et al. appropriate for retina image analysis.

The approach proposed by Odstrcilik et al. [31] utilized Matched Filtering and minimum error thresholding technique to extract a binary blood vessel tree. The method was improved to segment blood vessel with variations of vessel diameter.

By considering five width classes of retinal vessels, five different kernels were designed according to regular blood vessel cross-sectional profiles.

Wang et al. [32] proposed a novel vessel enhancement technique based on a matched filter with multiwavelet kernels. This method generated very competitive results and did not require a training phase. No thresholds were manually adjusted as the system depended on adaptive thresholding.

An algorithm based on a novel seeded multi-scale line-tracking procedure and morphological post-processing was proposed by Vlachos and Dermatas [33]. The tracking process started with an estimation of the confidence matrix obtained from a group of seeds that were extracted from an image's histogram. This tracking process terminated when a specific condition of cross sectional profile became invalid.

Thin and less visible vessel patterns were enhanced using 2-D Gabor wavelets in a study conducted by Akram et al. [34]. Blood vessels were sharpened using a sharpening filter before the vessel was extracted. Vessels edges were detected in sharpened images and the vessels segmentation binary mask was created by assigning a value of one to the pixels that belong to blood vessels and a value of zero for non-vessels pixels. Lastly, morphological dilation was applied to refine the vessel segmentation mask.

4 Discussion

Vessel tracking is efficient as it analyses a smaller number of pixels than the image dimension [35]. Patton et al. [7] stated that vessel tracking can provide very accurate measurements of vessel widths and tortuosity, but it may be confuse by the vessel crossings and bifurcations. When the contrast between the vessels and background is weak, tracking may be terminated. Human intervention is needed when the automatic detection for initial points is unavailable [25]. Morphological processing methods that use a priori features of the vasculature shape have the advantage of speed and noise resistance [28]. Neural network methods are efficient but their accuracy depends on the quality of the image [26]. However, with the accuracy for predictive inference, neural network methods have a potential to support clinical decision-making [27].

Supervised schemes result in better segmentation, however, they require a training procedure that depends on hand-labeled vessel segmentation, which is very time consuming. By contrast, unsupervised methods do not require a training set and they are usually faster and better suited to automatic vessel segmentation [10, 25]. Nonetheless, Bankhead et al. [10] stated that unsupervised methods require slight modifications according to image properties such as quality, type, and size, before they can be used. Unsupervised methods require these modifications because they often use a filter and operations that are only suitable for particular types of images. However, perfect segmentation is impossible if the quality of the image is poor or images with pathologies are analyzed, regardless all the algorithms [16, 35].

5 Conclusion

This paper provides a review on retinal vessel segmentation. Vessel segmentation is completed before further image analysis to detect any retinal abnormalities. Therefore, developing an efficient and fast algorithm that can achieve a high level of accuracy is important. The DRIVE and STARE datasets were used to evaluate the proposed algorithm. Retinal blood vessel segmentation can be completed using either a supervised or unsupervised method. Supervised method required training images that need difficult manual segmentation but they provide better segmentation results. By contrast, unsupervised methods are faster and do not require a training image. Future work should focus on improving the literature review regarding the details of the methods used in vessel segmentation and further analysis of retina vessels, such as changes in diameter, bifurcation angles, and tortuosity, should be included.

Acknowledgments This material is based upon work supported by Fundamental Research Grant Scheme (FRGS), under Vote No. R.J130000.7828.4F537 and Ministry of Higher Education (MOHE). Any opinions, findings, and conclusions or recommendations expressed in this material are those from the authors and do not necessarily reflect the views of the Universiti Teknologi Malaysia.

References

1. Ganguly, D., Chakraborty, S., Balitanas, M., Tai-hoon, K.: Medical imaging: a review. Security-enriched urban computing and smart grid. Commun. Comput. Inf. Sci. **78**, 504–516 (2010)
2. Deserno, T. M.: Fundamentals of Biomedical Image Processing. Biomedical Image Processing. Biological and Medical Physics, Biomedical Engineering. Springer, Berlin (2011)
3. Iftimia, N., Hammer, D.X., Brugge, W.R.: Introduction To Optical Imaging In Clinical Medicine. John Wiley & Sons, Inc., Hoboken (2011)
4. Abràmoff, M.D., Garvin, M.K., Sonka, M.: Retinal Imaging and image analysis. IEEE Trans. Med. Imaging. **3:**169–208 (2010)
5. Herbert, F.J., Michael, J.C.: Automated Image Detection of Retinal Pathology. Taylor and Francis Group, United States (2010)
6. Hooshyar, S., Khayati, R.: Retina vessel detection using fuzzy ant colony algorithm. In: Canadian Conference Computer and Robot Vision. IEEE (2010)
7. Patton, N., Aslam, T.M., MacGillivray, T., Deary, I.J., Dhillon, B., Eikelboom, R.H., Yogesan, K., Constable, I.J.: Retinal image analysis: Concepts, applications and potential. Prog. Retinal Eye Res. **25**(1), 99–127 (2006)
8. Fraz, M.M., Remagnino, P., Hoppe, A., Uyyanonvara, B., Rudnicka, A.R., Owen, C.G., Barman, S.A.: Blood vessel segmentation methodologies in retinal images—a survey. Comput. Methods Programs Biomed. **108**(1), 407–433 (2012)
9. Bhuiyan, A., Nath, B., Chua, J., Ramamohanarao, K.: Automatic detection of vascular bifurcations and crossovers from color retinal fundus images. In: Third International IEEE Conference on Signal-Image Technologies and Internet-Based System, pp. 711–718 (2007)
10. Bankhead, P., Scholfield, C.N., McGeown, J.G., Curtis, T.M.: Fast retinal vessel detection and measurement using wavelets and edge location refinement. PLoS ONE **7**(3), e32435 (2012)

11. Jegatha, R., Lakshmi, K.: Retinal blood vessel segmentation using gray-level and moment invariants-based features. J. Comput. Appl. **5**(EICA2012-3):271 (2012)
12. Che Azemin, M.Z., Kumar, D.K.: Estimating retinal vessel diameter change from the vessel cross-section. In: IFMBE Proceedings 5th Kuala Lumpur International Conference on Biomedical Engineering, vol. 35, pp. 655–658 (2011)
13. Garhöfer, G., Vilser, W.: Measurement of Retinal Vessel Diameters. Ocular Blood Flow, pp. 101–122. Springer, Berlin (2012)
14. Lupaşcu, C.A., Tegolo, D.: Automatic unsupervised segmentation of retinal vessels using self-organizing maps and k-means clustering. Computational intelligence methods for bioinformatics and biostatistics. Lecture Notes in Computer Science, vol. 6685, pp. 263–274 (2011)
15. Asad, A.H., Azar, A.T., Hassaanien, A.E.: Ant colony-based system for retinal blood vessels segmentation. In: Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012). Advances in Intelligent Systems and Computing, vol. 201, pp. 441–452 (2013)
16. Honale, S.S., Kapse, V.S.: A review of methods for blood vessel segmentation in retinal images. Int. J. Eng. Res. Technol. (IJERT). **1**(10):1 (2012)
17. Raja, J.B., Ravichandran, C.G.: blood vessel segmentation for high resolution retinal images. IJCSI Int. J. Comput. Sci. Issues. **8**(6):2 (2011)
18. Staal, J.J., Abramoff, M.D., Niemeijer, M., Viergever, M.A., Ginneken, B.V.: Ridge based vessel segmentation in color images of the retina. IEEE Trans. Med. Imaging **23**, 501–509 (2004)
19. Hoover, A., Kouznetsova, V., Goldbaum, M.: Locating blood vessels in retinal images by piece-wise threhsold probing of a matched filter response. IEEE Trans. Med. Imaging **19**(3), 203–210 (2000)
20. MESSIDOR: Methods for Evaluating Segmentation and Indexing techniques Dedicated to Retinal Ophthalmology. http://messidor.crihan.fr/index-en.php (2004)
21. Marín, D., Aquino, A., Gegúndez-Arias, M.E., Bravo, J.M.: A new supervised method for blood vessel segmentation in retinal images by using gray-level and moment invariants-based features. IEEE Trans. Med. Imaging **30**(1), 146–158 (2011)
22. Ardizzone, E., Pirrone, R., Gambino, O., Radosta, S.: Blood vessels and feature points detection on retinal images. In: 30th Annual International IEEE EMBS Conference Vancouver, British Columbia, Canada, pp. 2246–2249 (2008)
23. Malek, J., Tourki, R.: Blood vessels extraction and classification into arteries and veins in retinal images. In: 10th International Multi-Conference on Systems, Signals and Devices (SSD) Hammamet, Tunisia (2013)
24. Chanwimaluang, T., Fan, G.: An efficient algorithm for extraction of anatomical structures in retinal images. In Proceedings ICIP, pp. 1193–1196 (2003)
25. Li, H., Zhang, J., Nie, Q., Cheng, L.: A retinal vessel tracking method based on bayesian theory. In: 8th IEEE Conference on Industrial Electronics and Applications (ICIEA), pp. 232–235 (2013)
26. Kee, Y.P., Lila Iznita, I., Ahmad Fadzil, M.H., Hanung, A.N., Hermawan, N., Vijanth, S.A.: Segmentation of retinal vasculature in colour fundus images. In: Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA 2009). Monash University, Sunway campus, Malaysia (2009)
27. Manoj, S., Muralidharan S.P.M.: Neural network based classifier for retinal blood vessel segmentation. Int. J. Recent Trends Electr Electron Eng. **3**:44 (2013)
28. Fraz, M.M., Remagnino, P., Hoppe, A., Uyyanonvara, B., Rudnicka, A., Owen, C.G., Barman, V.: Retinal vessel segmentation using ensemble classifier of bagged decision trees. In: IET Conference on Image Processing (IPR 2012), pp. 1–6 (2012)
29. Yin, Y., Adel, M., Bourennane, S.: An automatic tracking method for retinal vascular tree extraction. In: Acoustics, IEEE International Conference on Speech and Signal Processing (ICASSP), pp. 709–712 (2012)

30. Franklin, S.W., Rajan, S.E.: Computerized screening of diabetic retinopathy employing blood vessel segmentation in retinal images. Biocybernetics Biomed. Eng. **34**(2), 117–124 (2014)
31. Odstrcilik, J., Kolar, R., Budai, A., Hornegger, J., Jan, J., Gazarek, J., Kubena, T., Cernosek, P., Svoboda, O., Angelopoulou, E.: Retinal vessel segmentation by improved matched filtering: evaluation on a new high-resolution fundus image database. IET Image Process. **7**(4):373–383 (2013)
32. Wang, Y.F., Ji, G., Lin, P., Trucco, E.: Retinal vessel segmentation using multiwavelet kernels and multiscale hierarchical decomposition. Pattern Recogn. **46**, 2117–2133 (2013)
33. Vlachos, M., Dermatas, E.: Multi-scale retinal vessel segmentation using line tracking. Comput. Med. Imaging Graph. **34**, 213–227 (2010)
34. Akram, M.U., Atzaz, A., Aneeque, S.F., Khan, S.A.: Blood vessel enhancement and segmentation using wavelet transform. In: International Conference on Digital Image Processing, pp. 34–38 (2009)
35. Tramontan, L., Ruggeri, A.: Automatic refinement of vascular tracking in retinal images: false vessels detection. In: 25th International Symposium on Computer-Based Medical Systems (CBMS), pp. 1–6 (2012)

Proposed Approach for Targeted Attacks Detection

Ibrahim Ghafir and Vaclav Prenosil

Abstract For years governments, organizations and companies have made great efforts to keep hackers, malware, cyber attacks at bay with different degrees of success. On the other hand, cyber criminals and miscreants produced more advanced techniques to compromise Internet infrastructure. Targeted attack or advanced persistent threat (APT) attack is a new challenge and aims to accomplish a specific goal, most often espionage. APTs are presently the biggest threat to governments and organizations. This paper states research questions and propose a novel approach to intrusion detection system processes network traffic and able to detect potential APT attack. This detection of APT attack is based on the correlation between the events which we get as outputs of our detection methods. Each detection method aims to detect one technique used in one of APT attack steps.

Keywords Cyber attacks · Targeted attacks · Advanced persistent threat · Malware · Intrusion detection system

1 Introduction

Nowadays the cost of cyber attacks, malicious activities on Internet infrastructures, is estimated to somewhere between 100 billion and 1 trillion US dollars annually around the world [1]. Targeted attack or advanced persistent threat (APT) attack is a new challenge and aims to accomplish a specific goal, most often espionage. APTs are presently the biggest threat to governments and organizations [2]. These APTs create a problem for existent detection methods because the current techniques are based on known signatures of attacks and APTs often use new security holes for

I. Ghafir (✉) · V. Prenosil
Faculty of Informatics, Masaryk University, 60200 Brno, Czech Republic
e-mail: ibrahim_ghafir@hotmail.com

V. Prenosil
e-mail: prenosil@fi.muni.cz

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_7

attacks. The financial losses due to a successful APT attack can be very high as it is confirmed through many previous research findings on APTs [3–5]. The expected economic effect of attacks is the major influence on investments in security measures [6].

In this paper we state our research questions and present our proposed approach for APT detection. In this research we are aiming to contribute to intrusion detection systems, particularly to APT attack detection. The goal of this work is to research a novel approach of intrusion detection processes network traffic and able to detect APT attack. The detection of APT attack is based on the correlation between the events which we get as outputs of our detection methods. We believe that the opportunity for using this approach in APTs detection is big and, to the best of our knowledge, still unexplored.

The remainder of this paper is organized as follows. State of the art is described in Sect. 2. We state our research questions in Sect. 3. Section 4 presents our proposed approach for APT attack detection and Sect. 5 concludes the paper.

2 State of the Art

In January 2011, Google published the first report about APT. Known as Operation Aurora [3]; the attack has begun in the second half of 2009. It was considerably large-scale and is mentioned to have targeted 34 companies, including Dow Chemical, Morgan Stanley, Northrop Grumman, Symantec, Northrop Grumman and Yahoo, as well as Google itself. Later on, other research findings on APT attack are reported in [4, 5, 7].

Some researches have been done on analyzing already identified of APTs. In [8], the authors showed how to detect the N most likely infected hosts of the attack; their approach is based on the knowledge of previous APT attacks. By improving the performance in terms of false positives and detection rate, they developed a search engine for APT investigators to quickly reveal the possible infected hosts based on the features of a known APT infected host. They made use of N-gram based mechanisms.

During the year 2011, many targeted attacks were identified by Symantec. This large corpus of APTs was analyzed in-depth by the authors in [9]. Based on advanced TRIAGE data analytics, they were able to attribute many of advanced persistent threats to attack campaigns quite likely accomplished by the same individuals. They analyzed the dynamics and features of those attacks and presented new ideas into the modus operandi of attackers engaged in those campaigns.

By using an undirected graph in [10], the authors showed that APTs against the same target could be correlated. In addition, it is possible to identify clusters and create a map of APT activity that may uncover the activities of single group of malware writers.

With regards to detect potential targeted attack, in [11] they proposed a novel system to detect possible APT attack based on the information gathered on the host's side. The system depends on clustering techniques to classify groups of hosts that have a similar behavior with respect to the suspicious resources they request (e.g., C&C servers, drive-by downloads or exploit kits). The system was called SPuNge and implemented in a working prototype. SPuNge correlates the sites and industry data in which those hosts run (e.g., government or gas and oil) to detect interesting attack activities.

An abridged version of initial Duqu analysis was presented in [12]. A European corporation was targeted by a new malware, Duqu, and valuable information was stolen. The authors described the Duqu detector toolkit, a set of heuristic tools that they developed to detect Duqu and its variants.

Given the related works presented above, most of the above works focus on analyzing already identified campaigns, while the scope of our work is to present a new approach that detect possible APT attacks, none of the related works address explicitly the problem of detecting potential APT attacks by means of monitoring network traffic and correlation between detection methods of possible techniques used in APT attack life cycle.

3 Research Questions

To achieve the goal of this work we should answer the following research questions:

Research question 1: What are the detection methods can be used for detecting possible techniques used through APT attack life cycle? To answer this question we have proposed 8 detection methods presented in Sect. 4. We will try to implement these methods in the first phase of our research. These proposed methods are not fixed; we can remove or suggest a new method based on the research progress.

Research question 2: How can we make the detection system resulting from our approach extensible and flexible? The attackers always try to find new techniques to perform APT attack, therefore each detection method should be independent from the other methods, so at any time we can add new method (for detecting new technique used in APT attack life cycle) to the system and correlate it with the other methods in the correlation framework. To achieve the flexibility, in the correlation framework it should be easy to remove or add a new rule for raising an alert on APT attack detection.

Research question 3: Is this approach able to handle the network traffic in the real-time? The detection system should support the real-time detection because if an attack, or an attempted attack, is detected quickly, then it can be much easier to trace back the attacker, minimize the damage and prevent further break-ins. To answer this question, in our approach and in the first phase, the detection methods should not depend on storing data and then analyzing it for detection. They should be able to process the network traffic in the real-time and submit their events to the next phase for correlation.

Research question 4: Is this approach effective? The effectiveness of the approach, which is its ability to detect APT attacks, should be high. This should be combined with a high accuracy resulting into a low number of false warnings. We expect that the chance at a false positive is lower when there is a direct link to other steps or correlation between the events. In order to achieve efficiency for our approach we should identify suitable rules for correlation between the events and this will depend on the evaluation of each detection method and will be done in the last phase of our research.

4 Proposed Approach

In this research we are aiming to contribute to intrusion detection systems, particularly to APT attack detection. The goal of this work is to research a novel approach of intrusion detection processes network traffic and able to detect APT attack. The detection of APT attack is based on the correlation between the events which we get as outputs of our detection methods. We believe that the opportunity for using this approach in APTs detection is big and, to the best of our knowledge, still unexplored.

We will implement our proposed approach on top of Bro intrusion detection system [13, 14]. Bro is a passive, open-source network traffic analyzer. It is primarily a security monitor that inspects all traffic on a link in depth for signs of suspicious activity. The most immediate benefit that we gain from deploying Bro is an extensive set of *log files* that record a network's activity in high-level terms. These logs include not only a comprehensive record of every connection seen on the wire, but also application-layer transcripts such as, e.g., all HTTP sessions with their requested URIs, key headers, MIME types, and server responses; DNS requests with replies; and much more.

Our proposed approach, as it is shown in Fig. 1, consists of two main phases:

In the first phase, we process the network traffic to detect possible techniques used in APT attack life cycle. To this end we have our detection methods; each

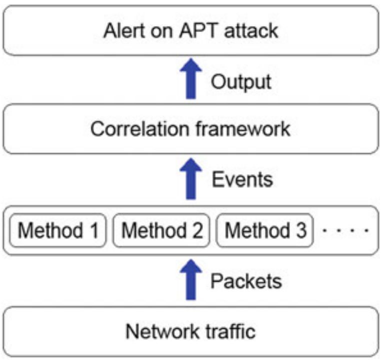


Fig. 1 Architecture of our proposed approach

detection method aims to detect one technique used in one of APT attack steps. Each detection method is independent from the other methods and should be evaluated before it is adopted. The outputs of these detection methods should be submitted to the second phase where correlated to raise an alert on APT attack.

In the second phase, we have the correlation framework, this framework takes the events (the outputs of our detection methods) as an input and correlates them according to rules specified by the user (we can specify those rules based on the evaluation of each method) to raise an alert on APT attack detection. The correlation method is based on voting between the detection methods to raise an alert on APT attack and the detection can be based on one, two or three events. We believe that this correlation will reduce the false positive rate of our detection system.

A key question in our research is: What are the detection methods can be used for detecting possible techniques used through APT attack life cycle. To answer this question taking into consideration the life cycle of APT attack, which is shown in Fig. 2 [15], we have 8 detection methods should be implemented in the first phase of our research:

1. Intelligence gathering: This initial phase aims to get information about the target, like its organizational structure, IT environment and even about people who are working for that target. For this purpose, the attacker can use public sources (LinkedIn, Facebook, etc.) and prepare a customized attack.
2. Initial compromise (Point of entry): Performed by use of social engineering and spear phishing, over email, using zero-day exploits. Another popular infection method was planting malware on a website that the victim employees will be likely to visit. The most common technique used for this step is spear phishing emails, which may contain link to malicious website, malicious attachment or link to malicious file.

Fig. 2 Typical steps of APT attack

Method 1, from previous findings on APTs, we have a list of exploited domain names (FQDNS) used in APT attack [5], we can analyze the traffic of possible protocols used in sending and retrieving emails and if there is a link to one of exploited FQDNS, we can detect spear phishing attack.
Method 2, detection of any connection to a malicious domains based on a blacklist of malicious domains [16–21].
In case of malicious attachment, executable files supposed to end in.exe are made to appear as simple document files (pdf, doc, ppt, excel).
Method 3, detecting if the content of the file is exe while the extension is not exe.
Method 4, for each new file, we can calculate MD5, SHA1 and SHA256 hash, and compare with a blacklist of file hashes (from previous findings on APTs) [5], if a match is found, we detect an attack.

3. Command and control (C&C) communication: After an organization's perimeter has been breached, continuous communication between the infected host and the C&C server should be preserved to instruct and guide the compromised machine. These communications are usually protected by Secure Sockets Layer (SSL) encryption, making it difficult to identify if the traffic directed to sites is malicious.
   Method 5, we have a blacklist of SSL certificates (from previous reports) [22, 23], so we can monitor SSL certificates and match with the blacklist, if a match is found, we detect an attack.
   Method 6, we have a blacklist of C&C servers [24–27], any connection to one of those servers is an attack.
   Another technique can be used in this step is domain flux technique [28]; an exploited host may try to connect to a large number of domain names which are expected to be C&C servers. The goal of this technique is to make it difficult or even impossible to shut down all of these domain names. This technique leads to many of DNS query failures because not all of these domains are registered.
   Method 7, domain flux detection based on DNS query failure.
4. Lateral movement: Once getting an access to the target's network, the attacker laterally moves throughout the target's network searching for new hosts to infect. Some techniques used: Brute force and pass the hash attacks.
5. Asset/Data discovery: This step aims to identify and isolate the noteworthy assets within the target's network for future data exfiltration.
   Since the traffic of the steps 4 and 5 are inside the compromised network, we cannot see it.
6. Data exfiltration: Data of interest is transmitted into external servers which are controlled by the attacker. There are some techniques used for data exfiltration like built-in file transfer, via FTP or HTTP and via the Tor anonymity network.
   Method 8, Tor connection detection based on Tor server list [29].

The blacklists of blacklist-based detection methods should be automatically updated each day and the detection by all methods should be in the real time.

5 Conclusion

In this paper we proposed a novel approach for detecting potential APT attack and presented the research questions which we should answer to achieve this goal. The detection of APT attack is based on the correlation between the detection methods of possible techniques used in APT attack life cycle. We believe that the opportunity for using this approach in APTs detection is big and, to the best of our knowledge, still unexplored.

Acknowledgments This work has been supported by the project "CYBER-2" funded by the Ministry of Defence of the Czech Republic under contract No. 1201 4 7110.

References

1. Kshetri, N.: The global cybercrime industry: economic, institutional and strategic perspectives. Springer, Berlin (2010)
2. Wood, P., Nisbet, M., Egan, G., Johnston, N., Haley, K., Krishnappa, B., Tran, T. K., Asrar, I., Cox, O., Hittel, S., et al.: Symantec Internet Security Threat Report Trends for 2011, vol. XVII (2012)
3. Tankard, C.: Advanced persistent threats and how to monitor and deter them. Netw. Secur. **2011**(8), 16–19 (2011)
4. Kaspersky Lab ZAO. Red October diplomatic cyber attacks investigation. http://www.securelist.com/en/analysis/204792262/Red_October_Diplomatic_Cyber_Attacks_Investigation. Accessed 10-11-2014
5. Mandiant Intelligence Center. Apt1: Exposing one of china's cyber espionage units. Technical report, Mandiant, Tech. Rep (2013)
6. Rakes, T. R., Deane, J. K., Rees, L. P.: It security planning under uncertainty for high-impact events. Omega **40**(1), 79–88 (2012)
7. Ronald, D., Rafal R.: Tracking ghostnet: Investigating a cyber espionage network. Inf. Warf. Monitor, p. 6 (2009)
8. Liu, S.T., Chen, Y. M., Lin, S. J.: A novel search engine to uncover potential victims for apt investigations. In: Network and Parallel Computing, pp. 405–416. Springer, Berlin (2013)
9. Thonnard, O., Bilge, L., O'Gorman, G., Kiernan, S., Lee, M.: Industrial espionage and targeted attacks: Understanding the characteristics of an escalating threat. In Research in Attacks, Intrusions, and Defenses, pp. 64–85. Springer, Berlin (2012)
10. Lee, M, Lewis, D.: Clustering disparate attacks: mapping the activities of the advanced persistent threat. In: Proceedings of the 21st Virus Bulletin International Conference, pp. 122–127 (October 2011)
11. Marco Balduzzi, Vincenzo Ciangaglini, and Robert McArdle. Targeted attacks detection with spunge (2013)
12. Bencsath, B., Pek, G., Buttyan, L., Felegyhazi, M.: Duqu: Analysis, detection, and lessons learned. In ACM European Workshop on System Security (EuroSec), vol. 2012 (2012)
13. Paxson, Vern: Bro: a system for detecting network intruders in real-time. Comput. Netw. **31**(23), 2435–2463 (1999)
14. Bro Project. The bro network security monitor. http://bro.org/. Accessed 10-11-2014
15. Trend Micro white paper. The custom defense against targeted attacks. http://www.trendmicro.fr/media/wp/custom-defense-against-targeted-attacks-whitepaper-en.pdf. Accessed: 10-11-2014

16. Blade defender. http://www.blade-defender.org/eval-lab/blade.csv. Accessed 10-11-2014
17. Malware domain list. http://www.malwaredomainlist.com/hostslist/hosts.txt. Accessed 10-11-2014
18. Malware domains. http://www.malware-domains.com/files/. Accessed 10-11-2014
19. Abuse.ch. Palevo domain blocklist. https://palevotracker.abuse.ch/blocklists.php?download=domainblocklist. Accessed 10-11-2014
20. Abuse.ch. Spyeye domain blocklist. https://spyeyetracker.abuse.ch/blocklist.php?download=domainblocklist. Accessed 10-11-2014
21. Abuse.ch. Zeus domain blocklist. https://zeustracker.abuse.ch/blocklist.php?download=domainblocklist. Accessed 10-11-2014
22. Abuse.ch. SSL blacklist a new weapon to fight malware and botnet. http://securityaffairs.co/wordpress/26672/cyber-crime/ssl-blacklist-new-weapon-fight-malware-botnet.html. Accessed 10-11-2014
23. Mandiant. Mandiant apt1 report appendix f update: SSL certificate hashes. https://www.mandiant.com/blog/md5-sha1/. Accessed 10-11-2014
24. Malware domain list. http://www.malwaredomainlist.com/hostslist/ip.txt. Accessed 10-11-2014
25. Abuse.ch. Palevo C&C ip blocklist. https://palevotracker.abuse.ch/blocklists.php?download=ipblocklist. Accessed 10-11-2014
26. Abuse.ch. Spyeye ip blocklist. https://spyeyetracker.abuse.ch/blocklist.php?download=ipblocklist. Accessed 10-11-2014
27. Abuse.ch. Zeus ip blocklist. https://zeustracker.abuse.ch/blocklist.php?download=ipblocklist. Accessed: 10-11-2014
28. Yadav, S., Reddy, A.K.K., Narasimha Reddy, A.L., Ranjan, S.: Detecting algorithmically generated domain flux attacks with DNS traffic analysis. IEEE/ACM Trans. Netw. **20**(5), 1663–1677 (2012)
29. Tor Network Status. CSV list of all current tor server ip addresses. http://torstatus.blutmagie.de/. Accessed 10-11-2014

Accelerating Turbo Similarity Searching on Multi-cores and Many-cores Platforms

Marwah Haitham Al-laila, Mohd Norhadri Hilmi and Nurul Hashimah Ahamed Hassain Malim

Abstract Turbo Similarity Searching (TSS) is a two phases searching procedure that has been proven by previous works as one of the best searching method on chemical databases. TSS however consumes lots of computation time due to the number of searches and fusion carried out in the second phase of its procedure. Hence, TSS would not be able to cater the increase in the chemical database size due to this limitation. With the emergence of the parallel technology, this research looks into accelerating TSS on the widely-used many-cores i.e. the Graphics Processing Unit (GPU) and multi-cores platform. This would not only solve the computational time issue but also the cost as GPUs can be obtained at a lower cost. Hence, the implementation of TSS will help the medicinal chemist to execute the virtual screening in an accurate and fast manner. This study investigates the best possible method to parallelize TSS via experimentation of three parallel designs; two designs were implemented on GPU platform using the Compute Unified Device Architecture (CUDA) API namely CUDA 1 and CUDA 2 whilst one design was implemented on multi-core platform using OpenMP API. The CUDA 1 design had shown tremendous speedup and low GPU-memory utilization as compared to CUDA 2 design. In general observation, the parallel CUDA 1 was 131 times faster than sequential and 51 times faster than parallel OpenMP. This leads to the conclusion that CUDA 1 design as the best parallel design for TSS.

Keywords Graphical processing unit (GPU) · Compute unified device architecture platform · Turbo similarity searching · Virtual screening · Open multiprocessing

M.H. Al-laila
University of Mosul, Mosul, Iraq
e-mail: marwa_allayla2000@yahoo.com

M.H. Al-laila · M.N. Hilmi · N.H.A.H. Malim (✉)
School of Computer Sciences, Universiti Sains Malaysia, 11800 Penang, Malaysia
e-mail: nurulhashimah@usm.my

M.N. Hilmi
e-mail: hadri.hilmi@gmail.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_8

1 Introduction

Chemoinformatics is a transdisciplinary science that joins the chemistry aspect to computational and informational methods. The ultimate aim of this field as described by [1] is to solve chemical problems. Most of Chemoinformatics applications have been widely used in the pharmaceutical industry to predict interaction between chemical entities, to seek through in house databases for similar compounds as alternative drugs etc. One of the main applications in Chemoinformatics is Virtual Screening (VS) which is an alternative approach to the traditional and costly high-throughput screening used by pharmaceutical industries to find candidates for drugs discovery [2]. VS is further divided into few approaches one of which is Similarity Searching (SS). SS is the simplest method and the most widely used to find similar compounds for a target drug. SS produced a list of most similar structure in the database to the target being sought based on their similarity scoring. This is based the concept of Similar Property Principle that states compounds with similar structure tends to exhibit similar property [3]. Hence, the higher the resemblance between compounds the likelihood of the exhibiting similar bio-activity also increases. This would allow the new structure found to be used as an alternative to the target structure. The effectiveness of SS has encouraged researches to look deeper into enhancing and improving it by incorporating more compounds as target [3–9]. This concept called multi-target searching uses multiple targets structure as input instead of using a solo target [5]. This approach is also known as multiple reference similarity searching.

There are two main approaches that are currently used to implement the concept of multi-target searching namely Group Fusion (GF) and Turbo Similarity Searching. GF simply combines ranked result produced by several similarity searches (usually ten) into a final ranking which determines the degree of resemblance between targets and database compounds. According to recent studies by [6, 7], the effectiveness of searching was significantly improved when multiple targets compounds were used in GF, currently only 10 targets compounds are used. However, group fusion is not effective when only a few targets are known. All this led the introduction of an alternative multiple targets similarity searching concept known as Turbo Similarity Searching (TSS) [3, 5]. SS and TSS had been discussed in detail by [3]. TSS had been proven to be effective at finding drug leads. However the computation of TSS is time consuming. The advancement of current technology, such as multi-core processors and Graphic Processing Unit (GPU), can be used to reduce the computation time of TSS.

Multi-core processors are shared-memory systems that could be programmed for parallelism using OpenMP as API. OpenMP as described by [10] is a set of compiler directives and callable runtime library routines to express shared-memory parallelism. OpenMP is looked as the alternative for the message passing programming model and a replacement of Pthreads on shared-memory system in high performance computing space [10]. GPUs have been proven as pervasive scientific parallel architectures [11]. They accommodates high-throughput for floating point operations (up to 1.03 TFLOP/s), which is a desirable advantage for any scientific computing applications. GPU consists of large numbers of simple processing cores (up to 512 in a single chip) which enables it to executes parallel workload entity steadily [11].

With the increasing number of chemical compound databases, the requirement of high-performance structure similarity search algorithm is growing rapidly for large scale database search, compound library clustering, virtual screening, and other Chemoinformatics applications. It is believed that parallel implementation of the TSS algorithm on different parallel platforms can provide acceptable speedup. The idea of implementing TSS calculations onto a GPU is completely new, but implementing SS calculations onto a GPU is not new. The evaluation of Tanimoto coefficient of binary fingerprints on three different architectures (a quad-core Intel architecture, a Cell architecture and a GPU architecture) by [12] show that that a single CPU thread run three times faster than a GPU for calculating a large number of floating point Tanimoto coefficients. This finding looks unusual and it is being caused by the proposed GPU implementation is not fairly optimized, making the entire results biased toward the quad-core and the Cell architectures.

A novel GPU accelerated algorithm for all-VS-all Tanimoto matrix calculation and nearest neighbor search was introduced by [13]. The algorithm is up to 39 times superior to the existing commercial software that runs on CPUs. Because of the utilization of intrinsic GPU instructions, this approach is nearly 10 times faster than existing GPU-accelerated sparse vector algorithm, when Unity fingerprints are used for Tanimoto calculation. Another GPU accelerated structure similarity algorithm in both GPU and CPU versions has been tested by [14] on seven NCI 60-cell lines HTS datasets using virtual progressive screening protocol for performance and generality. The results have shown that the algorithm is two orders of magnitude faster than the CPU version on a PC machine with one Tesla GPU card more than 120-times faster. However, the GPU acceleration is not significant for the similarity search on smaller (few thousands) datasets because the algorithm requires to pay the additional costs of communication between GPU and CPU. This algorithm is particularly good for virtual screening on large datasets (i.e. multimillion compounds).

A detailed analysis of the advantages of using many-core architectures for calculating five different commonly used similarity coefficients (Tanimoto, Dice, Cosine, Euclidean and Hamming/Manhattan) had been presented by [15]. Considering both binary fingerprints and floating point descriptors as molecule representation, they found that the proposed algorithm on different experimental setups, obtaining up to a 206-fold speed-up for fingerprints and up to a 328-fold speed-up for descriptors. Previous works implemented SS onto GPU but none of them indicates any existing TSS implementation on GPU or multi-core processor.

Hence in our study we intend to look into the possibility of parallelizing TSS which held better accuracy but more time consuming than SS on multicore processor and GPUs. Therefore, this paper presents the parallel design of TSS and their performance evaluation to find the best design that yields fastest execution time without jeopardizing the accuracy.

2 Background

Components of TSS comprises of the chemical database and descriptors used to represent the data, similarity coefficient, fusion rule and Nearest Neighbor. The chemical database that is widely used is MDDR (MDL Drug Data Report). The MDDR version used holds nearly 102, 540 biologically relevant compounds [16]. Compounds in this database are represented by the binary fingerprints. Binary fingerprints encode the presence and absence of chemical compounds by denoting the former as '1' and the latter as '0'. Hence a representation of a chemical compound would be in the form of bit-strings with only 0 and 1 s as shown in [3].

Similarity coefficient is defined as an equation that used to know the percentage of similarity between the compounds. Similarity coefficient is depending on three terms. For example Let A and B represent the compounds that will be compared by binary bits. The first term is to calculate the bits "on" in A denotes by a. The second term is by calculating bits "on" in B denotes by b. The third term is to calculate the bits "on" in both A and B denotes by c. The Tanimoto Coefficient (Tc) is widely used, and it is usually executed with 2 fingerprints and the its value ranges from 0 to 1 where 0 denotes dissimilarity whilst 1 denotes perfect similarity (i.e. identical). The similarity between A and B using Tc is calculated using the following equation where S is the similarity value [5]:

$$S_{A,B} = c/(a + b - c) \tag{1}$$

According to [3, 5] TSS procedure (as shown in Fig. 1) involves two main phases. In the first phase is the initial SS of given target compound is conducted and the result produced is ranked decreasing based on the Tc value calculated between the target ane each compound in the database. In this case, 102,540 compounds. The upper compounds of the ranking list are those that are most similar to the target. These are called Nearest Neighbour (NN). In the second phase the top 200 compound of this list are used as target compounds in separates SS (hence a total of 200 SS are being executed). Results (i.e. Ranking List) of each of the searched are then combined into one final list by selected the highest (MAX) Tc value of each compound recorded by any of the searches. The combination (also called fusion) is done gradually starting with 5, 10, 15, 30, 40, 50, 100, and 200 list. The final rank is then sorted decreasingly to get the most similar structure. A detailed description on TSS can be found in [3]. We present here only essential information on TSS to help reader gauge the concept of TSS.

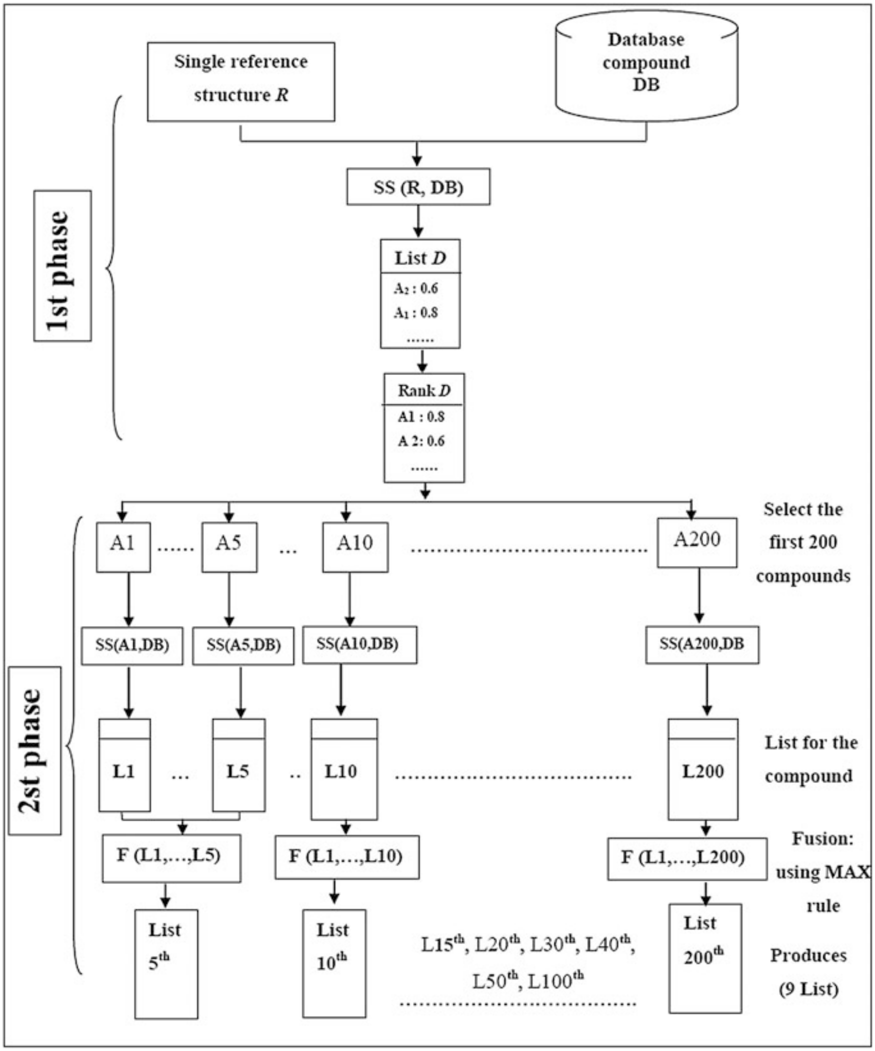


Fig. 1 High-level architecture for TSS where R is defined as target compound used to search a database DB with size N where N = 102,540 and A = 200

3 Parallelizing the Turbo Similarity Searching Algorithm

3.1 Parallelising TSS Using CUDA

Two designs have been implemented for parallelizing TSS on CUDA. The first design named CUDA 1 isolates the compounds (i.e. data) through the cores of GPU. The 1st phase of TSS requires a read from text file (prep-processing) and the

```
int n =  102540                         //number of compounds in the dataset
int block_size = 1024;                 // number of threads per block
int num_blocks = n/block_size + (n%block_size == 0 ? 0:1);
                                        //Number of blocks Per Grid here it is equal 101
```

Fig. 2 CUDA Kernel parameters for TSS

execution of the initial SS. Hence, the preprocessing is done on the CPU (Host) declared by the __host__ qualifier whilst the SS is executed on GPU (Device) and declared by the __global__ qualifier. The 2nd phase of TSS (SS for 200 NN and the gradual fusion of ranked list from 5 to 200 NN), are all executed on GPU hence being declared by the __global__ qualifier. The CUDA kernel provides variables that are dependent on the size of launching grid and blocks. All these parameters are set by CUDA automatically, based on the kernel launch parameters (the block and grid dimensions). In CUDA 1 design, the number of parallelism is expressed by the size of database compounds thus make the properties of threads and blocks in CUDA being defined as shown in Fig. 2. Whilst the high level architecture diagram that shows data movement between host and device is presented in Fig. 3.

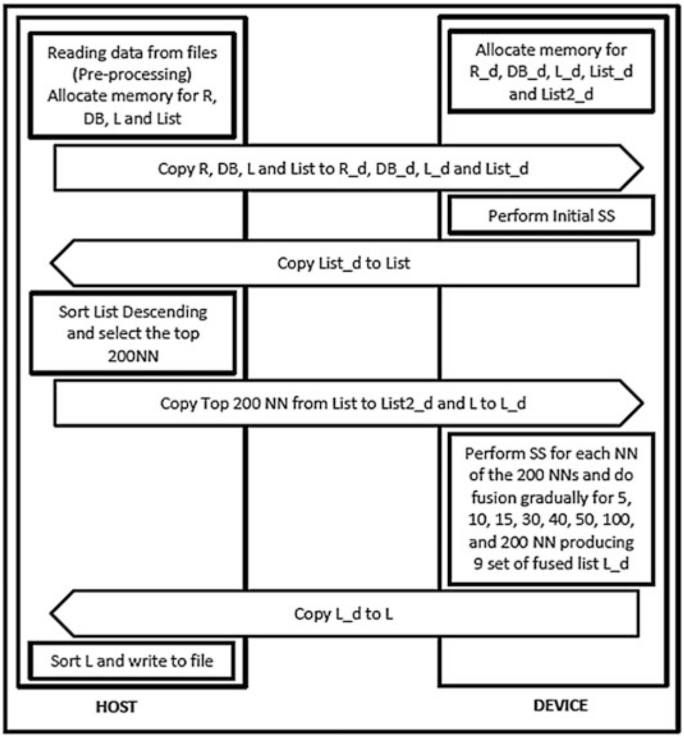


Fig. 3 High level architecture diagram for CUDA 1 design

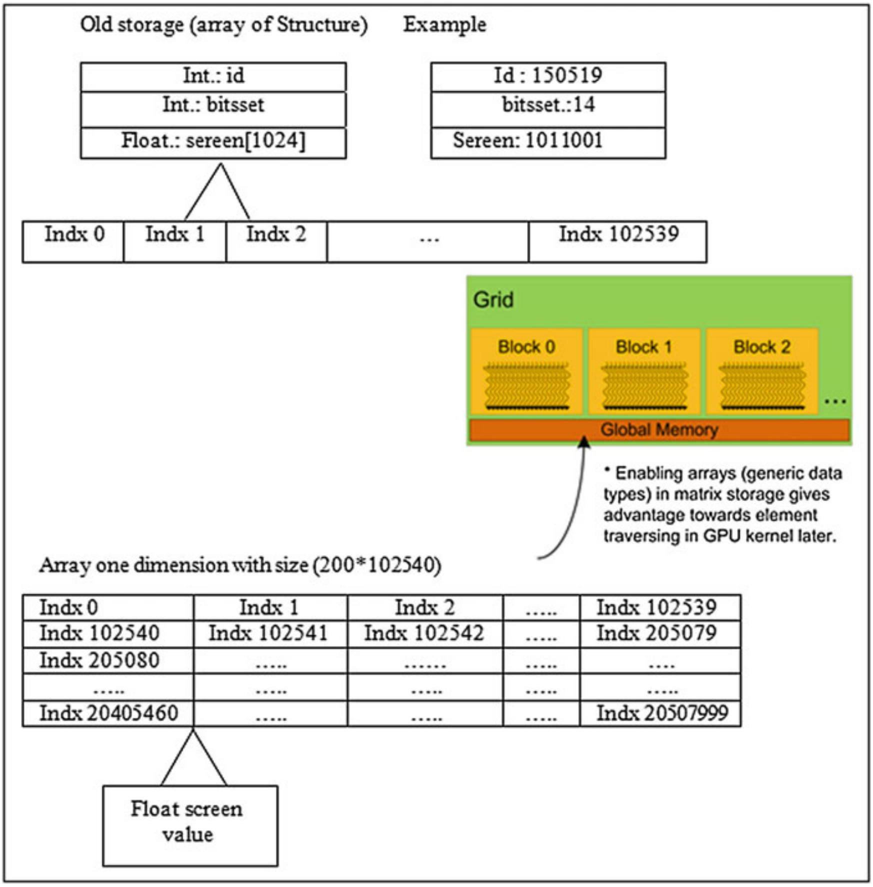


Fig. 4 The new data storage proposed in CUDA 2 design

Unlike in CUDA 1 design whose data storages are arrays of n size. CUDA 2 design took advantage of a method called Row-major Order Indexing. This method is a popular method for describing and storing multi-dimensional arrays in linear memory [17]. For this method to work, the current data storage that is used in CUDA 1 design is changed to a monolithic (1D) array. The proposed data storage is based on row-major order with matrix form of multi-dimensional arrays as shown in Fig. 4.

One of the advantages of newly proposed data storage is the existence of a generic data type. In the old data storage, one compound structure is defined as one structural element while in the new proposed data storage, the aspects of compound structures (id and score) are separately allocated as a linear array as a basic type (which is int and float). Therefore, the new data storage (matrix storage) is much more lightweight and suitable for use with lightweight task operation (or computation) and scalable towards either multiple reference compounds or expandable database compounds. This storage has allowed the GPU to hold and sort the initial SS list (List_d) instead of copying it back to the host for sorting and top 200 NN selection.

3.2 Parallelising TSS Using OpenMP

Parallelising TSS using OpenMP is much simpler. OpenMP has a loop-based work-sharing construct that implicitly parallelized for loop after the compiler directives #pragma omp parallel for is put before any for loop in the implementation code. Since our TSS implementation contains two main and independent for loops, this work-sharing constructs fits nicely to enforce parallel execution of TSS. Hence, the main loop of the algorithm is identified and we inserted the construct that forces the segment of algorithm to be executed in parallel.

4 Experimental and Results

The parallel implementations of TSS algorithms are executed on the Biruni Server (biruni.cs.usm.my) which is located in the Parallel and Distributed Laboratory, School of Computer Sciences, Universiti Sains Malaysia. The features of the CPU and the GPU of the Biruni server is displayed in Table 1.

The parallel implementations of TSS are evaluated by comparing the time taken to completely execute TSS on the sequential and parallel platforms. These timings did not include the preprocessing step in the implementation. Therefore the sequential runtime can be considered as the time measure from the beginning to the

Table 1 The CPU and the GPU features

CPU specifications	GPU specifications
Model: AMD PhenomTM II X4 810, Number of Cores: 4, Clock Speed: 2.6 GHz, Memory Size: 4 GB	Model: NVIDIA Tesla C2050, total amount of global memory: 2687 Mbytes, number of multiprocessors: 14, number of cores per multiprocessor: 32, warp size: 32, maximum sizes of each dimension of a grid: 65,535 × 65,535 × 65,535, maximum sizes of each dimension of a block: 1024 × 1024 × 64, maximum number of threads per block: 1024, CUDA capability major/minor version number: 2.0, total amount of shared memory per block: 49,152 bytes (∼50 Kb)

end of 1st phase and 2nd phase on the sequential platform. The parallel runtime measures the time for transferring the data from the CPU to the GPU, time for performing the searching process on the parallel platform, and finally time to transfer data from the GPU to the CPU. In order to harness the advantage of GPU, each cores of the GPU needs to get at least one block to execute. Therefore, the kernel parameters for the number of blocks per grid are set to 101, to keep all the cores on the GPU as busy as possible. Since there is a limit to the number of threads per block, all threads of a block are expected to reside on the same core and share the limited memory resources of that core. Biruni's block contains 1024 threads. Hence, the maximum number of threads per block is set to 1024 to reduce the cost of communication between threads.

4.1 Performance Comparison Between Sequential and Parallel Implementation of TSS

Table 2 illustrated the comparison of the execution time between the sequential, parallel OpenMP, parallel CUDA1 and parallel CUDA2 for phases involved in TSS.

The execution time of the parallel CUDA 1, CUDA 2 and parallel OpenMP of the 1st phase for TSS algorithm is not so much lower than the execution time of the sequential version of the same algorithm. This is because due to the small size of data being handled in SS which led to a less complex processing that did not take full advantage of the GPU and CUDA. The parallel CUDA programs executed with 102,540 (the number of database structures) threads together with only one target as compared to the 2nd phase in which involved SS execution on 200NN as targets to each of the 102,540 database compounds. This is an embarrassingly parallel problem which would really benefit from the coupling of GPU and CUDA. Hence, the execution time of the parallel CUDA 1 and CUDA 2 design of the 2nd phase of TSS is observed to be extremely lower compared to the execution time of the sequential version and parallel OpenMP version. Such performance is also conceivable due to the availability of a huge number of cores in Biruni (448 cores as we tested) and by using to the parallel reduction approach used for the fusion of the SS results.

Table 2 Overall execution time performance for TSS algorithm

TSS design implementation	Execution time (s)	
	1st phase (initial SS)	2nd phase (SS with fusion- after 200NN)
Sequential	1.1060	191.432
Parallel OpenMP	0.3366	75.283
Parallel CUDA 1	0.33078	1.464
Parallel CUDA 2	0.3308	1.392

Table 3 Timings for fusion procedures based on sequential and parallel implementation

Numbers of NN's SS results being fused	Execution time (ms)			
	Sequential	OpenMP	CUDA1	CUDA2
5	6346	4287.6	31.894	30.341
10	11,822	7363.8	66.792	63.385
15	17,260	10,888.8	98.846	93.320
20	22,602	14,083.4	131.548	124.234
30	32,656	19,314.8	199.846	189.193
40	42,732	24,291.4	267.704	253.419
50	52,592	29,010.4	339.664	322.185
100	99,314	45,289.8	703.630	671.054
200	192,064	75,285.4	1460	1392.35

The performance superiority of CUDA 2 seems to surpass CUDA 1 consistently even when we look deeper into the gradual fusion procedures with a nine set of NNs. Table 3 shows the execution time in milliseconds for each fusion procedure. When the speedup (i.e. sequential runtime divided by parallel runtime) for each design being calculated across all procedures, we perceive that the speedup for TSS using OpenMP is in the range of 1.49X to 2.55X (faster) than the sequential TSS. Whereas, CUDA1 TSS and CUDA2 TSS recorded speedup in the range of 1.32X to 1.99X and 1.38X to 209X, respectively. However, we notice that the increasing trend of speedup on the OpenMP TSS as more NN's lists are fused did not hold in both CUDA designs. Instead, the speedup decreases as the number of NN's list fused increases. Most of contribution to the better performance of parallel CUDA is coming from the optimization in both CUDA kernel functions of SS and fusion procedures. The underlying lightweight tasks had affected directly the computation while the threads were not overly used in order to reduce idle threads during execution. The decrease in speedup for both CUDA1 and CUDA2 is due to the execution of multiple branch (IF) statements for the purpose of fusion which added more complexity to the procedure. This scenario is the most feared by GPU programmers and its best to be avoided [18].

4.2 *Comparing the Two Different CUDA Designs*

The difference of general performance between CUDA 1 and CUDA 2 is very minimal (approximately 0.072 s). This is often considered insignificant difference in parallel platforms performance comparison. However, we perform another comparison in terms of GPU memory usage for both designs to find a design that not only has best speedup but also less GPU memory consumptions.

In the 1st phase, both designs used the same size of memory, but in the 2nd phase the total memory used by parallel TSS implementation of CUDA 2 seems to double

Table 4 Overall GPU memory used of parallel CUDA1 and CUDA2 according to 2nd phase

Numbers of NN's SS results being fused	GPU memory used for CUDA 1 (MB)	GPU memory used for CUDA 2 (MB)
5	444.20	805.39
10	444.98	807.35
15	445.76	809.30
20	446.54	811.26
30	447.32	815.17
40	448.11	819.08
50	448.89	822.99
100	449.67	842.55
200	450.45	881.67

the size used by CUDA 1. This is being portrayed in Table 4 where it is observed that CUDA 1 memory usage ranges from 444 to 450 MB; an increase of only 6 MB when more NNs are fused. CUDA 2, on the other hand, occupied memory in the range of 809 MB to 881 MB; an increase of 72 MB when more NNs are fused. Hence, in terms of GPU memory utilization, CUDA 1 is the better design.

5 Conclusion

This paper proposed three parallel designs for TSS. Two of which are based on the many-cores (GPU-CUDA) platforms whilst another is based on the multi-core platforms. These parallel designs focused on parallelizing the SS procedures on the 1st and 2nd phase of TSS as well as the fusion procedure in the 2nd phase of TSS. The performance of each design were benchmark against the sequential implementation where CUDA 2 design was found to be the best parallel implementation of TSS with 209X speedup. This is due to the embarrassingly parallel nature of the 2nd phase of TSS and the use of parallel reduction approach. The parallel design using OpenMP failed to perform due to the limited number of cores to cater due to the low usages of cores and the dependency of data in reduction operation (depicted shared memory issues). Adding to the presence of large computational needs (i.e. to perform screening with more than 100 thousands compounds) and the dependency of data in reduction operation (typical shared memory issues). In terms of memory utilization, the proposed data storage in CUDA 2 design seems to be taking a lot of memory space as compared to the array-based data storage used in CUDA 1. Hence, CUDA 1 is the best design as the parallel version of TSS considering the high speed and minimal space consumption. We are currently implementing the CUDA 1 design on a larger database and the result shall be reported in another publication. Overall, we noticed from the results obtained that the parallel version of TSS implemented on the GPU performs better than its sequential version and OpenMP version on the same data type. Thus, we may conclude that TSS can be accelerated by implementing and executing it on GPU. However in order to harness the GPU to its maximum, we have to use less flow control instructions as possible in the device code so that any thread which may has major effect on the execution of the program could be skipped.

Acknowledgments We would like to thank Dr. Shereena M Arif for providing us the MDDR database.

References

1. Gasteiger, J. (ed.): Handbook of Chemoinformatics (2003)
2. Johnson, M. A., Maggiora, G. M.: Concepts and Applications of Molecular Similarity (1990)
3. Zainal, A., Yusri, N., Malim, N., Arif, S. M.: The influence of similarity measures and fusion rules toward turbo similarity searching. In: International Conference on Electrical Engineering and Informatics (2013)
4. Whittle, M., et al.: Enhancing the effectiveness of virtual screening by fusing nearest neighbor lists: a comparison of similarity coefficients. J. Chem. Inf. Comput. Sci. **44**(5), 1840–1848 (2004)
5. Malim, N., Pei-Chia, Y., Arif, S. M.: New strategy for turbo similarity searching: implementation and testing. In: IEEE 2013 International Conference on Advanced Computer Science and Information Systems (2013)
6. Hert, J., Willett, P., Wilton, D.J.: Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. J. Chem. Inf. Comput. Sci. **44**, 1177–1185 (2004)
7. Whittle, M., Gillett, V.J., Willett, P., Loesel, J.: Analysis of data fusion methods in virtual screening: theoretical model. J. Chem. Inf. Model. **46**, 2193–2205 (2006)
8. Hert, J., et al.: Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbor information. J. Med. Chem. **48**, 7049–7054 (2005)
9. Whittle, M., Gillett, V.J., Willett, P., Loesel, J.: Analysis of data fusion methods in virtual screening: similarity and group fusion. J. Chem. Inf. Model. **46**, 2206–2219 (2006)
10. Dagum, L., Menon, R.: OpenMP: an industry standard API for shared-memory programming. IEEE Comput. Sci. Eng. **5**(1), 46–55 (1998)
11. NVIDIA. Nvidia's next generation cuda compute architecture: Fermi
12. Sachdeva, V., Freimuth, D., Mueller, C.: Evaluating the jaccard-tanimoto index on multi-core architectures. Comput. Sci. ICCS 2009, pp 944–953 (2009)
13. Ma, C., Wang, L., Xie, X.-Q.: GPU accelerated chemical similarity calculation for compound library comparison. J. Chem. Inf. Model. **51**(7), 1521–1527 (1998)
14. Yan, X., Gu, Q., Lu, F., Li, J., Xu, J.: GSA: a GPU-accelerated structure similarity algorithm and its application in progressive virtual screening. Mol. Divers. **16**(4), 759–769 (2012)
15. Maggioni, M., Santambrogio, M.D., Liang, J.: GPU-accelerated chemical similarity assessment for large scale databases. Procedia Comput. Sci. **4**, 2007–2016 (2011)
16. Southan, C., Várkonyi, P., Muresan, S.: Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds. J. Cheminform. **1**, 10 (2009)
17. Otoo, E. J., Rotem, D.: Parallel access of out-of-core dense extendible arrays. In: IEEE International Conference on Cluster Computing, pp. 31–40 (2007)
18. Huang, Q., Huang, Z., Werstein, P., Purvis, M.: GPU as a general purpose computing resource. Paper presented at the ninth international conference on parallel and distributed computing, applications and technologies, 2008

Correlated Topology Control Algorithm for Survival Network in MANETS

A.H. Azni, Rabiah Ahmad, Kamaruzzaman Seman, Najwa Hayaati Mohd Alwi and Zul Azri Mohamad Noh

Abstract This paper proposed a new correlated topology control (CTC) algorithm to achieve optimal survival networks topology under correlated node behavior. CTC algorithm enable every node to selects its own neighbors with highly correlated cooperative node dynamically. To select neighbors, measurement scheme of correlated node cooperativity is proposed to determine misbehave node and critical node. The scheme will check on correlated degree of neighboring nodes for any misbehave node and at the same time will check if the removal of node cause network to partition. In the case if nodes removal will partition the network, a new reconstruction scheme *LNL* is proposed to reconnect the broken link. The simulation results show that CTC algorithm can prolong node lifetime and maintain generated survival network topology. Besides that, the algorithm successfully prevented correlated event from misbehave node and network partitioning from isolated node. Further, it is also shown that as energy consumption decreased, network performance especially packet forwarding can be significantly improved and network survivability is optimally achieved.

Keywords Topology control · Survival network · Correlated behavior · MANETS

A.H. Azni (✉) · K. Seman · N.H.M. Alwi
Faculty Science and Technology, Universiti Sains Islam Malaysia (USIM), Nilai, Malaysia
e-mail: ahazni@usim.edu.my

K. Seman
e-mail: kamaruzzaman@usim.edu.my

N.H.M. Alwi
e-mail: najwa@usim.edu.my

R. Ahmad · Z.A.M. Noh
Faculty of Information Technology and Communication, University Technical Malaysia Melaka (UTeM), Durian Tunggal, Malaysia
e-mail: rabiah@utem.edu.my

Z.A.M. Noh
e-mail: zulazri@utem.edu.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_9

1 Introduction

Topology control is one of the most important techniques used in MANETs and sensor networks for reducing energy consumption and radio interference [1]. Usually, topology control can be determined by node behavior status. The main idea of topology control is that, instead of transmitting with the maximal power, nodes are collaboratively determined their neighboring node by forming the proper neighborhood relation under certain criteria, with the purpose of maintaining connectivity [2]. In this paper, topology control is proposed based on cooperative node selection and discard node that is highly potential of performing misbehave node referring to malicious, selfish and fail node. The nodes in the topology are maintained with high correlated cooperative degree to increase survivability. As nodes are mutually cooperative, a good node will affect other nodes' behaviors positively, while misbehavior node will jeopardize its neighbors. Misbehave node will be discarded because they caused extra energy consumption to neighboring node with their suspicious activity [3]. The paper is organized as follows: Sect. 2 discuss related work of topology algorithm, then, this paper proposed correlated topology control algorithm by measuring correlated node cooperativity and topology reconstruction in Sect. 3. Next, the performance of proposed algorithm is evaluated in extensive simulation experiments in Sect. 4 and followed by a conclusion in Sect. 5.

2 Related Works

Many topology control algorithms have been design for ad hoc networks. Most works are design to prolong network lifetime by minimizing per node energy consumption to increase network survivability. Such example includes clustering [4], adjusting nodes transmission power [5] and balancing energy consumption [6]. Clustering in MANETs divided nodes into different virtual groups based on some rules for different behavior of nodes to be included or excluded from the cluster [7]. Xu et al. [8], Wattenhofer and Li [9] are among the first to have proposed the used of clustering algorithm in topology control to reduce nodes' energy consumption. Work by Swaroop et al. [10], El-Hajj and Kountanis [11], Yu and Chong [12] proposed an adaptive clustering, deal with overlapping cluster structure and proposed non-overlapping cluster architecture without clusterheads to reduce network bottlenecks. Recent work by Sahin et al. [13] take advantage of bio inspired technique to design genetic algorithm for VANETs. Another techniques used are based on mobility [14] and movement control algorithm [15] to support mobile node.

Topology control in previous works deals with single node information to create a selection rules in topology formation. However, little works on neighborhood node information has been explored in topology control algorithm for the purpose of network survivability. In this paper, topology control is done in collaborative manner in which it can help to increase network survivability. The node selection is based on correlated degree between neighboring nodes to form a correlated set of cooperative network topology. As a result, the correlated cooperative neighborhood sets will generate a survival network topology against correlated node behavior in MANETs.

3 Correlated Topology Control (CTC) Algorithm

3.1 Overview

Correlated Topology Control (CTC) is a distributed and localized protocol in which every node can select cooperative adjacent nodes as its neighbors and excludes misbehave node from the network. The basic idea of CTC is to enhance the network survivability against correlated node behavior by connecting cooperative nodes dynamically via message exchanges with mutual neighbors. More specifically, every node selects its adjacent node with the highest cooperative correlated degree into neighborhood set and discard misbehave node that has a high potential to perform correlated event.

This research involves two phases of topology control process which are topology neighborhood discovery/selection and neighborhood topology updates. Phase 1 involves node initialization to obtain local information from neighboring nodes. This information is required to measure correlated cooperativity scheme. In Phase 2, topology update is needed to check on neighborhood list of any misbehave node which is potential to create correlated event. If node removal will cause network to partition, a new topology reconstruction must be performed to maintain connectivity from node removal.

3.2 Phase 1: Neighborhood Discovery

Neighborhood discovery is the process through which a node selects a set of neighbor node list from its one-hop neighborhood. This set of selected neighbor nodes is called its correlated cooperative set. To perform a neighborhood selection, each node broadcasts a “Hello” message which contains node id, residual energy information and packet forwarding ratio [16]. This local information is needed to measure *node cooperativity* to inquire new node to be added to its correlated cooperative set. Let *δ(u)* denote node *u* cooperativity, $n_{fwd}(u)$ and $n_{drop}(u)$ denote the number of packets should be forwarded and drop, respectively. Then $\delta(u)$ can define as:

$$\delta(uv) = \frac{n_{fwd}(u)}{n_{drop}(u)} \tag{1}$$

Equation (1) has been defined in [17] to calculate probability of forwarding and also in *Watchdog* [18] to detect node misbehavior. However, an extension not mention in previous works is to let nodes measure *correlated node cooperativity* of node *u* and node *v* as a correlated node behavior. Based on the measurement of both nodes cooperativity, *correlated node cooperativity* can be calculated as:

$$\omega_{uv}(t) = \delta_u * \delta_v \tag{2}$$

Equation (2) also refers to link connectivity of node *u* and node *v*. Let *G* be the weighted undirected graph of the topology *G* = (*V*, *W*), where *V* is the set of adjacent nodes, $\{v_1, v_2, v_3, \ldots v_N\}$ and *W* is the set of links. The adjacent matrix of weighted graph *G* denote the *N* × *N* adjacency node matrix at time *t*:

$$W(t) = \begin{bmatrix} \omega_{11}(t) & \omega_{12}(t) & \ldots & \omega_{1N}(t) \\ \omega_{21}(t) & \omega_{22}(t) & \ldots & \omega_{2N}(t) \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{N1}(t) & \omega_{N2}(t) & \ldots & \omega_{NN}(t) \end{bmatrix} \tag{3}$$

where

$$\omega_{uv}(t) = \begin{cases} \omega(u,v) \text{ if node } u \text{ and } v = \delta \\ \beta, \text{ otherwise} \end{cases} \tag{4}$$

δ refers to the *correlated node cooperativity* of node *u* and node *v* which comply all routing rules and forwarding all the data packets at its best, while β denotes either node is in misbehave state. Nodes in misbehave state may drop its packets to save energy or launch malicious attacks will be disconnected to prevent correlated event.

In CTC, each node calculates *correlated node cooperativity* to decide if it accepts a request as its neighbors. When node *u* receives a neighbor *Ngbr-Req* from adjacent node *v*, it compares ω_{uv} with its *cooperative threshold* value denote it by ω^*_{uv} If ω_{uv} is within *cooperative threshold* value, node *u* will send *Ngbr-Rep* and adds node *v* into neighbor set $W_cSet(u)$; otherwise node *u* discards this request. Node *u* will continue inquiries until at least *d* neighbors. Algorithm 1 summarizes the procedure of processing neighbors' discovery procedure (Table 1).

By correlated cooperative neighborhood status selection, nodes with high *correlated node cooperativity* are organized together and form a survival network topology without misbehaves nodes.

Table 1 Algorithm for neighbors' discovery procedure

Algorithm 1: Neighbors' Discovery Procedure	
Input: *d*, node *u*, and $Adj(u)$ **Output:** $W_cSet(u)$ 1. Initiate $W_cSet(u) := \varnothing$ Create a temp set $Temp(u) := \varnothing$, Created a counter $nReq := 0$, 2. $\forall v \in Adj(u)$, Measure ω_{uv} 3. **While** $(nReq < d$ **AND** $Temp(u) \neq Adj(u))$ **do** 4. send *Ngbr-Req* to *v* 5. $Temp(u) := Temp(u) + v$	6. **if** (Received *Ngbr-Req*) from $v \in Adj(u)$ 7. **then** **if** $\omega_{uv} \geq \omega^*_{uv}$ **AND** $v \notin W_cSet(u)$ **then** 8. Send *Ngbr-Rep* to *u* 9. $W_cSet(u) := W_cSet(u) + v$ 10. $nReq := nReq + 1$ 11. **else** 12. Discard *Ngbr-Req* 13. **end if** 14. **end if** 15. **end while**

3.2.1 Phase 2: Network Topology Update

To maintain survival network topology, an update of the neighborhood set is needed if there is a change in node behavior status and link connectivity. Node in MANETs is subject to change its behavior following state transition discusses in [19]. Node behavior status will reflect the changes in correlated node cooperativity value which is the value of link connectivity. The changes also may cause network to performed correlated event and also cause network to partition especially with misbehave nodes. To avoid that to happen, topology update is done periodically to check on misbehave node in the network. Whenever node is performing its routing update, it will check its neighborhood list of any misbehave node. Misbehave node is identified if *correlated node cooperativity* fall below threshold level. For topology update, the threshold value is set higher than node discovery threshold to allow reconstruction of new topology. Apart from misbehave node detection, the topology update also needed to identify node isolation which may cause network to partition. The node topology update algorithm can be performed periodically. The procedure for node topology update is repeated every time the node performs its routing update.

When misbehave node is identified, the topology update algorithm will check again if the removal of node will cause network to partition. Node is considered isolated if no other alternative path exists to establish a connection. If an alternative path exists, then the node is not critical because its removal will not disconnect the network. Thus, a topology modification scheme is needed to reconstruct a new connection so that network partition can be avoided. This work extend modification scheme Least Number of Link (*LNL*) proposed by Kim et al. [20] where the scheme randomly choose unconnected neighbor node to add additional connectivity when the critical node is removed (Table 2).

Table 2 Algorithm for network topology update procedure

Algorithm 2: Neighbors' Topology Update Procedure	**Algorithm 3: Modification Scheme (LNL)**
Input: node u, and $Adj(u)$	**Input:** node u, and $Adj(u)$
Output: $W_cSet(u)$	**Output:** $W_cSet(u)$
1. $\forall v \in Adj(u)$, Measure ω_{uv}	1. select $v \in Adj(u)$
2. **if** Received *Ngbr-Req* from $v \in Adj(u)$ **then**	2. **if** $v \notin W_cSet(u)$ **then**
3. **if** $\omega_{uv} \leq \omega_{uv}^{**}$ **AND** $v \in W_cSet(u)$ **then**	3. Compute λ of $L(t)$
4. Compute λ of $L(t)$	4. **If** $\lambda_2 > 1$ ***then***
5. **If** $\lambda_2 > 1$ **then** remove v	5. $W_cSet(u) := W_cSet(u) + v$
6. Recalculate $W_cSet(u)$	6. **else**
7. **else**	7. **Repeat step to 1**
8. **Go to Algorithm 3**	8. **end if**
9. **end if**	9. **end if**
10. **end if**	
11. **Repeat Step 1**	
12. **end if**	

4 Results and Analysis

In this section, CTC algorithm is examined based on the network performance improvement in energy consumption, load distribution and packets forwarding ratio. The performance of survival network topology is evaluated using NS-2 (v2.35) simulation and MATLAB (7.10a). Simulations are performed for a 1000 m × 1000 m^2 are consisting of 100 mobile nodes, distributed randomly over the network within 200 m communication radius. Initial energy was set to 100 W with mobility speed range between 2 and 4 m/s. The random ad hoc network topology is created 50 times and the results are averaged. Each node becomes misbehave according to 0.1, 0.3 and 0.5 value which means at about 10, 30 and 50 % of the nodes misbehave in the network.

4.1 Energy Consumption

The simulation experiments are designed to compute energy consumption using the proposed algorithm and validated by comparing them with standard AODV. Figure 1 shows average energy consumption of nodes in *t* time. At time *t* = 3000 s, energy consumption of nodes is up to 40 % when CTC algorithm is used compared to standard AODV where the energy level almost consumed 100 %.

The reduction in energy consumption is due to limited number of node degree specified by CTC algorithm. By keeping node degree low, the energy consumption will reduce since the routing activity has been disseminated to other nodes as well. Besides, the removal of misbehave node especially malicious node has helped the network reduced excessive energy consumption due to suspicious activity.

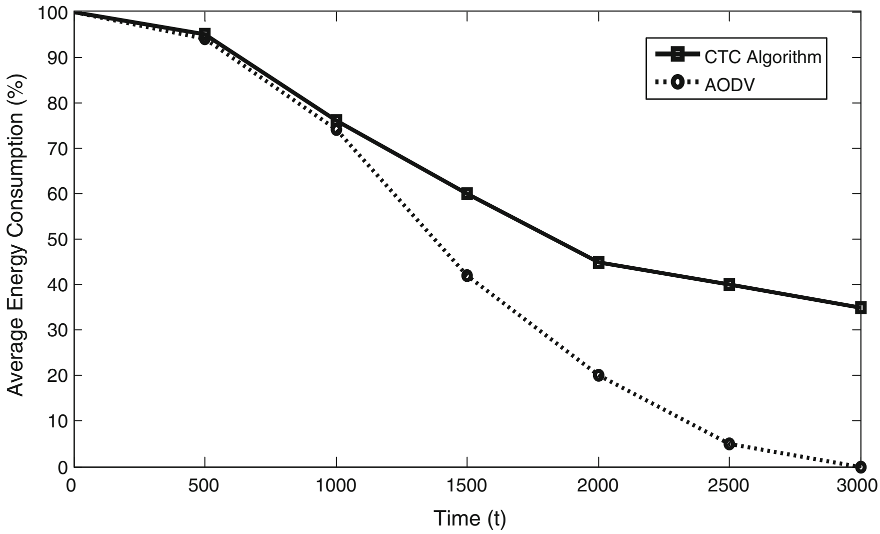


Fig. 1 Energy consumption

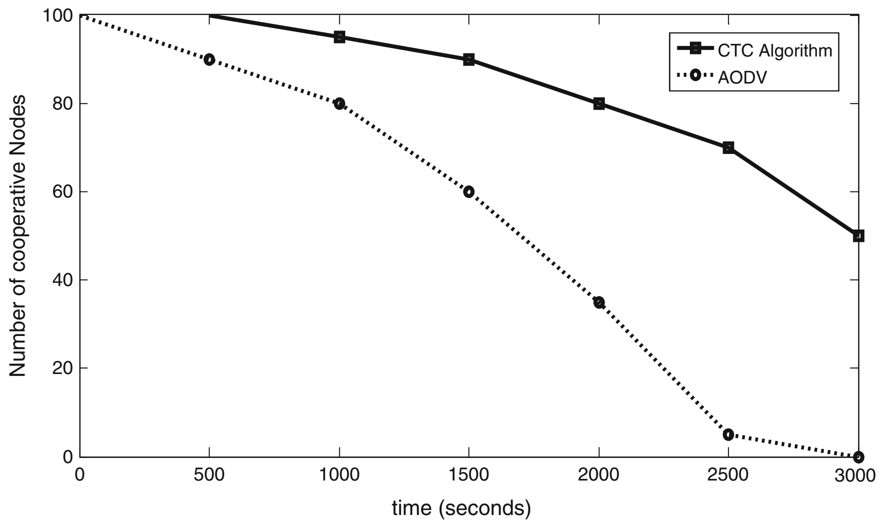


Fig. 2 Numbers of cooperative nodes

Furthermore, reconstruction of node topology prevents node from correlated event and network partitioning which gives a tremendous affect to network survivability. The effect of reduction in energy consumption directly increases the number of cooperative nodes. The numbers of cooperative nodes shown in Fig. 2 are much greater in CTC algorithm due to balance energy consumption among nodes which effectively prolong network lifetime.

4.2 Packets Forwarding Ratio

The probability of forwarding plays a decisive role in performance improvement as the node selection is dependent on it. Figure 3 illustrated packets forwarding ratio against misbehave node. It can be seen that packet forwarding ratio using standard AODV decrease drastically whenever misbehave rate is more than 0.2 or 20 % which also means higher drop ratio.

On the contrary, packets forwarding ratio are much better in topologies generated by CTC algorithm where it is still higher as infection rate is reaching 20 %. This observation indicates that when packets drop is less, node will behave cooperatively and the impact of network performance is tremendously improved. However, as infection rate is reaching to 60 %, both networks could not survive, even *LNL* scheme could no longer improve the performance. This observation is consistent with critical condition describe in [17] where at 60 % infection rate, misbehave node becomes an epidemic.

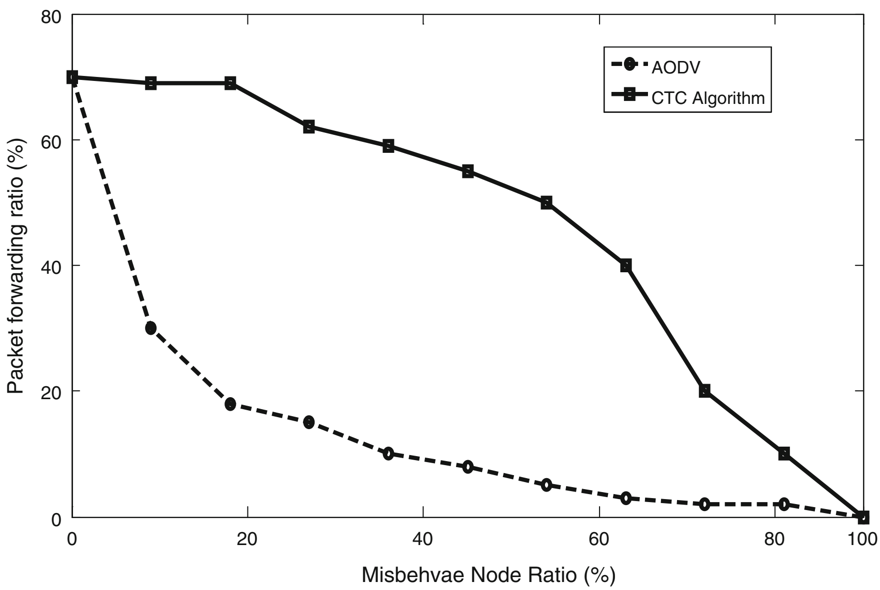


Fig. 3 Packet forwarding ratio versus misbehave node ratio

5 Conclusion

This paper has proposed a new CTC algorithm to achieve optimal survival networks topology under correlated node behavior. CTC algorithm enables every node to dynamically select its own neighbors with highly correlated cooperative node and check on correlated degree of neighboring nodes for any misbehave node. In the case where nodes removal will partition the network, a new reconstruction scheme *LNL* is proposed to reconnect the link. The simulation results confirmed that CTC algorithm could prolong node lifetime and maintain generated survival network topology. Besides that, the algorithm has successfully prevented correlated events from misbehave node and network partitioning from isolated node. For future work, the depth application of CTC algorithm needs to be investigated to other protocols and domains such as secure routing protocol (SRP) and SAODV. It is also proven that network topology without CTC algorithm consumes more energy than that with CTC algorithm. Thus, in the future, pilot test can be improved by applying a proposed design that tackles correlated node behavior problem effectively and increases the chances of survivability as proven in the simulation studies.

Acknowledgments The authors would like to express their gratitude to Universiti Sains Islam Malaysia (USIM) for the support and facilities provided. This research paper is supported by Fundamental research Grants No: FRGS/2/2014/ICT03/USIM/02/1.

References

1. Santi, P.: Topology control in wireless ad hoc and sensor networks. ACM Comput. Surv. **37** (2), 164–194 (2005)
2. Li, N., Hou, J.C.: Topology control in heterogeneous wireless networks: problems and solutions. In: INFOCOM 2004. Twenty-third Annual Joint Conference IEEE Computer Communication Society, vol. 1 (2004)
3. Azni, A.H., Ahmad, R., Noh, Z.: Survivability modeling and analysis of mobile ad hoc network with correlated node behavior. Procedia Eng. **53**, 435–440 (2013)
4. Yeung, C.: A Performance Comparison of Clustering Algorithms in Ad Hocnetworks. University of Central Florida (2006)
5. Arun Kumar, B.R., Reddy, L.C., Hiremath, P.S.: Mobile ad hoc networks : issues, research trends and experiments. IETECH J. Commun. Tech. **2**(2):57–63 (2008)
6. Hamed, R.G., Karimpour, J.: (EBHCR) Energy balancing and hierarchical clustering based routing algorithm for wireless sensor networks. Aust. J. Basic Appl. Sci. **5**(9), 1376–1380 (2011)
7. Yu, J., Chong, P.: A survey of clustering schemes for mobile ad hoc networks. IEEE Commun. Surv. Tutorials **7**(1), 32–48 (2005)
8. Xu, Y., Bien, S., Mori, Y., Heidemann, J., Estrin, D.: Topology control protocols to conserve energy in wireless ad hoc networks (2003)
9. Wattenhofer, R., Li, L.: Distributed topology control for power efficient operation in multihop wireless ad hoc networks. In: Proceedings of the IEEE INFOCOM 2001. Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies, vol. 3, pp. 1388–1397 (2001)

10. Swaroop, P., Sunitha, R., Aghila, G.: ASCM: an autonomous subtree-based clustering scheme for MANET. In IEEE Proceeding of International Conference on Methods and Models in Computer Science, ICM2CS 2009, pp. 1–5 (2009)
11. El-Hajj, W., Kountanis, D.: A fuzzy-based virtual backbone routing for large-scale MANETs. Int. J. Sens. Netw. **4**(4), 250–259 (2008)
12. Yu, J., Chong, P.: An efficient clustering scheme for large and dense mobile ad hoc networks (MANETs). Comput. Commun. **30**(1), 5–16 (2006)
13. Sahin, C., Urrea, E., Uyar, M.: Design of genetic algorithms for topology control of unmanned vehicles. Int. J. Appl. Decis. Sci. **3**(3), 221–238 (2010)
14. J. Wu and F. Dai, Wu, Jie, and Fei Dai. "Mobility-sensitive topology control in mobile ad hoc networks." Parallel and Distributed Systems, IEEE Transactions on 17, no. 6, pp. 522–535, 2006
15. Basu, P., Redi, J.: Movement control algorithms for realization of fault-tolerant ad hoc robot networks. In: IEEE Network, vol. 18(4):36–44 (2004)
16. Cartigny, J., Simplot, D., Stojmenovic, I.: Localized minimum-energy broadcasting in ad-hoc networks. In: IEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No.03CH37428), vol. 3, pp. 2210–2217 (2003)
17. Azni, A., Ahmad, R., Noh, M., Azri, Z., Correlated node behavior in wireless ad hoc networks: an epidemic model. In: IEEE 2012 International Conference For Internet Technology And Secured Transactions, pp. 403–410 (2012)
18. Liang, G., Vaidya, N.H.: When WatchDog Meets Coding (2010)
19. Azni, A., Ahmad, R., Noh, Z.: Correlated node behavior model based on semi markov process for MANETS. J. Comput. Sci. Issues **9**(1), 50–59 (2012)
20. Kim, T.H., Tipper, D., Krishnamurthy, P., Swindlehurst, A.L.: Improving the topological resilience of mobile ad hoc networks. In: 7th International Workshop on Design of Reliable Communication Networks, 2009. DRCN 2009, pp. 191–197 (2009)

A New System Call Classification for Android Mobile Malware Surveillance Exploitation via SMS Message

Madihah Mohd Saudi, Muhammad Zuhair Abd Rahman, Azirah Alfaizah Mahmud, Nurlida Basir and Yumn Suhaylah Yusoff

Abstract Nowadays, Android has become the most widely used platform for smartphones. Due to the active used of smartphones, the floodgates of mobile malware threats are open every single day. Mobile malware harms users by illegally disable a mobile device, allowing malicious user to remotely control the device and steal personal information stored on the device. One of the surveillance features that attackers could abuse to gain those benefits is by exploiting the SMS message. Therefore, this paper introduces a new system call classification for SMS exploitation using a covering algorithm. The new system call classification can be used as a guidance to defend against mobile malware attacks. 1260 malware samples related to SMS exploitation from the Android Malware Genome Project have been analysed. The experiment was conducted using the dynamic analysis and open source tools in a controlled lab environment.

Keywords Android · Malware · Surveillance · Exploitation · System call · SMS message · Strace · Smartphone

M. Mohd Saudi (✉) · M.Z. Abd Rahman · A.A. Mahmud · N. Basir · Y.S. Yusoff
Faculty of Science and Technology (FST), Universiti Sains Islam Malaysia (USIM), 71800 Bandar Baru Nilai, Negeri Sembilan, Malaysia
e-mail: madihah@usim.edu.my

M.Z. Abd Rahman
e-mail: zuhairabdrahman@gmail.com

A.A. Mahmud
e-mail: azirah.alfaizahm@gmail.com

N. Basir
e-mail: nurlida@usim.edu.my

Y.S. Yusoff
e-mail: suhaylah@usim.edu.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_10

1 Introduction

As the popularity of smartphone increased, it grew up together with lots of new functions [1]. Nowadays, smartphone is use to surf social networking sites and for online banking transaction. Users tend to keep their confidential information such as contacts, password, bank account number, credit card number and other valuable information in their smartphone. Unfortunately, as the usage of smartphone is increasing, so does the exposure to mobile malware infection. Therefore, smartphone becomes the main target for the attackers to steal confidential information. In 2013, 804 new families or variants were accounted from Android platform 50 % increase from previous year and the remaining 23 new threats were targeted against Symbian [2]. From Fig. 1, it can be concluded that mobile threats in Android and Symbian platforms are increasing. AndroidOS_Smszombie.A is an example where Android malware exploits china mobile via SMS payment. It exploits smartphone by camouflage itself as a wallpaper application [2].

Android currently is evolving as one of the most prominent open source platforms for mobile devices like smartphones, notebooks and tablets. It is not just an operating system but also a complete software stack that includes application framework, libraries and some key applications. Android architecture is made up of different components and layers, which consists of application, libraries, Android runtime and linux kernel. The objectives of this research paper are to examine and evaluate how mobile malware uses system call which is an interface showing request from kernel in exploiting SMS messages and to produce a system call classification that consist of a compilation of mobile surveillance exploitation using system call for SMS message.

This paper is organised as follows. Section 2 presents the related works with mobile malware detection techniques. Section 3 explains the methodology used in this research paper which consists of dynamic analysis, how the system call analysis was conducted and the lab architecture. Section 4 presents the research findings which consists the results of the proposed system call classification. Section 5 concludes and summarises the future work of this research paper.

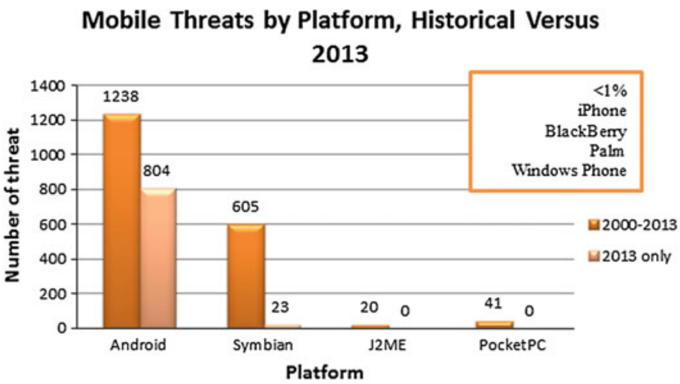


Fig. 1 Mobile threats by platform, historical versus 2013

2 Related Works

Malware is generally known as a malicious code and is able to do anything in other programs such as writing messages, modifying file, stopping a running program and other functions [3]. Mobile malware is known as malicious software that is specifically built to attack mobile phone or smartphone systems in order to damage or disrupt the device. To infiltrate and spread out the mobile malware, attackers used certain techniques to infect users' smartphones. Papers by [4–6] highlighted some of the techniques which are commonly used by attackers to install the mobile malware which are the repackaging technique, update attack technique and drive by download technique. Repackaging is a technique where attackers disassemble benign or popular applications, insert malicious code into these applications and reassemble them back. Then these malicious applications will be uploaded in Google Play (Android) or other third party applications market. Update attack technique is quite similar with the repackaging technique, except that when the update attack technique only includes an update component in its package, which is relying on the Internet connection. Once the Internet connection is available, users will manually or automatically fetch or download malicious payload at runtime and avoid the anti-virus static scanning. In contrast with drive by download attack technique, it persuades users to download these applications via spams or advertising. User will download the malicious application once they visit the site. Therefore, based on these mobile malwares infection techniques, a few techniques have been developed to detect these threats.

Previously, several approaches for mobile malware detection techniques that are related and similar to this research paper have been done by [7–9]. In these papers, they had highlighted some of the common techniques used for mobile malware detection as summarised in Table 1. According to Tchakounte and Dayang [6], system call is a native code calls used to directly invoke native functionalities of the Kernel. There are thousand of system call per application, but this research paper is focusing on system call that generates bad activities in the background.

Blasing et al.'s technique in detection is good as the measurement is diverse and is not dependent on parameters but only on proof of concept level [8] while Burguera and Zurutuza implemented dynamic analysis and manage to detect and differentiate genuine application compared to malicious application [7]. However, the solution is prone to false-positive detection if the application make used of less system call.

Surveillance can be defined as close observation by a suspected criminal or spy. In this situation, victims are being watched in every action taken [10]. Android smartphone is one of the devices where attackers could exploit its surveillance features. There are five surveillance features of smartphone that attackers can be exploited which are; audio, camera, call logs, location and SMS messages. Audio and camera features can be exploited when malicious applications are launched in Android smartphone. It may turn on the microphones and camera without users realizing it [11]. Meanwhile, call logs feature is being exploited when the

Table 1 Related works of malware detection on android

Related work	Detection analysis	Key feature	Detection technique	Challenges for improvement	Application/classification created
Blasing et al. [8]	Static and dynamic analyses	System call	Monitors system call that logs the return value of each system call independently to the parameters	Low detection accuracy	AASandbox
Burguera and Zurutuza [7]	Dynamic analysis	System call	Obtain traces of application's behaviour	Small dataset for training	CrowDroid
Wu et al. [9]	Static analysis	Manifest file and API calls	Detects malware through the manifest file and trace the API calls	Limited to certain android malwares	DroidMat
This research paper	Dynamic analysis	System call	Based on 5 main system call used for surveillance and rule induction using covering algorithm	Focusing on SMS messages	New classification for system call for SMS exploitation prior application formation

smartphone can make calls without users' even trigger it. On the other hand, SMS messages are being exploited when messages are sent without user's consent. As for location, it can be exploited when attackers could trace user location based on Global Positioning System (GPS) location.

For this research paper, it has some similarities with [7], which used dynamic analysis to analyse the behaviour of Android applications and overcome the challenges for work done by Blasing et al. [8] and Wu et al. [9]. Furthermore, this research is focusing on the system call used in exploitation of SMS messages only.

3 Methodology

In order to produce a new system call classification for SMS exploitation, the researchers has conducted few experiments and researches. A controlled laboratory environment is created to conduct the experiment as illustrated in Fig. 2 and almost 80 % of the software used in this testing is an open source or available on a free basis as summarised in Table 2. By using Strace tool, system call is generated from Android samples and is examined to outline system call that may exploit SMS message features. It allows learning application behaviour effectively through system call on Android as it is based on Linux [12]. All of the system call generated from the dataset is collected and tabulated. Malicious dataset is installed in the Android emulator. The dataset is downloaded from Android Malware Genome Project [1] which consists of different types of mobile malwares. The dataset

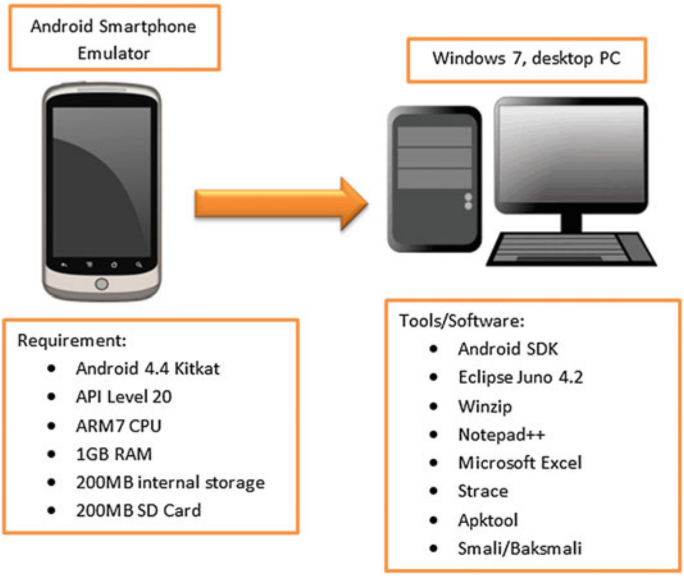


Fig. 2 Controlled laboratory architecture

Table 2 Software used in the lab

Software	Function
Notepad++	To conduct the dynamic analysis
Android SDK	To conduct the dynamic analysis
Winzip	To unzip compressed file
Strace	To learn application behaviour effectively through systems call
Eclipse Juno 4.2 Apktool	To conduct the dynamic analysis
Smali/Baksmali	To decompile apk resource file into a folder
	To disassemble the dex files

consists of 1260 Android malware samples that come from 49 different malware families.

The process of analysis is built on the following steps:

- Start the Android Virtual Device from the SDK.
- Install the binaries (*adb install xxxxx.apk*).
- Use of Android Debug Bridge (ADB) to emulate the device (*adb shell*).
- Identify and retrieve the parent process of the application installed (*ps*).
- Entry point to trace system call of the process (*strace—p pid*).

The result of occurrence for each system call is tabulated and labelled as 1 and 0. 1 indicate the presence of the system call in the application sample while 0 indicates that the system call is absent in the application sample. The result obtained earlier is calculated and the percentage of occurrence of each system call is recorded to

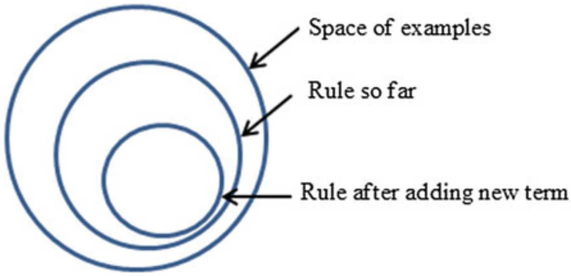


Fig. 3 The instance space during operation of a covering algorithm

compare their existence in each application sample. It is easier to summarize the data because their patterns can be seen and analysed easily.

To collect the system call from the dataset, Covering Algorithm has been used as the basis. Covering Algorithm operates by adding tests into the rules that is under construction to create a rule with maximum accuracy [13]. It involves finding an attribute to split on but the criterion for the best attribute is an attribute-value pair to maximize the probability of the desired classification.

Referring to Fig. 3, the space containing all the instances, a partially constructed rule, and the same rule after a new term has been added. The new term restricts the coverage of the rule: The idea is to include as many instances of the desired class as possible and exclude as many instances of other classes as possible [13]. For this research, the *specific to general rule* is used and summarized as the following.

- The instances are being picked up and generalised by repeatedly dropping condition.
- Let R be a new rule, if all instances covered by R then removed it and continue until all instances are covered.

In this research paper, the covering algorithm is used to develop a new classification of the system call for the SMS exploitation.

4 Findings

Based on the testing conducted, thousands of system call has been retrieved. It is not feasible to enumerate all of the system call but the focus is on the system call that generates bad activities in the background. There are 68 system calls that have been captured from the execution as summarized in Fig. 4.

By examining the list of the system call generated, there are several system call which could be used by the attackers to exploit the SMS messages. From this exploitation, attackers can gain money as victim might send SMS message without noticing it. The victim would be subscribed to the premium-rate SMS message which could lead to loss of money. By reviewing each function of the system call (Linux documentation), two system call are used to send message from a socket

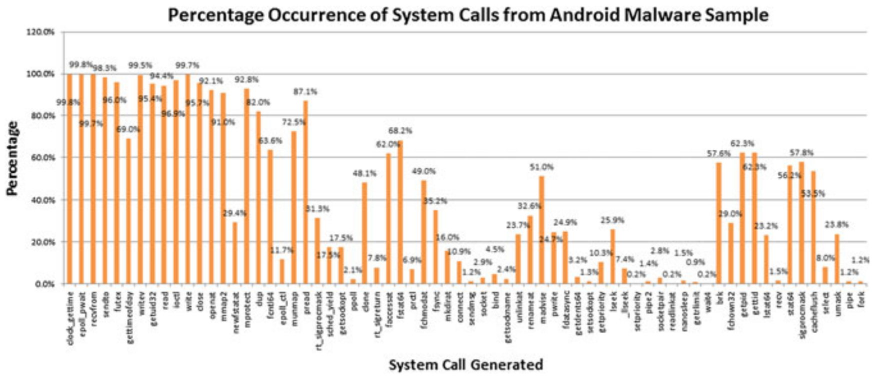


Fig. 4 Percentage occurrence of system call from android malware sample

which are *sendto()* and *sendmsg()*. However, *sendto()* system call has percentage of occurrence 98.3 % which is almost 100 %. This means *sendto()* system call exists in every malware samples examined.

Besides, this system call will be generated when user interacts with an application at most of the time. Therefore, this system call is not intended to exploit SMS message. However, *sendmsg()* system call has only 1.2 % of occurrence, which is hardly to be found in the malware samples. This system call is the most likely system call that attackers could use in exploiting SMS message. From the 1260 malware samples executed, there are only 15 samples contained this system call.

For evaluation, further inspection on those samples are conducted to prove that *sendmsg()* system call is used in exploiting SMS message. Based on Fig. 5, there

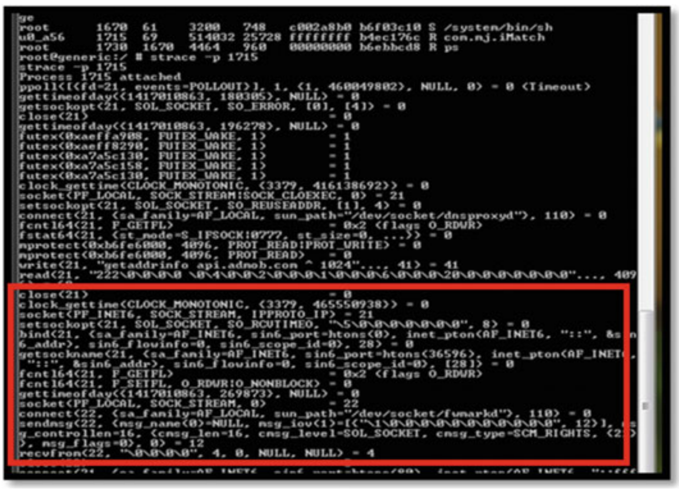


Fig. 5 System call generated from iMatch application

are several system call generated to exploit SMS messages from the iMatch application. The system call generated are *clock_gettime()*, *socket()*, *setsockopt()*, *bind()*, *getsockname()*, *fcntl64()*, *gettimeofday()*, *connect()*, *sendmsg()*, and *recvfrom()*. It can be concluded that these system call are used to exploit the SMS messages because these system call are executed at one time within system call *close(21)* and *close(22)*. As stated earlier, *close()* system call is use to close a file descriptor. Therefore, system call executed within *close()* system call frame is one complete process of an application. The process of exploiting SMS message starts at first by creating an endpoint of communication at socket 21 followed by binding a name to socket 21 and getting socket 21 name. After that, it initiates a connection to socket 22 and send message from that socket and by right, the message is received from socket 22. This is the flow on how the exploitation of SMS message features in iMatch application occurred.

While from Battery Saver application, system call generated to exploit SMS message are *clock_gettime()*, *socket()*, *bind()*, *getsockname()*, *connect()*, *sendmsg()*, and *recvfrom()* only, as illustrated in Fig. 6. The process of exploiting SMS message starts by creating an endpoint of communication at socket 29 followed by binding a name to socket 29 and getting socket 29 name. After that, it initiates a connection to socket 30 and send message from that socket and then the message is received from socket 30. This is the flow on how the exploitation of SMS message features in Battery Saver application occurred.

Based on the analysis obtained from the two applications earlier, two patterns of system call list are identified as shown in Fig. 7. From the two patterns identified, a main list of bad system call attackers used in exploiting SMS messages is developed. This main list is generated by eliminating the same system call found in all classifications and in the normal system call where the percentages of occurrence is

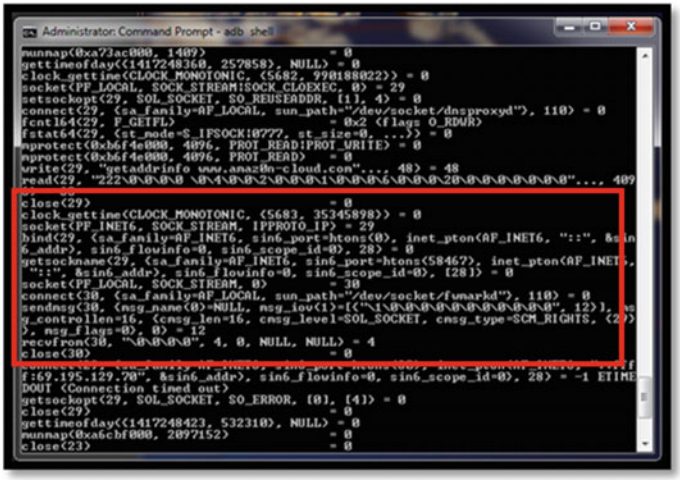


Fig. 6 System call generated from battery saver application

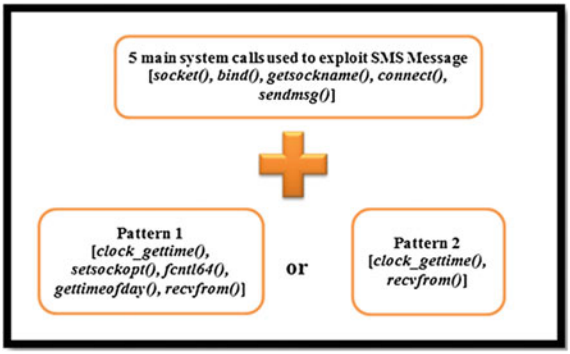


Fig. 7 Classification of system call for exploiting SMS messages

more than 50 % in the malicious sample. Therefore, the main list is the system call that attackers would invoke in any kind of application in order to exploit SMS message features.

The above classification is based on concept of the covering algorithm (specific rule induction) as explained in Sect. 3. Therefore, in order to determine whether an application is being exploited its SMS message, it should have either one of these combinations of system call to prove them.

(1) *socket(), bind(), getsockname(), connect(), sendmsg().* **or**
(2) *[socket(), bind(), getsockname(), connect(), sendmsg()*] + *clock_gettime(), setsockopt(), fcntl64(), gettimeofday(),* and *recvfrom*(). **or**
(3) *[socket(), bind(), getsockname(), connect(), sendmsg()*] + *clock_gettime()* and *recvfrom().*

5 Conclusion

As a conclusion, this research has managed to produce a new system call classification for SMS exploitation. The result can be used as a reference and comparison by other researchers with the same interests. For future work, the dataset will be transformed into nominal data for further analysis using different machine learning algorithms. This paper is part of a larger project to build up an automated mobile malware detection model. Ongoing research will include other mobile malware system call classification and the development of software to automate the mobile malware detection.

Acknowledgements The authors would like to express their gratitude to Universiti Sains Islam Malaysia (USIM) and Islamic Science Institute (ISI), USIM for the support and facilities provided. This research paper is supported by Ministry of Higher Education Malaysia under FRGS grant [FRGS/1/2014/ICT04/USIM/02/1] and Universiti Sains Islam Malaysia grant [PPP/FST/SKTS/30/12712] and also [USIM/RAGS/FST/36/51013].

References

1. Zhou, Y., Jiang, X.: Dissecting android malware: characterization and evolution. In: IEEE Symposium on Security and Privacy (SP), pp. 95–109 (2012)
2. Pan, B.: Android malware exploits china mobile SMS payments. http://blog.trendmicro.com/trendlabs-security-intelligence/android-malware-exploits-china-mobile-sms-payments/. Last Accessed 28 Jan 2015
3. Ramu, S.: Mobile malware evolution, detection and defense. EECE 571B, Term Survey Paper, April 2012, University of British Columbia (2012)
4. Zhou, Y., Wang, Z., Zhou, W., Jiang, X.: Hey, you, get off of my market: detecting malicious apps in official and alternative android markets. In: Proceedings of the 16th Network and Distributed System Security Symposium, NDSS (2012)
5. Abdullah, Z., Saudi, M.M., Anuar, N.B.: Mobile malware detection: proof of concept. In: 3rd International Conference of Software Engineering and Computer Systems 2013 (ICSECS'13), Universiti Malaysia Pahang (2013)
6. Tchakounte, F., Dayang, P.: System calls analysis of malwares on android. Int. J. Sci. Technol. **2**(9) (2013)
7. Burguera, I., Zurutuza, U.: Crowdroid: behavior-based malware detection system for android. In: Proceedings of the 1st ACM workshop on Security and privacy in smartphones and mobile devices (SPSM '11) (2011)
8. Blasing, T., Batyuk, L., Schmidt, A. D., Camtepe, S. A., Albayrak, S.: An android application sandbox system for suspicious software detection. In: 2010 5th International Conference on Malicious and Unwanted Software (MALWARE), pp. 55–62 (2010)
9. Wu, D.J., Mao, C.H., Wei, T.E., Lee, H.M., & Wu, K.P.: Droidmat: android malware detection through manifest and API calls tracing. In: 2012 Seventh Asia Joint Conference on Information Security (Asia JCIS), pp. 62–69 (2012)
10. Wong, C. (2014). Smartphone location-based services in the social, mobile, and surveillance practices of everyday life
11. Techdirt. (2011). Url:https://www.techdirt.com/blog/wireless/articles/20110417/21485513927/smartphone-apps-quietly-using-phone-microphones-cameras-to-gather-data.shtml, Last Accessed 28 Jan 2015
12. DieNet. (2014). Strace linux man page. Url:http://linux.die.net/man/1/strace. Last Accessed 28 Jan 2015
13. Witten, I.H., Frank, E.: Data Mining Practical machine Learning Tools and Techniques, 2nd edn. (pp. 105–111) (2009)

A Systematic Review Analysis of Root Exploitation for Mobile Botnet Detection

Hasan Al-Banna Hashim, Madihah Mohd Saudi and Nurlida Basir

Abstract Nowadays, mobile botnet has become as one of the most dangerous threats for smartphone. It has the capabilities of committing many criminal activities, such as remote access, Denial of Service (DoS), phishing, spreading malwares, stealing information and building mobile devices for illegitimate exchange of information and it is crucial to have an efficient mobile botnet detection mechanism. Therefore, this research paper presents a systematic review analysis of root exploitation for mobile botnet detection and a proof of concept how the mobile botnet attacks. This proof of concept includes analysis of mobile botnet sample using reverse engineering technique and static analysis.

Keywords Mobile botnet · Root exploitation · Reverse engineering · Android

1 Introduction

The communication device has undergone rapid development in telecommunication industry, from basic cellular phone to advance device called smartphone. Smartphones are able to store some users' confidential information from various services like gaming, social networking, internet browsing and online banking. Therefore, hackers have taken this opportunity to gain and exploit users' information whereby security issues are taken less seriously by the owner of the devices.

H.A.-B. Hashim · M. Mohd Saudi (✉) · N. Basir
Faculty of Science and Technology (FST), Universiti Sains Islam Malaysia (USIM), 71800 Bandar Baru Nilai, Negeri Sembilan, Malaysia
e-mail: madihah@usim.edu.my

H.A.-B. Hashim
e-mail: hasanhashim.fst@gmail.com

N. Basir
e-mail: nurlida@usim.edu.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_11

There are many types of malware that attack and threaten smartphone users such as viruses, Trojans, worms and botnet [1]. Botnet is known as the most dangerous malware as they pose serious threats to smartphones and networks [2, 3]. The term bot refers to an infected device, while botnet is a group of infected bots. Command and Control (C&C) server is usually used for the communication between botnet and bot-master. The main difference between botnet and other threats lies in the fact that they are dynamically controlled by a sophisticated bot-master [4]. Figure 1 shows the mobile botnet architecture.

According to the F-Secure survey, the mobile malware threat families have risen up to 26 % in year 2013. The survey also stated that for every five malwares threats, there is one mobile botnet threat. Relatively, mobile botnet is a new research domain which constitutes a number of problems [5].

Recently, a number of mobile botnets have evolved and capable of exploiting root privilege of Android based device. DroidDream is a botnet that masquerade as legitimate applications which exploit root privilege [6], while RootSmart botnet evade the detection by using the same Android default setting application icon before it exploit root privilege [7]. Similarly, GingerMaster is repackaged into

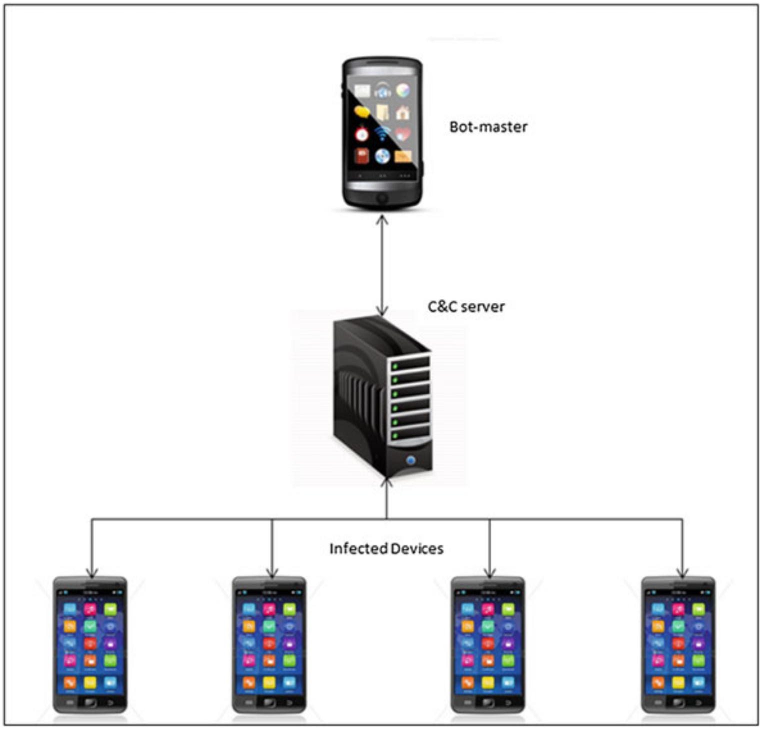


Fig. 1 Mobile botnet architecture

popular legitimate applications which contains root exploit to attract user downloads [8].

In year 2014, billions of Android devices from previous to KitKat version are vulnerable to privilege escalation attacks. The security vulnerability (CVE-2014-7911) allows attacker to bypass the Address Space Layout Randomization (ASLR) defence and execute arbitrary code of their choice on a target under certain circumstances [9]. Another disclosed vulnerability (CVE-2014-3153) which also present in most versions of Android could give attackers the ability to acquire root access on affected devices [10]. This vulnerability exists in Android 4.4 and earlier. The objective of this research paper is to study mobile botnet mechanism in exploiting root privilege and send the information to the C&C server.

This paper is organised as follows. Section 2 presents the related works with Android security mechanism, root privilege exploitation, mobile botnet detection and apoptosis. Section 3 explains the methodology used in this research paper which consists of controlled laboratory architecture, tools and dataset used in research. Section 4 presents the research finding which consists of proof of concept of mobile botnet implication. Section 5 concludes and summarises the challenges and future work of this research paper.

2 Previous Works

Android operating system is a Linux-based system where programs and systems are executed within their own processes. Various functions provided by Android such as user and group control, efficient memory management, and pre-emptive multitasking. Furthermore, Android application privilege mechanism prevents Java application from abusing the system and resources. Developers are forced to sign their application programs to distribute them [11].

Android prevents users from using root privilege directly to protect the system against reckless user activities. The protection mechanism implemented in Android is based on standard Linux protection mechanism. This mechanism includes assigning user-ID (UID) or group-ID (GID) to the application. Further protection is implemented via permission framework that restricts specific operation of processes. During application development, the permissions are specified in AndroidManifest.xml file. Once the development is done, these permissions cannot be changed. Users will be informed regarding these permissions during installation process. If users agree then the application is installed and works [12].

However, recent attacks show that Android security framework is not sufficient for transitive policy enforcement, allowing privilege escalation attacks [13]. One of the techniques for acquiring privilege is through rooting. Rooting is defined as unlocking the operating system that allows user to install any unofficial application, modify the operating system, replace firmware and overclock the processor [14]. According to Vidas et al. [15] rooting is a method of exploiting the vulnerability of Android platform to gain administrative privilege over the system. It includes

Table 1 Related works for mobile malware detection

Related work	Detection analysis	Key feature	Detection technique	Challenges for improvement
Bläsing et al. [16]	Static and dynamic analysis	System calls	Monitors system calls that logs that return value of each system call independently to the parameters	Low detection accuracy
Wu et al. [17]	Dynamic analysis	Manifest file and API Calls	Detects malware through API calls and manifest file	Cannot detect malware that use update attack technique
Feizollah et al. [18]	Dynamic analysis	Network anomaly	Obtain the network traffic of mobile botnet.	Small dataset for training
This research paper	Static and dynamic analysis	Manifest file, API calls and system calls	Based on the parameters in key feature and rule induction using apoptosis	Focusing on root exploitation

attacks on Linux kernel, daemons and services. Normal user can obtain root privilege once the device is rooted. The security mechanisms in Android become useless after the root exploitation is done.

Previously, several studies on mobile botnet detection have been conducted by other researchers [16–18]. In these papers, they had highlighted several common techniques used for mobile botnet detection as summarised in Table 1.

In this research, apoptosis mechanism will be implemented for the detection process. Apoptosis is a process of common self-destruction mechanism where the cells have the ability to conduct self-destruction in biological systems [19]. Other researchers have investigated the Apoptotic Computing model and Apoptosis concept in computing system. Mohd Saudi et al. [20] have developed a system that employed apoptosis which control and isolate a computer attacked by worm called STAKCERT. The STAKCERT framework operates in two phases; phase one utilises a set of algorithms to detect and identify a worm, whilst phase two utilises apoptosis to separate the computer from the network to prevent the worms from spreading. Thus, the system follows the apoptosis metaphor of self-destruct process. In Jones master's dissertation, the researcher has implemented apoptotic self-destruct and stay-alive signalling specifically investigating memory requirements of inheritance vs an abstract oriented approach (AOP) [21]. Besides that, a research by Lv et al. [19] has proposed a self-destruction model which inspired by apoptosis. The model is described by Markov Regeneration Stochastic Petri Net (MRSPN) to provide an approach of quantification analysis.

Several studies have been done previously by other researchers which implement biological mechanism [22–24]. Table 2 summarised the comparison between apoptosis and other algorithms (Table 3).

Table 2 Comparison between apoptosis with other algorithms

Related work	Algorithm used	Strength	Weakness
Yusof and Jantan [22]	Genetic algorithm	The detection accuracy increased using AMS classifier	The class that authors chose for classification is overlapped which caused bottleneck in memory usage
Zhou et al. [23]	Artificial immune system	High detection rate	Implemented in custom modified android only
Zhou et al. [24]	Support vector machine (SVM)	High detection rate with low false negative rate	High battery consumption
This research paper	Apoptosis	High detection accuracy rate and low battery consumption	–

In this research paper, the mobile botnet sample was reverse engineered and followed by static analysis, prior the formation of new mobile malware detection model. Furthermore, this research is focusing on the root exploitation.

3 Methodology

Lab architecture with an isolated environment (no outgoing connection) was set up for reverse engineering process. Figure 2 shows the controlled lab environment. In this lab environment, open source applications were used as displayed in Table 3. The training dataset in this research consists of different types of mobile malware and was downloaded from Android Malware Genome Project [25, 26] while for the testing and evaluation, this experiment used dataset from Drebin [27]. The static analysis was conducted in this lab environment where, the files and its content associated with the mobile botnet were inspected without running the application. In this research paper, MisoSMS sample is used for the proof of concept as a preliminary study. Static analysis was conducted on the sample to examine the files and its content associated with botnet activities.

Table 3 Software used in the lab

Software	Function
VMWare	To build up virtual operating systems in a computer
ApkTool	To decompile the APK file
Dex2Jar	To convert DEX file to Jar file
JD-GUI	To view the application code
Leafpad	To conduct the static analysis
Android SDK	To run Android simulation

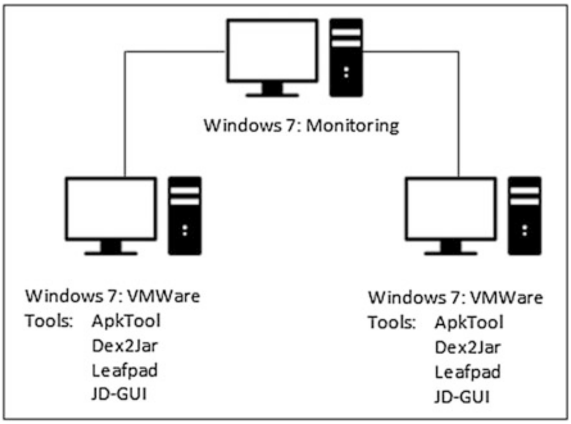


Fig. 2 Mobile malware controlled lab architecture

4 Preliminary Study

A case study using MisoSMS sample was carried out to study the way mobile botnet exploit Android root privilege and SMS (Short Message Service) service. Reverse engineering process and static analysis were applied to analyse the code using the same architecture in methodology. Reverse engineering is a process where technological principles of system or device are discovered by analysing its structure, function and operation. The advantages of reverse engineering are the ability to understand the function of program and learn the files that target the program access. Thus, the reverse engineering is beneficial to analyse malware characteristics and behaviours.

Based on the experiment conducted, the finding shows that this mobile botnet has the ability to steal SMS messages and send it to the C&C server. The application is disguised as a legitimate Google setting application but slightly different in its name “google Vx”. A popup will appear that ask user for administrative permission when the user click on the application as shown in Fig. 3. Once user click “Activate” button, the application will get the root privilege without user consent. The author of this malware uses social engineering technique to obtain the root privilege.

Further analysis revealed that this mobile botnet request many sensitive permissions in AndroidManifest.xml as shown in Fig. 4. The main activity of this malicious application is to record all SMSs’ in user device and send it to the C&C server. Normally, the attacker would use SMS forwarding as a mechanism to send the messages. However, MisoSMS uses different technique to send messages to the attacker which using webmail account through SMTP connection. Figure 5 shows the source code how this application send recorded SMS to the C&C.

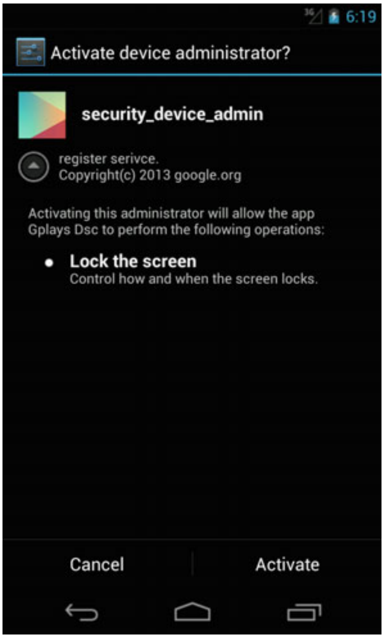


Fig. 3 Administrative permission request

```
<uses-sdk android:minSdkVersion="8" android:targetSdkVersion="17" />
<uses-permission android:name="android.permission.RECEIVE_BOOT_COMPLETED" />
<uses-permission android:name="android.permission.RECEIVE_SMS" />
<uses-permission android:name="android.permission.WRITE_SMS" />
<uses-permission android:name="android.permission.READ_SMS" />
<permission android:name="android.permission.READ_SMS" />
<uses-permission android:name="android.permission.SEND_SMS" />
<uses-permission android:name="android.permission.INTERNET" />
<uses-permission android:name="android.permission.ACCESS_WIFI_STATE" />
<uses-permission android:name="android.permission.CHANGE_WIFI_STATE" />
<uses-permission android:name="android.permission.WRITE_EXTERNAL_STORAGE" />
<uses-permission android:name="android.permission.MOUNT_UNMOUNT_FILESYSTEMS" />
<uses-permission android:name="android.permission.READ_PHONE_STATE" />
<uses-permission android:name="android.permission.ADD_SYSTEM_SERVICE" />
<uses-permission android:name="android.permission.RECEIVE_MMS" />
<uses-permission android:name="android.permission.RECEIVE_WAP_PUSH" />
<uses-permission android:name="android.permission.READ_CONTACTS" />
<uses-permission android:name="android.permission.ACCESS_NETWORK_STATE" />
```

Fig. 4 Sensitive permissions requested by application

Based on the experiment outcome, the mobile botnet activities were identified using reverse engineering and static analysis. Therefore, more samples are needed to be investigated to form a new mobile botnet classification before developing a mobile botnet detection mechanism.

```
public class MisoSendMessage {
  public MisoSendMessage() {
    super();
  }

  public void Action() {
    MisoData.request_data.rec_phone_list.clear();
    MisoData.request_data.data_head.sms_rec_phone_count = 0;
    MisoData.request_data.data_head.sms_task_id = MisoData.replay_data.data_head.sms_task_id;
    int v0;
    for(v0 = 0; v0 < MisoData.replay_data.data_head.sms_task_rec_phone_count; ++v0) {
      RequestStruct_RecPhoneInfo v2 = new RequestStruct_RecPhoneInfo();
      Object v1 = MisoData.replay_data.rec_phone_list.get(v0);
      this.sendSms(((ReplayStruct_RecPhoneInfo)v1).rec_phone_num, MisoData.replay_data.data_head.sms_task_content);
      v2.recphoneid = ((ReplayStruct_RecPhoneInfo)v1).rec_phone_id;
      v2.reccode = 3;
      MisoData.request_data.rec_phone_list.add(v2);
      MisoData.request_data.data_head.sms_rec_phone_count = ((short)(MisoData.request_data.data_head.sms_rec_phone_count + 1));
      SystemClock.sleep(500);
    }
  }

  private void sendSms(String paramString1, String paramString2) {
    try {
      Iterator v7 = SmsManager.getDefault().divideMessage(paramString2).iterator();
      while(v7.hasNext()) {
        SmsManager.getDefault().sendTextMessage(paramString1, null, v7.next(), null, null);
      }

      return;
    }
```

Fig. 5 Reverse engineered of MisoSMS code

5 Conclusions and Future Work

The popularity and functionality of mobile devices not only attract user but also attacker. Mobile devices such as smartphones can be infected by malware and turn this devices into botnet which can be manipulated for cyber-criminal activities. The current solutions for mobile botnet detection are still lacking in term of detection accuracy and continual improvement is needed. Therefore, it is important to conduct more research on mobile botnet classification and detection mechanism. The motivation to pursue research in this area is to provide high accuracy and efficient mobile botnet detection model. This paper presented how root privilege being exploited by mobile botnet and is part of larger project to confront mobile botnet attacks. Ongoing research includes mobile botnet classification and the development of software base detection for mobile botnet attacks.

Acknowledgments The authors would like to express their gratitude to Universiti Sains Islam Malaysia (USIM) and Institute Science Islam (ISI) for the support and facilities provided. This research paper is supported by Ministry of Education grants [USIM/FRGS/FST/32/50114] and [USIM/RAGS/FST/36/51013].

References

1. Eslahi, M., Salleh, R., Anuar, N.: Bots and botnets: an overview of characteristics, detection and challenges. In: 2012 IEEE International Conference on Control System, Computing and Engineering (2012)
2. La Polla, M., Martinelli, F., Sgandurra, D.: A survey on security for mobile devices. IEEE Commun. Surv. Tutorials **15**(1), 446–471 (2013)
3. Pieterse, H., Olivier, M.: Android botnets on the rise: trends and characteristics. In: Information Security for South Africa (2012)
4. Hachem, N., Ben Mustapha, Y., Granadillo, G., Debar, H.: Botnets: lifecycle and taxonomy. In: Conference on Network and Information Systems Security (2011)
5. Karim, A., Ali Shah, S. Salleh, R.: Mobile botnet attacks: a thematic taxonomy. In: Rocha, Á., Correia, A., Tan, F.B. Stroetmann, K.A. (ed) New Perspectives in Information Systems and Technologies, vol. 2, 1st edn., pp.153–164. Springer International Publishing, Portugal. (Online) Available at: http://www.springer.com/gp/book/9783319059471 (2014). Accessed 9 Apr 2015
6. Strazzere, T.: Update: Android Malware DroidDream: How it Works|Lookout Blog. (online) Blog.lookout.com. Available at: https://blog.lookout.com/blog/2011/03/02/android-malware-droiddream-how-it-works/ (2011). Accessed 9 Apr 2015
7. Symantec.com: Android.Bmaster|Symantec. (online) Available at: http://www.symantec.com/security_response/writeup.jsp?docid=2012-020609-3003-99 (2015). Accessed 9 Apr 2015
8. Jiang, X.: GingerMaster. (online) Csc.ncsu.edu. Available at: http://www.csc.ncsu.edu/faculty/jiang/GingerMaster/ (2015). Accessed 9 Apr 2015
9. Wei, W.: Billions of Android Devices Vulnerable to Privilege Escalation except Android 5.0 Lollipop. (online) The Hacker News|Biggest Information Security Channel. Available at: http://thehackernews.com/2014/11/billions-of-android-devices-vulnerable.html (2015). Accessed 9 Apr 2015
10. Donohue, B.: Android root access vulnerability affecting most devices|Threatpost|The first stop for security news. (online) Threatpost.com. Available at: https://threatpost.com/android-root-access-vulnerability-affecting-most-devices/106683 (2014). Accessed 9 Apr 2015
11. Enck, W., Ongtang, M., McDaniel, P.: Understanding android security. IEEE Secur. Priv. Mag. **7**(1), 50–57 (2009)
12. Park, Y., Lee, C., Lee, C., Lim, J., Han, S., Park, M., Cho, S.: RGBDroid: a novel response-based approach to android privilege escalation attacks. In: 5th USENIX Workshop on Large-Scale Exploits and Emergent Threats, p. 9 USENIX, USA. (online) Available at: https://www.usenix.org/system/files/conference/leet12/leet12-final14_0.pdf (2012). Accessed 9 Apr 2015
13. Bugiel, S., Davi, L., Dmitrienko, A., Fisher, T., Sadegi, A., Shastry, B.: Towards taming privilege-escalation attacks on android. In: NDSS. The Internet Society (2012)
14. Broida, R.: How to Easily Root an Android Device—CNET. CNET. (online) Available at: http://www.cnet.com/how-to/how-to-easily-root-an-android-device/ (2013). Accessed 9 Apr 2015
15. Vidas, T., Zhang, C., Christin, N.: Toward a general collection methodology for android devices. Digit. Invest. **8**, S14–S24 (2011)
16. Bläsing, T., Schmidt, A., Batyuk, L., Camtepe, S., Albayrak, S.: An android application sandbox system for suspicious software detection. In: 5th International Conference on Malicious and Unwanted Software (MALWARE), pp. 55–62. IEEE, Nancy (2010)
17. Wu, D., Mao, C., Wei, T., Lee, H., Wu, K.: DroidMat: android malware detection through manifest and API calls tracing. In: Seventh Asia Joint Conference on Information Security (Asia JCIS), pp. 62–69. IEEE (2012)
18. Feizollah, A., Anuar, N., Salleh, R., Amalina, F., Ma'arof, R., Shamshirband, S.: A study of machine learning classifiers for anomaly-based mobile botnet detection. Malays. J. Comput.

Sci. **26**(4), pp. 251–265. (online) Available at: http://mjcs.fsktm.um.edu.my/document.aspx?FileName=1449.pdf (2014). Accessed 9 Apr 2015
19. Lv, H., Wang, H., Feng, G., Zhu, Q.: Modeling and analysis of apoptosis-inspired software self-destruction. In: Qu, X., Yang, Y. (ed) Information and Business Intelligence, 1st edn., pp. 606–613. Springer, Berlin (2011)
20. Mohd Saudi, M., Cullen, A., Woodward, M.: Efficient STAKCERT KDD processes in worm detection. World Acad. Sci. Eng. Technol. **5**(7), 344–348. (online) Available at: http://scim.brad.ac.uk/staff/pdf/ajcullen/v79-89.pdf (2011). Accessed 9 Apr 2015
21. Jones, D. (2010). Implementing biologically-inspired apoptotic behaviour in digital objects: an aspect-oriented approach. Master Degree. Open University
22. Yusoff, M., Jantan, A.: Optimizing decision tree in malware classification system by using genetic algorithm. Int. J. New Comput. Archit. Appl. (IJNCAA) **1**(3), 694–713 (2011)
23. Zhao, M., Zhang, T., Ge, F., Yuan, Z. (2012). RobotDroid: A lightweight malware detection framework on smartphones. J. Netw. **7**(4), 715–722. (online) Available at: http://ojs.academypublisher.com/index.php/jnw/article/download/jnw0704715722/4621 (2012). Accessed 9 Apr 2015
24. Zhao, M., Zhang, T., Wang, J., Yuan, Z.: A smartphone malware detection framework based on artificial immunology. J. Netw. **8**(2), 469–476 (2013)
25. Xiang, C., Binxing, F., Lihua, Y., Xiaoyi, L., Tianning, Z.: Andbot: towards advanced mobile botnets. In: LEET '11 Proceedings of the 4th USENIX Conference on Large-Scale Exploits and Emergent Threats, p. 11. USENIX, California. (online) Available at: https://www.usenix.org/legacy/event/leet11/tech/full_papers/Xiang.pdf (2011). Accessed 9 Apr 2015
26. Xie, L., Zhang, X., Seifert, J., Zhu, S.: pBMDS. In: Proceedings of the Third ACM Conference on Wireless Network Security—WiSec '10, pp. 37–48. (online) Available at: http://www.cse.psu.edu/~szhu/papers/pBMDS.pdf (2010). Accessed 9 Apr 2015
27. Bose, A., Hu, X., Shin, K., Park, T.: Behavioral detection of malware on mobile handsets. In: MobiSys '08 Proceedings of the 6th International Conference on Mobile Systems, Applications, and Services, pp. 225–238. ACM, New York. (online) Available at: http://cse.ucdenver.edu/~tamvu/courses/fall2013/p225-bose.pdf (2008). Accessed 9 Apr 2015

Automatic Test Case Generation from UML State Chart Diagram: A Survey

Yasir Dawood Salman and Nor Laily Hashim

Abstract The need for developing high-quality systems and applications with the minimum faults and error in them has been increased recently. In addition, a question of time and expenses that should be as low as possible always is concerned. Thus, it needs to be updated with the testing techniques that are more structured and automated which are used during the analysis and design phase. The significant role of automated testing techniques is that it helps speeding up the delivery of services of the products to the market with little chances of loss, and increases the software value. If the purpose were to decrease the expenses and getting closer to technology, then testing automation would be a crucial choice. The aim of this survey is to improve the understanding of UML diagram based testing techniques. Test case generation from state chart has been have focused on. Also, classify the various research approaches to their methods. The issues of test coverage associated with these methods have been discussed also.

1 Introduction

Testing is considered an essential part of the today's software development and has proven to be a useful tool to enhance the programming coding quality. Testing can help to detect the software bugs which compilers are not usually able to detect [1]. However, whether a program is correct or not, its correctness can be guaranteed through using the testing in order to enhance the quality of its code. The task of testing in software is so complicated, that it includes the software evaluation to show whether it meets the needs. Theoretically or practically, this is usually a difficult task.

Y.D. Salman (✉) · N.L. Hashim
Universiti Utara Malaysia, Kedah, Malaysia
e-mail: yasir.dawod@gmail.com

N.L. Hashim
e-mail: laily@uum.edu.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_12

Come as no surprise that for the past decade a great amount of research work has been conducted over automatic test case generation [2–8]. Therefore, a great amount of different techniques to generate the test case has been explored intensively and propelled. In contrast, software systems have gotten to be progressively complex, for instance, with components developed by diverse vendors, utilizing distinctive techniques within diverse programming languages and actually running on diverse platforms.

State chart diagrams in UML can be used to construct the dynamic aspects of a system. This diagram consists of transitions, states, actions, and events [9] and by emphasizing the flow of control from state to state as it shows a state machine. State chart is comprehensive Finite State Machine (FSM) with concurrency, hierarchy, and communication, and these extensions allow small diagrams to express complex behavior in a modular method [10].

The purpose of generating test case using UML state chart diagram is to verify the relations between the behavior, state transition, state, action, and event. This technique is used to determine if one can fulfil the system specifications through the state based motion of the system. In the state based system, there are three reasons for the fault. The first is when the state diagram cannot accurately transfer the system function specification. Secondly is when the state chart diagram configuration is wrongly or unreliable. The final is when converting form a state chart diagram to programmable code [11]. Several of surveys and reviews study have been conducted on test case generation [12–18], where they cover in general the generation of test cases a variety of inputs. This paper on the other hand focuses on varies methods of test case generation from UML state chart diagram, where the commonalties and trends in the methods used are explored. The following section presents test case generation approaches using state chart diagram.

The paper is structured as follows. Section 2 a review related work of automatic test case generation using UML state chart diagram. Section 3 presents the trends and gaps identified from the survey. Finally, Sect. 4 outlines conclusion and future research work.

2 Test Generation Approaches Using UML State Chart Diagram

Researchers such as [19, 20] have paid considerable attentions to automatic test case generation from UML diagrams. At the same time, there were more researchers who work on generating test cases from UML state chart diagrams [21–23].

Kansomkeat and Rivepiboon [20] develop a transformation method from state chart diagrams into intermediate diagrams that are used to generate test sequences. The test cases are generated automatically from state chart diagrams created by the Rational Rose tool. The testing coverage criterion is used to guide the generation of test cases and to produce the intermediate model Testing Flow Graph (TFG) from the all-state coverage and all transition coverage. Based on their fault detection abilities, their test cases measured the effectiveness of test case generation. From the generated test cases, results of simple test experiments revealed high effectiveness in the test case generation. However, usually more than one object often participates in the execution of a use case. Therefore, it will be difficult to test using this approach with the chance of such behavior occurred. In addition, it does not generate multi test data due to the lack of coverage choices.

Offutt et al. [24] developed a method to automatically generate the test cases from state chart diagrams, by changing events for Boolean class attributes. The developments of many useful coverage criteria that were centered on the state chart diagrams were effective. Class-level testing is the aim of their approach. This method attains transition-pair coverage, transition coverage, and full predicate coverage.

Gnesi et al. [19] offered a formal test case generation by providing a mathematical basis for conformance testing and automatic test case generation for state chart diagrams that was established on an operational semantic. With transitions labelled by input/output pairs, they proposed a formal conformance testing relation for input enabled transition systems. In order to succeed in the specified requirements, testing software is identified as conformance testing. Considering the formal specification, a conformance relation defines the accuracy criterion of the implementation. However, in order to practice this technology proper test selection strategies are needed to use the test generation algorithm practically.

Briand et al. [25] focused on creating a methodology using state chart diagram to define the system state required for each event or transition, where parts of the paths, input values for the parameter for all actions and events associated with transitions have to be defined. Their work generates a test case specification involving a possible sequence of transitions. A requested sequence tree is constructed and then it will be used to develop the test restraints for the transition sequences to test it, to get the interactions among state dependent objects in their work.

Li and Lam [26] presented an approach to generate test sequences from state chart diagrams using ant colony optimization. A UML state chart diagram is transformed into intermediate model called a directed graph. By exploring the directed graph by a group of ants cooperatively, the test sequences are generated. From this generation they achieve the all-state coverage in the coverage criteria.

Santiago et al. [27] presented a method to automate test case generation from UML state chart diagrams using a software specification model. This method converts the UML state chart diagrams model into an XML-based language table, and by using the Perform Charts tool, and generates intermediate model from state based on control flow. Their indication is to determine that by using a higher-level technique, such as UML state chart diagrams; there will be possibility to represented complex software with clarity and rich details. UML state chart diagrams can enable to model a complex system more realistically and provide hierarchy and parallelism for it. Although these conditions are not enough in order to guarantee that a test case generation approach is successful, but still they show an improvement when they have been compared with the use of the Condado as an unconnected tool with a FSM specification. In addition, the Condado implements the switch cover method for the control part. A switch is a transition-to-transition pair, and their method generates test cases to cover all pairs of transitions in the model in the coverage criteria.

Murthy et al. [28] suggested a new foundation for generating test cases using the UML state chart diagram as a base model of behavior. They also defined a test ready state chart diagram, which indicates that the model is ready with data for a test generator to generate test scripts automatically from it. To generate the paths, they start from the start node with a state transition and reconnoitering each next node subsequent of its state transitions, if any along a state transition is satisfied, it provides guard condition. They solve the problem of generating test case from UML state chart diagram by defining all the sentential forms derivable from an equivalent extended context free grammar model. Additionally, coverage criteria that were achieved are path coverage and basic path coverage.

Ali et al. [21] have projected a method for state-based integration testing. Their work forms an intermediate test model named State Collaboration Test Model (SCOTEM) from the corresponding state chart diagrams and UML collaboration diagrams. SCOTEM copies all possible paths for object state changes that message sequences may cause. Then SCOTEM produces test paths centered on several coverage criteria. For them revealing the state-dependent interaction errors is the goal behind the generated test cases. Their work reflects the analysis of all possible states of cooperating levels in an interface.

Santiago et al. [29] presented an environment name automated Generated Test case based on State Charts (GTSC) which allows a test designer to generate test cases based on state charts test criteria and FSM methods. This interesting characteristic allows test sequence generation from both state chart and FSM techniques based on the same FSM. However, other comparisons needs to be made namely all-paths-k-C0-onfiguration of the state chart Coverage Criteria Family (SCCF) as well as the round-trip route testing offered by Binder [30] and all-paths-k-configurations. Similarly, there can be more comparisons between the latest FSM-based methods, such as state counting, and some SCCF criteria. Such an analysis will be enabled with the help of mutation testing by GTSC in applying these test criteria methods.

Kosindrdecha and Daengdej [22] proposed a new method to generate and prepare both test data and test case based on state chart diagram, called "TGfMMD" method. This method has been developed to verify the state chart diagram before generation of both test cases, and test data from extended state chart diagram. However, this method has not yet been tested with a complex state chart diagram.

Swain et al. [31] proposed a novel technique to generate test cases automatically from UML state chart diagram and activity diagram. They construct it based on the model an intermediate representation, that they named it state-activity diagram (SAD). They generate the test case from the use of SAD generation, Depth First Search (DFS) and mutation analysis. In addition, in order to detect harmonization of state chart diagram as well as activity diagram faults within a use case of the system exercise, an activity synchronization in the context of multiple state combinations has been used. They also achieves transition coverage and state/activity path coverage. For the testing, they have implemented a prototype tool based on their approach. However, in their work the tester should select the test data for each test case manually.

Shirole et al. [32] have also worked on the automatic generation of test case using UML state chart diagram. They used the Genetic Algorithm (GA) as medium for their tool by combining information from state chart diagram in it. They propose a search-based approach to handle infeasible paths and test data generation. They use the following steps to generate the test cases. First is to transform the UML Specifications into Extended Finite State Machine (EFSM), secondly to transform the EFSM into Extended Control Flow Graph, third is to generate test sequences using GA and DFS, and finally, select the test cases using data-flow techniques. In the coverage criteria, they focused on state cover, transition cover, all-definition cover, and all du-path. However, the state chart diagrams that they considered are very simple, what will lead to less coverage when dealing with scenarios that are more complex. Also because of using DFS and fitness function, all path coverage is not fully obtained.

Li et al. [33] presented a test case generation approach, which takes UML state chart diagrams as inputs. They first construct the state chart diagram conforming to system requirement. Then, analyze the .mdl file of state chart diagram, extract the main information of the state chart diagram and convert it to a directed graph. Finally, they designed an algorithm to construct the Euler circuit based on a directed graph and generate test cases automatically by Euler circuit algorithm. Their specified test coverage criteria are the state coverage and the transition coverage of state chart diagram, also to minimize the number of test cases.

In an earlier study Swain et al. [23] proposed an approach to generate test cases from UML state chart diagram. They have named their approach, Automatically Generating Test cases from State Chart Diagram (AGeTeSC). First, they have constructed the state chart diagram for a given object. Then the state chart diagram is traversed, conditional predicates are selected and these conditional predicates are transformed into source code. Then, the test cases are generated and stored by using function minimization technique. From the state chart diagram, they perform a DFS to select the associated predicates. After selecting the predicates, they guess an initial dataset. They have generated test predicate conditions from a state chart diagram, which are used to generate test cases. Their technique accomplishes little coverage in test case like transition pair coverage, state coverage, action coverage, and transition coverage. It also achieves fully predicate coverage by generating a test data for each conditional clause. Besides that, it can handle transitions with guards and achieves transition path coverage. Here the quantity of test cases is minimized and they reach transition path coverage in testing the limitations decided by simple predicates, but the test case needs to be optimized.

Additionally Swain et al. [34] proposed an approach for test cases generation named, Test Generation and Minimization for O-O software with State Charts

(TeGeMiOOSc). It starts by analyzing the system, which is going to be tested and accepted by user, then build the state chart diagram. After that, they convert the given UML state chart diagram into an intermediate model, that they named it a state transition graph. DFS is used to form test sequences and generating all the possible paths. Then obtain all the valid sequences of the application until final edge is reached. Finally, they minimize a set of test cases by calculating node coverage for each test sequence. In the same year a work of Swain et al. [35] has performed a similar experiment to generate test case from UML state chart diagram and they have named it, Generation and Minimization of test cases from State Charts (GeMiTefSc). Their approach at first build a state chart diagram model for SUT, next, they conjugated state transition graph from state chart diagram. Then, by using the graph, all the required information is extracted. Then, by applying Wang's algorithm [36] they generated the test cases. Finally, they minimized the set of test cases by calculating node coverage for each test case and this help them to determine which test case are covered by other test cases. However in their works, after creating the intermediate graph they rely on the DFS to generate the paths, what will lead to less in coverage when the state chart diagram have loops and feedbacks in it. Also by using minimization, they minimize a set of test cases what will cause to overlap or ignoring some of the important data therefore will have less coverage.

Chimisliu and Wotawa [37] in their earlier work have proposed a method for generating test cases automatically aiming at achieving transition coverage and state coverage of the model. Their proposed approach presents an automatically transformation of the system composed of communicating state charts diagram into a Language of Temporal Ordering Specification (LOTOS). They also showed how to generate test cases in semi-automatically way by making use of an input from the user as explanations on the UML diagram. In their work, generated test cases coverage criteria did not contain any reject transitions. Thus, the generation process was not as efficient as in the case when the user provides explanations that can be used as refuse transitions in the test purpose.

In their more recent work, Chimisliu and Wotawa [38] and Chimisliu and Wotawa [39] proposed an improved tool for test case generating from UML state chart diagram by using control, data and communication dependencies. They generated the test cases by using the Test Generation with Verification (TGV) technology [40], a test case generator from the analysis and the construction of distributed processes toolbox. For the coverage criteria, their generation technique aimed at achieving transition coverage only. Therefore, the lack of coverage will indicate to the need to enhance this method or obtaining a novel one.

3 Findings from the Survey

The following section presents the trends, gaps and commonalities found from the 21 studies that conducted test cases generation from UML start Chart diagram.

Table 1 reviews the past decade researchers and their studies. The input model column shows that the current researchers use state chart diagram or combinations of other diagrams and also the method they used to generate the test cases. Furthermore, the intermediate models that are generated as intermediary between

Table 1 Test case generated methods using UML state chart diagram

Author(s)	Input model	Method	Intermediate model	Coverage criteria
Kansomkeat and Rivepiboon [20]	State chart	Parsing TFG, mutation analysis	Testing flow graph (TFG)	State and transition
Offutt et al. [24]	State chart	Spec test	Specification graph	Transition coverage Full predicate coverage Transition-pair coverage Complete sequence
Gnesi et al. [19]	State chart	Input/output label transition systems (IOLTSs), random test selection	–	–
Briand et al. [25]	State chart	Normalization and analysis of operation contracts and transition guards	Invocation sequence tree (IST)	All transitions, all transition pairs, full predicate, and all round-trip paths
Li and Lam [26]	State chart	Ant colony optimization	Directed graph	All states
Santiago et al. [27]	State chart	Perform charts and Condado	FSM	All pairs of transitions
Murthy et al. [28]	State chart	Extended UML state chart model	Context free grammar model	Path coverage, Basic path coverage
Ali et al. [21]	Collaboration diagrams and state chart	State collaboration test model (SCOTEM)	Testing flow graph (TFG)	Single-path coverage All-transition coverage, N-path coverage All-path coverage

(continued)

Table 1 (continued)

Author(s)	Input model	Method	Intermediate model	Coverage criteria
Santiago et al. [29]	Finite state Machines and state Charts	Switch cover, distinguishing Sequence and unique Input/output methods	FSM	All transitions
Kosindrdecha and Daengdej [22]	State chart	TGfMMD method	Sketch diagram-based technique	All nodes
Swain et al. [31]	State chart and activity chart	SAD generation, DFS, mutation analysis	State activity diagram (SAD)	Transition coverage and activity path coverage
Shirole et al. [32]	State chart	Genetic algorithm	Extended control flow graph	State cover, Transition cover, All-definition cover, and All du-path
Li et al. [33]	State chart	Euler circuit algorithm	Directed graph	State coverage criteria, Transition coverage criteria
Swain et al. [23]	State chart	Depth first search (DFS), Model J unit	State chart graph	State coverage, Transition coverage, Transition pair coverage
Swain et al. [34]	State chart	Test generation and Minimization for O-O software with state charts (TeGeMiOOSc)	State graph	State coverage, Action coverage, Transition coverage, Transition path coverage, Condition coverage
Swain et al. [35]	State chart	Generation and minimization of test cases from State Charts (GeMiTefSc)	State graph	State coverage, Action coverage, Transition coverage, Ppath coverage, Condition coverage
Chimisliu and Wotawa [37]	State chart		Test purpose	Transition coverage, State coverage

(continued)

Table 1 (continued)

Author(s)	Input model	Method	Intermediate model	Coverage criteria
Chimisliu and Wotawa [38]	State chart	Test generation with Verification technology (TGV)	Test purpose	Transition coverage
Chimisliu and Wotawa [39]	State chart		Test purpose	Transition coverage

the input model and the generated paths, and coverage criteria are also clarified in this table.

These studies illustrate the importance of integrating of UML state chart diagram with other intermediate model to generate the test cases. The conclusion from these studies describe that most of them need to translate UML state chart diagram into another description, such as a graph or a table (intermediate model), which will be used to derive the test cases. Furthermore, many studies worked on DFS algorithm [23] to generate the test paths. However, this algorithm will lead to loss of paths, especially for the loops [31]. Therefore, there is a need to redefine the path coverage criterion through loop path coverage and generate an enhanced DFS algorithm or other algorithm to generate the paths [8].

In generating test case using UML state chart diagrams very few studies reveal their proposed algorithms or their testing implementation conducted during the testing. Among the studies are [20, 22, 27, 29, 41]. This scenario will lead to difficulties in updating their work or do enhancement on them. Furthermore, it is hard to implement it in a larger scale or produce it to generate test case in a fully automated way.

In rationalizing the generation of test case, the quality or the adequacy of test cases is often described with coverage criteria. From these studies, the most common coverage criteria are path coverage, transition coverage, and state coverage, what will be necessary to cover in future studies.

There are many approaches, like GA, model checking, or graph search algorithms, that are used to cover such coverage criteria for graph based models. However, there are few problems with the existing UML state chart diagrams test generation approaches. One of them is selecting the right input graph that has enough complexity to generate the accurate coverage percentage. Studies from [23, 34, 35] used very simple state chart diagrams. In addition, there are approaches by [37–39] that selected very few coverage criteria, where they just cover the transaction coverage or unnecessary one.

Many of the test case tools were not integrated. The one that have been used for test case generation, they demand several effort from the software testers since all the testing steps require manual interference in order to make appropriate adjustments on the output of a tool to be used as input to another tool [29].

4 Conclusion and Future Work

UML recently become a focused model in the field of software testing. New techniques and methods for the generation of test case from these UML diagrams needs to be identified. To identify them in this paper, a literature survey on generating test cases from UML state chart diagram has been conducted.

From this survey, it shows that existing methods on state chart test case generation methods mostly concentrate on the DFS algorithm, in which some do not work for maximum test coverage, while some methods prepare and generate a significant number of tests with less test coverage. In the future, we have planned to develop a test case generation method that minimizes the size of tests, time and cost, while preserving maximum test coverage using Modified Condition/Decision Coverage criterion. In addition, this method will prove that this technique is more capable of detecting more number of faults than compared to current existing techniques.

References

1. Patwa, S., Malviya, A.K.: Impact of coding phase on object oriented software testing. Covenant J. Inform. Commun. Technol. (CJICT) **2**, 57–67 (2014)
2. Cartaxo, E.G., Neto, F.G.O., Machado, P.D: Test Case Generation by Means of UML Sequence Diagrams and Labeled Transition Systems. In SMC, pp. 1292–1297 (2007)
3. Mingsong, C., Qiu, X., Xu, W., Wang, L., Zhao, J., Li, X.: UML activity diagram-based automatic test case generation for Java programs. Comput. J. **52**, 545–556 (2009)
4. Javed, A.Z., Strooper, P.A., Watson, G.: Automated generation of test cases using model-driven architecture. In: Automation of Software Test, AST'07. Second International Workshop on, pp. 3–3 (2007)
5. Kim, H., Kang, S., Baik, J., Ko, I.: Test Cases Generation from UML Activity Diagrams. In: Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, SNPD, pp. 556–561 (2007)
6. Kundu, D., Samanta, D.: A novel approach to generate test cases from UML activity diagrams. J. Object Technol. **8**, 65–83 (2009)
7. Lilly, R., Uma, U.G.: Reliable Mining of Automatically Generated Test Cases from Software Requirements Specification. IJCSI, pp. 87–91 (2010)
8. Mingsong, C., Xiaokang, Q., Xuandong, L.: Automatic test case generation for UML activity diagrams. In: Proceedings of the 2006 International Workshop on Automation of Software Test, pp. 2–8 (2006)
9. Rumbaugh, J., Jacobson, I., Booch, G.: Unified Modeling Language Reference Manual. Pearson Higher Education, New York (2004)
10. Utting, M., Legeard, B.: Practical Model-Based Testing: A Tools Approach. Morgan Kaufmann, US (2010)
11. Kim, W.Y., Son, H.S., Kim, R.Y.C.: A study on test case generation based on state diagram in modeling and simulation environment. In: Advanced Communication and Networking. Springer, Berlin, pp. 298–305 (2011)
12. Karambir, Kuldeep, K.: Survey of software test case generation techniques. Int. J. Adv. Res. Comput. Sci. Softw. Eng., pp. 937–942 (2013)

13. Anand, S., Burke, E.K., Chen, T.Y., Clark, J., Cohen, M.B., Grieskamp, W., et al.: An orchestrated survey of methodologies for automated software test case generation. J. Syst. Softw. **86**, 1978–2001 (2013)
14. Shirole, M., Kumar, R.: UML behavioral model based test case generation: A survey. ACM SIGSOFT Softw. Eng. Notes **38**, 1–13 (2013)
15. Rafi, D.M, Moses, K.R.K., Petersen, K., Mäntylä, M.V.: Benefits and limitations of automated software testing: Systematic literature review and practitioner survey. In: Proceedings of the 7th International Workshop on Automation of Software Test, pp. 36–42 (2012)
16. Prasanna, M., Sivanandam, S., Venkatesan, R., Sundarrajan, R.: A survey on automatic test case generation. Acad. Open Internet J. **7**:1–6(2005)
17. Pahwa, N., Solanki, K.: UML based test case generation methods: A review. Int. J. Comput. Appl. **95**, 1–6 (2014)
18. Shamsoddin-Motlagh, E.: A review of automatic test cases generation. Int. J. Comput. Appl. **57** (2012)
19. Gnesi, S., Latella, D., Massink, M.: Formal test case generation for UML statecharts. In: Proceedings of Ninth IEEE International Conference on Engineering Complex Computer Systems, pp. 75–84 (2004)
20. Kansomkeat, S., Rivepiboon, W.: Automated generating test case using UML statechart diagrams. In: Proceedings of the 2003 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on Enablement Through Technology, pp. 296–300 (2003)
21. Ali, S., Briand, L.C., Rehman, M.J.-U., Asghar, H., Iqbal, M.Z.Z., Nadeem, A.: A state-based approach to integration testing based on UML models. Inf. Softw. Technol. **49**, 1087–1106 (2007)
22. Kosindrdecha, N., Daengdej, J.: A test generation method based on state diagram. JATIT pp. 28–44 (2010)
23. Swain, R.K., Panthi, V., Behera, P., Mohapatra, D.: Automatic test case generation from UML state chart diagram. Int. J. Comput. Appl. pp. 26–36 (2012)
24. Offutt, J., Liu, S., Abdurazik, A., Ammann, P.: Generating test data from state-based specifications. Softw. Test. Verification Reliab. **13**, 25–53 (2003)
25. Briand, L.C., Labiche, Y., Cui, J.: Automated support for deriving test requirements from UML statecharts. Softw. Syst. Model. **4**, 399–423 (2005)
26. Li, H., Lam, C.P.: An ant colony optimization approach to test sequence generation for state-based software testing. In: Fifth International Conference on Quality Software, (QSIC 2005), pp. 255–262 (2005)
27. Santiago, V., do Amaral, A.S.M, Vijaykumar, N., Mattiello-Francisco, M.F., Martins, E., Lopes, O.C.: A practical approach for automated test case generation using statecharts. In: 30th Annual International Computer Software and Applications Conference, 2006 (COMPSAC'06), pp. 183–188 (2006)
28. Murthy, P., Anitha, P., Mahesh, M., Subramanyan, R.: Test ready UML statechart models. In: Proceedings of the 2006 International Workshop on Scenarios and State Machines: Models, Algorithms, and Tools, pp. 75–81 (2006)
29. Santiago, V., Vijaykumar, N.L., Guimarães, D., Amaral, A.S., Ferreira, É.: An environment for automated test case generation from statechart-based and finite state machine-based behavioral models. In IEEE International Conference on Software Testing Verification and Validation Workshop, 2008 (ICSTW'08), pp. 63–72 (2008)
30. Binder, R.V.: Testing Object-Oriented Systems: Models, Patterns, and Tools. Addison-Wesley Professional, Boston (2000)
31. Swain, S.K., Mohapatra, D.P., Mall, R.: Test case generation based on state and activity models. J. Object Technol. **9**, 1–27 (2010)
32. Shirole, M., Suthar, A., Kumar, R.: Generation of improved test cases from UML state diagram using genetic algorithm. In: Proceedings of the 4th India Software Engineering Conference, pp. 125–134 (2011)

33. Li, L., He, T., Wu, J.: Automatic test generation from UML statechart diagram based on euler circuit. Int. J. Digit. Content Technol. Appl. **6** (2012)
34. Swain, R.K., Behera, P.K., Mohapatra, D.P.: Minimal Test Case Generation for Object-Oriented Software with State Charts. arXiv preprint arXiv:1208.2265 (2012)
35. Swain, R.K., Behera, P.K., Mohapatra, D.P.: Generation and Optimization of Test cases for Object-Oriented Software Using State Chart Diagram. arXiv preprint arXiv:1206.0373 (2012)
36. Linzhang, W., Jiesong, Y., Xiaofeng, Y., Jun, H., Xuandong, L., Guoliang, Z.: Generating test cases from UML activity diagram based on gray-box method. 11th Asia-Pacific Presented at the Software Engineering Conference (2004)
37. Chimisliu, V., Wotawa, F.: Model based test case generation for distributed embedded systems. In: IEEE International Conference on Industrial Technology (ICIT), pp. 656–661 (2012)
38. Chimisliu, V., Wotawa, F.: Improving test case generation from UML statecharts by using control, data and communication dependencies. In: 13th International Conference on Quality Software (QSIC), pp. 125–134 (2013)
39. Chimisliu, V., Wotawa, F.: Using dependency relations to improve test case generation from UML statecharts. In: IEEE 37th Annual Computer Software and Applications Conference Workshops (COMPSACW), pp. 71–76 (2013)
40. Claude, J., Thierry, J.: TGV: Theory, principles and algorithms: A tool for the automatic synthesis of conformance test cases for non-deterministic reactive systems. Softw. Tools Technol. Transf. **7**, 297–315 (2002)
41. Hartmann, J., Imoberdorf, C., Meisinger, M.: UML-based integration testing. In: ACM SIGSOFT Software Engineering Notes, pp. 60–70 (2000)

Performance Analysis of Multilevel Inverter Topologies for a Standalone Photovoltaic System

M. Lakshmi and S. Hemamalini

Abstract The selection of inverter topology is the major constraint for the grid-connected system to meet the power quality issues. In this paper, performance analysis of multilevel inverter topologies for a standalone Photovoltaic (PV) system is presented. Comparison of the Voltage Source Inverter (VSI), Cascaded H-bridge multilevel inverter (CHB-MLI) and Multilevel DC link inverter (MLDCLI) is done based on harmonic contents, switch count and the filter selection. The variation in real power, reactive power and power factor of the system is also analyzed for changes in temperature, irradiation and load conditions. In this work, the load is an induction motor whose performance is analyzed for different inverter topologies. The system is simulated using MATLAB/Simulink.

Keywords Photovoltaic (PV) array · DC-DC boost converter · MLI · Filter and induction motor

1 Introduction

In recent years, most of the industrial and residential loads are connected to the power line through cost effective power converters, which enhance the overall system efficiency, performance and reliability [1]. Traditionally, two level and three level inverters are used to convert the DC voltage to AC voltage for ease of conversion with less number of switches. The disadvantage is the high switching frequency for the switches in the inverters to reduce total harmonic distortion (THD) which results in high switching losses and higher electromagnetic interference (EMI).

M. Lakshmi (✉) · S. Hemamalini
School of Electrical Engineering, VIT University, Chennai, India
e-mail: lakshmi.m@vit.ac.in

S. Hemamalini
e-mail: hemamalini.s@vit.ac.in

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_13

Multilevel inverters such as CHB-MLI and MLDCLI are used in standalone PV application due to the advantages of low EMI, low switching frequency and low THD. The output filter size also reduces due to lower THD. Thereby, MLI is cheaper, lighter and more compact [2, 3]. For high power and high voltage applications, MLI topologies are usually used. Renewable energy sources such as solar, wind and fuel cells are used as an input DC voltage source for MLI topologies. These energy sources are environment friendly and inexhaustible in nature [4–6]. PV is one of the most important DC voltage sources that have no pollution, no fuel cost, no noise and less maintenance compared with other alternative sources [7]. These DC voltage sources can be controlled individually, which improves reliability even in case of failure of any one of the DC sources [8–11]. Furthermore, an increase in the number of levels 'L' in MLI topology reduces the voltage stress on the switching device. Thereby, low rated DC link capacitors are chosen which reduces the cost of capacitors [12–15].

In this paper, performance analysis of three level VSI, seven level CHB-MLI and seven level MLDCLI for standalone PV system is done. In the PV array, the variation in irradiation generates variable output voltage, which is regulated by the Maximum power point tracking (MPPT) controller. A three-phase standalone PV system with various inverter topologies is simulated using MATLAB/Simulink. The output results are obtained under different irradiation conditions of PV array and for variable load conditions. For the above-mentioned inverter topologies, the THD content and the selection of filters are analyzed. The power and power factor variation of the induction motor drive system is also studied and a suitable inverter topology for the induction motor drive system is suggested.

2 System Description

Figure 1 shows the overall configuration of a standalone PV system with different inverter topologies. The rating of the solar array is so chosen as to provide the required power to drive the motor under different irradiation and temperature conditions. The DC-DC boost converter is designed to regulate the voltage from the PV array using MPPT control. In addition, independent voltage control is possible

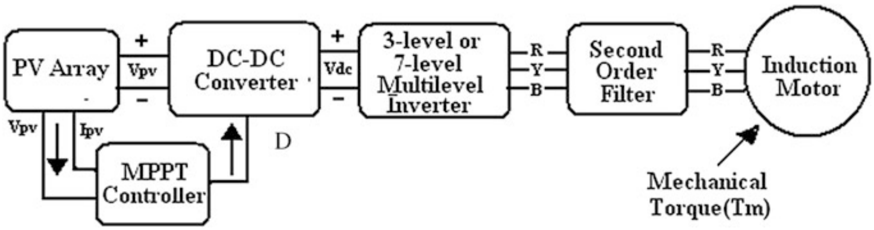


Fig. 1 Block diagram of overall system configuration

in MLI owing to separate DC links. Thus, power generated from PV systems can be maximized through individual MPPT control in each level [16]. A study of multilevel inverter topology is carried out with a second order filter in order to reduce the THD below 5 % as per IEEE 519 standard. The performance of the induction motor under variable loading conditions is also analyzed.

2.1 PV Array

PV array is a combination of multiple PV modules connected in series and parallel. The equivalent circuit of a PV cell is represented as a DC current source (I_L) with anti-parallel diode, shunt resistance (R_p) and series resistance (R_S) as shown in Fig. 2. The photo current depends on the ambient temperature (T) and solar irradiation (S) [17].

The output current of a PV cell is as given in Eq. (1).

$$I = n_p I_{ph} - n_p I_{rs}\left[e^{\left(\frac{q}{KTA}*\frac{V}{n_s}\right)} - 1\right] \tag{1}$$

where n_s and n_p are the number of series and parallel cells, I_{ph} is the photo current, I_{rs} is the reverse saturation current Eq. (3), V is the cell voltage, K is the Boltzmann constant, q is the charge of an electron and A is the diode ideality factor.

$$I_{ph} = [I_{scr} + \lambda(T - T_r)] * \frac{S}{1000} \tag{2}$$

where I_{scr} is the short circuit current at the reference temperature and irradiation, λ is the temperature coefficient and T_r is the reference temperature.

$$I_{rs} = I_{rr} * \left[\frac{T}{T_r}\right]^3 * \left[e^{\left(\frac{qE_G}{KA}*\left[\frac{1}{T_r}-\frac{1}{T}\right]\right)} - 1\right] \tag{3}$$

where I_{rr} is the reverse saturation current at reference temperature and E_G is the energy band-gap given in Eq. (4).

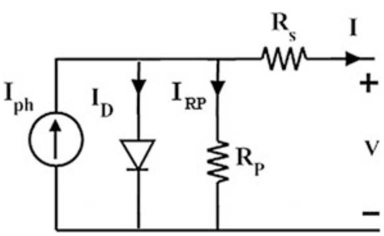


Fig. 2 Equivalent circuit for PV cell

Table 1 Parameters of PV module SPSM250 at STC

Peak power (P_{max})	250 W_p
Open circuit voltage (V_{oc})	43.21 V
Short circuit current (I_{sc})	7.63 A
Voltage at max. power (V_{mp})	35.5 V
Current at max. power (I_{mp})	7.04 A
No. of cells (n_s)	72

$$E_G = E_G(0) - \frac{\alpha T^2}{T + \beta} \tag{4}$$

where $E_G(0)$ is the initial energy band-gap, α and β are the current and voltage temperature coefficients.

The PV array V-I characteristics is dependent on temperature and irradiation [18]. The MPPT is used to extract maximum power from the PV array to the load. The MPPT controller operates the solar PV array near MPP to improve the system efficiency [19]. The parameters and description of solar PV panel are shown in Table 1.

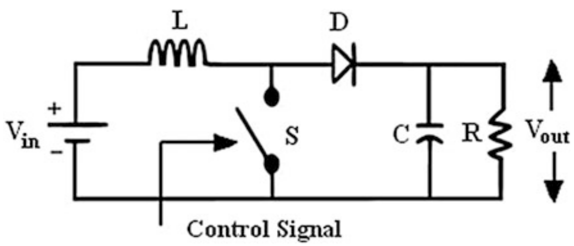
2.2 DC-DC Boost Converter

The output voltage of the PV array is generally low. Thereby, a DC-DC boost converter shown in Fig. 3 is required to step up the voltage for practical utilization. The converter acts as an interface between the PV array and the inverter. The input to the boost converter from a PV array is an unregulated DC voltage due to variation in irradiation and temperature. By adjusting the duty cycle (D), the output voltage of the converter is regulated.

The voltage gain offered by the converter is given in Eq. (5).

$$\frac{V_{out}}{V_{in}} = \frac{1}{1 - D} \tag{5}$$

Fig. 3 DC-DC boost converter



where V_{in} is the input DC voltage, V_{out} is the output DC voltage and D is the duty cycle.

The value of critical inductance and capacitance [20, 21] are calculated using Eqs. (6) and (7).

$$L = \frac{D\,(1 - D)^2 R}{2f} \tag{6}$$

$$C = \frac{D V_{out}}{f\;V_r R} \tag{7}$$

where f is the switching frequency, V_r is voltage ripple factor and R is the load resistor. The number of levels in an inverter decides the regulated DC voltage source requirement.

2.3 *Multilevel Inverter*

In general, three level VSIs are used for medium power applications. The emerging trend in renewable energy system employs MLI topology for high power applications. Several MLI topologies are developed, among which the CHB-MLI and MLDCLI are quite popular [22].

2.3.1 Three Level Voltage Source Inverter

Three-phase VSIs are used in medium to high power applications that provide a controlled voltage and frequency at the output using sinusoidal pulse width modulation (SPWM) technique. The number of switches required in a three phase three level VSI is obtained from the expression given in Table 2. For the conventional three-phase VSI shown in Fig. 4, the number of switches used is six. The switches in the same leg of the inverter cannot be switched on at the same instant, as it creates short circuit at the DC link voltage source [23].

Table 2 Comparison of different inverter topologies

Components	Converter type		
	VSI	CHB-MLI	MLDCLI
Main switching device	2 × L	(L − 1) × 2	L + 3
Main diode	2 × L	(L − 1) × 2	(L + 3)
Clamping diode	0	0	0
DC bus capacitors	(L − 1)/2	(L − 1)/2	(L − 1)/2
Balancing capacitor	0	0	0

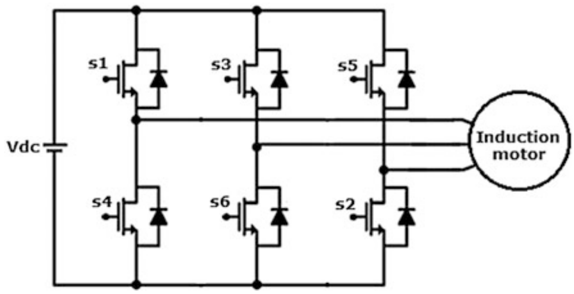


Fig. 4 Three phase three level voltage source inverter

The inverter has eight switching states; where in two state produces zero line voltage at the output. Under this condition, the line currents circulate through either the upper or lower switching elements. The continuing states produce non-zero output line voltages. The changes in the switching states generate a required voltage waveform. Hence the resultant discrete values of voltages like $-V_{dc}$, 0, and $+V_{dc}$ are obtained as AC output line voltages.

2.3.2 Seven Level Cascaded H Bridge—Multi Level Inverter

In three phase CHB-MLI shown in Fig. 5, each phase consists of 'n' series connected H-Bridge converters and every H-Bridge DC link is fed by a PV array. The three output voltage levels in each H-bridge modules are $-V_{dc}$, 0 and V_{dc}. The cascaded MLI with 'n' input sources provides (2n + 1) levels to synthesis AC output waveform [24]. The number of DC sources in each phase is considered as three. In the three phase seven level inverter topology, 36 switches are used and are calculated using the expression in Table 2.

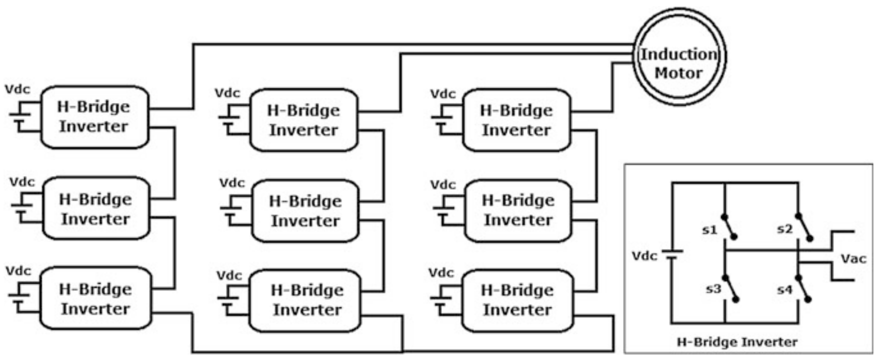


Fig. 5 Three phase seven level cascaded H-bridge inverter

2.3.3 Seven Level Multilevel DC Link Inverter

To overcome the problems with Cascaded H-bridge multilevel inverter MLDCLI is used. Figure 6 shows a schematic diagram of the MLDCLI topology, which consists of a multiple DC source and a single-phase full-bridge (SPFB) inverter. This topology requires (L + 3) active switches for a given number of voltage level 'L'. The DC source is designed by connecting several half-bridge cells in series. Each cell is having a voltage source, controlled through two switches. The switches in the cells operate at double the fundamental frequency of the output voltage. In the SPFB inverter, the four switches S1 and S4, S2 and S3 always work in pairs at the fundamental frequency of the output voltage. The MLDCL delivers staircase DC bus voltage that approximates the rectified waveform of the reference sinusoidal voltage. The SPFB inverter changes the voltage polarity to produce an AC voltage of a staircase shape [14].

The MLDCLI topology significantly reduces the switch count by increasing the number of levels 'L' above five as compared with the CHB-MLI. Due to the reduction in switch count, the cost is minimized and results in smaller size and volume [15]. As the number of level increases, the number of component requirement reduces.

2.3.4 Inverter Control Strategy

The inverter control strategy is used to regulate the output voltage of the inverter as well as speed of the induction motor. The PI controller is used to minimise the error between the induction motor speed and the reference speed. The three phase reference waveform is generated from the output of the PI controller with appropriate phase delay. The pulses are generated by comparing reference waveform with the carrier waveform to control the inverter output.

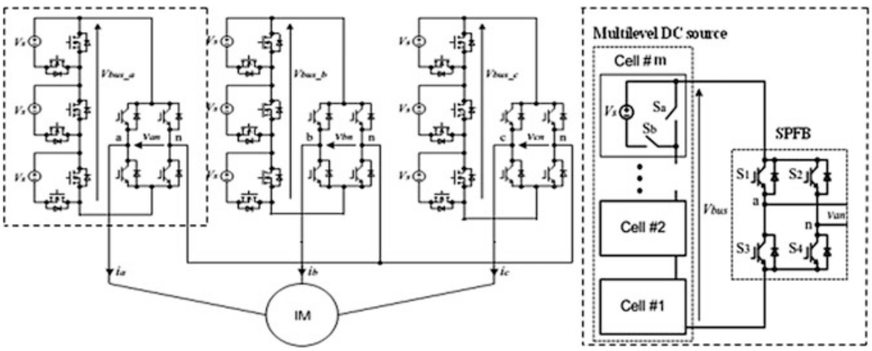


Fig. 6 Three phase seven level multilevel DC link inverter

2.4 Filter Design

A first order filter is generally used to suppress the current harmonics. These filters require higher damping, as it is designed for line frequency. The second order filter replaces the first order filter because of its poor dynamic response. In this paper, a second order filter is designed for a suitable damping factor (ξ) and the natural frequency (ω_n) to eliminate the higher magnitude lower order harmonic contents [25].

In general, the transfer function of a second order system is given by

$$\frac{V_o}{V_{in}} = \frac{\omega_n^2}{S^2 + 2\varsigma\omega_n S + \omega_n^2} \tag{8}$$

where

$$\xi = \frac{R}{2}\sqrt{\frac{C}{L}} \tag{9}$$

The transfer function of a second order LCR filter is given in Eq. (10).

$$\frac{V_o}{V_{in}} = \frac{1/LC}{S^2 + R/LS + 1/LC} \tag{10}$$

where R is the damping resistor, L and C are the filter inductance and capacitance. The filter size of the three phase three level VSI increases due to higher voltage magnitude and lower order harmonic contents present at the output. Owing to this drawback, MLI is used to eliminate lower order harmonics by increasing the number of levels.

2.5 Induction Motor Drive

Over the past decades, DC machines are employed in the drive system because of its variable speed applications. Due to the major drawback of the existence of mechanical commutator and brush assembly, DC drives have become outdated in industrial applications. The toughness, low cost, better performance and easy maintenance makes the asynchronous motor advantageous in many industrial or general applications [26]. Squirrel cage induction motors (SCIMs) are widely used and are cheaper in cost as compared to slip ring induction motors.

3 Results and Discussion

The MATLAB/Simulink model of the three phase three level inverter with closed loop control is shown in Fig. 7. The system parameters and specifications are detailed in Table 3.

In Fig. 8a, the power output of a PV array for variation in solar irradiation is shown. The regulated DC link voltage of the converter for variable input voltage is shown in Fig. 8b. The number of voltage sources used in each phase of a three phase CHB-MLI and MLDCLI is computed using (L − 1)/2, where L is the number of levels. Thereby, in each phase of CHB-MLI and MLDCLI three voltage sources of 120 V each are connected in series.

Based on the oscillatory response; the damping required for the VSI based standalone PV system is 0.762 and the THD obtained after filtering is 0.19 %. Similarly, the damping required for CHB-MLI and MLDCLI systems are 0.432 and 0.55. The corresponding THD values are 0.05 and 0.07 %.

The harmonic contents before and after filter is represented in Fig. 9. The CHB-MLI has better harmonic profile as compared to the other two inverter topologies. In the case of MLDCLI, the difference in THD compared to CHB-MLI is 0.02 %. However, it can be compromised with the cost of the inverters due to the

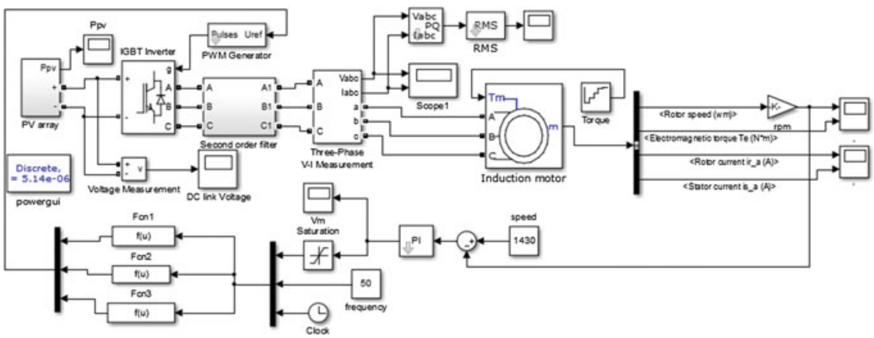


Fig. 7 MATLAB/Simulink model and control scheme of the inverter topology

Table 3 System Parameters

PV array	P = 5 KW, $N_s = 10$, $N_p = 2$, $V_{oc} = 432.1$ V, $I_{sc} = 15.26$ A, $V_{mp} = 360$ V, $I_{mp} = 13.876$ A
Boost converter for VSI	$V_{in} = 360$ V, $V_{out} = 720$ V, $f_s = 100$ kHz, L = 642 µH, C = 470 µF
Boost converter for MLI	$V_{in} = 72$ V, $V_{out} = 120$ V, $f_s = 100$ kHz, L = 100 µH, C = 330 µF
Inverters	$V_{out(inv)} = 400$ V, $f_s = 2$ kHz, V_{dc} (VSI) = 720 V, V_{dc} (MLI) = 360 V
Induction motor	SCIM, 5.4 hp, Frequency, f = 50 Hz, Speed, N = 1430 rpm

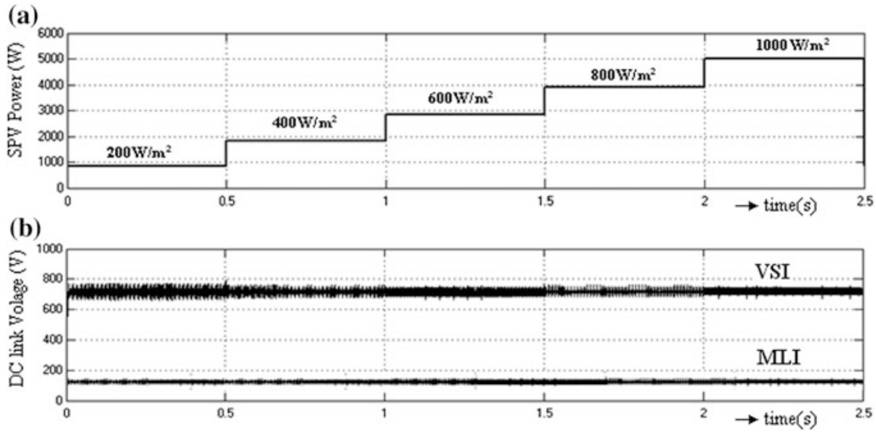


Fig. 8 **a** Power generated by the PV array for variation in irradiation. **b** Regulated DC link voltage

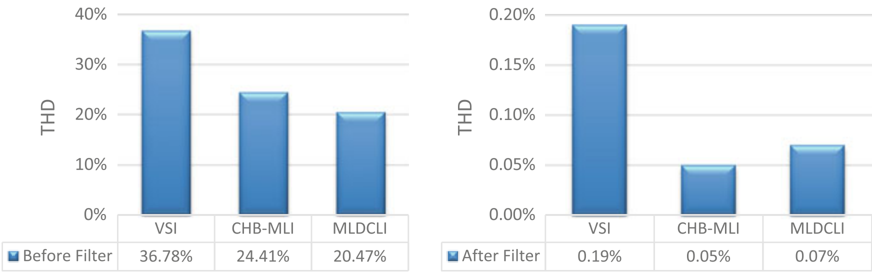


Fig. 9 THD analysis of inverter topologies before and after filter

less number of switches in seven-level MLDCLI. The induction motor load characteristics such as the speed, torque, stator current and the rotor current for all the closed loop inverter topologies are approximately the same. Hence, the simulated output of VSI along with the induction motor drive system is presented in Fig. 10 under variable load conditions.

The real and reactive power transfer from the source to load is analyzed under different loading condition and is given in Table 4. While running the motor in half load condition, the power factor is low and the reactive power is more than the real power. Whereas real power is more than the reactive power and the power factor is maintained at 0.8 while running the motor in full load condition. The power factor varies based on the power delivering capability of PV system and loading of induction motor.

In Fig. 11a, the induction motor is half loaded from 0.5 to 1 s and from 2 to 2.5 s it is operated at full load condition. The graphical representation of the power factor analysis is shown in Fig. 11b. For different loading on the induction motor drive

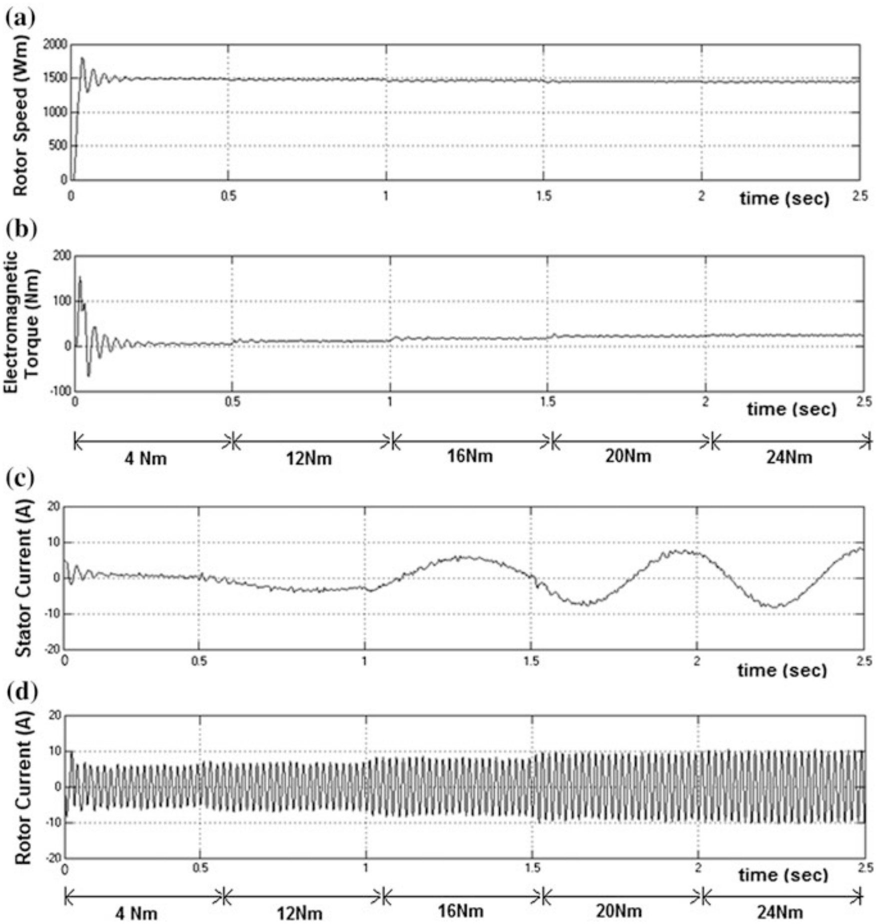


Fig. 10 Induction motor performance under variable load condition. **a** Rotor speed. **b** Electromagnetic torque. **c** Rotor current. **d** Stator current

Table 4 Real and reactive power drawn by the induction motor

Type of inverter	Load torque (Nm)	Real power (W)	Reactive power (VAR)	Power factor
VSI	Half load (12)	1931	3032	0.537
	Full load (24)	3877	2885	0.802
CHB-MLI	Half load (12)	1873	2726	0.566
	Full load (24)	3820	2995	0.798
MLDCLI	Half load (12)	2081	2658	0.616
	Full load (24)	3978	2886	0.809

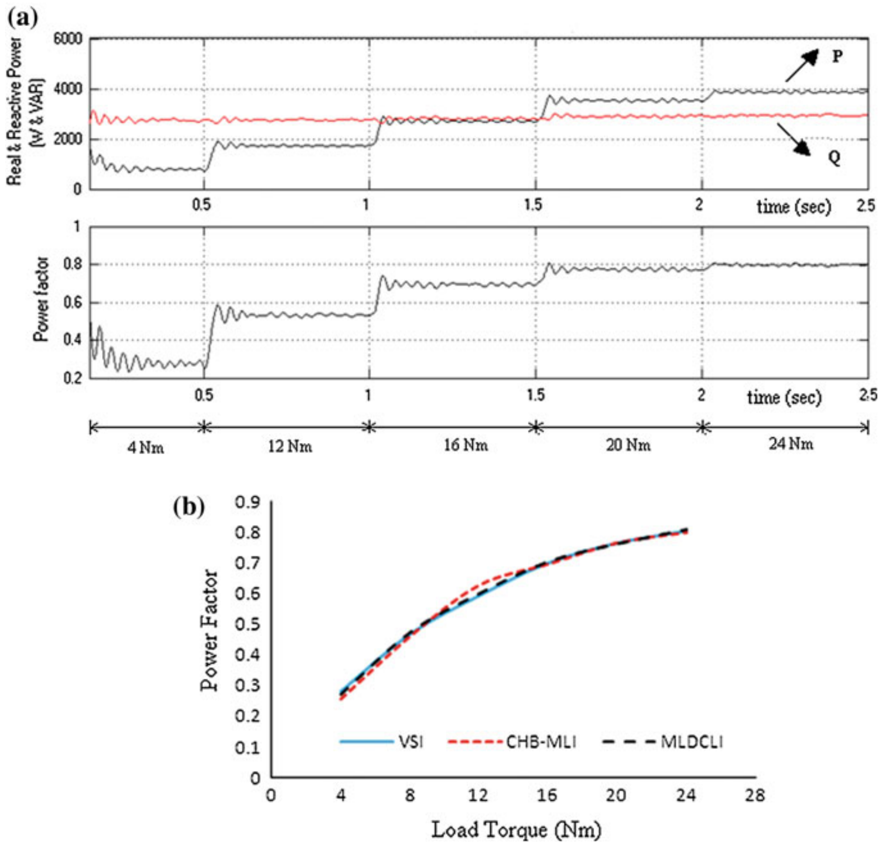


Fig. 11 **a** Real power-P, reactive power-Q and power factor under variable load. **b** Power factor analysis for inverter topologies

system, variation in the real power, reactive power and power factor is observed for the different inverter topologies. Solar powered MLDCLI fed induction motor drive system has better power factor while running at half load as well as full load condition compared with the other two inverter topologies.

4 Conclusion

The performance analysis of inverter topologies for a standalone PV fed induction motor drive system is carried out in this paper. The PV array, boost converter and three topologies of inverters are modeled and the complete system is simulated in MATLAB/Simulink. The performance of the inverter topologies are analyzed with respect to THD, number of switches and the filter selection. From the analysis, it is observed that MLDCLI has more advantages compared to other inverter topologies. The voltage THD at the output of MLDCLI is slightly greater than CHB-MLI. However, this is compromised with the reduced number of switches used in MLDCLI. Hence, the initial cost of the inverter is reduced and the switch losses are greatly minimized. The filter size is reduced because of the lower value of THD and an improvement in power factor of the induction motor is observed for varying load conditions and for variation in irradiation conditions. The analysis of simulation results shows that the three phase MLDCLI topology is effective for the standalone PV fed induction motor drive system. In future, the MLDCLI topology can be implemented for a grid-connected photovoltaic system to analyse the power quality issues like harmonics and to improve the power factor.

References

1. Rodriguez, J., Lai, J.S., Peng, F.Z.: Multilevel inverters: Survey of topologies, controls, and applications. IEEE Trans. Ind. Appl. **49**(4), 724–738 (2002)
2. Hinga, P.K., Ohnishi, T., Suzuki, T.: A new PWM inverter for photovoltaic power generation system. In: 25th Annual IEEE Power Electronics Specialists Conference (PESC'94), vol. 1, pp. 391–395 (1994)
3. Cheng, Y., Qian, C., Crow, M.L., Pekarek, S., Atcitty, S.: A comparison of diode-clamped and cascaded multilevel converters for a STATCOM with energy storage. IEEE Trans. Ind. Electron. **53**(5):1512–1521 (2006)
4. Lai, J.S., Peng, F.Z.: Multilevel converters-a new breed of power converters. IEEE Trans. Ind. Appl. **32**(3):509–517 (1996)
5. Tolbert, L.M., Peng, F.Z., Habetler, T.G.: Multilevel converters for large electric drives. IEEE Trans. Ind. Appl. **35**(1):36–44 (1999)
6. Kuo, Y.C., Liang, T.J, Chen, J.F.: Novel maximum-power-point-tracking controller for photovoltaic energy conversion system. IEEE Trans. Ind. Electron. **48**(3):594–601 (2001)
7. Mariethoz, S., Rufer, A.: Resolution and efficiency improvements for three-phase cascade multilevel inverters. In: IEEE 35th Annual Power Electronics Specialists Conference (PESC'04), vol. 6, pp. 4441–4446 (2004)
8. Tolbert, L.M., Peng, F.Z., Cunnyngham, T., Chiasson, J.N.: Charge balance control schemes for cascade multilevel converter in hybrid electric vehicles. IEEE Trans. Ind. Electron. **49** (5):1058–1064 (2002)
9. Corzine, K.A., Wielebski, M.W., Peng, F.Z., Wang, J.: Control of cascaded multilevel inverters. IEEE Trans. Power Electron. **19**(3):732–738 (2004)
10. Leon, J.I., Vazquez, S., Kouro, S., Franquelo, L.G., Carrasco, J.M., Rodriguez, J.: Unidimensional modulation technique for cascaded multilevel converters. IEEE Trans. Ind. Electron. **56**(8), 2981–2986 (2009)
11. Hua, C.C., Wu, C.W., Chuang, C.W.: A digital predictive current control with improved sampled inductor current for cascaded inverters. IEEE Trans. Ind. Electron. **56**(5):1718–1726 (2009)
12. Vazquez, S., Leon, J.I., Franquelo, L.G., Padilla, J.J., Carrasco, J.M.: DC-voltage-ratio control strategy for multilevel cascaded converters fed with a single DC source. IEEE Trans. Ind. Electron. **56**(7), 2513–2521 (2009)
13. Cecati, C., Ciancetta, F., Siano, P.: A multilevel inverter for photovoltaic systems with fuzzy logic control. IEEE Trans. Ind. Electron. **57**(12), 4115–4125 (2010)
14. Su, G.J.: Multilevel DC-link inverter. IEEE Trans. Ind. Appl. **41**(3), 848–854 (2005)

15. Villalva, M.G., Gazoli, J.R., Filho, E.R.: Comprehensive approach to modeling and simulation of photovoltaic arrays. IEEE Trans. Power Electron. **24**(5), 1198–1208 (2009)
16. Xiao, B., Shen, K., Mei, J., Filho, F., Tolbert, L.M.: Control of cascaded H-bridge multilevel inverter with individual MPPT for grid-connected photovoltaic generators. In: IEEE Energy Conversion Congress and Exposition (ECCE'12), pp. 3715– 3721 (2012)
17. Aghazadeh, H., Kojabadi, H.M., Yazdankhah, A.S.: Stand-alone PV generation system with maximum power point tracking. In: 9th International Conference on Environment and Electrical Engineering (EEEIC'10), pp. 549–552 (2010)
18. Pandey, A., Dasgupta, N., Mukerjee, A.K.: Design issues in implementing MPPT for improved tracking and dynamic performance. In: 32nd Annual Conference on IEEE Industrial Electronics (IECON'06), pp. 4387–4391 (2006)
19. Rivetta, C.H., Emadi, A., Williamson, G.A., Jayabalan, R., Fahimi, B.: Analysis and control of a buck DC-DC converter operating with constant power load in sea and undersea vehicles. IEEE Trans. Ind. Appl. **42**(2), 559–572 (2006)
20. Mohan, N., Undeland, T.M., Robbins, W.P.: Power Electronics: Converters, Applications, and Design, Book. Wiley, New York (1989)
21. Lanjewar, S.M., Ramsha, K.: Design of control scheme and performance improvement for multilevel DC link inverter fed PMBLDC motor drive. Int. J. Electr. Electron. Eng. (IJEEE) **1** (3):2231–5284 (2012), ISSN (PRINT)
22. Rashid, M.H.: Power Electronics Handbook: Devices, Circuits, and Applications. Academic Press is an imprint of Elsevier (2007)
23. Xiao, B., Hang, L., Riley, C., Tolbert, L.M., Ozpineci, B.: Three-phase modular cascaded H-bridge multilevel inverter with individual MPPT for grid-connected photovoltaic systems. In: Twenty-Eighth Annual IEEE Applied Power Electronics Conference and Exposition (APEC'13), pp. 468–474 (2013)
24. Steinke, J.K.: Use of an LC filter to achieve a motor-friendly performance of the PWM voltage source inverter. IEEE Trans. Energy Convers. **14**(3), 649–654 (1999)
25. Kim, S.H., Kim, Y.H., Seo, K.M., Bang, S.S., Kim, K.S.: Harmonic analysis and output filter design of NPC multi-level inverters. In: 37th IEEE Power Electronics Specialists Conference (PESC '06), pp. 1–5 (2006)
26. Zhang, Y., Jiang, Z., Yu, X.: Indirect field-oriented control of induction machines based on synergetic control theory. In: IEEE Power and Energy Society General Meeting—Conversion and Delivery of Electrical Energy in the 21st Century, pp. 1–7 (2008)

Development of Massive Open Online Course for University Course (A Case Study Approach)

Yohannes Kurniawan, Taruna Diyapradana, Fina Shabrina Sutarto and Elian Wira Sena

Abstract The aim of this research is to create a Massive Open Online Course (MOOC) for XYZ University to be opened for public use and education. The method of this MOOC was covered in Horton's book E-Learning by Design, and the analysis method will use Unified Process (UP). What was gained from this analysis and development is an MOOC site that can be used to deliver introductory courses in Information System to the general public for free and can be accessed easily, be it from computer or mobile devices. What can be concluded from this development are courses in Information System (IS) area to decide whether or not potential students want to pursue an education in the field, students mostly turn to books, internet searches, and non-MOOC online courses. This MOOC provides a platform where they can learn from IS lecturers so that they do not have to learn unassisted, which gives them a stronger understanding of the basics of IS to make an informed decision on pursuing an education in IS.

Keywords Online course · Development · E-learning · Information system

1 Introduction

In the globalization era, information technology is evolving rapidly and gives a big impact almost in every spheres of life. Information technology helps people to make their activities become easier.

Learning method used to be done in traditional and conventional way, it's called face to face between the teacher and the student. But now, it has changed, there's no need for physical meetings or face to face anymore between teacher and student to learn. The teacher can give the material through electronic media so that can give

---

Y. Kurniawan (✉) · T. Diyapradana · F.S. Sutarto · E.W. Sena
Information Systems Department, Bina Nusantara University,
11480 Jakarta Barat, Indonesia
e-mail: Kurniawan_yohannes@yahoo.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_14

flexibility in learning process. This was discussed in Gundry [1] where e-learning is stated to take place “any place, any time”. Even though in reality it isn’t that simple, it is true that e-learning does help learners to study in places they otherwise cannot study at.

The development of information technology in learning process can stimulate more learning activities because as Gundry discussed, there is no time limitation and don’t need a high cost. So it can change the old paradigm about traditional and conventional way to learn that require the student to meet the teacher physically.

E-learning is one of the method of using electronic media and communication in learning process. With a good implementation of e-learning, it can make the users multiply, increasing value, decreasing the learning cost, and also preserve the old users so they can continue to learn through electronic media.

One of the e-learning methods is called MOOC (Massive Open Online Class). Pappano [2] described MOOC as a free, credit-less, and massive online course, as opposed to traditional online courses that charge tuition, carry credit, and limit enrollment. Because of this, anyone with an Internet connection can apply, therefore increasing the number of potential students.

With all those considerations, this paper about the use of e-learning was formulated, which XYZ University can use to be their community services activity.

The MOOC is an e-learning facility designed for the XYZ University to be used by high school students and the general public, aimed to introduce the field of information systems to the learners. The MOOC will be programmed with the PHP programming language. The development of the MOOC will be done in the Information Systems Laboratory of XYZ University.

The main issue that will be addressed in this paper is the development for the School of IS MOOC, how the research is done and who is the target audience, and the features to be implemented in the MOOC.

The purpose of this research is to analyze and devise an e-learning platform that allows its users to follow learning activities without the constraints of space and time. This will help provide a new alternative for potential students of XYZ University to learn about the world of information systems, where they can understand about what information systems is and learn the basics about it before deciding whether they want an education in the field or not.

2 Methods

The method used in creating this MOOC project is the approach detailed by William Horton in the book E-Learning by Design. The MOOC project will consist of: (1) Absorb-Type Activities: Activities in which learners extract and comprehend knowledge from a given source of information; (2) Do-Type Activities: These activities transform the information gained during the absorb activities into knowledge and skills; (3) Connect-Type Activities: These activities bridge gaps between learning and real-life application. Learners are made to apply learning in situations they will encounter on the job or their daily lives; (4) Tests: Tests are created to indicate how well learners meet learning objectives and provides feedback on performance of an objective.

3 Results and Discussion

According to Horton [3], E-learning is the use of electronic technologies to create learning experiences. The definition is left open-ended because the experiences vary and are not limited to classes.

According to Joseph and Nath [4], MOOC as a term stands for Massive Open Online Course, in which *Massive* refers to scale giving opportunity for connections among participants, *Open* means open access, open syllabi, and self-directed learning outcomes, *Online* means making materials available on internet in abundance, and *Course* refers to the structure of the online course.

Some of the most famous MOOC providers are Coursera, Open2Study, and Canvas. These sites offer online courses for free, working together with various universities from around the world. Some of these websites also offer "Signature Track" options where members can pay to have the courses they take be recognized as class credit if it fulfills certain conditions.

3.1 Advantages and Challenges of MOOC

As with all methods of learning, MOOC has its own advantages and disadvantages in its use. Barnett et al. [5] discussed in their paper that some of MOOC's advantages are its accessibility, increased potential for student engagement, and expanded lifelong learning opportunities.

- Accessibility: MOOCs are accessible because they are usually free and people from everywhere, even countries that has limited access to education facilities.
- Increased student engagement: Increased student engagement means that the learning outcomes and development of the student will be enhanced, and so is the reputation and performance of the institution.
- Lifelong learning opportunities: Lifelong learning opportunities are expanded by allowing participants to pursue a particular interest with classes offered by the MOOC and provides underprivileged populations the opportunities for education so they can pursue lifelong learning.

However, Barnett et al. [5] also discussed some challenges associated with using MOOCs. Some of the most common are individual instruction, student performance assessment, and long-term administration and oversight.

3.2 The Current System

Currently, if people want an introductory look into information systems, they have to read books about the topic, search the Internet, or participate in online classes. Of course, these methods have their own advantages and disadvantages. These methods also have their own absorb, do, and connect activities associated with them. The table below details the absorb, do, and connect activities related to each learning medium (Table 1).

A recent development in the field of online learning is the rise of MOOCs. MOOCs allow students from all over the world to learn at the same time, mostly for free, though several sites offer college credit for an optional fee. In MOOCs, fellow students can collaborate through forums, chats and other communication features the site may have, and grade each other assignments.

Before the development of this MOOC project, an attempt to research several MOOC providers by participating in a class and trying its features in order to gain an understanding on how the user experience of an MOOC feels like was made. The sites that were tried are Open2Study, Belajar Daring, XL Future Leaders, and Coursera.

The key differentiator in this MOOC, use of videos will be focused on more over readable material such as text and downloadable presentations. Researchers from the MOOC platform edX, in Guo et al. [6] found that content-wise, presentation slideshows are less engaging than informal "talking head" and Khan-style tablet drawing format videos.

The sampling method used was the cluster sampling method, with the institutions as the individual clusters. And 207 people answered the questionnaire, but 6 of them did not complete it and thus were deemed invalid responses. In total, 201 valid respondents has been gained from these institutions, and the following charts will interpret their responses regarding the questionnaire. More than half (51 %) of the respondent's final completed education was junior high school, which would currently put them in high school. About a quarter (26 %) respondents are high school graduates, which would currently place them in a bachelor program. This places the bulk of the respondents as Generation Y (1981–1999 births, will be covered later in this chapter), the primary targets of the MOOC. The results are detailed in the chart below. These results could advise future content contributors to make courses in the fields that potential students are interested in the most (Fig. 1).

Table 1 Absorb, do, connect activities on each learning medium

	Internet	Books	Online class
Absorb	Article contents	The contents of the book itself	Video lectures, presentations
Do	Online quizzes	Exercises, if available	Quizzes, assignments
Connect	Case studies available on the internet	Case studies, if available	Case studies and exams, if applicable

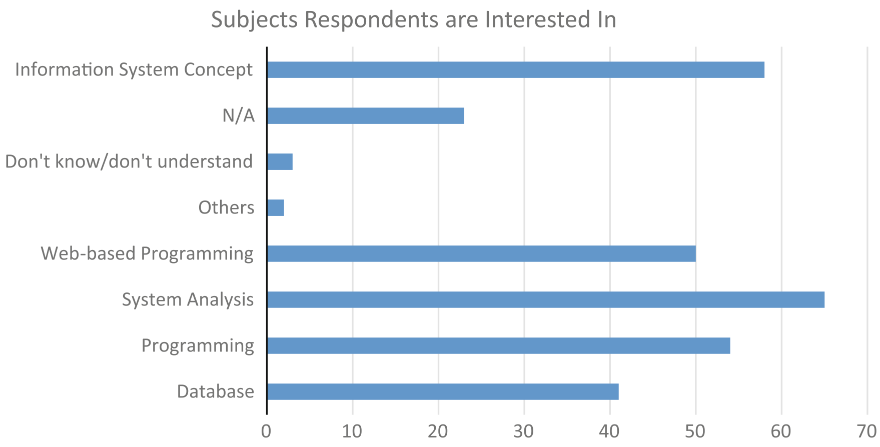
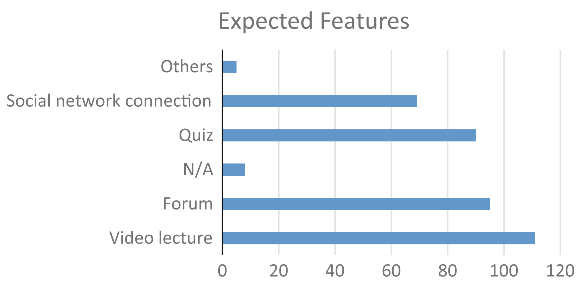


Fig. 1 Subjects of interest questionnaire results chart

Fig. 2 Expected features questionnaire results chart



Respondents were also asked about the features they expect to be in an MOOC. Like the previous question with the subjects, this is a multiple answer question. 111 respondents answered video lectures, 95 expected forums, 90 wanted quizzes, and 69 are interested in social network connections. The other answers written in by respondents are certificates, college tips, mini educational games, practice exercises, and Q&A with lecturers and experts. Eight respondents did not answer this question. The chart detailing the answers to this question is below (Fig. 2).

3.3 Suggested New System

After going over the flow of activities of learning about IS from the Internet, non-MOOC online classes and books along with their respective advantages and disadvantages and interpreting the data gathered from the questionnaire, a massive open online course is proposed be built for the School of Information Systems to help prospective students experience studying Information Systems before they decide to pursue an education in the field. This will also foster interest in the field. This MOOC will provide introductory courses to help the learners know more about the world of information systems.

In this MOOC, use of videos will be focused on more over readable material such as text and downloadable presentations. Researchers from the MOOC platform edX, in Guo et al. [6] found that content-wise, presentation slideshows are less engaging than informal "talking head" and Khan-style tablet drawing format videos. The enthusiasm of lecturers and a more personal feel helps engagement further, as is motion and visual flow for tutorial videos. Length-wise, shorter videos are much more engaging, as the study showed that the median engagement time is at most six minutes, regardless of video length. It was also shown that students often make it less than halfway through videos longer than nine minutes. It is recommended that videos are segmented into chunks shorter than 6 min. This shows that in order to make the most engaging videos, a more personal, enthusiastic style is essential to the production and so is the length, which should be kept short as to not bore the viewer.

The reason that video is more concentrated on is that the primary target for this MOOC is the Generation Y, potential students of XYZ University. This generation spans births from 1981 to 1999, and make up most of XYZ University's potential students. Honore et al. [7] stated in their research that Generation Y has a low boredom threshold, are mostly visual learners, and has little tolerance for delays. This is the reason why concentrating on video lectures will help Generation Y learners to invest in the MOOC more. It is also stated that Generation Y does a lot of multitasking, which can be done better with video than with downloaded text or presentations. They can run the video in the background as they do other things and they would still gain something from it. Reilly [8] also supports the point by stating in his article that Generation Y are visual learners. Reilly also stated that Generation Y reads less and less well, which would hinder the effectiveness of giving reading material. Therefore, a video that could capture their attentions without much text would be useful in delivering course materials to Generation Y learners.

3.3.1 System Definition

The MOOC being developed will be used and maintained by XYZ University, consisting of two parts:

- Front-End: This part concerns most of the users of the MOOC, where students can register into the MOOC site, enroll and take classes, discuss in the forums, and do quizzes.
- Back-End: This part is where the lecturers can put up the material for the classes and edit them. Lecturers can also put up quizzes for material assessment. This is also where the admin can edit, delete, or add new material or forum posts.

Table 2 Absorb, do, and connect activities of the MOOC

	MOOC
Absorb	Video lectures
Do	Online quizzes in every material
Connect	Discussion forums where students can discuss about the subject freely with classmates and lecturers

3.3.2 Proposed Absorb, Do, Connect Activities

As with the current learning system that was discussed in the previous chapter, this MOOC will have its own absorb, do, and connect activities in order to facilitate learning. The absorb, do, and connect activities will be detailed in Table 2.

3.3.3 Deployment Environment

In this MOOC project, the plan is to use a two-layer architecture as detailed by Burd et al. [9]. The architecture will consist of the user interface layer and the domain layer.

The user interface layer will be where the actors (namely students, admins, and lecturers) interact with the website. Data input by the customers are taken from this layer and the data from the website are displayed on this layer. In the domain layer, the user's information will be captured and used to update the database. This layer acts as both the domain and data access layers, where the database is located.

For the network, this MOOC will be published to the Internet, so that the general public can easily access it. The site will be hosted on XYZ University servers.

This diagram below details how the components in the MOOC are placed. The application is made using the CodeIgniter framework on the server side, where this framework is using the Model View Controller (MVC) architecture pattern. MVC is a pattern where the controller makes an activity to be done by the model. Then, the results are displayed by means of the controller calling on the model to interact with the MySQL Database. Then, the model is displayed on the view to be sent to the client's browser. The model is sent to view using JavaScript, HTML, CSS, JQuery, and Bootstrap to be displayed on the client's browser (Fig. 3).

The use case and use case descriptions will be divided into the front-end and the back-end use cases, detailed in the following use case diagram (Fig. 4).

The entity relationship diagram below depicts the data in terms of entities and relationships described by the data (Fig. 5).

The following screenshots are taken from the School of IS MOOC website that has been developed, showing the pages that exist inside it and walking through several of the website's functionalities (Figs. 6, 7, and 8).

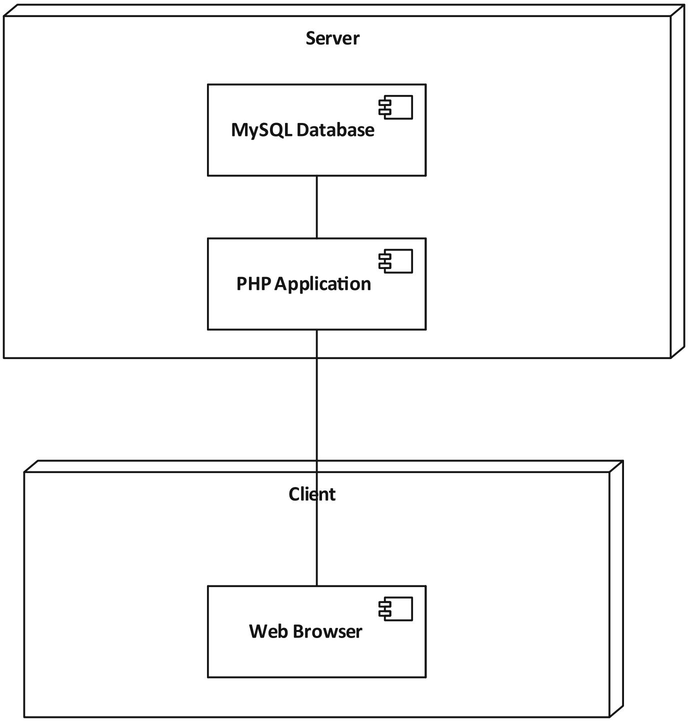


Fig. 3 School of IS MOOC deployment diagram

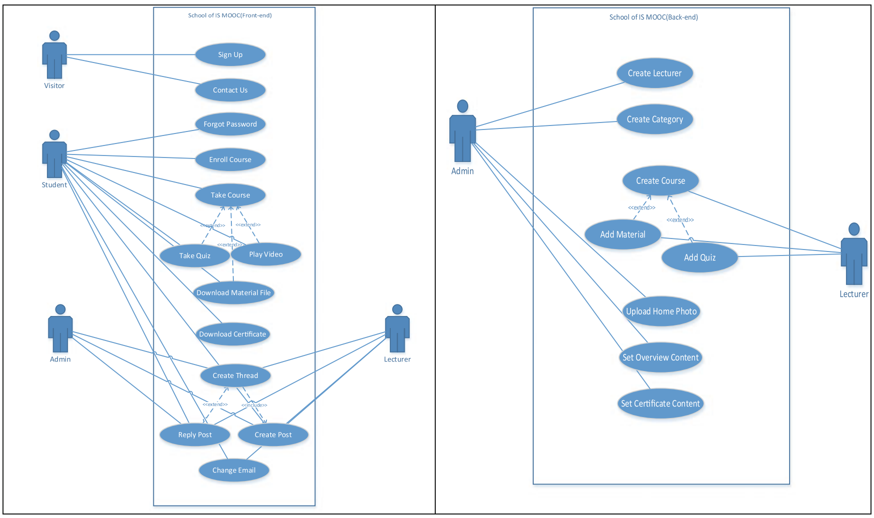


Fig. 4 Use case diagram for school of IS MOOC’s

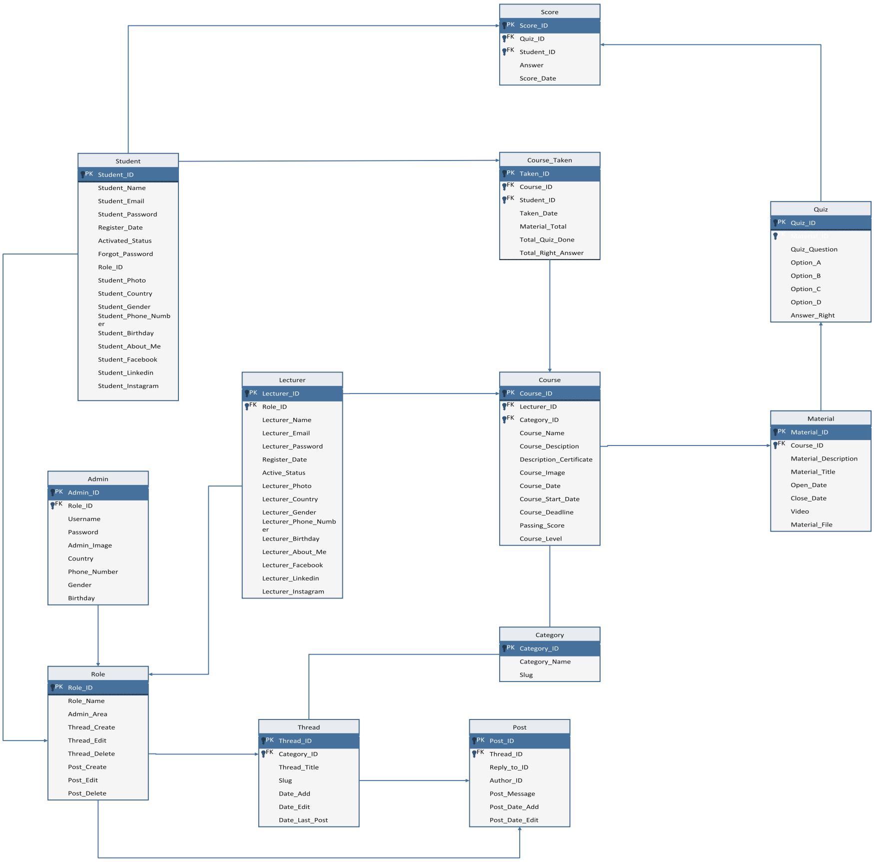


Fig. 5 School of IS MOOC entity relationship diagram

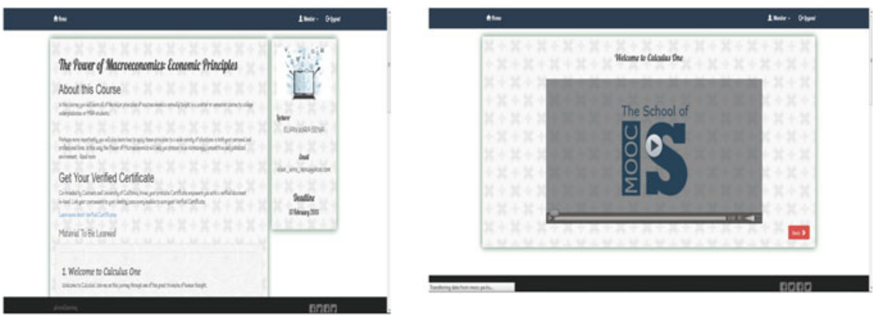


Fig. 6 Material page

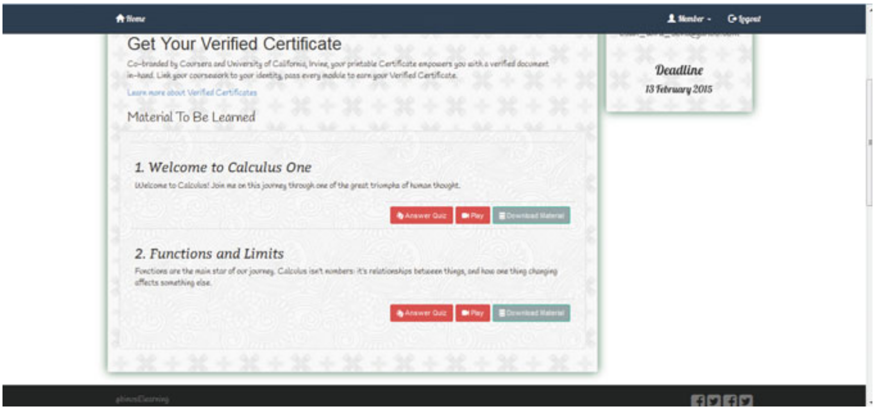


Fig. 7 Choose quiz

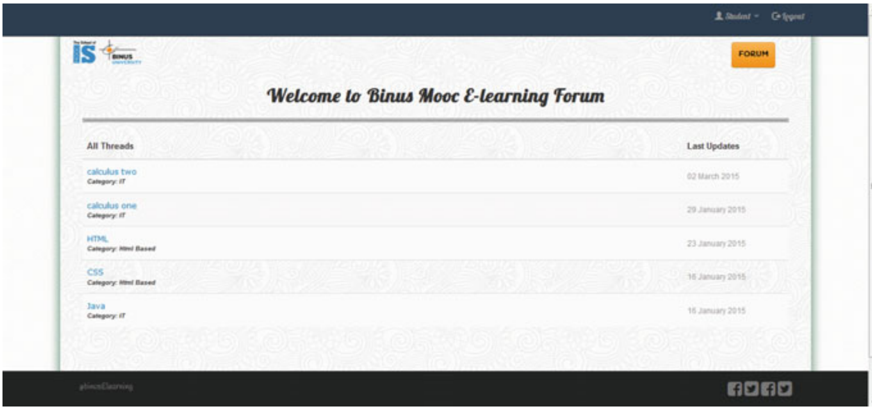


Fig. 8 Forum homepage

4 Conclusion

After developing the MOOC platform, it can be concluded that for an introductory course in IS to decide whether or not they want to pursue further education in the field, potential students mostly turn to books, internet searches, and non-MOOC online courses.

This MOOC will hopefully help to make it easier for them to learn about information systems, by providing a platform where they can learn from experienced IS lecturers and experts, rather than learning on their own unassisted. This way, they can learn more effectively and gain a stronger understanding of the basics of IS to make an informed decision on whether to pursue an education in the IS field.

With this MOOC, interest in the field of IS and knowledge of XYZ University can be fostered in potential students regardless of region, as it is able to reach anywhere with Internet access. This would potentially bring in new IS students for xyz University from all over the country.

For future development, this platform can potentially be implemented in XYZ University's other faculties, so that they can also attract potential students from all over the country and foster interest in their respective fields. In the end, this will also make XYZ University more famous, as the university's name will be known throughout the country.

A mobile application dedicated to the School of IS MOOC could also be created, so that students can access the MOOC at any time without resorting to the mobile web. Social media integration is also something that is being considered, seeing that most of Generation Y are social media-savvy, so it can be used to further promote the School of IS MOOC.

References

1. Gundry, J.: How Flexible is E-Learning? Retrieved from Knowledge Ability. http://www.knowab.co.uk/elflexible.html (2003)
2. Pappano, L.: Massive Open Online Courses Are Multiplying at a Rapid Pace. Retrieved from NYTimes.com. http://www.nytimes.com/2012/11/04/education/edlife/massive-open-online-courses-are-multiplying-at-a-rapid-pace.html?pagewanted=all&_r=1& (12 Nov 2012)
3. Horton, W.: E-Learning by Design. Pfeiffer, San Francisco (2012)
4. Joseph, A.I., Nath, B.A.: Integration of Massive Open Online Education (MOOC) System with in-Classroom Interaction And Assessment and Accreditation: An extensive report from a pilot study. Kolkata: Department of Computer Science, St. Xavier's College (Autonomous) (2013)
5. Barnett, D.R., Chen, X., Stephens, C.: Fad or future: The Advantages and Challenges of MOOCs. Research-to Practice Conference in Adult and Higher Education (2013)
6. Guo, P. J., Kim, J., Rubin, R.: How Video Production Affects Student Engagement: An empirical study of MOOC videos. ACM
7. Honore, S., Schofield, C.P.: Generation Y and learning. Ashridge J. 26–32 (2010)
8. Reilly, P.: Understanding and teaching generation Y. Engl. Teach. Forum 2–11 (2012)
9. Satzinger, J.W., Jackson, R.B., Burd, S.D.: Systems Analysis and Design in a Changing World, 5th edn. Cengage Learning, Boston (2008)

Multi-resolution Analysis of Linear Prediction Coefficients using Discrete Wavelet Transform for Automatic Accent Recognition of Diverse Ethnics in Malaysian English

M.A. Yusnita, M.P. Paulraj, S. Yaacob, M. Nor Fadzilah and Z. Saad

Abstract Accent is a major cause of variability in speaker-independent automatic speech recognition (ASR) systems. Under certain circumstances, this behavioral factor introduces unsatisfactory performance of the systems. Thus, accent analyzer in the preceding stage of the ASR system becomes a promising solution. This paper proposes a multi-resolution approach which applies discrete wavelet transform (DWT) to conventional linear prediction coefficients (LPC) to optimize the extraction of accent from speech utterances in Malaysian English. This paper introduces a multi-numbered LPC (dyadic DWT-LPC) using a defined scale named as level dyadic division scale and an equal-numbered LPC (uniform DWT-LPC) approaches. Using the extracted features, accent models based on K-nearest neighbors were developed. Experimental results showed that the proposed multi-resolution dyadic DWT-LPC and uniform DWT-LPC features surpassed the conventional LPC by significant increases of classification rate of 12.7 and 17.0 % respectively. The promising results of 93.4 % and 88.5 % were achieved using the proposed methods.

M.A. Yusnita (✉) · M. Nor Fadzilah · Z. Saad
Faculty of Electrical Engineering, Universiti Teknologi MARA Malaysia,
13500 Permatang Pauh, Penang, Malaysia
e-mail: yusnita082@ppinang.uitm.edu.my

M.P. Paulraj
School of Mechatronic Engineering, Universiti Malaysia Perlis,
Pauh Putra Campus, 02600 Arau, Perlis, Malaysia

S. Yaacob
Universiti Kuala Lumpur Malaysian Spanish Institute, Kulim Hitech Park,
09000 Kulim, Kedah, Malaysia

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_15

1 Introduction

Numerous studies have attempted to use short-time Fourier transform (STFT) domain to extract accent features in methods such as filter-bank analysis, Mel-frequency cepstral coefficient (MFCC), perceptually linear predictive [1–3], linear prediction coefficients (LPC) and formant analysis [4–6]. Others employed temporal features such as pitch contour and energy [5, 7]. STFT has substantial use in automatic accent and speech recognition (ASR) but the drawback of this method is limited precision fixed by its window size. Even if short-time speech frames of 20–30 ms are considered quasi-stationary, in reality they contain several different phonemes information where a particular accent trait can be embedded. The shift towards multi-resolution paradigm enables discrete wavelet transform (DWT) to be useful in speech and accent recognition. An attempt for phoneme classification using Mel-scale like wavelet packet tree structure was proposed by Farooq and Datta [8] outperformed the standard features, MFCC using linear discriminant analysis. Abdalla and Ali [9] determined MFCC from wavelet sub-bands for speaker identification using hidden markov models to increase robustness in noisy environment. A comparison of MFCC, wavelet packet and perceptually bark scaled wavelet was reported by Tohidypour et al. [10] for robust ASR. Nehe and Holambe [11] demonstrated new features based on prediction coefficients derived from DWT and wavelet packet sub-bands for improved accuracy of isolated-word ASR on the NIST T1-45 database.

To date, there has been a limited number of applications of DWT to the field of accent recognition in the aforementioned literature. It is the aim of this paper to propose a multi-resolution approach which applies DWT to conventional LPC to optimize the extraction of Malaysian English (MalE) accents from two types of speech modes. A new level dyadic division scale (LDDS) is introduced for different sub-bands in order to produce multi-numbered LPC coefficients. It is hypothesized that different DWT sub-band levels require different amount of prediction coefficients to encode the reduced frequency precisions as the level increases. Another approach is to extract equal-numbered LPC coefficients at fourfold cost of feature size. These applications of DWT overcomes the drawback of uniform weighting to the whole spectrum [12] of the conventional LPC. The idea of multi-numbered sub-band coefficients was previously utilized by Tufekci and Gowdy [13] to derive their proposed discrete wavelet coefficients calculated from Mel-scaled log filter bank energies while Nehe and Holambe [11] proposed a uniform amount of LPC-based coefficients generated from DWT and wavelet packet sub-bands.

This paper is organized as follows. In Sect. 2, a brief description about experimental setup and speech database is presented. Section 3 describes the methodology of the proposed multi-resolution techniques using DWT to extract LPC as accent features. The results are discussed in Sect. 4. Lastly, Sect. 5 concludes the important findings of this paper.

Table 1 Distribution of speakers in Malaysian English accents database in terms of ethnic group, gender and types of utterance

Ethnic	Gender	No of speakers	No of utterances	
			IWs	STs
Malay	Male	16	1440	738
	Female	22	1620	918
	Total	38	3060	1656
Chinese	Male	19	1705	849
	Female	15	1350	765
	Total	34	3055	1614
Indian	Male	16	1440	816
	Female	15	1350	765
	Total	31	2790	1581
Total	Male	51	4585	2403
	Female	52	4320	2448
	Total	103	8905	4851

2 Speech Database and Experimental Setup

For the purpose of this research, a new MalE accents database was initiated. Several recording sessions were conducted to elicit speech from MalE speakers of three main ethnics i.e. the Malays, Chinese and Indians. The volunteers, aged between 18 and 40 years were originated from various north, south, west and east regions of the country and as such they were also influenced by their regional accents. The tasks consisted of two modes of speaking namely, isolated words (IWs) and sentences (STs) using reading speaking style. The source of speech materials can be referred from [14, 15]. Each target word was replicated five times while each target sentence was replicated three times of each speaker. The recording was carried in a semi-anechoic acoustic chamber having background noise of approximately 22 dB using a condenser, supercardioid and unidirectional microphone and a laptop computer sound card with MATLAB program. The sampling rate and bit resolution were set to 16 kHz and 16 bps respectively. The number of utterances and their distributions in term of accent, gender and type of utterances are described in Table 1.

3 Methodology

The discrete wavelet transform (DWT) is an alternative to STFT which provides multi-resolution analysis for analyzing different frequencies more accurately. The continuous wavelet transform (CWT) is defined as a summation over all time of the original signal multiplied by successive scaled, shifted versions of the wavelet function and is expressed mathematically as in Eq. (1).

$$X_{WT}(\tau, c) = \frac{1}{\sqrt{|c|}} \int x(t) \cdot \psi * \left(\frac{t-\tau}{c}\right) \mathrm{d}t \quad (1)$$

where $x(.)$ is the signal to be analyzed, $\psi(.)$ is the mother wavelet or the basis function, τ is the shifting parameter which relate to the position along the original signal, thus correlates to time information and c is the scaling parameter which corresponds to the frequency information.

All the wavelet functions used in the transformation are derived from the mother wavelet through translation (shifting) and scaling (dilation or compression) operations. The DWT particularly chooses scales and positions that are based on power of two or well known as dyadic fashion [16–19]. An efficient way to implement this is through the implementation of quadrature mirrors filters (QMF). The most popularly and successfully used wavelet is Daubechies, named after the inventor i.e. Ingrid Daubechies. It is a compactly-supported orthonormal wavelets family [18].

The proposed system for feature extraction process is depicted in Fig. 1 and the specific algorithm to extract the DWT-derived LPC feature vector in frame basis to produce l-dimensional feature database is described below.

Step 1 Initially silence parts of a speech signal are removed. The signal is frame-blocked into a frame length of 32 ms with a frame shift of 16 ms and using fuzzy voiced-unvoiced segmentation, only voiced frames are taken for further processing.

Step 2 Pre-emphasis filtering is applied to the voiced frames using first-order FIR with emphasis coefficient of 15/16 to compensate the attenuation in the spectral energy by 6 dB per octave. Then, Hamming-windowing is applied to smooth out the signal transition at both edges of a 50 % overlapping frame.

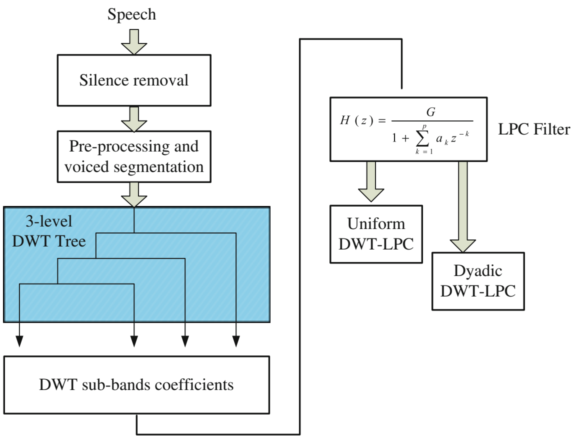


Fig. 1 Block diagram of the proposed feature extraction of hybrid DWT-LPC approach using uniform LPC and dyadic LPC extraction

Step 3 For each frame, DWT is used to decompose a signal into a frame of three-level sub-bands. The wavelet function used in this experiment is Daubechies of third-order.

Step 4 The signal constituents of each frame signal are transformed into DWT coefficients i.e. approximation level-3 (cA3), detail level-3 (cD3), detail level-2 (cD2) and detail level-1 (cD1) and taken as sub-bands for further extraction.

Step 5 LPC coefficients are calculated from each sub-band and all the sub-bands' LPC coefficients are concatenated as input feature vector.

Step 6 Repeat Step 1–5 for the other samples of speech signals and form a matrix of feature vectors with assigned class attribute to each feature vector.

Since each subsequent decomposition level halves the frequency samples of the original signal to produce dyadic sub-bands, for dyadic DWT-LPC, the proposed number of linear prediction coefficients to be extracted from each sub-band is the result of the original LPC-order divided by a method called level dyadic division scale (LDDS). Table 2 describes the proposed LDDS applied to the sub-bands decomposition level.

As the size of the sub-band decreases in the subsequent level, the time resolution increases to better split into lower resolution components of the signal subsequently [18]. Technically, the number of LPC required also should be lessened in dyadic manner in parallel with the lesser resolution of the signal components.

Meanwhile, for uniform DWT-LPC, regardless of their contents, all sub-bands' information is extracted into an equal number of coefficients. Either type of the DWT-LPC features are obtained by concatenating starting from the deepest level of the decomposition tree. Table 3 gives the detail description of the number of LPC parameters in each sub-band of the dyadic type. In higher level of the

Table 2 The proposed level dyadic division scale (LDDS) for dyadic DWT-LPC extraction

Level	LDDS
1	$2^1 = 2$
2	$2^2 = 4$
3	$2^3 = 8$

Table 3 Contribution of LPC coefficients in multi-resolution sub-bands using LDDS for dyadic DWT-LPC features

p-order	Number of prediction coefficients (#LPC)				
	cA3	cD3	cD2	cD1	Total
8	1	1	2	4	8
10	1	1	3	5	10
12	2	2	3	6	13
14	2	2	4	7	15
16	2	2	4	8	16
18	2	2	5	9	18

decomposition tree, there are lesser frequency samples. This is the reason for the reduced number of LPC parameters required as in the pre-stated hypothesis.

Finally, the l-dimensional DWT-LPC feature vectors can be generally represented as in Eq. (2).

$$x^k(l) = [d_1, d_2, d_3, \ldots, d_l]^{\mathrm{T}} \tag{2}$$

where $x(.)$ is the feature vector, $d(.)$ is the DWT-LPC parameter derived either using uniform or dyadic of dimension $l = 1, 2, 3, \ldots$, L evaluated for the kth frame and T denotes transposition. The uniform will have four times as much as the original LPC size.

4 Results and Discussion

In order to complete the design of automatic accent recognition (AAR) system, K-nearest neighbors (KNN) models were developed using LPC and the hybrid of DWT-LPC features. All evaluations of performance utilized independent test samples method by partitioning the feature database into 60–40 % of training and testing datasets respectively. To average out any sources of randomness in the learning methods, ten runs per experiment were conducted for a particular set of parameters and the average of classification rates (CRs) was taken as the objective measures for accuracy performance.

In this experiment, the LPC method was tested against six different orders between 8 and 18 coefficients to investigate the effect to the AAR performance for varying LPC-order on four speech test scenarios by fixing K to 2 and using cosine

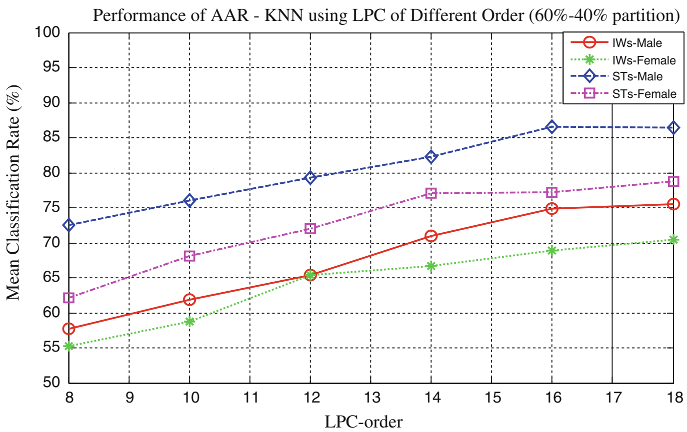


Fig. 2 KNN performance of conventional LPC for IWs and STs speech modes of different genders for varying LPC order

distance metric as parameters. The results are shown in Fig. 2. It is worth noting from the graph that the mean CRs were incredibly increasing as the LPC order increased from 8 to 16 and there was not much increment after that. The increment of mean CRs was between 14.0 and 17.7 % for the male speaker datasets of both speech modes while that of the female speaker datasets was between 15.2 and 16.6 % for varying LPC order from p = 8 to 18. With reference to the LPC order of 16, the best accuracies were yielded for both STs speech mode i.e. 86.5 % for the STs-male and 77.3 % for the STs-female, followed by 74.9 % for the IWs-male and 68.9 % for the IWs-female using this generalized KNN models. It can be concluded that, this achievements were still poor (<80 %) especially for the IWs speech mode.

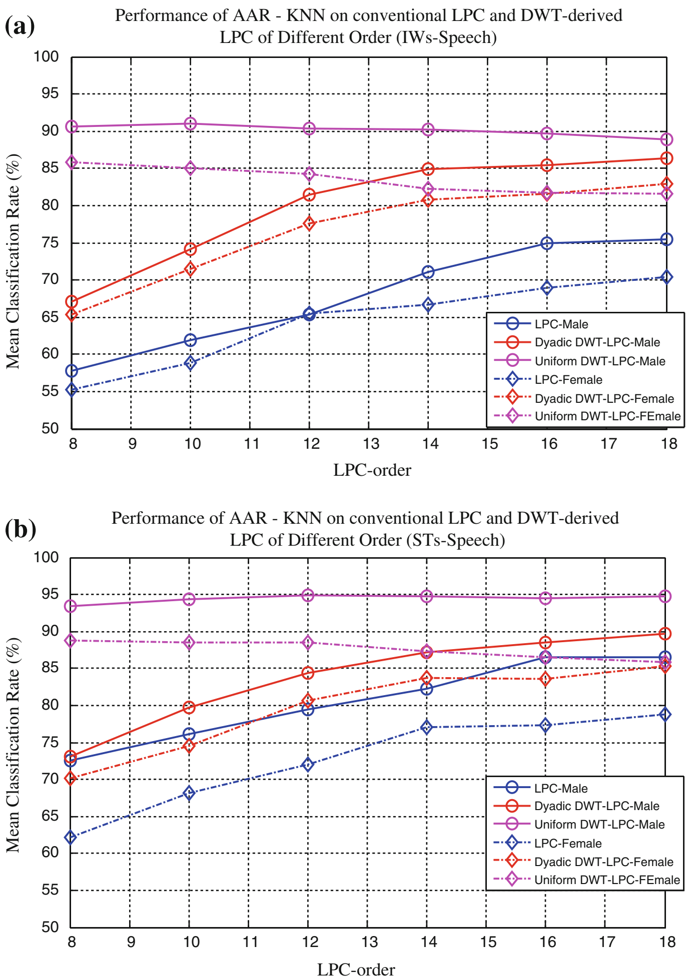


Fig. 3 Performance comparison of conventional LPC and two methods of DWT-derived LPC for **a** IWs and **b** STs speech modes of different genders for varying LPC-order

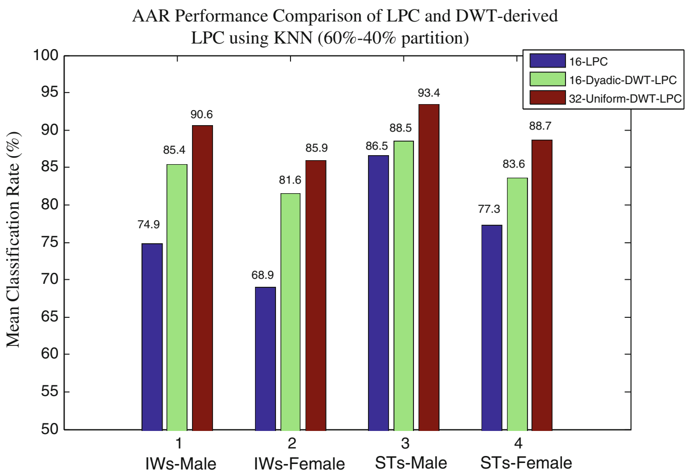


Fig. 4 Performance of the 16-LPC, 16-dyadic-DWT-LPC and 32-uniform DWT-LPC features of four speech test scenarios using KNN

Figure 3a, b compare the performance of the conventional LPC and DWT-derived LPC across varying LPC-order for the IWs speech and STs speech of different genders. The results suggested that either using dyadic or uniform of DWT-LPC methods, the mean CRs could be improved across all orders for both speech modes and genders. It can be observed that the dyadic DWT-LPC performance followed the same trends as the LPC performance curves and surpassed the latter consistently across different orders and genders.

On the other hand, the uniform DWT-LPC displayed a fairly constant performance for the male datasets across the varying orders. However a bit deterioration of performance was observed for the female datasets as the order increased. On top of that, the proposed uniform DWT-LPC yielded the best performance among the three feature sets.

In order to compare percentage of improvement that had been made using DWT-derived features with respect to the conventional LPC, Fig. 4 depicts the best selected LPC-order performance for all speech test scenarios. It is worth noting from the bar chart that the mean CRs had increased by 10.5–12.7 % for the IWs speech mode of both genders whilst 2.0–6.3 % for the STs speech mode of both genders when comparing the dyadic DWT-LPC to conventional LPC. In the same manner, the CRs had increased by 15.7–17.0 % for the IWs speech of both genders whilst 6.9–11.4 % for the STs speech mode of both genders when comparing the uniform DWT-LPC to conventional LPC. The best accuracies were yielded for the STs-male i.e. 93.4 % followed by the IWs-male i.e. 90.6 %. The poor performance issue of the IWs speech mode using the conventional LPC had been resolved by gaining higher accuracy rates of 90.6 and 85.9 % for the male and female datasets respectively and also for the STs-female dataset to obtain an accuracy of 88.7 %.

5 Conclusions

The current work has presented automatic accent recognition of diverse ethnics in Malaysian English (MalE) using KNN models. Extracting a set of features using multi-resolution approach of DWT-derived LPC to correlate with different ethnical accents in MalE has not been done in the past research. Two methods of multi-resolution techniques using DWT were proposed to yield the LPC coefficients namely, dyadic DWT-LPC and uniform DWT-LPC to overcome the weakness of the conventional LPC. The efficacy of the proposed methods were tested on two speech modes i.e. isolated words (IWs) and continuous speech (STs), elicited from male and female volunteers. It was found that the dyadic DWT-LPC yielded better performance than the conventional LPC by as much as 6.3 and 12.7 % for the STs and IWs speech modes respectively. The achievements were more significant for the uniform DWT-LPC by as much as 11.4 and 17.0 % respectively for the STs and IWs speech modes. In the nutshell, we conclude that the multi resolution analysis of LPC with DWT application provides promising results of 93.4 and 88.5 % accuracy of the three-class accent problem in the MalE accents database using these two proposed multi-resolution techniques.

References

1. Arslan, L.M., Hansen, J.H.L.: Language accent classification in American English. Speech Commun. **18**(4), 353–367 (1996)
2. Dupont, S., Ris, C., Deroo, O., Poitoux, S.: Feature extraction and acoustic modeling: an approach for improved generalization across languages and accents. In: 2005 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 29–34. IEEE, San Juan (2005)
3. Picone, J.W.: Signal modeling techniques in speech recognition. Proc. IEEE **81**(9), 1215–1247 (1993)
4. Deshpande, S., Chikkerur, S., Govindaraju, V.: Accent classification in speech. In: 4th IEEE Workshop on Automatic Identification Advanced Technologies, pp. 139–143. IEEE (2005)
5. Liu, W.K., Fung, P.: Fast accent identification and accented speech recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 221–224. IEEE, Phoenix (1999)
6. Yusnita, M.A., Paulraj, M.P., Yaacob, S., Shahriman, A.B, Saidatul, A.: Malaysian English accents identification using LPC and formant analysis. In: 2011 IEEE International Conference on Control System, Computing and Engineering, pp. 472–476. IEEE, Penang (2011)
7. Hou, J., Liu, Y., Zheng, T.F., Olsen, J., Tian, J.: Multi-layered features with SVM for Chinese accent identification. In: 2010 International Conference on Audio, Language and Image Processing, pp. 25–30. IEEE, Shanghai (2010)
8. Farooq, O., Datta, S.: Mel filter-like admissible wavelet packet structure for speech recognition. IEEE Signal Process. Lett. **8**(7), 196–198 (2001)
9. Abdalla, M.I., Ali, H.S.: Wavelet-based mel-frequency cepstral coefficients for speaker identification using hidden markov model. J. Telecommun. **1**(2), 16–21 (2010)
10. Tohidypour, H.R., Seyyedsalehi, S.A., Behbood, H.: Comparison between wavelet packet transform, Bark Wavelet & MFCC for robust speech recognition tasks. In: 2nd International Conference on Industrial Mechatronics and Automation, pp. 329–332. IEEE, Wuhan (2010)

11. Nehe, N.S., Holambe, R.S.: New feature extraction methods using DWT and LPC for isolated word recognition. In: IEEE Region 10 Conference (TENCON 2008), pp. 1–6. IEEE, Hyderabad (2008)
12. An introduction to front-end processing and acoustic features for automatic speech recognition. http://www.nada.kth.se/~rosell/courses/rosell_acoustic_features.pdf
13. Tufekci, Z., Gowdy, J.N.: Feature extraction using discrete wavelet transform for speech recognition. In: The IEEE Southeastcon 2000, pp. 116–123. IEEE, Nashville (2000)
14. Yusnita, M.A., Paulraj, M.P., Yaacob, S., Yusuf, R., Shahriman, A.B.: Analysis of accent-sensitive words in multi-resolution mel-frequency cepstral coefficients for classification of accents in Malaysian English. Int. J. Automot. Mech. Eng. **7**, 1053–1073 (2013)
15. Yusnita, M.A., Paulraj, M.P., Yaacob, S., Mokhtar, N.F., Shahriman, A.B.: Acoustic analysis of formants across genders and ethnical accents in Malaysian English using ANOVA. Procedia Eng. **64**, 385–394 (2013)
16. Cohen, A., Daubechies, I., Feauveau, J.C.: Biorthogonal bases of compactly supported wavelets. Commun. Pure Appl. Math. **45**(5), 485–560 (2006)
17. Mallat, S.: A Wavelet Tour of Signal Processing. Academic Press, San Diego (1998)
18. Misiti, M., Misiti, Y., Oppenheim, G., Poggi, J.M.: Wavelet Toolbox: For Use with MATLAB. The MathWorks Inc, Natick (1996)
19. Weeks, M.: Digital Signal Processing Using MATLAB and Wavelets. Infinite Science Press, Massachusetts (2006)

Brain Cancer Cell Detection Optimization Schemes Using Image Processing and Soft Computing

Chudapa Thammasakorn, Chakchai So-In, Wiyada Punjaruk, Urachart Kokaew, Boonsup Waikham, Songyut Permpol and Phet Aimtongkham

Abstract This paper introduces a novel methodology to automatically measure a number of brain cancer cells using optimized image processing and soft-computing for classification. The former approach is used to prepare the cell image from the medical laboratory, such as background removal, image adjustment, and cell detection including noise reduction. Then, Gabor filter is applied to retrieve the key features before feeding into different soft-computing techniques to identify the actual cells. The results show that the performance of Fuzzy C-Mean with image processing optimization is outstanding compared to neural networks, genetic algorithms, and support vector machines, i.e., 96 % versus less than 90 % in precision, in addition to the superior computational time of around two seconds.

Keywords Brain cancer cell · Cell counting · Cell detection · Image processing · Machine learning · Soft computing

C. Thammasakorn · C. So-In (✉) · U. Kokaew · B. Waikham · S. Permpol · P. Aimtongkham
Applied Network Technology (ANT), Department of Computer Science,
Faculty of Science, Khon Kaen University, Khon Kaen, Thailand
e-mail: chakso@kku.ac.th

C. Thammasakorn
e-mail: t_chudapa@kkumail.com

U. Kokaew
e-mail: urachart@kku.ac.th

B. Waikham
e-mail: boonsup@kku.ac.th

S. Permpol
e-mail: songyut_p@kkumail.com

P. Aimtongkham
e-mail: phet@kkumail.com

W. Punjaruk
Department of Physiology, Faculty of Medicine, Khon Kaen University,
Khon Kaen, Thailand
e-mail: pwiyad@kku.ac.th

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_16

1 Introduction

Brain cancer originates from abnormal neuronal cells, which have aberrant genetic mutation. Different types of brain cancer originate from different types of neuronal cells. Its symptoms are various and depend on the location of the brain tumor including headache, seizure, abnormal vision, vomiting, and mental changes as well as unconsciousness. The brain tumor can be detected by either MRI or CT scan as the main investigation for examining the site and size of brain tumor providing information for surgical approach. Additionally, pathological findings of brain tumor specimens are essentially required for the definite diagnosis [1].

Currently, many types of brain cancers provide poor prognosis with short overall survival rate and patients suffer from side effects of chemotherapy and radiation. One of the major causes of treatment failure is multidrug resistance resulting from cancer stem cells. Therefore, researches in this field to improve treatment outcome of brain tumor patients are necessarily required. To study the characteristics of brain tumors, the experiments are standardly conducted in vitro study using brain tumor cell lines.

In the current studies, the on-going research of brain cancer is very important, especially focusing on the role of cancer stem cells in tumor aggressiveness, multidrug resistance, and tumor recurrence. The research is generally performed towards a methodology study of cell biology to get a better understanding of the cellular and molecular mechanisms involved in brain tumor initiation and growth, such as cell growing, cell proliferation, cell apoptosis, and drug response. This key information will open a new window for brain tumor treatment targeting cancer stem cells to provide better treatment outcome and prognosis [2].

It is worth noting that the quality of the cell line, cell category, and cell counting can be detected primarily by microscope, and then analyzed by researchers or expert doctors. This process may consume too much time as well as the risk of human error, and in particular, when the interpretation relies mostly on their knowledge, expertise, and experience.

Recently, the advances in computational biology and biomedical engineering have played important roles in medical and diagnostics including cell recognition [3]. Many computational techniques have been used to aid doctors including the visual interpretation, one of which is image processing that deals with the recognition from the scan image of patients [4]. In addition, there are several efforts making use of soft-computing techniques as the optimization solver based on their distinctive characteristics [5, 6], appropriately used for imprecision and uncertainty, one of the applications is in the area of cell recognition.

For example, in 2007, Phukpattaranon et al. [7] proposed the algorithm to segment breast cancer cells based on color contents in microscopic images using neural networks (NN) including noise removal and shape simplification. Its main limitation is the classification speed of NN. A year later, Malek et al. [8] focused on the use of fuzzy logic classification with features derived from wavelet transforms including cell detection and segmentation of breast cancer nuclei.

In addition, in 2012, Han et al. [9] investigated how to apply support vector machines (SVM) as a classification technique based on Laplace edge features to detect cell nuclei. This proposal is outstanding; however, its performance may be limited with cell tint. In the same year, similarly, Arteta et al. [10] also evaluated a machine learning method for cell detection based on the SVM framework including dynamic programming techniques to select a non-overlapping region. However, this method may be interfered with the intensity and density of cell images.

There are limitations with those approaches, e.g., no detailed discussion or low classification accuracy; thus, the objective of this study is twofold: first, to investigate the possibility to apply various baseline soft-computing techniques using traditional image processing specifically for brain cancer cell detection; then, second, to propose a novel methodology applying the optimization of the image processing with Fuzzy C-Mean to identity the actual cells. Note that with the extra image processing steps, several image preparations were performed using Gabor filter to retrieve the key features as a representation of the cell before applying Fuzzy C-Mean for final brain cancer cell detection.

This paper is organized as follows: Sect. 2 briefly provides an overview of baseline methodology of image processing and soft-computing applied in the context of brain cancer cell recognition. Then, Sect. 3 presents our novel methodology including a detailed description. The comparative performance of different models and approaches are illustrated and discussed in Sect. 4. Finally, Sect. 5 contains our conclusion and possible future work.

2 Image Processing and Soft Computing for Cell Detection

This section states an overview of applying image processing and soft-computing based classification into cell detection.

2.1 Image Processing

Image processing is one of the computational processes to automatically recognize or interpret the image given its characteristics, such as size, shape, and direction, one of these is applied in the area of cell recognition as one of the pattern recognition techniques. In general, as discussed in the research by Al-tarawneh [11], there are four main methods for image processing as follows:

(a) Gray Scale Convertor: this step is used to convert *RGB* into gray (I) scales. An example of this conversion is stated in the equation below [12].

$$I = \frac{1}{3}(R + G + B) \tag{1}$$

(b) Image Segmentation: this process is used to divide an image into multiple parts. This is typically used to identify objects or other relevant information in digital images. There are different ways to perform image segmentation such as Morphology—to approximate the background (*Bg*) including illumination adjustments using opening operation (Open) given input image (*I*) and template (*T*) over erosion and dilation.

(c) Background Subtraction: to increase the difference between the actual cell and its background, this step (*F*) is used to cut-off the background (*Bg*) from the gray scale (*I*) of two dimensional images as stated in the equation below [13].

$$[F_{ij}]_{mxn} = [I_{ij}]_{mxn} - [Bg_{ij}]_{mxn} \tag{2}$$

(d) Noise Removal: this step is used to remove unwanted noise such as any pixels that are not significant given the threshold [14].

2.2 *Soft Computing*

Soft computing techniques are generally used to resolve a problem which involves an uncertainty as a non-linear solver [5]. There are several classes of soft computing, e.g., NN, Fuzzy Logic (FL), SVM, and Evolutionary Computation (EC) such as Genetic Algorithm (GA) [6], all of which are used as our comparative study.

2.2.1 Neural Network (NN)

Neural Network [15] is one of the techniques to imitate the human brain behavior to instruct the pattern recognition framework, and here, multi-layer perceptron (MLP) was selected for the purpose of brain cancer cell detections. The general framework consists of input (*N*), hidden layer (*M*), and output (*J*). In this research, the input data is 1440 nodes (based on the matrix from Gabor filter feature extraction [16]). The number of hidden nodes are varied from 1 to 64 given 400 as a maximum number of iterations (*R*) and an error threshold (*e*) as 0.01; and here, the random weight (*w*) is in range of [−1 1].

The learning mechanism will be re-computed given the estimated average error in each round comparing the resulting output to the threshold of cell identification, and here is 0.9 for cell and −0.9 for non-cell [17] resulting in the proper weight. For testing, this particular weight will then be computed with the testing image given the threshold for cell identification decision. Note that these parameter selection criteria (also with other soft-computing techniques) also depend on the data dimensionality.

2.2.2 Genetic Algorithm (GA)

Genetic Algorithm (GA) is one of the evolutionary computational methods [6]. Generally, GA imitates the way to solve the problem to reach the optimum point using genetic transformation, such as mutation and crossover. The optimal solution starts with the initial population (*P*) randomly selected, and then feeds into the genetic process as one generation in order to retrieve the local solution, which is then used to compare with the global solution given a particular threshold.

In this research, GA is used to generate a proper threshold for binary image identification against the image matrix generated from image processing (after the noise removal state) in addition to the black and white to gray conversion. The final decision of cell detection is via region properties [17]. Considering the GA stage, the random value between 0 and 255 was selected as the initial population (here is 100). The population will be ranked based on the fitness value stated in Eq. 3. Here, $W0$ is the summation of the intensity values of the processed image via two dimensional histogram computation of size s and t (each of which is the value from the selected genes). $H0$ is the summation of the multiplication between the intensity value and logarithm of its inverse. Hst is similar to $H0$ but with the size of the actual processed image [18].

Next, some of the genes which have high fitness values will be selected to apply genetic operators in a binary operation (bitwise) to create a new population [18]. This procedure will be re-computed given the iteration threshold, i.e., 500. The gene with the highest fitness value will be used as the final threshold for the final cell detection.

$$Fitness = \log(W0 \times (1 - W0)) + \left(\frac{H0}{W0}\right) + \left(\frac{Hst - H0}{1 - W0}\right) \tag{3}$$

2.2.3 Fuzzy System

The fuzzy C-Mean (FCM) algorithm for detecting the brain cancer cell image is used to compare the gray scale value of every pixel with the value of the cluster center. The main difference is used to determine if the cluster center of the pixel should belong to; it assigns a value between 0 and 1 for each cluster. Here, the fuzzy rule states that the summation of the membership value of a pixel to all clusters must be 1. The higher the membership value, the more likely that pixel belongs to that cluster. The FCM clustering is obtained by minimizing an objective function shown in Eq. 4 [19]:

$$J = \sum_{k=1}^{c} \sum_{i=1}^{n} \mu_{ik}^{m} ||p_i - v_k||^2 \tag{4}$$

In this equation, J is the objective function; n is the number of pixels in the brain cancer cell image (after the noise removal stage in addition to the black and white to gray conversion) in size of $M \times N$. c denotes the number of clusters, and here is 2. μ is the fuzzy membership value. m denotes a fuzziness factor, and here, its value is greater than 1. p_i is the ith pixel in the image; v_k is the centroid of the kth cluster. $\| \|$ is the inner product from pixel point p_i to the kth cluster center (v_k) (e.g., Euclidean distance).

Note that the selection of the centroid of the kth cluster is achieved using Eq. 5. In addition, Eq. 6 shows the selection criterion of fuzzy membership. In addition, after the computation for pixel clustering (resulting in two groups), there is another step to identify the brain cancer cell from the total image given the threshold, i.e., the average of the maximum and minimum of the first and second groups, respectively.

$$v_k = \frac{\sum_{i=1}^{n} \mu_{ik}^{m} p_i}{\sum_{i=1}^{n} \mu_{ik}^{m}} \tag{5}$$

$$\mu_{ik} = \frac{1}{\sum_{l=1}^{c} \left(\frac{p_i - v_k}{p_i - v_l} \right)^{\frac{2}{m-1}}} \tag{6}$$

2.2.4 Support Vector Machine (SVM)

Support Vector Machine [20] is one of the classification techniques, and generally applies a binary classification using the linear equation to separate the data into two, i.e., between cancer cell and non-cancer cell. The kernel function is also used to convert the data into higher dimensions in order to be suitable for linear classification. Note that to construct a hyperplane in a feature space, SVM has to transform the n-dimensional input vector x into N-dimensional feature vector through a choice of an N-dimensional vector function (ϕ) as stated in equation below.

$$\phi : \mathbb{R}^n \rightarrow \mathbb{R}^N \tag{7}$$

It is worth noting that given the training data set $D = \{(x_i, y_i), i = 1, 2, \ldots, n\}$ in that $x_i \in \mathbb{R}^n, i = 1, 2, \ldots, n$ where each value of x_i belongs to one of the two classes that is identified by the class label $y_i \in \{-1, 1\}$. For cell detection, x_i is a representation of a fixed size image (18 × 18 pixels being a 1440 dimensional integer valued vector) containing an instance of the cancer cell ($y_i = +1$) or non-cancer cell ($y_i = -1$) pattern (See Eq. 8). Here, the kernel function $K(x, x_j) = \emptyset(x)\emptyset(x_j)$ and additional weight (w) are used to link the feature space in dimension j into outer space; b is the bias or threshold which was set to 0.8. In this research, three different main kernels were investigated, namely, Linear, Polynomial, and Quadratic.

$$f(x) = sign\left\{ \sum_{j=1}^{n} w_j \phi_j K(x, x_j) + b \right\} \tag{8}$$

3 Brain Cancer Cell Detection Using Image Processing and Soft Computing

Three main steps are used to count the brain cancer cells excluding the training and testing processes as follows: image processing, feature extraction, and classification.

3.1 Brain Cancer Cell Image Processing Stages

In addition to the baseline image processing technique for cell recognition during image processing, previously discussed in Sect. 2, we propose a modification of that technique by introducing five extra steps, nine main steps in total, to prepare the cell image for the next step (or feature extraction) as follows. Note that each step can be visually illustrated in the Fig. 1a–i.

(a) Gray Scale Convertor: this step is used to convert *RGB* into gray (*I*) [12].
(b) Morphology: this step is used to approximate the background using morphological opening operation for (non-uniform) illumination adjustments.

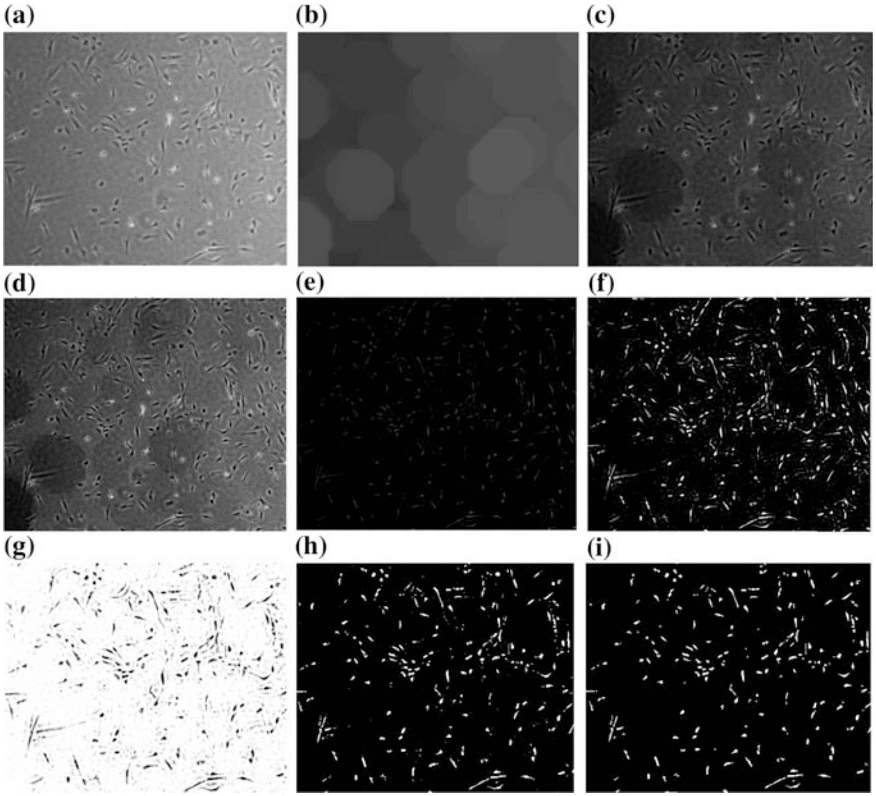


Fig. 1 Image processing steps

(c) Background Subtraction: to increase the difference between the actual cell and its background, this step (F) is used to cut-off the background (Bg) from the gray scale (I) of two dimensional images [13].
d) Matrix Conversion: this step is used to change the image matric to normalize gray scale images such that their pixels (G) will be in range of 0.00–1.00.
(e) Image Filter Subtraction: this step (Z) is used to remove the gray portion (G) and constant (C) from the two dimensional filter (K) images as stated in the equation below.

$$\left[Z_{ij}\right]_{mxn} = \left[K_{ij}\right]_{mxn} - \left[G_{ij}\right]_{mxn} - \left[C_{ij}\right]_{mxn} \tag{9}$$

(f) Image Expansion: to improve the brightness and clarity for image adjustment, this step is used to expand the image from the narrow-band gray scale image to make it wider in normalized 256 levels of color range.
(g) Image Completion: in the complement of a binary image, at this step, zeros become ones and ones become zeros; black and white are reversed. In the complement of a variety of intensity (between 0.00 and 1.00), each pixel value is subtracted from the maximum pixel value supported by the class (or 1.0 for double-precision images) and the difference is used as the pixel value in the output image, i.e., dark areas become lighter and light areas become darker.
(h) Color Generation: this step is used to convert the image back to the original one using *not* operator such that the actual cell will be in white and black for the background using a specific threshold.
(i) Noise Removal: this step is used to remove unwanted noises such as any pixels that are not significant given the threshold [14].

3.2 Brain Cancer Cell Feature Extraction

To achieve feature extraction, there are four main steps as follows. First, the cell image will be divided into 18 × 18 pixels and we then apply histogram equalization to improve the image clarity [21]. After that, the results will be fed into the Gabor Filter [16] to create the image matrix of (5 × 18, 8 × 18), and here, the scaling factor selection (M) is 5 and direction (N) is 8. Finally, with the compression ratio of three (90/3, 144/3), this matrix will be converted again into an image vector of 1440 (30 × 48) in size.

3.3 Brain Cancer Cell Classification

After the previous step, the main features of each image will be generated. As stated in Sect. 2, the evaluation is over four baseline soft-computing techniques, namely, NN (MLP), EC (GA), Fuzzy Logic in terms of Fuzzy C-Mean, and SVM.

4 Performance Evaluation

In this section, the performance of the methodology in various soft-computing based approaches was evaluated including the optimization of image processing techniques.

4.1 Simulation Configurations

To comparatively show the performance of each proposal, the simulation test bed is a standard library and models in Matlab [17] on a Windows 7 64-bit machine with 2.10 GHz, 8 GB DDR-SDRAM, and 1 TB 7200 ppm hard disk. The microscopic brain cancer cell images were provided by the department of Physiology, Faculty of Medicine, Khon Kaen University. The cell images including noise and background were in forms of JPEG files (web.kku.ac.th/chakso). The trained images will be divided into two classes: either brain cancer cell or not as the image size of 18 × 18 pixels. For testing, a whole cell image will be used in JPEG format such that the resolution is less than 600 × 700.

To validate the performance, a standard K-fold cross validation technique was performed in that K is 4, i.e., the percentage of training is 75 and 25 for the testing. Due to the limitation of timing constrain for each experiment, there is 200 images in total. Two main metrics were applied: mean and standard deviation in terms of detection precision (% accuracy) and computational time (seconds).

Since there are also configurations of each soft computing technique, our investigation then evaluated each accordingly. Thus, in this research, there are four main scenarios to illustrate the performance. First, with the limitation of human justification, the effect of number of hidden nodes of NN was evaluated by ranging from 1, 8, 16, 32, and 64, respectively. Second, three main kernel functions of SVM were evaluated, i.e., Linear, Polynomial, and Quadratic. Third, the configurations providing the best solution of NN and SVM were used to compare with GA (500 iterations with 100 in populations) and FCM accordingly. Finally, our optimized image processing methodology applying FCM, called FCM O-Image, will be also evaluated against the others.

4.2 Simulation Results

Considering the first scenario, Table 1 shows the effects when varying the number of hidden nodes of NN. There is a significant difference among the detection precision, i.e., around 75–86 %; with 64 nodes being the best, but with the highest computational time—more than 20 min. In general, the execution time of all nodes is high.

Table 1 Detection performance: NN with various numbers of hidden nodes (1–64)

Hidden nodes	Accuracy (%)	Execution time (s)
1	75.02	781.67
8	80.79	844.40
16	81.02	938.72
32	83.73	1115.75
64	86.01	1296.24

Table 2 Detection performance: SVM with various kernel models

Type	Accuracy (%)	Execution time (s)
Linear	60.84	488.29
Polynomial	61.76	437.48
Quadratic	61.17	478.43

Table 2 shows the performance in terms of accuracy and computational complexity of different SVM kernels. It is observed that different kernel functions have no significant effects on the detection precision (around 60–61 %); however, using Polynomial as the kernel function achieves the lowest computational time, i.e., around 6 min compared to the others, i.e., 7–8 min.

In addition, with the best performance of NN (64 hidden nodes) and SVM (polynomial), Table 3 shows the comparative performance against the other two soft-computing approaches, i.e., GA and FCM. Generally, FCM results in the best approach in terms of detection precision—almost 90 % and the accuracy is in the sequence of NN, GA, and SVM, respectively. However, the computational time of GA is the lowest (around 3.85 s) and in the sequence of FCM, SVM, and NN, respectively.

In the last scenario, our optimized image processing with FCM (FCM O-Image) was evaluated to improve the accuracy further. Table 3 also shows that the detection precision of the optimization is around 95.69 % while taking the computational time of just only 2.56 s. Note that in our intensive evaluation, the image processing optimization was also integrated to other soft-computing techniques and the accuracies are 91.21, 79.06, and 86.37 % with 1429.61, 454.57, and 4.77 s in computational time, respectively.

Table 3 Detection performance: NN, SVM, GA, FCM; and FCM O-Image

Type	Accuracy (%)	Execution time (s)
NN	81.02	938.72
SVM	61.76	437.48
GA	76.17	3.85
FCM	87.27	2.16
FCM O-Image	95.69	2.56

5 Conclusions

A methodology study of cell biology to gain knowledge of cellular and molecular mechanisms in brain tumors plays an important role in the understanding of brain cancer and this requires the expertise of researchers and doctors. The computational biology and bio-engineering were invented for aiding and especially mitigating human errors. Thus, this research focuses on the image processing with soft-computing approaches for brain cancer cell detection. Various models of soft computing were comparatively evaluated, i.e., NN, GA, SVM, and FCM.

It should be noted that although NN can yield high accuracy, but with the limitation of time complexity. There are many steps of GA probably affecting the outcome. SVM may not classify the tainted images. However, FCM yields the highest accuracy and most importantly with its optimization of image processing, called FCM O-Image, the result can be improved up to 96 % while maintaining less time computational complexity. However, although the methodology can achieve performance improvement, more investigation is needed in other scenarios and constraints including a wide variety of cell images and an extensive image database, and all of these are for future work.

References

1. Lyer, V., Lee, S.: MRI, CT, and PET/CT for ovarian cancer detection and adnexal lesion characterization. Am. J. Roentgenol. **194**, 311–321 (2010)
2. Stephen, C.F.: Cross-linking of matrix polymers in the growing cell walls of angiosperms. Ann. Rev. Plant Physiol. **37**, 165–186 (1986)
3. Sarai, A., Siebers, J., Selvaraj, S., Gromiha, M.M., Kono, H.: Integration of bioinformatics and computational biology to understand protein-DNA recognition mechanism. J. Bioinform. Comput. Biol. 83–169 (2005)
4. Faggiano, E., Lorenzi, T., Perotto, S.: TV-H^{-1} variational inpainting applied to metal artifact reduction in CT images. Comput. Vis. Med. Image Process IV. **4**, 277–282 (2013)
5. Jang, H., Topal, E.: A review of soft computing technology applications in several mining problems. Appl. Soft Comput. **22**, 638–651 (2014)
6. Zhang, J., Zhan, Z., Lin, Y., Chen, N., Gong, Y., Zhong, J., Chung, H.S.H., Li, Y., Shi, Y.: Evolutionary computation meets machine learning: a survey. IEEE Comput. Intell. Mag. **6**(4), 68–75 (2011)
7. Phukpattaranon, P., Limsiroratana, S., Boonyaphiphat, P., Kayasut, K.: Automated breast cancer cell image segmentation. In: International Conference on Biomedical Engineering, pp. 241–244. Springer, Malaysia (2006)
8. Malek, J., Sebri, A., Mabrouk, S., Torki, K., Tourki, R.: Automated breast cancer diagnosis based on GVF-snake segmentation, wavelet features extraction and fuzzy classification. J. Sig. Process Syst. **55**, 49–66 (2008)
9. Han, J., Breckon, T.P., Randell, D.A., Landini, G.: The application of support vector machine classification to detect cell nuclei for automated microscopy. Mach. Vis. Appl. **23**(1), 15–24 (2010)

10. Arteta, C., Lempitsky, V., Noble, J.A., Zisserman, A.: Learning to detect cells using non-overlapping extremal regions. In: Medical Image Comput and Computer-Assisted Intervention. pp. 348–356, Springer, France (2012)
11. Al-tarawneh, M.S.: Lung cancer detection using image processing techniques. Leonardo Electron. J. Practices. Technol. **11**, 147–158 (2012)
12. Bagley, J.D.: The behavior of adaptive systems which employ genetic and correlation algorithms. Doctoral dissertation (1967)
13. Hoppner, F., Klawonn, F., Kruse, R., Runkler, T.: Fuzzy Cluster Analysis, pp. 36–43. Wiley, New York (1999)
14. Othman, A.: Generalised object detection and semantic analysis: casino example using matlab. Clin. Orthop. Relat. Res. (2011)
15. Corinna, C., Vladimir, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
16. Prasad, N., Domke, J.: Filter Visualization. Technical Report. University of Maryland (2005)
17. Matlab R2014a (www.mathworks.com)
18. Lihongyan.: Using genetic algorithms for image segmentation of the source. (2006)
19. Waleed, A., Siti, A., Shahnorbanun, H.: MRI brain segmentation via hybrid firefly search algorithm. J. Theor. Appl Inf. Technol. **61**(1), 73–90 (2014)
20. Cristianini, N., Taylor, J.H.: An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, Cambridge (2000)
21. Omprakash, P., Yogendra, P.S., Maravi, S., Sanjeev, S.: A comparative study of histogram equalization based image enhancement techniques for brightness preservation and contrast enhancement. Sig. Image Process. **4**, 11–25 (2013)

Robust Iris Segmentation Through Parameterization of the Chan-Vese Algorithm

Gugulethu Mabuza-Hocquet, Fulufhelo Nelwamondo and Tshilidzi Marwala

Abstract The performance of an iris recognition system relies on automated processes from the segmentation stage to the matching stage. Each stage has traditional algorithms used successfully over the years. The drawback is that these algorithms assume that the pupil-iris boundaries are perfect circles sharing the same center, hence only use circle fitting methods for segmentation. The side effect posed by the traditional rubber sheet model used for normalization is; one cannot work backwards to place the discriminative features on the original image. This paper proposes a different approach to each stage using algorithms different from the traditional ones to address the above issues. Bresenham's circle algorithm to locate and compute pupil-iris boundaries. Chan-Vese algorithm with pre-defined initial contour and curve evolution parameters for accurate segmentation. Preprocessing techniques to enhance and detect iris features for extraction. Labeling features of strongest pixel connectivity and using Harris algorithm for feature extraction and matching.

Keywords Iris segmentation · Chan-Vese algorithm · Sobel edge detector · Harris corner detector · Feature matching

G. Mabuza-Hocquet (✉) · F. Nelwamondo
Modelling and Digital Science, Council for Scientific and Industrial Research, Pretoria, South Africa
e-mail: GMabuza@csir.co.za

F. Nelwamondo
e-mail: FNelwamondo@csir.co.za

G. Mabuza-Hocquet · F. Nelwamondo · T. Marwala
Engineering and the Built Environment, University of Johannesburg, Johannesburg, South Africa
e-mail: TMarwala@uj.ac.za

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_17

1 Introduction

The iris is regarded as the most reliable biometric trait, [1, 2]. The high accuracy achieved during identification and verification technology is due to its unique and intricate pattern for each person for each eye. An iris recognition system comprises of four classical stages prevailing after image acquisition. These are (i) segmentation; separates the iris from the rest of the eye. (ii) Normalization; caters for image size variations and pupil dilations, is achieved using Daugman's rubber sheet model. (iii) Feature extraction stage extracts features of interest from the normalized iris. Lastly feature comparison to match a reference iris template to a query image [1–3]. Each of the stages are automated processes that already constitute traditional algorithms to achieve results. However, when the original eye image has been captured with non-sophisticated equipment under uncontrolled environments with varying image quality, it becomes almost impossible to even perform the first process of segmentation with the use of the traditional segmentation algorithm [4]. The work done here establishes the use of different methods from the classical ones to still attain accurate results regardless of any irregularities that come with the original image and its quality. The proposed method also achieves robust and affordable results from the segmentation stage to a high matching accuracy at a faster rate.

2 Problem Definition

The processes of an iris recognition system have had classical algorithms used successfully for each stage over the years. Like any other method or algorithm, they still need to be designed such that they are suitable for varying needs and environments. In iris segmentation, for instance, the use of the traditional integro differential equation followed by the Hough Transform (HT) [1–3], is the norm. However, this algorithm rigidly assumes that the pupil and iris boundaries share the same center and hence only models iris segmentation as a perfect circle when this is not generally the case. Due to this, segmentation becomes unsuccessful especially when the location of the original image is not around the camera's centered gaze. Also, the 2005/6 Iris Challenge Evaluation's (ICE) experimental results demonstrated that most iris recognition algorithms have high false rejection rates (FRR), [4, 5] and a concerning low acceptance rate. These factors highlight opportunities for improvement and development of novel algorithms to still attain accurate results under any circumstances.

The objective of this paper is to approach each stage of iris recognition using different algorithms from the traditional ones to achieve desired results given any image quality by; (i) Using Bresenham's circle algorithm to localize and detect the pupil-iris boundaries. (ii) Parameterizing and implementing the Chan-Vese algorithm using the parameters obtained in (i), that readily define the initial contour and will stop the evolving curve at the desired location to achieve accurate segmentation of the iris. (iii) Using preprocessing techniques to perform feature enhancement on segmented images. (iv) Using the Sobel detector to highlight iris features for extraction. (v) Using pixel connectivity to label only the articulated iris features. (vi) Lastly, using Harris algorithm to extract, and match corresponding iris global features of interest between a reference feature template and its query image.

3 Related Work on Iris Segmentation

Automatic segmentation is one of the basic, crucial and most challenging tasks [6], that an iris recognition system needs and relies on. Improved, fast and accurate segmentation methods lead to accurate feature extraction, inducing a positive impact not only on the matching subsystem but the performance of the entire iris recognition system. Classical approaches to iris segmentation have been the use of the traditional integro-differential operator, an algorithm introduced by J. Daugman [1]. The gradient-based binary edge map construction followed by Hough Transform (HT) algorithm proposed by Wildes [3], uses circular boundaries to localize the iris. The commonality in these algorithms is that (i) they are either intensity-based methods or region-based methods. They only assume that the pupil-iris boundary is concentric hence a perfect circle and therefore only make use of circle fitting methods in order to locate a non-circular pupil-iris boundary. When the above methods are used and the iris is not located at the center of the image, these approaches prove to be unsuccessful. In summary, these methods do not necessarily consider the non-concentric geometric contours that brand the pupil and iris boundaries where features are most distinctive.

A technique proposed and developed by Kass [7] for image segmentation is known as active contour models or snakes. It is a method that finds a curve that will suitably separate an object of interest from the rest of the image by moving through the spatial domain of an image to minimize an energy functional. For instance, let Ω be a bounded open subset of $\mathbb{R}^2$, with $\partial\Omega$ its boundary. Let u_0: be a given image and $C(s)$:[0 1] $\rightarrow \mathbb{R}^2$ be a parameterized curve starting around the object to be detected. Depending on the gradient of the image (u_0), active contour models make use of an edge-detector to stop an evolving contour on the desired object boundary and defined by: $\inf_C J_1(C)$ [8], where:

$$J_1(C) = \alpha \int_0^1 |C'(s)|^2 ds + \beta \int_0^1 |C''(s)| ds - \lambda \int_0^1 |\nabla u_0(C(s))|^2 ds. \tag{1}$$

where α, β, λ are positive parameters, C' and C'' denote the first and second derivatives of *C(s)* with respect to *(s)*.The first two terms referred to as the internal energy; control the smoothness of the contour. The third term attracts the contour towards the object in the image (external energy), [8]. The drawback with such models is that if the discrete gradients are bounded, the stopping function will not be zero on the edges and this may cause the evolving curve to simply pass through the object boundary. Chan and Vese [8], state that by minimizing the energy (1), the curve can be located at the points of maxima hence acting as an edge detector but keeping smoothness in the object boundary.

3.1 The Chan-Vese Algorithm in Summary

The method of Chan and Vese [8] is "the minimization of energy based segmentation", where they minimize a fitting term with an addition of regularizing terms, (2). They start by defining an evolving curve C in Ω, as the boundary of an open subset ω of Ω. The region inside and outside the curve is denoted by ω and $\Omega \backslash \varpi$ respectively. Their method assumes that an image u_0 is formed by two regions of distinct values $u^i{}_0$ and $u^o{}_0$ having piecewise-constant intensities. With $u_0 \approx u^i{}_0$ and $u_0 \approx u^o{}_0$ falling inside and outside the object boundary (C_0) respectively. They then add regularizing terms; the length of the curve C and the area of the region inside C, which when introduced to the energy functional becomes:

$$F(c_1, c_2, C) = \mu.Length(C) + v.Area(inside(C)) + \lambda_1 \int\limits_{inside(C)} |u_0(x, y - c_1)|^2 dxdy + \lambda_2 \int\limits_{outside(C)} |u_0(x, y) - c_2|^2 dxdy. \tag{2}$$

where $\mu \geq 0$, $v \geq 0$, λ_1, $\lambda_2 > 0$ are fixed parameters, constants c_1 and c_2 depend on C, are respective averages of the image inside and outside C. Considering a case of arbitrary dimension, the isoperimetric inequality can be used which makes *(Length (C))*$^{N/N-1}$ to be comparable with the Area (inside (C)). For this work the proposed regularizing terms are pre-defined by first determining the diameter of the pupil and iris respectively. This is done to (i) compute the radii and respective centers of the pupil and iris. (ii) To pre-determine the boundaries C_0 of the initial contours and (iii) to influence the stopping of the evolving curve C, before segmentation is performed. The model of Chan and Vese is that of a minimal partition problem which can be formulated and solved using level set method. The variable of the evolving curve is replaced by the zero level set of a Lipschitz function $\phi{:}\Omega \rightarrow \mathbb{R}$. This simplifies the Mumford-Shah functional; which is responsible for segmentation. Using the formulation of the level set, the solution image u, constants c_1 and c_2 can be expressed can as a function of ϕ by:

$$c_1(\phi) = \int_{\Omega} u_0(x,y)H(\phi(x,y))\,dxdy \bigg/ \int_{\Omega} H(\phi(x,y))\,dxdy, \tag{3}$$

$$c_2(\phi) = \int_{\Omega} u_0(x,y)(1 - H(\phi(x,y)))dxdy \bigg/ \int_{\Omega} (1 - H(\phi(x,y)))dxdy \tag{4}$$

where H is the Heaviside function. Parameterizing the descent direction by some artificial time $t \geq 0$ and deducing the associated Euler-Lagrange equation, leads to the active contour model, [8]:

$$\frac{\partial\phi}{\partial t} = \delta_{\varepsilon}(\phi)\left[\mu\, div(\frac{\nabla\phi}{|\nabla\phi|}) - v - \lambda_1(u_0 - c_1)^2 + \lambda_2(u_0 - c_2)^2\right] = 0\, in\, (0,\infty) \times \Omega, \tag{5}$$

$$\phi(0,x,y) = \phi_0(x,y)\ in\ \Omega\ \text{ and }\ \frac{\delta_{\varepsilon}(\phi)}{|\nabla\phi|}\frac{\partial\phi}{\partial\vec{n}} = 0\ on\ \partial\Omega, \tag{6}$$

where $\phi(0,x,y) = \phi_0(x,y)$ defines the initial contour, $\vec{n}$ is the exterior normal to the boundary $\partial\Omega$ and $\partial\phi/\partial\vec{n}$ denotes the normal derivative of ϕ at the boundary. The principal steps of the algorithm:

- Initialize ϕ^0 by ϕ_0, $n = 0$.
- Compute $c_1(\phi^n)$ and $c_2(\phi^n)$ by (3) and (4)
- Solve partial differential equation in ϕ from (5), to get ϕ^{n+1}
- See if solution is stationery, otherwise, n = n+1 and repeat, [8].

The segmentation approach that has been used here closely follows that of Yahya and Nordin [9], and Yanto et al. [10]. In their work [9] adopt the AdaBoost Cascade Detector to detect the iris region and to determine that the eye is not closed. They apply the Chan-Vese method to find the inner and outer boundaries of the iris. Their approach results in a rectangular block as the initial contour. The work done by [10], proposes an improvement of the active contours without edges model by Chan-Vese to perform iris segmentation. Their algorithm formulates the energy function defined by Chan-Vese as a Bayesain optimization problem. They incorporate a prior probability in the energy function to achieve robust segmentation of the iris. Another iris segmentation algorithm is one proposed by Vatsa et al. [4]. Their algorithm performs iris segmentation by (i) the detection of an approximate elliptical boundary using intensity thresholding. (ii) The modification of the Mumford–Shah functional to a narrow band over the above estimated boundaries to compute the exact inner and outer boundaries of the iris. The challenge with this approach is that if the initial contour of the iris is not placed exactly on the iris boundaries, the algorithm does not accurately segment the iris.

4 Proposed Approach

The block diagram demonstrates the proposed approach to each process of the iris recognition system to ultimately obtain highly accurate matching comparison results (Fig. 1).

4.1 Image Acquisition

The proposed algorithm is tested on two different databases. Non-ideal images from the University of Beira Interior (UBIRIS). Independently collected eye images from consenting participants using the Vista EY2 dual iris and face camera.

4.2 Segmentation Stage

This paper adopts the notion of Chan-Vese [8] algorithm by (i) adding pre-defined length comparable to the area of the pupil and iris as regularizing terms to the energy functional. (ii) Using active contour models in relation to the Mumfor-shah functional to perform segmentation. This model (i) detects and segments contours with or without gradient and (ii) objects that have very smooth and discontinuous boundaries. As the regularizing terms this work proposes the use of the Bresenham's circle algorithm, [13] which is based on the midpoint circle algorithm in order to localize the pupil-iris boundaries (C_o) prior segmentation and to stop the evolving curve on the area designated by the computed pupil and iris radii respectively. Bresenham's algorithm determines points that are needed to draw a circle. From the original image $u_0(x, y)$, the proposed method starts by independently determining respective end points $u_0(x_1, y_1)$ and $u_0(x_2, y_2)$ of both the iris and pupil diameter to locate their respective centers *(h, k),* using the midpoint formula. This is elaborated by Fig. 2:

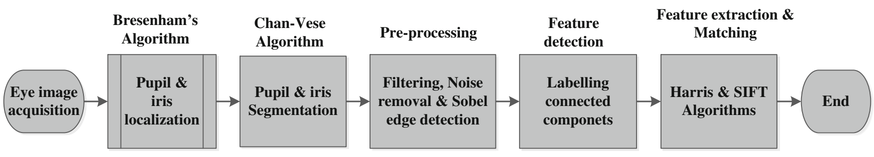


Fig. 1 Model of proposed approach

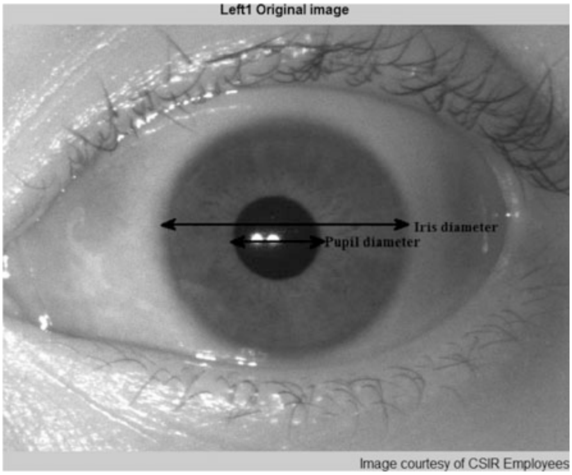


Fig. 2 Computation of iris and pupil diameters

With the midpoint formula to determine each center (*h, k*) of both the pupil and iris given by:

$$u_0(x)\left(\frac{x_1+x_2}{2}\right),\ u_0(y)\left(\frac{y_1+y_2}{2}\right) \Rightarrow (h,k). \tag{7}$$

The respective radii of the pupil and iris are calculated using the distance formula:

$$d=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2} \quad \text{and} \quad r=d/2 \tag{8}$$

The midpoint circle algorithm starts with an equation of a circle and the objective is to find a path through a pixel grid by using pixels with solutions as close to $x^2 + y^2 = r^2$ as possible. The value of r is only computed once, during initialization. Therefore, in relation with the Mumford-Shah functional the segmentation is represented by, [9]:

$$F^{MS}(u,C)=\mu.Length\,(C)+\lambda\int\limits_{\Omega}|u_0(x,y)-u(x,y)|^2dxdy+\int\limits_{\Omega/C}|\nabla u(x,y)|^2dxdy \tag{9}$$

where $u_0 : \overline{\Omega} \rightarrow \mathbb{R}$ is a given image, μ and λ are positive parameters. The length of the curve is parameterized by the radii determined by (8) for both pupil and iris. Another objective is so to evolve the contour such that it starts from the interior or center of the object to be segmented and moves towards the exterior to stop on the desired pupil and iris boundaries respectively. The re-definition of the above parameters also sought to address the challenge of having to place the initial contour exactly on the iris boundaries. With the Chan-Vese algorithm, interior contours are automatically detected and the initial curve can be anywhere in the image.

4.3 Image Enhancement and Edge Detection

Enhancing an image is to process an image so that the end result is more suitable than the original image for use in specific applications. This includes filtering, adjusting image contrast and detecting edges. Edge detection is a crucial step towards understanding image features since with accurate mapping of edges in an image, properties such as curves, shapes, holes, lines etc. can be identified, extracted and measured [11]. The segmented iris is first enhanced by filtering and adjusting its contrast using bottom hat filtering and contrast limited adaptive histogram equalization (CLAHE). The CLACHE operates on small data regions found within the iris rather than the entire image. The result is an enhanced image with visually appealing features of interest.

4.4 Feature Extraction and Matching

Once iris features are extracted, they have to be matched with another iris template within the database. While referred to as a corner detector, Harris algorithm does not just select corners; but any location that has large gradients in all directions [12]. Traditional algorithms transform a segmented iris from the Cartesian plane to log-polar (rubber-sheet method) prior feature extraction. This method introduces changes to the geometrical structure and original location of iris features. Also one cannot work backwards to locate the iris features on the original image. The methods proposed in this paper for each stage are able to holistically bypass the above mentioned drawbacks whilst providing accuracy in results. The failure of statistical independence is the main key to iris recognition. Traditionally, the Hamming distance (HD) is used as a measure of dissimilarity between any two iris templates, [1, 3]. Any two different irises are guaranteed to pass this test. However, if any two images fail the test by producing a HD ≤ 0.32, it will mean that the two images are from the same iris [1].

4.5 Labelling of Connected Components for Feature Detection

This is where regions of adjacent or neighboring pixels sharing the same intensity values are identified and grouped together. The most connected feature labels are then separated and displayed in an RGB colour scheme for feature articulation as shown in Fig. 4d.

4.6 Principal Steps of Proposed Algorithm

```
Inputs: u0=input grayscale image,
      :(h,k)=get pupil-iris centers by (11),
      :rp,ri= get pupil and iris radius by (7),
Output: φ0=segmented iris.
Initialization: localize pupil and iris boundaries.
Step 1: Get mask = pupil and iris initial contours by:
x^2 + y^2 <= r^2.
Step 2: Segment pupil and iris by parameterizing Chan-
Vese algorithm with calculated diameters and radii,
Step 3: Enhance segmented image by CLACHE,
Step 4: Detect edges using Sobel,
Step 5: Label connected components,
Step 6: Detect Harris strongest features,
Step 7: Match detected Harris features.
```

5 Experimental Results

5.1 Analysis and Discussion of Results

Figure 3a, b demonstrate the performance of the proposed method by using Bresenham's algorithm to independently compute the pupil-iris diameter and radii. The resulting values are used to parameterize the initial contour and the evolving curve of the Chan Vese algorithm; by localizing the pupil-iris boundaries and

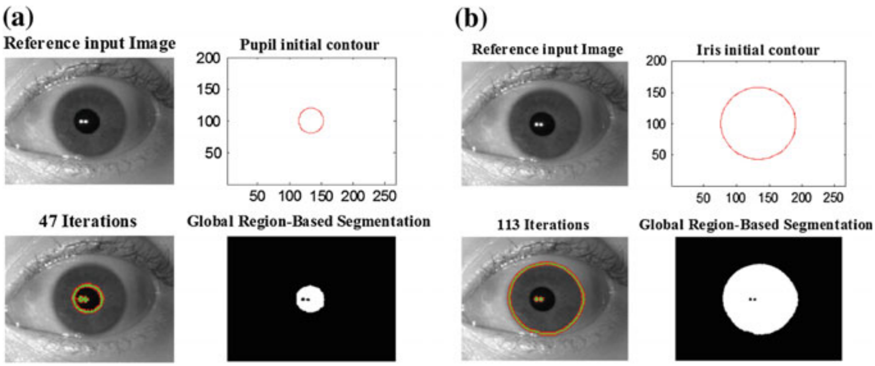


Fig. 3 **a**, **b** Pupil and iris localization with respective parameterized initial contours to evolve the curve for accurate segmentation

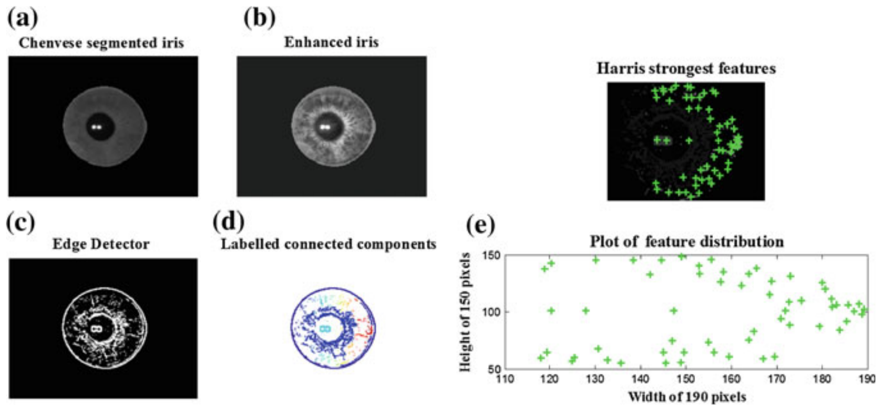


Fig. 4 **a** Segmented iris, **b** Enhanced iris, **c** Iris features highlighted with Sobel detector, **d** Different colors showing pixel connectivity of iris features, **e** Extracted Harris features

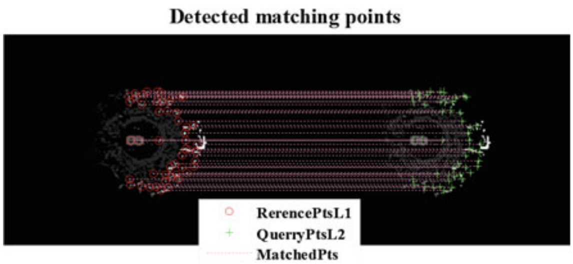


Fig. 5 Match points between reference and query image

influencing the number of iterations needed to stop the evolving curve on the desired boundary locations to achieve accurate segmentation. Figure 4a shows accurate iris segmentation, with preprocessing techniques on Fig. 4b, c to highlight unique features obtained by labelling pixels of strongest connectivity in Fig. 4d. Discrete features extracted using Harris algorithm for template creation is shown in Fig. 4e. Figure 5 shows a match of extracted feature points between a reference (L1) template and its query image (L2). The Hamming distance between L1 and L2 was calculated to be 0.0588 hence the failure of statistical independence, representing a perfect match. Figure 6a uses an image from UBIRIS database. The proposed approach still demonstrates accurate and robust iris localization and segmentation in Fig. 6a, b. Figure 7b, c show the extracted Harris features of two right eye images from UBIRIS and CSIR respectively. On Fig. 7b, c, Harris features are detected on both images with 70 and 55 key feature points respectively. On Fig. 7d the proposed method attempts to match the different images but detects no match. The calculated Hamming distance is 0.4664, passing the test of statistical independence, representing a non-match between the different images.

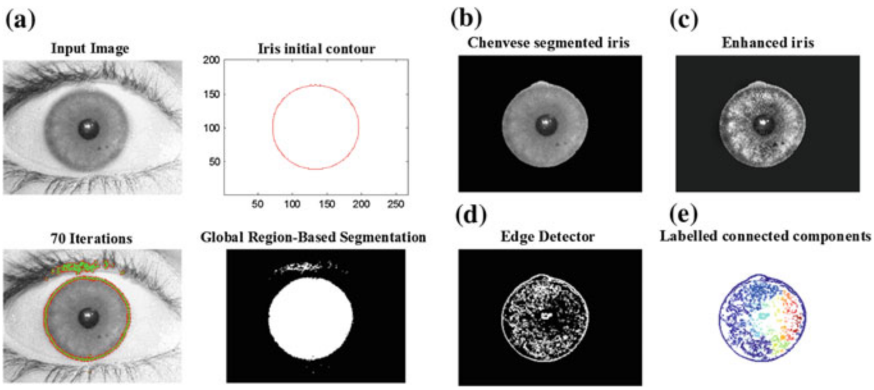


Fig. 6 UBIRIS image: **a** Parameterized initial contour with curve evolution. **b**, **c** Segmented and enhanced iris, **d** Iris features highlighted with Sobel detector, **e** Different colors showing pixel connectivity of iris features

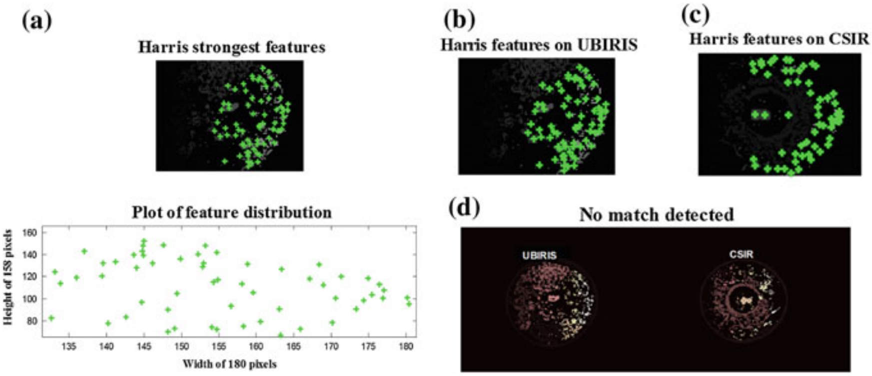


Fig. 7 **a** Harris features detected on UBIRIS image, **b**, **c** Demonstration of Harris features on images from different databases, **d** Shows no match between the two iris images

6 Conclusion

The experiment was carried out using non-ideal iris images from different databases to evaluate the performance of the iris recognition system using a different algorithmic approach at each stage. The proposed approach demonstrates robustness in that the system (i) produces accurate segmentation results, (ii) provides the exact global localization of extracted features and matching thereof. It also offers as much reliability at a lesser cost than the traditional methods.

Acknowledgments To UBIRIS for their iris database. The other iris images used were collected from willing participants within the Council for Scientific and Industrial Research (CSIR), South Africa.

References

1. Daugman, J.: How Iris recognition works. IEEE Trans. Circuits Syst. Video Technol. **14**(1), 21–30 (2004)
2. Poursaberi, A., Araabi, B.N.: Iris recognition for partial occluded images: methodology and sensitivity analysis. EURASIP J. Adv. Signal, 12 (2006), Article ID: 36751
3. Wildes, R.P.: Iris recognition: an emerging biometric technology. Proc. IEEE **85**(9), 1348–1363 (1997)
4. Vatsa, M., Singh, R., Noore, A.: Improving Iris recognition performance using segmentation, quality enhancement, match score fusion and indexing. IEEE Trans. Syst. Man Cybernet. Part B: Cybernet **38**(4), 1021–1035 (2008)
5. Liu, X., Bowyer, K.W., Flynn, P.J.: Experiments with an improved Iris segmentation algorithm. In: Proceedings of the 4th IEEE Workshop Automatic Identification Advanced Technologies, pp. 118–123 (2005)
6. Crandall, R.: Image Segmentation using the Chan-Vese algorithm. ECE 532 Project. Fall. (2009)
7. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: active contour models. Int. J. Comput. Vis. **1**, 321–331 (1988)
8. Chan, T., Vese, L.: Active contour models without edges. IEEE Trans. Image Process. **10**(2), 266–277 (2001)
9. Yahya, A.E., Nordin, M.J.: A new technique for Iris localization in iris recognition system. J. Comput. Sci. **6**(5), 527–532 (2010)
10. Yanto, G., Jaward, M.H., Kamrani, N.: Bayesian Chan-Vese segmentation for iris segmentation. In: IEEE Transactions on Visual Communication and Image Processing, pp. 1–6 (2013)
11. Maini, R., Aggarwal, H.: A comprehensive review of image enhancement techniques. J. Comput. **2**(3), 8–12 (2010)
12. Vincent O.R., Folorunso, O.: A descriptive algorithm for Sobel image edge detection. In: Proceedings of Informing Science and IT Education Conference, pp. 97–107 (2009)
13. Bresenham, J.: A linear algorithm for incremental display of circular arcs. Commun. ACM **20** (2), 100–106 (1977)

Adaptive and Intelligent Controller for Protection in Radial Distribution System

O.V. Gnana Swathika and S. Hemamalini

Abstract In this paper, a Field Programmable Gate Array (FPGA) based intelligent controller for overcurrent (OC) protection of a radial distribution system is realized. This controller monitors the radial distribution network continuously and when a fault is detected in the network, the FPGA based adaptive and intelligent controller performs the OC relay coordination and trips the appropriate circuit breakers of the system. The FPGA based digital prototype relay for overcurrent protection is realized with Atlys Digilent Spartan-6 FPGA kit. The performance of the FPGA based overcurrent protection technique for a 4-bus radial distribution system is compared with that of the conventional dual simplex algorithm. The proposed FPGA based overcurrent protection algorithm is also tested for an IEEE 33-bus radial distribution network.

Keywords Overcurrent relay · FPGA application · Radial distribution · Overcurrent protection

1 Introduction

Distribution Systems are the largest portion of the power system network that constitutes an array of radial feeders. The radial feeders of the distribution network are highly prone to faults, due to adverse weather conditions, equipment failure and accidents. A proper relay coordination technique helps to disconnect a minimum portion of the network and isolate the faulty part of the network using suitable overcurrent protection devices.

O.V. Gnana Swathika (✉) · S. Hemamalini
School of Electrical Engineering, VIT University, Chennai, India
e-mail: gnanaswathika.ov@vit.ac.in

S. Hemamalini
e-mail: hemamalini.s@vit.ac.in

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_18

Fault location in the distribution system is highly complex due to uncertainty of load, phase unbalance, non-homogeneity of line and fault resistance [1]. Various research and working models are derived for accurate fault location in the distribution system in order to maintain high reliability factor of the system. The complexity level of the system keeps varying and thus a specific constrained algorithm is not sufficient to serve the purpose of fault identification. The measurement units are not widely deployed and are mostly limited to the substation; which in turn causes lack of monitoring and tracking the critical system data. At distribution level, there is always load inconsistency and thus there is a drastic variation in the number of generators that serve the network. The phase unbalance is predominant in radial distribution network. These are certain challenges faced in the process of fault identification. The basic process for the fault location incorporates voltage and current measurement blocks that facilitate the calculation of the impedance using the fundamental component [2] or harmonics [3]. An additional calculation burden like recalculation of the voltage and current at each node [2, 4, 5] is required for the compensation of the characteristics unique to the distribution system. A Fuzzy approach to identify the most possible fault location in a distribution system is suggested in [6]. It is necessary to identify the exact fault location instantaneously and take necessary actions to resume supply to customers quickly. This includes isolating the faulted section from the healthy section appropriately. To meet out the above requirements, Distribution Automation System (DAS) comes into play to efficiently perform fault handling. The most important counterpart for this system is the availability of reliable measurement database from SCADA systems. Distribution Automation (DA) analytical tools include various application functions [7, 8]. New algorithms were then deduced using Artificial Intelligence (AI) [9], wherein all the short circuit analysis is done offline, and the fault is located online faster than conventional methods.

Overcurrent protection is a protection scheme that a distribution system demands when there is short circuit or overload condition in the network. Transients in voltage or current may force a relay to mal-trip or fail-to-trip [10]. Frequently occurring switching transient data is updated in a database. The immunity of the relay to such transients is tested and developed. The key challenge in this scheme is to map all the switching transient data to the database. Reliability of a distribution system [11] is improved by locating and rectifying faults even before a crisis occurs by using intelligent distribution fault identification algorithms. In this, the distribution fault anticipation technique makes the distribution network to handle any stress that maybe caused due to increasing loads or bad environmental conditions. This technology has scalability issues due to reasons like huge data that is difficult to analyze. In [12] a central controller is used with communication to multiple measurement units of a realistic distribution system. Communication based digital relays identify high impedance faults in the network. This setup requires switching devices and relays on each feeder and is hence expensive. The overcurrent relay (OCR) is realized using DSP processor [13]. However, this method does not provide accurate results, as no mathematical algorithms are involved. Fuzzy controller [14] is used to improve the performance of the overcurrent relay, but it is applicable only for medium voltage distribution networks. For optimum coordination of OCR, continuous genetic algorithm [15] is employed in a ring main system. The knowledge of the breaker operating time and coordination time interval is essential, without which this algorithm will not operate. The directional overcurrent coordination is analyzed for different topologies of the network and is solved using genetic algorithm [16]. OCR coordination for a distribution system with distributed generators is demonstrated using evolutionary algorithm [17]. The drawback of the algorithm is that the Time Multiplier Setting (TMS) variables of relays increase in complex network. Due to the unpredictable behavior of the distribution network, it is very crucial to identify the faulted section and isolate it from the healthy portion of the network quickly.

In this paper, an FPGA based intelligent controller with overcurrent protection capability for a distribution system is proposed. The intelligent controller monitors the various feeder currents of the system continuously. If any abnormality is identified in any section of the system, suitable OC relay coordination is employed immediately to ensure the safety of the network. The performance of the FPGA based controller is proved to be faster than the conventional Dual Simplex Method.

2 Directional Overcurrent (OC) Relay Coordination

The conventional overcurrent relay coordination algorithms are classified as trial and error method [18], topological analysis method [19, 20] and optimization method [21]. Dual simplex algorithm falls under the category of optimization method.

2.1 Problem Formulation

The directional OC relay coordination problem in a power system network is stated as an optimization problem, wherein the sum of the operating times of the relays in the system is minimized for near end fault. The objective function of the problem is given in (1).

$$\min z = \sum_{i=1}^{m} a_i TMS_i = \sum_{i=1}^{m} t_{i,i} \tag{1}$$

$$a_i = \frac{\lambda}{(PSM_i)^{\lambda} - 1} \tag{2}$$

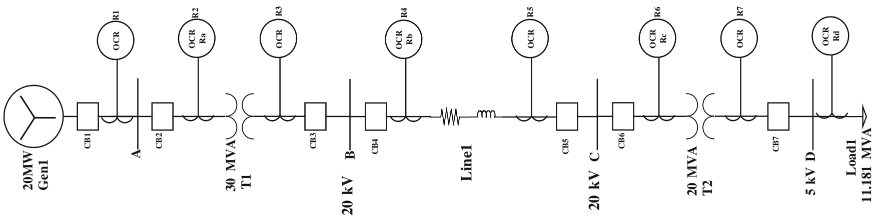


Fig. 1 4-bus radial distribution system

where

a_i	weight assigned for operating time of ith relay R_i
m	number of relays just beyond 'b' buses
TMS_i	Time Multiplier Setting of ith relay R_i
PSMi	Plug Setting Multiplier of ith relay R_i
λ	Constant of Relay
$t_{i,i}$	Operating time of primary relay at *i* for near end fault

The minimization problem is subjected to the following constraints:

(i) Coordination criteria

$$t_{b,i} - t_{i,i} \geq \Delta t \tag{3}$$

where

$t_{b,i}$	Operating time of backup relay for the same near end fault
Δt	Coordination time interval

(ii) Relay operating time

$$t_{i,i,max} \geq t_{i,i} \geq t_{i,i,min} \tag{4}$$

The dual simplex method is employed for optimum coordination of overcurrent relays in a radial test system [21, 22] using C programming. For the test system shown in Fig. 1, the normal IDMT relay with constants γ of 0.02 and λ of 0.14 are chosen. Minimum operating time for each relay is considered as 0.2 s and the coordination time interval (CTI) is taken as 0.3 s. The calculated values of a_i using (2) for the test system in Fig. 1 are shown in Table 1.

Table 1 Calculation of a_i

Fault position	Relay			
	R_a	R_b	R_c	R_d
Just beyond bus A	2.15	–	–	–
Just beyond bus B	2.77	2.15	–	–
Just beyond bus C	2.77	2.15	2.15	–
Just beyond bus D	1.75	1.47	1.47	2.15

2.2 Algorithm for Dual Simplex Method

1. Start.
2. Convert the minimization problem to maximization problem.
3. The greater than or equal to type constraints are converted to less than or equal to type constraints and surplus variables are added.
4. Create the dual simplex table by considering original variables as non-basics and slack variables as basics.
5. Form the cost coefficient $(C_j - \sum c_i \cdot e_{ij})$ row [22].
6. (a) Check if any element in this row is positive. If YES then go to step 13.
   (b) Check if all the elements in the row are non positive. If YES, then check if all the elements in the column of 'b' are non negative. If YES then optimal solution is reached. Go to step 12.
   (c) Check if all the elements in the row are non positive. If YES, then check if at least one element in the column of 'b' is negative. If YES then further optimization is possible. Go to step 7.
7. Identify the row which has the maximum negative value of b. This is considered as key row.
8. (a) There is no feasible solution for the problem, if all the coefficient values (aij) in key row are non negative. Now go to step 13.
   (b) If some a_{ij} in the key row are less than 0, then find the ratio for those columns:

$$\text{Ratio} = \left(C_j - \sum c_i \cdot e_{ij}\right) / \left(a_{ij}\right)$$

9. Identify the column having the smallest (positive) ratio. This is considered as the key column.
10. Identify the pivot element by mapping the key column and key row. Start the formation of next dual simplex table.
11. Go to step 5.
12. Print results.
13. Stop.

3 Proposed FPGA Based Adaptive and Intelligent Protection Control

In this paper, FPGA based adaptive and intelligent protection controller is developed to efficiently provide overcurrent protection for a radial distribution network. FPGA constitutes an array of configurable logic blocks, I/O Blocks and Block RAMs. The configurable logic block (CLB) contains slices. Each slice contains look-up tables (LUT), carry and control logic and registers. These slices are used independently or together for wider logic functions. There are buffers associated with each CLB, which are accessed by all the outputs of a CLB.

In FPGA based systems, signal processing functions are implemented with higher degree of freedom and low cost. Moreover they are easily reprogrammable for any changes in the functionality [23]. This eliminates the redesigning costs. While contrasting FPGA with microprocessor or DSP, the performance of FPGA is not tied to the clock rate. Thus highly parallel structures are easily constructed for processing data using FPGA [24]. The FPGA based systems perform at higher speed when compared with Microcontroller or DSP counterparts. Due to these predominant features, FPGA based systems are currently employed to perform power flow monitoring, fault identification [25] and protection in complex distribution systems. The power flow data of a power system network are transmitted to FPGA using wireless data monitoring cards [26]. To handle huge data from the network, more number of channels with higher sampling rate is required. Hence, FlexRIO is incorporated for data acquisition and real time data processing [27]. Due to the above features, FPGA devices are used in control platforms in various real time applications such as wireless telecommunications, image and signal processing, robotics and renewable energy systems [28].

Adaptivity of relays in a radial distribution system enables the protective system to act in accordance with the upcoming events. The proposed FPGA based adaptive controller performs two functions. The primary function is to perform fault monitoring, where the time stamped current data from all the feeders of the test system are captured and fed to the central protection center. The second function is to perform fault handling i.e. once a fault is identified in a feeder segment, relay coordination is done to isolate the faulted section from the healthy portion of the radial distribution network.

The directional overcurrent relay is employed in the radial distribution system for overcurrent protection. The normal current in all the feeders of the system are calculated to fix the CT ratio in all the feeders. The threshold current is computed using (5).

$$Threshold\ Current = \frac{Full\ Load\ Current\ in\ a\ Specific\ Feeder}{CT\ Secondary\ Current} \tag{5}$$

The current sensors are deployed at all feeders of the radial distribution network to sense the current that flows in that feeder at any instant of time. In the test system shown in Fig. 1, current sensor is placed in each feeder segment. These currents are captured continuously into a text file in the PC. The purpose of the controller is to monitor these feeder currents by reading the text file and comparing them with their respective threshold currents to ensure that they do not exceed the permissible current limit. If the current sensed in each feeder of the radial system is not more than their respective threshold current, the system is said to be in the normal state. If the above condition is violated, then the system is said to be in faulted state. The time of operation for the relays placed in the respective feeders is dynamically calculated using (6–9).

$$Pick - up\ current = Setting\ from\ relay\ catalogue * CT\ phase\ current \tag{6}$$

$$Time\ Multiplier\ Setting(TMS) = Setting\ as\ per\ relay\ catalogue \tag{7}$$

$$Current\ setting(CS) = \frac{Fault\ Current(IF)}{(1.5 * Pick - up\ current)} \tag{8}$$

$$Time\ of\ operation\ of\ relay = \frac{0.14 * TMS}{[(IF/CS)0.02 - 1]} \tag{9}$$

In the faulted state, the fault indicator available within the intelligent overcurrent protection center immediately indicates the abnormal behavior of the system. The faulted feeder details are transmitted through the Universal Asynchronous Receiver and Transmitter (UART) and displayed on the HyperTerminal running on the PC. If the fault in that feeder is not cleared within a stipulated amount of time (9), then the backup protection comes into play.

4 Results and Discussion

For the 4-bus radial distribution test system shown in Fig. 1, a FPGA based intelligent overcurrent protection controller is designed. Normal Inverse Overcurrent Relays (OCR) [29] R1 through R7 and CTs are fixed at seven feeders starting from upstream to downstream. Both current and time grading are assumed. This test system is subjected to load flow analysis. Feeders Gen1-A, A-T1, T1-B, B-line1, Line1-C, C-T2 and T2-Load1 have 8-bit digital equivalent current signals notated as *isa, iat1, it1b, ibm, imc, ict2 and it2l* respectively.

4.1 Test System Monitoring and Feeder Current Extraction

The feeder currents are periodically captured in a text file. The normal current in various feeder segments are listed in Table 2. In Gen1-A feeder, the full load current is 0.5672 kA. Hence a 600/5 CT is used in Gen1-A feeder. Similarly the CT's are fixed for the remaining feeders.

The threshold current for all the feeders of the 4-bus systems are tabulated in Table 3. A three-phase fault is initiated in the test system between T2 and Load. The fault currents in the various feeders are as shown in Table 2. Once the feeder current exceeds the threshold current shown in Table 3, the relay coordination ensures

Table 2 Normal and 3-phase fault current in a 4-bus radial system

Line segment	Normal current, kA	Fault current, kA
Gen1-A	0.5672	2.808
A-T1	0.2836	2.808
B-C	0.2836	1.404
C-D	1.1345	5.616

Table 3 Threshold current for relays

Relay number	Threshold current (rms), A
R1	120
R2	120
R3	60
R4	60
R5	60
R6	60
R7	240

which relay should trip to isolate the faulted section from the healthy portion of the network.

4.2 FPGA Based Intelligent Overcurrent Protection Controller

The FPGA based overcurrent protection scheme is implemented using Atlys circuit board which is a complete, ready-to-use digital circuit development platform based on a Xilinx Spartan-6 LX45 FPGA with speed grade of minus 3. The Spartan-6 LX45 is optimized for high-performance logic. It has 6822 slices, each of which contains four numbers of 6 input LUTs and eight flip-flops. The memory specification is 2.1 Mbits of fast block RAM. 58 DSP slices are also available. The clock speed is 500 MHz+. The VHDL modules used for the implementation of the OCR coordination algorithm are as shown in Fig. 2.

4.2.1 Overcurrent Relay Coordination Algorithm

Text_read.vhd

The purpose of text_read.vhd module is to read the text file located in the PC into the FPGA kit. This file contains the values of all feeder currents of the test system at any instant of time. As there are 7 feeder segments, 7 set of feeder currents will be available in that file. The output of this VHDL file is labeled as dataout (0:6) which represent the 7 individual instantaneous feeder currents. Each dataout port is an 8-bit value.

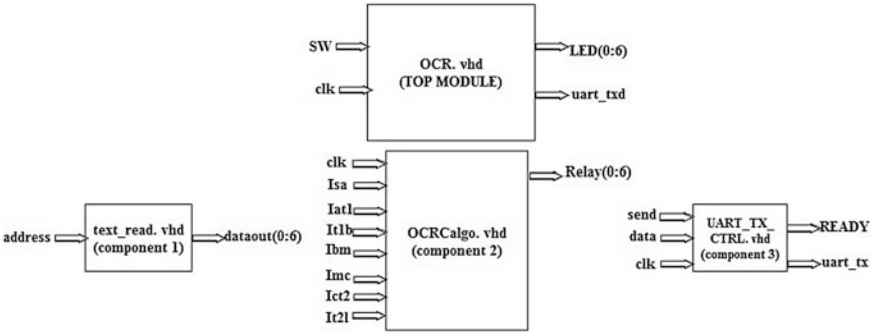


Fig. 2 VHDL modules for OCR coordination algorithm

OCRCalgo.vhd

The input to OCRCalgo.vhd module are the feeder currents namely isa, iat1, it1b, ibm, imc, ict2 and it2l, along with the clk signal. Dataout (0:6) are mapped to these currents. This module is responsible for identifying if there is an OC fault in the feeder and also to trip the appropriate circuit breakers taking into account the calculated time of operation of individual relays. Hence the primary and backup protection is triggered from this block in the event of occurrence of OC fault. The output of this relay is connected to the associated breakers of individual relays R1 through R7.

UART_TX_CTRL.vhd

The Atlys kit includes an EXAR USB-UART bridge to allow PC applications to communicate with the board using a COM port. UART onboard of the FPGA kit is used to establish asynchronous serial communication. The UART serial module has three sub-modules: the baud rate generator, receiver module and transmitter module. The local clock signal is provided by the baud rate generator. Once the baud-rate is fixed, both the transmitter and the receiver's internal clock are set to the same frequency. The UART receiver module enables the reception of serial signals at RXD and converts them into their respective parallel data. The UART transmit module with the help of frame format converts the data bytes into serial bits and transmits those bits through TXD. The UART_TX_CTRL.vhd module contains three input ports namely 'send', 'data' and 'clk'. The two output ports are 'READY' and 'uart_tx'. The string definition is such that the values stored at each index are the ASCII values of the indicated character [30]. This module is used to display the string, which is the error message indicating details of faulted feeder, on the HyperTerminal.

OCR.vhd

OCR.vhd is the top module which has three components namely Text_read.vhd, OCRCalgo.vhd, UART_TX_CTRL.vhd which are mapped to it. SW and clk serve as inputs to this top VHDL module. LED (0:6) and uart_txd are the outputs of this module. The user constraint file pins.ucf is associated with the top module at the implementation stage. This file matches equally named nets to equally named constraints. Since there are 7 feeders, 7 LEDs are rigged up in the kit. When SW = 1, the OCR coordination algorithm is triggered for the network that is being monitored. The UART_TXD is connected to UART port of the kit, which communicates with the HyperTerminal in PC. In order to incorporate the appropriate time of operation for relays R1 through R7, the clock of 96 MHz is utilized.

4.2.2 Procedure to Perform the OCR Coordination Using FPGA Kit

Make the connections and simulate the top module namely OCR.vhd. If a fault is initiated in the feeder T2-Load1, the simulated output for the relay coordination using FPGA is as shown in Fig. 3. The plot indicates that the relay operating time strictly follows the time of operation given in Table 4. The time of operation of the upstream relay is 2.9 ms. Double click on HyperTerminal and give a name for the new connection. Identify the COM port to which the FPGA kits' UART port is connected. Select that COM port detail in the dropdown menu. Click on CONFIGURE and set the following parameters: Bits per second: 9600; Data bits: 8; Parity: None; Stop bits: 1; Flow Control: None. The HyperTerminal is now ready to use. Synthesize and implement the top module OCR.vhd. Double click on Generate Programming file and this will generate the appropriate.bit file. Double click on "Configure Target Device". This triggers the IMPACT software to open. Double click on Boundary Scan. Right Click and select "Initialize Chain" and establish the connection between the PC and the FPGA kit. Assign the.bit file to the FPGA kit and thereby dump the bit file on the FPGA. If faulted feeder is T2L and if SW = 0, the fault in the feeder is indicated by glowing the appropriate LED. In this example

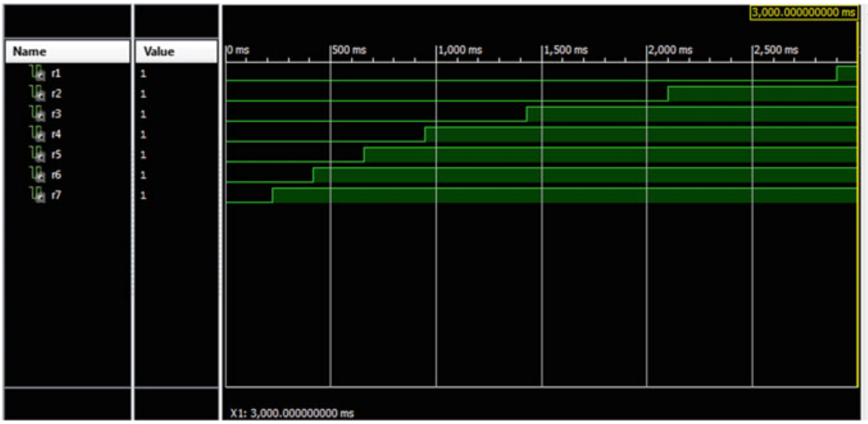


Fig. 3 FPGA based relay coordination 4-bus radial distribution system

Table 4 Time of operation set for the relays

Relay number	Time of operation, ms
R1	2900
R2	2100
R3	1430
R4	947
R5	660
R6	418
R7	223

LED6 associated with Relay 7 (downstream relay) will glow. If faulted feeder is T2L and if SW = 1, the OCR coordination algorithm will be triggered. The faulted feeder details are displayed on the HyperTerminal as:

"OC Fault in Feeder t2l"

LED6 glows to indicate fault in feeder T2L. If the fault is uncleared i.e. if the circuit breaker associated with R7 does not trip within the stipulated time of operation, then LED6 continues to glow and LED5 glows to indicate that backup protection is enabled in the network. This procedure continues from downstream to upstream until the fault is cleared as shown in Fig. 4.

An IEEE-33 bus system shown in Fig. 5 is considered for analysis. Let r_{i_j} indicate the relay notation for a relay placed between bus i and bus j. For a fault initiated in feeder connecting bus 31 and bus 32, the FPGA based overcurrent relay coordination is as shown in Fig. 6. The time of operation for the upstream relay in

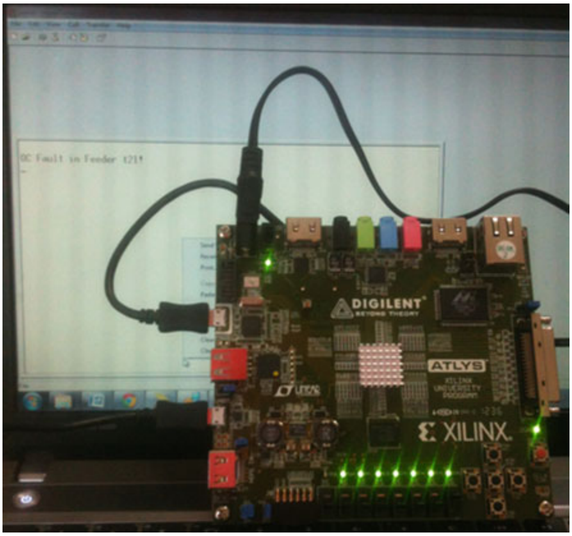


Fig. 4 FPGA based intelligent protection control system for 4-bus radial distribution system

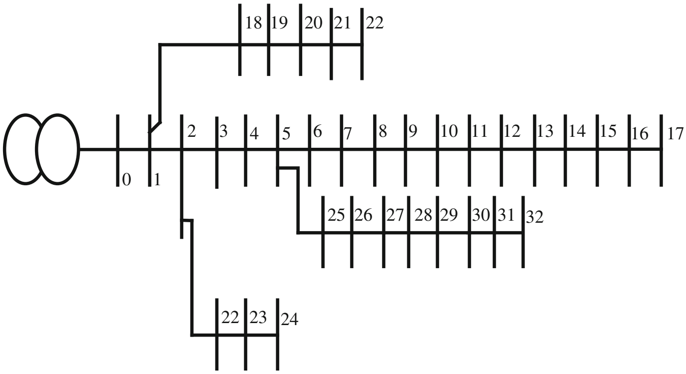


Fig. 5 IEEE 33-bus radial distribution system

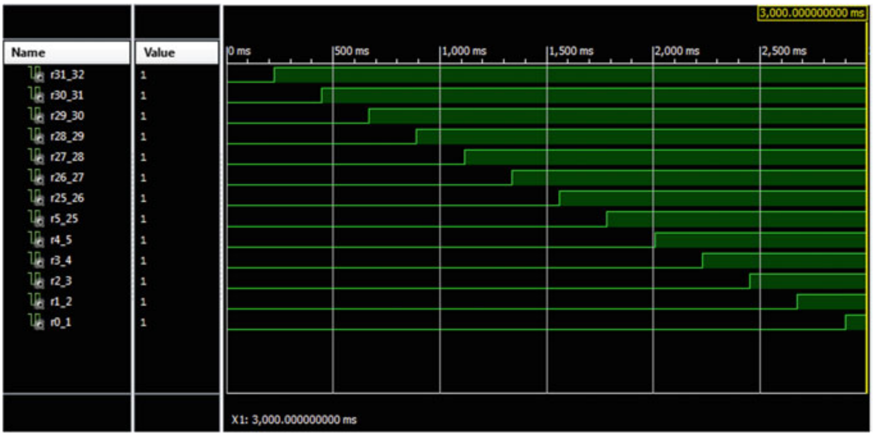


Fig. 6 FPGA based intelligent protection control system for the IEEE 33-bus radial distribution system

the FPGA based overcurrent relay coordination for the 4-bus and 33-bus system is at 2.9 s; whereas it is 3.1 s using conventional algorithms.

The resources used for the FPGA implementation of the overcurrent protection system for the 4-bus test system is shown in Table 5. Table 6 compares the time of operation of upstream relay for the 4-bus radial system for both FPGA based and conventional dual simplex based directional OC relay coordination [31, 32]. It is observed that there is a significant improvement in speed of clearing the fault, using the FPGA based adaptive and intelligent protection control system in comparison with the conventional algorithms for directional OC relay coordination. This algorithm is applicable to any complex distribution system and requires almost same number of bonded IOBs irrespective of the size of the network.

Table 5 Device utilization summary

FPGA resources	Used	Available	Utilization (%)
Number of slice register	181	54576	1
Number of slice LUTs	623	27288	2
Number of occupied slices	207	6822	3
Number of MUXCYs used	456	13644	3
Number of bonded IOBs	10	218	4

Table 6 Time of operation analysis

Time of operation (s)	FPGA based directional OC relay coordination	Dual simplex algorithm based directional OC relay coordination
	2.9	3.453

5 Conclusion

The proposed adaptive and intelligent protection control system is realized using FPGA technology. The FPGA based system identifies the fault instantly on the test system and appropriate OC relay coordination technique is applied to isolate the faulted section of the network from the healthy portion. The overcurrent relay coordination is completely centralized and based on the state-of-the-art technology. The FPGA based directional overcurrent relay coordination is proved to be faster than conventional optimization technique like Dual Simplex Method. This work maybe enhanced and implemented in large complex distribution systems with renewable energy sources.

Appendix

Generator (Gen1): 20 MW, 25MVA, 10 kV
Transformer (T1): 10 kV/20 kV, 30 MVA
Line Impedance: 2 + j4 Ω
Transformer (T2): 20 kV/5 kV, 20 MVA
Load1: 10 MW, 5 kV, 11.181 MVA

References

1. Lee, S.J., Choi, M.S., Kang, S.H., Jin, B.G., Lee, D.S., Ahn, B.S., Yoon, N.S., Kim, H.Y., Wee, S.B.: An intelligent and efficient fault location and diagnosis scheme for radial distribution systems. IEEE Trans. Power Deliv. **19**(2), 524–532 (2004)

2. Girgis, A.A., Fallon, C.M., Lubkeman, D.L.: A fault location technique for rural distribution feeders. IEEE Trans. Ind. Appl. **29**, 1170–1175 (1993)
3. Hami, M.E., Lai, L.L., Daruvala, D.J., Johns, A.T.: A new travelling-wave based scheme for fault detection on overhead power distribution feeders. IEEE Trans. Power Deliv. **7**, 1825–1833 (1992)
4. Das, R., Sachdev M. S., Sidhu, T. S.: A fault locator for radial sub-transmission and distribution lines. In: Proceedings of IEEE Power Engineering Society Summer Meeting, vol. 1, pp. 443–448 (2000)
5. Zhu, J., Lubkeman, D.L., Girgis, A.A.: Automated fault location and diagnosis on electric power distribution feeders. IEEE Trans. Power Deliv. **12**, 801–809 (1997)
6. Jarventausta, P., Verho, P., Partanen, J.: Using fuzzy sets to model the uncertainty in the fault location process of distribution networks. IEEE Trans. Power Deliv. **9**, 954–960 (1994)
7. Supply Restoration in Power Distribution Systems—A Benchmark for Planning Under Uncertainty. Available http://discus.anu.edu.au/~thiebaux/benchmarks/pds/ (1996)
8. Shirmohammadi, D., Liu, W.H.E., Lau, K.C., Hong, H.W.: Distribution automation system with real-time analysis tools. IEEE Comput. Appl. Power **9**(2), 31–35 (1996)
9. Lu, C.N., Tsay, M.T., Hwang, Y.J.: An artificial neural network based trouble call analysis. IEEE Trans. Power Deliv. **9**, 1663–1668 (1994)
10. Wang, F., Bollen, M.H.J.: Classification of component switching transients in the viewpoint of protective relays. Electr. Power Syst. Res. **64**(3), 197–207 (2003)
11. Don Russell, B., Carl Benner, L.: Intelligent systems for improved reliability and failure diagnosis in distribution systems. IEEE Trans. Smart Grid, 1(1), 48–56 (2010)
12. Sortomme, Eric, Venkata, S.S., Joydeep, M.: Microgrid protection using communication-assisted digital relays. IEEE Trans. Power Deliv. **25**(4), 2789–2796 (2010)
13. Goh, Y.L., Ramasamy, A.K., Nagi, F.H., Abidin, A.A.Z.: DSP based overcurrent relay using fuzzy bang-bang controller. Microelectron. Reliab. **51**, 2366–2373 (2011)
14. Goh, Y. L., et al.: DSP based fuzzy and conventional overcurrent relay controller comparisons. Microelectron. Reliab. 53, 1029–1035 (2013)
15. Prashant Bedekar, P., Sudhir Bhide, R.: Optimal coordination of overcurrent relay timing using continuous genetic algorithm. In: Expert System with Applications—Elsevier ScienceDirect, vol. 38, pp. 11286–11292 (2011)
16. Noghabi, A.S., Sadeh, J: Considering different network topologies in optimal overcurrent relay coordination using a hybrid GA. IEEE Trans. Power Deliv. 24(4), 1857–1863 (2009)
17. Yang, H., Wen, F., Ledwich, G.: Optimal coordination of overcurrent relays in distribution systems with distribution generators based on different evolution algorithm. In: Eur. Trans. Electr. Power (2011)
18. Urdaneta, A.J., Nadira, R., Luis Jimenez, G.P.: Optimal coordination of directional relays in interconnected power systems. IEEE Trans. Power Deliv. 3(3), 903–911 (1988)
19. Urdaneta, A.J., Restrepo, H., Marquez, S., Sanchez, J.: Coordination of directional relays timing using linear programming. IEEE Trans. Power Deliv. **11**(1), 122–129 (1996)
20. Abhyaneh, H.A., Al-Dabbagh, M., Karegar, H.K., Sadeghi, S.H.H., Khan, R.A.J.: A new optimal approach for coordination of directional overcurrent relays in interconnected power systems. IEEE Trans. Power Deliv. **11**(1), 122–129 (1996)
21. Prabhakar Bedekar, P., Sudhir Bhide, R.: Optimal coordination of directional overcurrent relays using hybrid GA-NLP approach. IEEE Trans. Power Deliv. 26(1), 109–119 (2011)
22. Prabhakar Bedekar, P., Sudhir Bhide, R.: Optimum coordination of overcurrent relays in distribution system using dual simplex method. In: Second International Conference on Emerging Trends in Engineering and Technology, IEEE, New York (2009)
23. Chekired, F., Mellit, A., Kalogiron, S.A., Larbes, C.: Intelligent maximum power point trackers for photovoltaic applications using FPGA chip: a comparative study. In: Solar Energy, vol. 101, pp. 83–99, Elsevier ScienceDirect (2014)
24. Simi Valsan, P., Shanti Swarup, K.: High-speed fault classification in power lines: theory and FPGA-based implementation. IEEE Trans. Ind. Electron. 56(5), 1793–1800 (2009)

25. Monmasson, E., Idkhajine, L., Wissem Naouar, M.: FPGA based controllers. IEEE Ind. Electron. Mag. 14–26 (2011)
26. Ozdemir, M.T., Sonmez, M., Akbal, A.: Development of FPGA based power flow monitoring system in a microgrid. Int. J. Hydrog. Energy 39, Elsevier ScienceDirect, pp. 8596–8603 (2014)
27. Zheng, W., Liub, R., Zhang, M., Zhuang, G., Yuan, T.: Design of FPGA based high-speed data acquisition and real-time data processing system on J-TEXT tokamak. Fus. Eng. Design 89, Elsevier ScienceDirect, pp. 698–701, (2014)
28. Chettibi, N., Mellit, A.: FPGA-based real time simulation and control of grid-connected photovoltaic systems. In: Simul. Model. Pract. Theory 43, Elsevier ScienceDirect, pp. 34–53 (2014)
29. IEEE standard inverse-time characteristic equations for over current relays IEEE Std C37.112–1996
30. ASCII Table http://www.asciitable.com/
31. Waleed Najy, K.A., Zeineldin, H.H., Wei L.W.: Optimal protection coordination for microgrids with grid-connected and islanded capability. IEEE Trans. Ind. Electron. 60(4) (2013)
32. Noghabi, A.S., Mashhadi, H.R., Sadeh, J.: Optimal coordination of directional overcurrent relays considering different network topologies using interval linear programming. In: IEEE Trans. Power Deliv. 25(3) (2010)

Challenges of Digital Note Taking

Mogeeb A.A. Mosleh, Mohd Sapiyan Baba, Sorayya Malek and Musaed A. Alhussein

Abstract There are world efforts to make technology act with education field for better learning achievements. Technology tries to replace the traditional learning environments, media, and activities into digital age. However, slow progress has been achieved to transfer the note taking activities into digital era. In this study, we explored current note taking tools which developed to bridge the gap between paper-based and technology-based notes. We tried to identify key specific problems and challenges that prevent note taking from existing in the digital age. This study is providing extensive investigation with systematic analysis about the impacts of current note taking tools in learning to identify constrains and limitations of typical note taking systems. Unfortunately, we agreed with similar previous studies that current tools are still inadequate and inefficient to be used for replacing the traditional note taking due to several issues. We found that developing a successful note taking applications is challenges because of four main issues, complexity, technology learning dilemma, integrity, and inefficiency issues. This study discusses the main implications to shape the future of digital notes.

Keywords Applications in note taking · Evaluation of CAL systems · Digital Note-Taking · Technology support learning · Education technology · Technology learning dilemma

M.A.A. Mosleh (✉) · M.S. Baba · S. Malek
Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia
e-mail: MogeebMosleh@um.edu.my; MogeebMosleh@yahoo.com

M.S. Baba
e-mail: pian@um.edu.my

S. Malek
e-mail: sorayya@um.edu.my

M.A. Alhussein
Faculty of Computer and Information Science, King Saud University, 51178 Riyadh, Kingdom of Saudia Arabia
e-mail: musaed@ccis.ksu.edu.sa

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_19

1 Introduction

Learning is process of acquiring and understanding knowledge. Human learning process occurs as part of education or personal development which involves with several activities. Note-taking is one of these activities performed to acquire knowledge, and improve learning outcomes. It is a task of recording information synthesis from a transient source, such as reading material and attending lecture. Note taking tasks assist learners in the process of concentration, thinking, memorizing, recalling process, and enhancing performance [1, 2]. It is a complex human behaviour related to personal information management with a variety of underlying mental processes, and cognitive interactions [3]. Note taking research had begun early on 1920s, when Crawford performed experimental study to examine the impacts of taking notes during lecture on student's performance [4]. Education and learning research reported that note taking process has two essential functions for supporting learner activities, encoding and external storage. Encoding improves learning by affecting learner cognitive process and working memory. While, external storage is the produced notes which used to record information for reviewing purposes [5]. Moreover, taking notes improves the ability of learn, integrate, and capture knowledge [6]. In addition, note taking improve students learning achievement and their academic performance [7], where about 99 % of students are writing notes, and 96 % of them consider note taking as an essential activity of their academic tasks [8]. Recently, technology application in education is evolving, and pedagogy is beginning to change the educators teach and students learn styles. Substantial evidence indicated that current technologies are promising, introducing better ways to teach and acquire knowledge. Technology offered special devices essentially to improve education and learning methods via developing various system and applications to facilitate learning activity [9]. There is a global effort to improve learning environments; hence, the idea that most devices would be integrated with standard note-taking capabilities using pen-based technology to replace traditional note-taking in the future is conceivable. Similarly, technology has begun to produce new ways to support education by developing new environments, such as using projectors to replace blackboards; slides are presented from the computer instead of writing on the blackboard; microphone, digital pen, laser pointer, and web-based courses [10]. Although we are in the digital age, note-taking as an education tool still struggles to exist in a traditional way. The lack of support for note-taking in digital format would increase the gap between traditional and digital learning tools in the next decades because most information and knowledge are transformed into digital representations. Challenges on the usability of traditional notes clearly appear via information management tasks because of the pervasiveness of current digital technology [11]. Furthermore, people are expected to manage a large amount of information with different formats from varying resources to complete their academic tasks. Traditional note-taking was unable to meet these challenges and encouraged the development of electronic note-taking applications. The digital document has also more advantages compared to paper documents, such as storability, transportability, computability, reproducibility, legibility, search ability, printability, and security. Since the digital learning materials afford new functions to enhance learning achievements. Thus, digital note taking advantages were the essential reason to perform such research and encouraged both researchers and developers to facilitate digital note-taking developments. This article is structured as follows. Section 2 provides historical review and systematic analysis for classification of current digital note taking tools. Section 3 addresses the main challenges that prevent note taking tasks to be existed in digital media. Section 4 argues the rationale for this study. Section 5 presents our conclusion within the expected scope and the limitations of digital note taking research.

2 Review of Note Taking

There are many hardware and software tools designed to facilitate note-taking activities. Existing note-taking tools were varied from simple tool to full applications. Studies focused on the note-taking functionality such as handwriting and highlighting, whereas other studies concentrated on the advantages of taking notes in digital media. We investigated the existing note taking tools, and tried to classify them based on learning principles as follows.

2.1 Note Taking as Active Learning Tool

Active learning is about building knowledge in different ways based on prior background [12]. Active learning is responsible for learning by creating activities, and doing things. Recent education technology research has focused more on the application that can support active learning. Several note taking tools were developed mainly to support active learning theory.

StuPad: it's a tool developed to support student recording streams of information, such as personal notes, video and audio stream, and related topic. It designed to organize and manage different types of information with two interfaces, one for capturing and recording information, and the other is for access and review of information. StuPad has a simple interface and supports a pen-based interface for writing notes, as shown in Fig. 1. This system can be improve by deploying typical infrastructure to support active learning in the classroom [13].

NoteTaker: designed to solve the problem of using computers for taking class notes inside classroom. It designed to adapt note-taking tasks by using the appropriate computer function, such as using a pen for drawing, using keyboard for text input, and using a pointing device for positioning and selecting. Developers reported the possibility to develop digital note taking application if hardware and software limitations resolved. NoteTaker evaluated to describe the note-taking

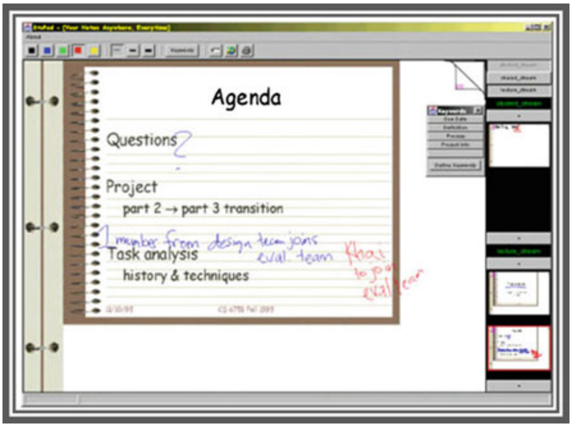


Fig. 1 StuPad system interface

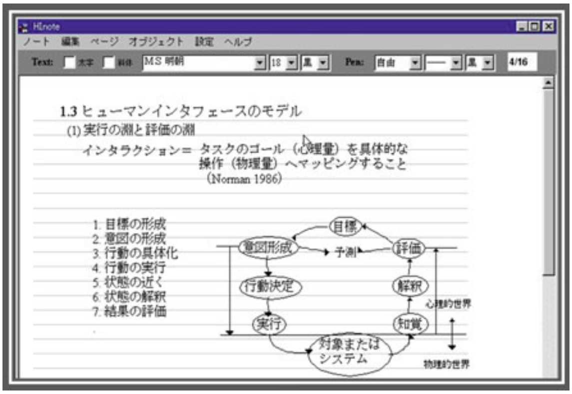


Fig. 2 NoteTaker system interface

application functional requirements, such as personal natures, combination of graphic and text, and time constraints [11] as shown in Fig. 2.

Classroom Presenter (CP): developed at the University of Washington in 2005. It supported students and instructors with enhanced tools to facilitate the learning process inside classroom. It supports instructions with the ability to collect, review, and monitor students feedback. It also supports students in taking notes and sharing their own works, but with limited functionality for later access and revision [14]. CP supported active learning using the materials presented, and integrated with specific functions such as flexibility to view materials and interaction mechanisms in classroom. Initial deployments of CP indicated that instructors could exploit this technology to achieve a wide range of instructional goals, with collaborative environment as shown in Fig. 3 [15].

Ubiquitous Presenter (UP): is an extension of the CP, it was developed at UC San Diego to support both pen-based and typed student submissions on the web. UP

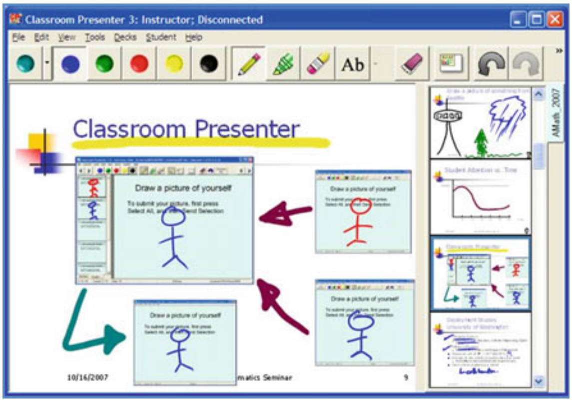


Fig. 3 Classroom presenter interface

includes extra functions to support interaction process using any web-enabled devices such as laptops, notebooks, and smart phones. UP system is designed based on a web-server architecture, in which the server acts as data repository for instructor and student materials such as lecture slides, instructor link, and student responses [16]. CP and UP could be used as interactive tools between students and instructors.

DyKnow: is commercial software developed for classroom management and interactive education. It provides students with several functions of note-taking, such as student response, content delivery, class capture, record notes, and notes review. DyKnow likewise allows the instructor to broadcast to students' screens to stimulate discussion, transmits prepared contents to student, allows students to poll assessment in real time. It supports note-taking tasks such as creating, annotating, audio recording, and saving materials on central server for later access. DyKnow has two different interfaces. The DyKnow for teachers allows them to control the digital classroom. DyKnow vision includes student response tools, note-taking functionality, class capture functionality, collaborative learning tools, and anytime-anywhere access. It require extensive experiments to evaluate its impact on learning achievement and performance [17].

Microsoft Tablet PC: Tablet PC was the first step in using pen as input device to computer, which was preferred by most people. Tablet PC promoted some features such as handwriting recognition. In addition, ink strokes in the tablet PC are stored differently from text and images as native data type. Tablet PC offered an alternative method for input data by using pen rather than other input device such as mouse and keyboard. Later on, Microsoft released the OneNote software for Tablet PC application, which was designed for taking notes as shown in Fig. 4 [18]. The Windows Journal application was also included on the tablet PC for Windows XP, Vista, and Windows 7. This application allows the user to create and organize handwritten notes and drawings using pen or mouse to compose handwritten note. According to research, this application is insufficient to meet the note-taking

Fig. 4 Microsoft Pc tablet system

requirements because of a few disadvantages, such as the inaccurate handwriting recognition, which affects the cognitive response of students [19].

Evernote: is commercial software designed to support the note-taking activity based on server-client architecture. It designed to assist people to capture idea, and to record information for later access and review. Evernote is available as a web application with full-feature desktop and mobile clients designed to allow users easily capture and find information content in any environment as shown in Fig. 5. This software supports users with various functions to capture texts, snapshots, digital ink, or audio. All information is automatically synchronized between the network and local devices. Evernote had not been evaluated to measure its efficiency on learning process [20].

E-Notes: developed by Wirth in 2003 to provide an electronic form of lecture notes that can be printed and annotated inside the classroom. E-notes evaluation indicated an improvement in student achievements. Experimental results showed that 96 % of

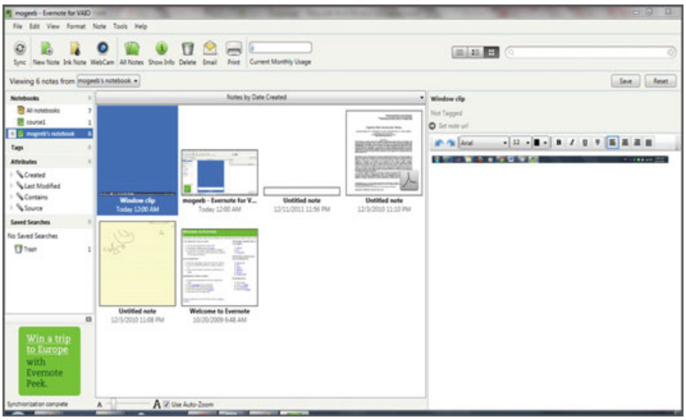


Fig. 5 Evernote interface

students found E-notes viable for use as note-taking application. Students reported that E-notes assisted them in concentrating for absorbing and understanding the material. It supports the delivery of notes to students before the lecture, as well as the annotation tool. However, a limited note-taking functions were implemented in E-notes [21].

NoteLook: was developed by Chiu et al. in (1999), which allows students to integrate notes and digital video by supporting automatic snapshots. This system uses a classroom camera to capture the screen and allows students to annotate snapshot images. This system has significant infrastructure with complex interface, which may hamper note-takers [22].

2.2 Note Taking as Active Reading Tool

Writing notes and comments while reading media content to elaborate a specific topic are the most common note-taking activities. Several tools have been designed to support these activities while reading materials, such as annotations, highlighting, underling, and linking [23, 24].

DigitalDesk: is designed based on the fact that people prefer traditional note-taking. It is designed to enhance traditional note-taking with computation technology instead of replacing paper and pen. It is an early attempt to connect between the interactions of physical documents in the digital world. DigitalDesk is a physical desk with a video camera that points to the desk, which captures the user interactions with paper documents, as shown in Fig. 6. It includes various tools that facilitate the interaction of physical paper such as paper paint, which allows users to select any part of a paper to be processed as digital documents, and collaboration environment that allows users to view the works of others [25].

Paper Augmented Digital Documents (PADD): is designed to break the gap between taking notes in the physical and digital media. Digital pen is used to connect with traditional paper documents. Digital paper is designed as a normal paper with printed invisible infrared dots. An infrared camera inside the digital pen is designed to detect the dot pattern for recording the location of the ink stroke. The digital pen can also capture and annotate the document. This application is a good note-taker because it allows ease of editing, navigation, annotation sharing, and archiving of the digital world, as shown in Fig. 7 [26].

XLibris: is a tool designed based on tablet PC to support note-taking tasks during reading, such as underlining, highlighting, and adding comments. It designed to perform limited note-taking tasks such as annotation, page turning, and handwriting. XLibris use an active digitizer behind the screen, which is controlled by a small electromagnetic field to replace textbooks with EBooks [27].

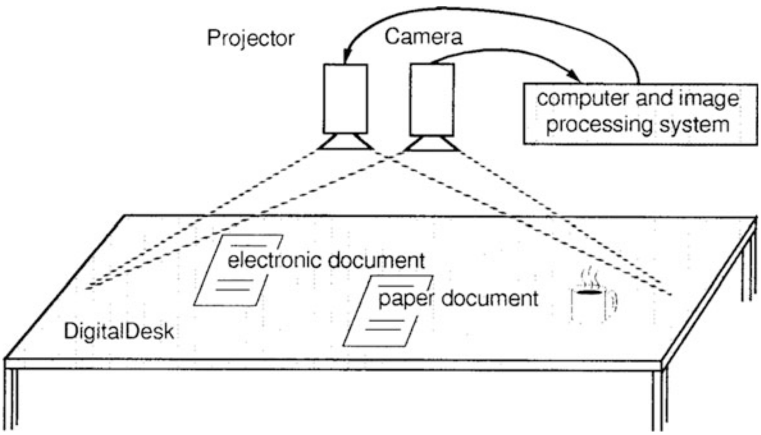


Fig. 6 Digital desk system architecture

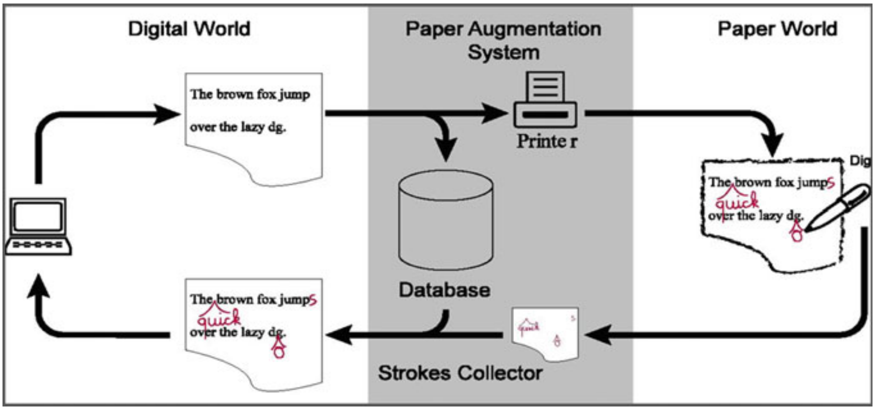


Fig. 7 PADD systems

PapierCraft: is an improved version of PADD system designed to support extra function on paper documents using Gestured with direct commands. These commands can specify digital features, whereas working in paper expands the range of possible interactions available in the handwriting interface. PapierCraft created a novel method for mediation between the object and the subject without effecting the note taking activities [28].

Sony Reader and Amazon Kindle: These two devices are designed to support active reading by replacing traditional paper, and textbooks with electronic books. Sony Reader was designed in early of 2006, and the Amazon Kindle was designed in late of 2007. Both devices use electronic paper display, comprising two

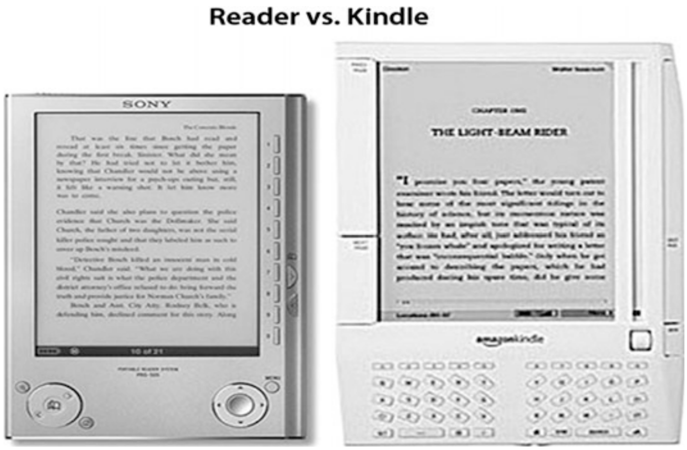


Fig. 8 Sony Reader and Amazon Kindle device

transparent silicone sheets for displaying sheet images. Electric paper can mimic the appearance and functions of normal paper with additional advantages such as being easily changed, small power consumption, non-backlight dependence, and more environment-friendly. However, these types of devices lack freeform annotation, and the ability to write notes. Several differences exist between these two devices. Kindle does not support many file formats, whereas the Sony Reader has no connectivity to the Internet. Sony Reader is simple, clean, looks more similar to a book, and is cheaper than Kindle. Figure 8 shows both devices.

InkSeine: is a prototype of ink application developed by Microsoft Research Center. The key idea behind InkSeine is to trigger search function for related content. InkSeine is designed with excellent user interface and includes most digital pen functions. This tool was not tested for learning purposes, but it used widely for taking notes for business meetings, discussions, and other tasks. InkSeine requires more integration tools to support electronic note-taking [29].

Paper-Top–Interface (PTI): is developed on 2010, to mix normal paper with digital technology. a visual marker based on augmented reality (AR) technique is used in designing PTI, with projector to display the materials in a classroom desk, as shown in Fig. 9. PTI allows students to view notes and write on paper using pencils. PTI has some advantages such as intuitive interaction, easy annotation, flexible spatial layout, and quick navigation. Initial evaluation results indicated that PTI is not in conflict with traditional learning style and can be efficiently used to take notes in classrooms [30].

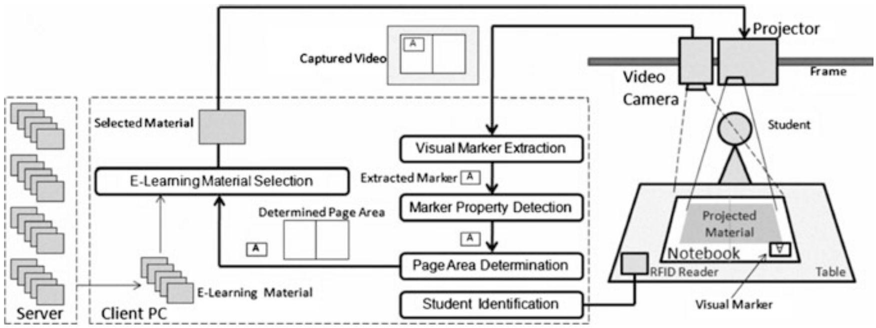


Fig. 9 Paper-top–interface (PTI) system overview

2.3 Note Taking as Collaboration Learning Tool

Current note taking applications can be also categorize according to their group targets, either individual or collaborative usage. Previous tools were considered as individual note taking tools and below some examples about sharing and collaboration tools.

Tivoli: is note taking tool support small group collaboration by using Xerox liveboard with pen-based techniques. It included pen and gestured commands for editing, printing, and importing backgrounds images. It allows only one user to collaborate at any given time. It does not constrain users, and allows anyone to change the whiteboard contents, which is physically constrained in the classroom [31].

Livenotes: is facilitated the cooperative and augmented task during lectures. It has a shared whiteboard that supports real-time interaction between small groups. The system includes wireless communication with a computer tablet to facilitate material sharing. It was designed to enable group members to interact during lecture, and presented material in the background of shared board to enhance student note-taking and annotation tasks. This system was specifically designed to augment note-taking interaction between students with no cooperative consideration for reviewing after class [32].

MicroNotes: designed for small notes, such as list of topics, an address of an interesting website related to a current topic, and question reminders. This application allows the posting of notable information and receipts between group members using any handheld devices. This system is designed to share special notes between group members [33].

NotePals: supports collaborative note-taking for recording and sharing notes. NotePals allows easy access for group member's notes, where each member can upload their notes in a shared repository. Group members can view the notes of other members by retrieving notes from the repository using topic context. It is mainly designed to support note sharing during meetings or discussions instead of note-taking purposes. The main disadvantage of the system is the lack of awareness in student's notes, direct communication is not allowed between users, and the lack of handwriting recognition to parse and search within the notes [34].

CoScribe: is designed to derive the implication of note-taking system in E-Learning. It was developed to support the collaboration between students by using there handwritten annotations on printed lecture slides. It is a paper-based tool for tagging and sharing student annotations and slides. It enabled students to create handwritten annotation, classify notes based on semantic structure, tag documents for easy access, and to share notes and collaborate with other students. It provides users with two interface views, a single-user and multi-user. It is similar to the annotation in traditional note, wherein technology remains in the background. The evaluation indicated that it efficiently supports student annotation [35].

2.4 Note Taking Tool for Wireless Handheld Devices (WHD)

Recently technology produced several devices, equipment, and tools essentially used for personal information management (PMI), such as PDA, Smart Phone, iPod, iPad, and so on. Incredible opportunity exists in using these devices in education if a suitable application is rationally developed to cover educational tasks and activities [36]. As the number of devices rapidly increases and networking infrastructures expand, society moves toward ubiquitous computing with technological advances and personalization of these tools to be used for media-based learning styles. Several tools were developed to support the learning activities on WHDs devices. Most of these tools are focused on information annotation, collaboration, indexing, and later access. Thus, these devices have limited capacity storage, and most data are stored in the web server [37].

Uchihashi and Wilcox in 1999 proposed an indexing system of digital data by clustering ink strokes automatically. These tools are used for limited searching by matching ink strokes rather than the recognized text [38]. Luchini et al. in 2003 designed a system to support learners in creating concept maps using hand-held devices such as pocket PC [36].

Nakabayashi et al. [39] described the development of self-learning environment by using both mobile phones and personal computers as client terminals. They extended the system functionality to enable offline learning, sharing course structure, and learner tracking information for learning activities using mobile phones and personal computers [39]. In 2006 and 2007, Dieterle tried to determine the impacts of wireless handheld devices on variety of learning settings. They reported that WHDs devices can enhance learning and teaching activities, and that ideal note-taking and information retrieval environments should be developed [40]. Varadarajan et al. [41] proposed an intelligent system with simple interface on PDA devices to allow fast indexing for digital notes in document repositories. Their system supports information query in inter- and intra-document indexing using latent semantic indexing. They reported that the system highly enhanced the student learning experience [41].

3 Discussion

In summary, most of these tools introduced note-taking into digital media. Some of these tools include specific feature and function, which are considered essential requirements for developing a useful note-taking application. For example, StuPad and CP support individual student annotation on lecture slides. However, LiveNotes allows students to take their own notes and to view the annotations of a small group of their peers without providing an explicit division or management of space conflicts. NotePals allows students to take small notes during lectures, juxtapose these notes with the lecture slides, and share with the entire class. In contrast, students who use LiveNotes and NotePals are unaware of the notes of other students during a lecture, which minimizes space and content conflicts but may result in duplicated effort.

Although note-taking software increased over the last decade, a few applications can be classified as a note-taking system, whereas, most of the developed tools only provide some functionality for limited note-taking tasks. Current note-taking applications do not specifically meet learning criteria. Moreover, common note-taking software does not only fail in supporting note-taking functions, but existing tools suffer in usability, mentality, knowledge capture, negative impact learning, as well as difficulties in retaining and retrieving information. Existing tools have several major learning deficiencies that negatively affect the learning process. In this study, we tried to identify the main challenges of digital note taking.

3.1 Current Challenges of Digital Notes

We found that current note-taking tools have four main critical issues that prevent note taking from being converted into digital device. Development of a successful note-taking application with currently available technology is very challenging. In order to simplify the problem, we classified the issues of existing tools into four main groups, where each group have several issues as shown in Fig. 10.

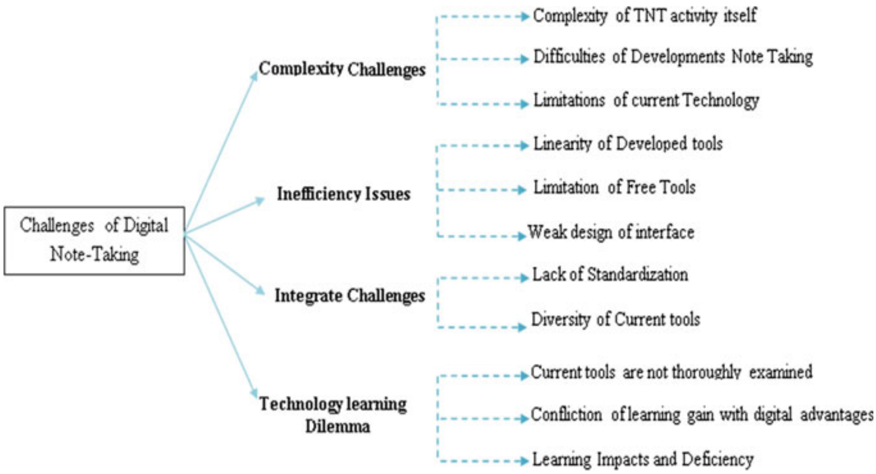


Fig. 10 Classification of digital note taking challenges

3.1.1 Complexity Challenges

The term *complex challenges* encompasses different types of note-taking issues, such as complexity in selecting appropriate tools based on learning theories and in implementing these tools and their interfaces. These difficulties occur in transferring realistic tasks into the digital worlds, and the limitation of current technology for implementing these tools.

I. **Complexity of Traditional Note Taking tasks and activities**
Note-taking activity is considered as a complex process because it needs several steps from comprehension and selection of information to written production. Note taking is a complex activity in terms of its functionality, components, and effects on learning behaviour and outcomes. Difficulties in representing traditional note-taking tasks as digital functions are major challenges because of numerous theoretical constraints, such as functional requirements, strategies, and working memory of note taking. Thus, a complex traditional activity is always more difficult to be represented in the digital world.

II. **Difficulties in Development Note-Taking tools**
Several challenges are faced by developers during the development of note-taking tools such as analysing the traditional note taking tasks. These challenges occur from the initial stage of identifying system requirement to the last stages of evaluation and obtaining user feedback. These challenges exist because note-taking tools must be implemented with special consideration to the pedagogical practice and the educational benefits of note taking. In addition, the challenges increase during the evaluation of these tools because no

research has yet discussed the evaluation criteria of these tools, such as learning outcomes, student behaviours, achievements, and performance.

III. **Limitation of Current Technology**
Limitation of current hardware and software tools in imitating the note-taking process, along with economic and social issues, increases the complexity of developing a note-taking application and prevents the adoption of electronic note taking. The complexity of using a computing device for a note-taking activity can be cited as a reason for the limitation of digital note taking [8]. In addition, a number of constraints and limitations on functionality, availability, and performance for developing a suitable note taking tools have been observed clearly.

3.1.2 Inefficiency Issues

Inefficiency of digital note tools is one of the main reasons for keeping to traditional note taking. Current tools of note taking are inefficient and inadequate to take notes in digital devices because of their linearity, limitations of the free form tools, and weak interface design. Time consuming and cognitive overload are the main factors that produced inefficiency issues in current note taking applications. Existing note-taking tools are still inefficient in the digital form because of the amount of time and activity required for the note taker.

I. **Linearity Issue**
Notes are mainly taken without linear consideration, and are written on a different page area with different position without limitation. However, note-taking tools are designed to use text editors for creating and editing notes as a digital document. As a result, delivered digital notes are linear and follow the analogy of a typewriter. Thus, linearity has broken the role of free form where linear tools exhibit efficiency in editing text without freedom option [11]. Most note-taking tools use the keyboard for typing notes, which is recommended because the use of keyboard for text input is faster than handwriting [11]. However, using the keyboard and mouse to input notes is a waste of time due to the extra time required for selecting the tool, choosing the font type and colour, and selecting the desired location for typing [42]. Moreover, research reported that input device can affect the note-taking function and strongly impact the learning cognition [43]. This limitation of linear representation makes users prefer the traditional note taking instead of switching into digital notes. Other effects of linearity include its tendency to allow text editors to insert whitespaces in the documents and the user-selected position. Note takers reported that linear application is considered useless during the note taking process [39, 44].

II. **Limitation of current free form tools**
A few tools were recently designed to support note takers in a nonlinear form with free-form canvas such as OneNote, NoteTaker, and LiveNotes. User feedback on such tools was generally positive, however, it has increased the note taker cognitive loads, and the processing time remains inefficient [45, 46]. Users of nonlinear applications should perform some pre-steps before drawing diagrams by specifying the area for drawing, selecting drawing tool, moving the mouse to the desired area, and clicking the mouse to start drawing. These pre-steps require extra time, which creates a critical learning problem for the note-takers. Research recommended a free form to support users write their notes in a flexible and efficient matter [30, 43]. Encounter problems exists in balancing between time efficiency and the freedom to enter information, whereas linear tools are reported as efficient in editing text.

III. **Weak interface Design**
Designing an appropriate interface for note-taking tools is a critical factor for the successful development of a note-taking system. Interface plays an important role for user acceptance in replacing digital note taking tools. Note-taking tools are supported different activity tasks such as classroom interaction, collaborative, material review, and monitoring and controlling events tools. Thus, tools are varying in terms of functionality and learning support, which leading to different types of interfaces elements. Current note-taking tools were designed with various interfaces, which results in the accomplishment of tasks in various views. Current tools interface mostly distract users, and reduce users attention and focusing [11].

3.1.3 Integrability Challenges of Note Taking

According to learning theories, note-taking applications should include several features, functions, and components to integrate as a single application for better learning. Several tools have been developed to achieve various note-taking functions; however, most of these tools have been built with limited functionality [8, 47]. Integration current tools in single application are challenges because of several issues.

I. **Wide Diversity of Current Tools**
Existing tools are diverse in interface components, system platforms, and hosted devices. Existing tools are insufficient for supporting all note taking activities and research efforts required to integrate note taking in a single application. By contrast, numerous tools and functions can be derived for note-taking applications. Furthermore, this challenge is increased because various note-taking systems have been built using different techniques without learning consideration, in which several techniques consider specific functionality, thus leading to variations in tools, functions, and target groups.

II. **Lack of Standardization**
Overall, numerous tools need to be developed, different criteria need to be considered, various disciplines should be involved during the development process of digital note. In addition, type of platforms and devices should be considered, and numerous constraints and limitations need to be sufficiently addressed. However, there are no studies that describe the roles and identify any standard criteria for the development of a typical note-taking application. No research has yet investigated, defined, or suggested any framework, standards, or criteria for the process of note taking developments. Guidelines are needed to simplify the process of developing a note-taking system based on the characteristics, activities, and features of traditional notes.

3.1.4 Technology Learning Dilemma

Education researchers have reported that using technology to support learning affects learning behaviours, styles, and outcomes [46, 48]. Researchers on note taking have agreed with the proposition that using technology in inappropriate ways has a negative impact on the learning process [42, 49]. Numerous tools have been developed to facilitate the learning process in note taking, however, most of these tools are not appropriate for achieving learning goals [50, 51]. We identified such problems as a term "Technology Learning Dilemma". It can be existed because of three main issues: (1) the negative effects of tools or their deficiency in terms of learning behaviours, styles, and outcomes; (2) the conflict between the benefits of using technology tools and learning theories; and (3) the leak of evaluation for current tools with respected to learning aims.

I. **Learning deficiencies**
We found that most of current note taking tools have several learning issues. For example, a copy–paste function has a negative effect on learning process because it allows the note taker to copy the text without reading it. Using the copy–paste function on note-taking applications decreases the ability of the learner to memorize knowledge, and promotes less retention. Copy–paste function reduces note-takers retention because it allows them to recode notes without reading them or focusing on what they have recorded [42, 51]. Another example of learning deficiency is sharing ability, which can change the behaviour of note takers and sometimes, can cause them to depend on the notes of others instead of writing their own notes. Research revealed that sharing function affects the encoding function of the note-taking process and exerts negative influence on learning outcomes [52, 53]. In addition, several developed systems allow students to compare notes during class and provide them with the possibility to discuss and post questions, which however, can negatively affect user concentration on the presented material [54].

II. **Confliction with digital advantages**
Linearity and free form entry options lead to conflictions between analogy of traditional notes and key advantages of digital note taking. Linear tool supports the advantages of writing readable and legible notes; however they impact current user practice in writing free notes. Vice versa, users are habitually taking notes in a free form approaches which afford learning features by reducing user time and cognitive load, and also support user familiarity, however free form tools is caused to loss the advantages of representing note digitally [55]. In contrast, this confliction in selecting the main tool for creating notes leads to delay in transferring notes into digital forms. In addition, several note-taking functions conflict with some advantages of digital notes, such as the handwriting function conflicting with the ability to edit and search digital notes. The handwriting function is an essential function for note taking because it supports learning by graphic familiarity and free form. Most note-taking applications designed to use the keyboard as main entry for taking notes because of the difficulties in developing handwriting tools and the advantages of electronic notes. However, graphic familiarity and free form as gain factors for note taking as a learning tool are lost if the keyboard is allowed to be used for the creation of note [8, 56]. This issue can exist in many circumstances with different scenarios of using technology to support learning. We argue that research groups focused on gaining several feature advantages without considering learning factors. For instance, the first group that supported note typing using a keyboard considered acquiring the advantages of the digital notes feature in terms of text typing [8, 11, 53]. However, this group neglected the note-taking constraints of free form as well as the learning role of the cognitive load and note familiarity. The second group gave a high priority to learning advantages instead of focusing on gaining the powerful tools of technological improvements [8, 43, 49].

III. **Existing tools are not thoroughly examined**
Technology learning dilemma also happens because note-taking tools are designed without considering learning theories. Current tools are developed without evaluating their effects on users because design decisions are mostly made by developers who are not well-versed in educational theories [21, 57]. Existing note-taking tools are not examined in terms of their effect on the behaviour and cognitive effort of the learner. A few studies have systematically evaluated the impact of note-taking tools on user behaviour instead of concentrating on satisfaction and motivation [17, 49]. In other words, if a note-taking application is built for educational purposes, it should be designed to maintain the learning benefits achieved through traditional note-taking, wherein each tool must be evaluated to ensure its suitability for educational purposes. Thus, tool requires evaluation in two targets, to obtain user feedback on the usability, and its impact on learning outcomes [20, 22, 43].

This research had confirmed that digital devices become ubiquitous, available largely with people such as PDA, iPhone, Tablet PC, and iPad. Experimental studies of current tools of note taking showed great interested for replacing the digital devices to take notes instead of using traditional paper and pen. The overall progress of technology in this area showed the possibility of transferring this task into digital environment if appropriate system was developed with usable and useful features. However, a slow progress was performed to achieve this objective because current note taking are suffering from several issues which elaborated previously. Successful application for digital note taking can be designed if extensive research performed to adapt the current challenges in a standard framework with a mediation approaches [58]. Thus, to develop a successful note-taking application, the previous issues need to be solved individually by establishing a theoretical note taking framework with mediation approaches. The framework is required to initialize the necessary guidelines for developments process of note taking tools. Furthermore, intelligent mediator is required to address technology learning dilemma by adapting note taking tasks into digital environments without distraction user current practice, and without impacts learning and cognitive theories.

4 Conclusions

We have argued that successful developments of student-oriented technologies are feasible but not an easy task. In addition, we found that there are several issues in sticking with pen and paper, rather than using digital tools for note taking including mainly complexity, inefficiency, integrability, and technology learning dilemma issues.

- **Complexity** is related to the difficulties in carrying out the various tasks of traditional note taking.
- **Inefficiency** is related to the time of achieving tasks and user cognition terms, where inappropriate design leads to unnecessary actions and cognitive over loads.
- **Integrability** is due to the wide diversity of current technology in both hardware and software requirements for development note taking tools, and the difficulties to integrate the current tools within typical note taking application.
- **Technology learning dilemma** describes the effects of technology in development of education tools, where we found that improper usage of technology can lead to negative impact on learner negatively leading to learning deficiency. It is an important factor when addressing the confliction between the gain of traditional and digital advantages of note taking.

Extracted problems of the current tools provided us with clues to the critical question of why people still use pen and paper to take notes. This research can be used to assist developers for establishing a design solution based on pedagogical opportunities of digital note-taking. Research can be used this study to shape the technology future of digital note taking. Additional work should be performed to analysis the functional and non-functional requirements of digital note taking to design a successful application of digital note taking. We argue here that development of digital note taking application is visible if theoretical and practical solutions propose to resolve these main challenges.

Acknowledgments Authors would like to thank Taiz and Malaya universities for their supports. This study was funded by UMRG grant of University of Malaya UMRG RG370-15AFR.

References

1. Di Vesta, F.J., Gray, G.S.: Listening and note taking. J. Educ. Psychol. **63**, 8–14 (1972)
2. Kiewra, K.A., Frank, B.M.: Encoding and external-storage effects of personal lecture notes, skeletal notes, and detailed notes for field-independent and field-dependent learners. J. Educ. Res. 143–148 (1988)
3. Piolat, A., Olive, T., Kellogg, R.T.: Cognitive effort during note taking. Appl. Cognit. Psychol. **19**, 291–312 (2004)
4. Crawford, C.C.: Some experimental studies of the results of college note-taking. J. Educ. Res. **12**, 379–386 (1925)
5. Peper, R.J., Mayer, R.E.: Generative effects of note-taking during science lectures. J. Educ. Psychol. **78**, 34 (1986)
6. Palmatier, R.A., Bennett, J.M.: Notetaking habits of college students. J. Read. 215–218 (1974)
7. Van Meter, P., Yokoi, L., Pressley, M.: College students' theory of note-taking derived from their perceptions of note-taking. J. Educ. Psychol. **86**, 567–576 (1994)
8. Kim, K., Turner, S.A., Pérez-Quiñones, M.A.: Requirements for electronic note taking systems: a field study of note taking in university classrooms. Educ. Inf. Technol. **14**, 255–283 (2009)
9. Abowd, G.D., Atkeson, C.G., Feinstein, A., Hmelo, C., Kooper, R., Long, S., Sawhney, N., Tani, M.: Teaching and learning as multimedia authoring: the classroom 2000 project. In: Proceedings of the Fourth ACM International Conference on Multimedia, pp. 187–198. ACM, New York (1997)
10. Kiewra, K.A., Benton, S.L., Kim, S.-I., Risch, N., Christensen, M.: Effects of note-taking format and study technique on recall and relational performance. Contemp. Educ. Psychol. **20**, 172–187 (1995)
11. Ward, N., Tatsukawa, H.: A tool for taking class notes. Int. J. Hum Comput Stud. **59**, 959–981 (2003)
12. McConnell, J.J.: Active learning and its use in computer science. In: ACM SIGCSE, pp. 52–54 (1996)
13. Truong, K.N., Abowd, G.D.: StuPad: integrating student notes with class lectures. In: CHI'99 Extended Abstracts on Human Factors in Computing Systems, pp. 208–209. ACM, New York (1999)
14. Anderson, R., McDowell, L., Simon, B.: Use of classroom presenter in engineering courses. In: FIE'05 35th Proceedings of Frontiers in Education, T2G-13. IEEE, New York (2005)
15. Anderson, R., Davis, P., Linnell, N., Prince, C., Razmo, V., Videon, F.: Classroom presenter: enhancing interactive education with digital ink. Computer **40**, 56–61 (2007)
16. Wilkerson, M., Griswold, W.G., Simon, B.: Ubiquitous presenter: increasing student access and control in a digital lecturing environment. ACM SIGCSE Bull. **37**, 116–120 (2005)
17. Berque, D.: An evaluation of a broad deployment of DyKnow software to support note taking and interaction using pen-based computers. J. Comput. Sci. Coll. **21**, 204–216 (2006)
18. Mock, K.: Teaching with Tablet PC's. J. Comput. Sci. Coll. **20**, 17–27 (2004)
19. Pittman, J.A.: Handwriting recognition: tablet pc text input. Computer **40**, 49–54 (2007)

20. Geyer, F., Reiterer, H.: Experiences from employing evernote as a tool for documenting collaborative design processes. In: Proceedings of the DIS'12 on Supporting Reflection in and on Design Processes (2012)
21. Wirth, M.A.: E-notes: using electronic lecture notes to support active learning in computer science. ACM SIGCSE Bull. **35**, 57–60 (2003)
22. Chiu, P., Kapuskar, A., Reitmeier, S., Wilcox, L.: NoteLook: taking notes in meetings with digital video and ink. In: Proceedings of the Seventh ACM International Conference on Multimedia. (1999) 149–158
23. Bothin, A., Clough, P.: Participants' personal note-taking in meetings and its value for automatic meeting summarisation. Inf. Technol. Manag. **13**, 39–57 (2012)
24. Weibel, N., Fouse, A., Hutchins, E., Hollan, J.D.: Supporting an integrated paper-digital workflow for observational research. In: Proceedings of the 16th International Conference on Intelligent User Interfaces, pp. 257–266. ACM, New York (2011)
25. Wellner, P.: Interacting with paper on the DigitalDesk. Commun. ACM **36**, 87–96 (1993)
26. Guimbretière, F.: Paper augmented digital documents. In: Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology, pp. 51–60 (2003)
27. Wilcox, L.D., Schilit, B.N., Sawhney, N.: Dynomite: a dynamically organized ink and audio notebook. In: Proceedings of the SIGCHI on Human Factors in Computing Systems, pp. 186–193. ACM, New York (1997)
28. Liao, C., Guimbretière, F., Hinckley, K.: PapierCraft: a command system for interactive paper. In: Proceedings of the 18th ACM Symposium on User Interface Software and Technology, pp. 241–244 (2005)
29. Hinckley, K., Zhao, S., Sarin, R., Baudisch, P., Cutrell, E., Shilman, M., Tan, D.: InkSeine: in situ search for active note taking. In: Proceedings of SIGCHI Conference on Human Factors in Computing Systems, pp. 251–260. ACM, New York (2007)
30. Mitsuhara, H., Yano, Y., Moriyama, T.: Paper-top interface for supporting note-taking and its preliminary experiment. In: IEEE International Conference on Systems Man and Cybernetics (SMC), pp. 3456–3462. IEEE, New York (2010)
31. Pedersen, E.R., McCall, K., Moran, T.P., Halasz, F.G.: Tivoli: an electronic whiteboard for informal workgroup meetings. In: Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems, pp. 391–398. ACM, New York (1993)
32. Kam, M., Wang, J., Iles, A., Tse, E., Chiu, J., Glaser, D., Tarshish, O., Canny, J.: Livenotes: a system for cooperative and augmented note-taking in lectures. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 531–540. ACM, New York (2005)
33. Lin, M., Lutters, W.G., Kim, T.S.: Understanding the micronote lifecycle: improving mobile support for informal note taking. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 687–694. ACM, New York (2004)
34. Davis, R.C., Landay, J.A., Chen, V., Huang, J., Lee, R.B., Li, F.C., Lin, J., Morrey III, C.B., Schleimer, B., Price, M.N.: NotePals: lightweight note sharing by the group, for the group. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 338–345. ACM, New York (1999)
35. Steimle, J., Gurevych, I., Mühlhäuser, M.: Notetaking in university courses and its implications for elearning systems. e-Learning Fachtagung Informatik, pp. 45–56 (2007)
36. Luchini, K., Quintana, C., Soloway, E.: Pocket PiCoMap: a case study in designing and assessing a handheld concept mapping tool for learners. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 321–328. ACM, New York (2003)
37. Parker, V.K., Kauffman, L., McManus, C., Buchman, S., Miller, C., Vance, R.: Use of hand-held devices with first year associate degree nursing students. Am. J. Health Sci. (AJHS) **2**, 39–44 (2011)
38. Uchihashi, S., Wilcox, L.: Automatic index creation for handwritten notes. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 3453–3456 (1999)
39. Nakabayashi, K., Hoshide, T., Hosokawa, M., Kawakami, T., Sato, K.: Design and implementation of a mobile learning environment as an extension of SCORM 2004

specifications. In: Seventh IEEE International Conference on Advanced Learning Technologies (ICALT), pp. 369–373 (2007)
40. Dieterle, E., Dede, C., Schrier, K.: Neomillennial" learning styles propagated by wireless handheld devices. In: Ubiquitous and Pervasive Knowledge AND Learning Management: Semantics, Social Networking and New Media to Their Full potential, pp. 35–66 (2007)
41. Varadarajan, A., Patel, N., Maxim, B., Grosky, W.I.: Analyzing the efficacy of using digital ink devices in a learning environment. Multimed. Tools Appl. **40**, 211–239 (2008)
42. Bauer, A., Koedinger, K.R.: Note-taking, selecting, and choice: designing interfaces that encourage smaller selections. In: Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 397–406. ACM, New York (2008)
43. Reimer, Y.J., Brimhall, E., Cao, C., O'Reilly, K.: Empirical user studies inform the design of an e-notetaking and information assimilation system for students in higher education. Comput. Educ. **52**, 893–913 (2009)
44. Schilit, B.N., Golovchinsky, G., Price, M.N.: Beyond paper: supporting active reading with free form digital ink annotations. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 249–256. ACM Press/Addison-Wesley Publishing Co., New York (1998)
45. Katayama, A.D., Robinson, D.H.: Getting students "partially" involved in note-taking using graphic organizers. J. Exp. Edu. **68**, 119–133 (2000)
46. Makany, T., Kemp, J., Dror, I.E.: Optimising the use of note-taking as an external cognitive aid for increasing learning. B. J. Edu. Technol. (2008)
47. Gathercole, S.E., Alloway, T.P.: Working Memory and Learning: A Practical Guide for Teachers. SAGE Publications, Thousand Oaks (2008)
48. Zhai, S., Kristensson, P.-O., Smith, B.A.: In search of effective text input interfaces for off the desktop computing. Interact. Comput. **17**, 229–250 (2005)
49. Bauer, A., Koedinger, K.: Evaluating the effect of technology on note-taking and learning. In: CHI'06 extended abstracts on Human Factors in Computing Systems, pp. 520–525. ACM, New York (2006)
50. Grabe, M., Christopherson, K.: Evaluating the advantages and disadvantages of providing lecture notes: The role of internet technology as a delivery system and research tool. Intern. High. Edu. **4**, 291–298 (2005)
51. Morgan, M., Brickell, G., Harper, B.: Applying distributed cognition theory to the redesign of the 'Copy and Paste' function in order to promote appropriate learning outcomes. Comput. Educ. **50**, 125–147 (2008)
52. Badger, R., White, G., Sutherland, P., Haggis, T.: Note perfect: an investigation of how students view taking notes in lectures. System **29**, 405–417 (2001)
53. Crooks, S., Katayama, A.: Effects of online note-taking format on the comprehension of electronic text. Res. Schools **9**, 22–23 (2002)
54. Kaptelinin, V., Nardi, B.A.: Acting with technology. Mit Press, Cambridge (2006)
55. Larson, R.B.: Enhancing the recall of presented material. Comput. Educ. **53**, 1278–1284 (2009)
56. Marton, F., Tsui, A.: Classroom Discourse and the Space of Learning. L. Erlbaum Associates (2004)
57. Vega, L.C., Dickey-Kurdziolek, M., Shupp, L., Perez-Quinones, M.A., Booker, J., Congleton, B.: Taking notes together: Augmenting note taking. In: Collaborative Technologies and Systems, pp. 16–23 (2007)
58. Mosleh, M.A., Baba, M.S.: A SmartInk: mediator solution of technology learning dilemma for digital note application. Int. J. Sci. Eng. Res. **4**, 61–67 (2013)

Cooperative Control of Multi-robot Systems with a Low-Degree Formation

Krishna Raghuwaiya, Bibhya Sharma and Jito Vanualailai

Abstract The utilization of team of robots working in a cooperative manner has huge benefits in moving large payloads. To perform such tasks, a multi robot structure or formation is necessary to coordinate the motions of the robots in a well planned manner. In this paper, the formation control problem of multi car-like mobile robots have been studied. The purpose is to get a swarm of mobile robots in a certain formation pattern to track a desired trajectory to accomplish a set objective. A set of artificial potential field functions is proposed using the Direct Method of Lyapunov for avoiding inter-robot, inter-formation and obstacle collisions and attraction to their designated targets. The effectiveness of the proposed nonlinear acceleration control laws is demonstrated through computer simulation results which prove the efficiency of our control technique and also demonstrates scalability for a large group of robots.

Keywords Formation · Lyapunov · Nonholonomic mobile robots · Low-degree

1 Introduction

Multi-Robot Systems (MRSs) have great advantages over single robot systems in terms of efficiency, redundancy, flexibility, robustness, cost benefits [1]. There are various areas which have benefited from the use of MRSs. Some of these include underwater and space exploration, hazardous environments, service robotics, entertainment, military rescue and reconnaissance, controlling formations of satellites and autonomous vehicle [2]. Cooperative control of MRSs have received considerable attention due to the fact that several collaborative tasks can be performed more efficiently and robustly using MRSs which may not be possible with

K. Raghuwaiya (✉) · B. Sharma · J. Vanualailai
School of Computing, Information and Mathematical Sciences,
The University of the South Pacific, Suva, Fiji
e-mail: Raghuwaiya_k@usp.ac.fj

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_20

individual robots acting alone. The benefits become apparent when considering distributed tasks, dangerous or hazardous tasks, tasks which require redundancy [3], and can also provide flexibility to task execution and robustness [1]. In many cases, the execution of the task requires formation control [4], and the accomplishment of the overall operation depends on each mobile robot operating in a desired manner. Cooperative Control problem discussed in this paper is referred to the problem of MRSs to coordinate their motions on executing tasks cooperatively in a given workspace. From recent literature, one of the most critical applications of cooperative behavior is in payload transportation [5]. In such applications, motion coordination with minimal error is required. Control strategies for formation or cooperative control of mobile robots can be roughly categorized into virtual structure, behavior-based and leader follower schemes [6]. Authors have proposed different schemes using different control methodologies for coordinating behaviors of MRSs. These control methods include centralized, completely decentralized or hybrid [7]. Guo et al. [8] proposed an adaptive scheme using graph theory together with Lyapunov based techniques to enclose a moving target and attain a desired inter robot formation. Hybrid systems comprising of distributed smooth-time varying feedback control laws together with a reactive control framework is proposed in [4, 9] while synchronization in the context of Lagrangian systems control was proposed in [10] for cooperative robot control. Mas and Kitts [5] proposed a cluster space control to control a group of four nonholonomic wheeled robots for object transportation.

We propose a leader-follower scheme to develop a cooperative formation controller for coordinating groups of MRSs. In this control methodology, we assign a leader in each MRS, who takes the responsibility to specify the objective of the task and one who also dictates a geometric path from some initial configuration to some final configuration. The control strategy formulates a low degree formation which allows for slight distortions in the formation structure of the leader and follower. These distortions would normally appear if the group encounters an obstacle. As in [11], this low degree formation structure allows for slight but temporary distortions for the formation to avoid obstacles. Based on artificial potential fields, the Direct Method of Lyapunov is then used to derive continuous acceleration-based controllers which render our system stable.

The remainder of this paper is structured as follows: in Sect. 2, the robot model is defined; in Sect. 3, the artificial potential field functions are defined under the influence of kinodynamic constraints; in Sect. 4, the acceleration-based control laws are derived, while in Sect. 5, stability analysis of the robotic system is carried out; in Sect. 6, we demonstrate the effectiveness of the proposed controllers via computer simulations which guide the follower robot to track the leaders reference path while maintaining a low-degree formation; and finally, Sect. 7 concludes the paper and outlines future work in the area.

2 Vehicle Model

In this section, we derive a kinematic model for multiple formations of car-like mobile robots.

We will consider h, $h \in \mathbb{N}$ formations with n, $n \in \mathbb{N}$, car-like mobile robots in the Euclidean plane, where the ith agent in the hth formation is denoted by $\mathcal{A}_{hi}$. We denote the hth formation as $\mathcal{A}_h$.

Without loss of generalization, we let $\mathcal{A}_{h1}$ represent the leader in the hth formation structure while the others in $\mathcal{A}_h$ take the role of followers. With reference to Fig. 1 and for $h = 1, \ldots, M$, $i = 1, \ldots, n$, (x_{hi}, y_{hi}) represents the Cartesian coordinates and gives the reference point of $\mathcal{A}_{hi}$.

Moreover, θ_{hi} gives the orientation of $\mathcal{A}_{hi}$ with respect to the z_1-axis of the z_1z_2-plane while ϕ_{hi} gives the steering angle with respect to its longitudinal axis of $\mathcal{A}_{hi}$.

L_{hi} represents the distance between the centers of the front and rear axles and l_{hi} is the length of each axle of $\mathcal{A}_{hi}$.

The mobile robots have velocity level nonholonomic constraint. We assume no slippage (i.e., $\dot{x}_{hi}\sin\theta_{hi} - \dot{y}_{hi}\cos\theta_{hi} = 0$) and pure rolling (i.e., $\dot{x}_{hi}\cos\theta_{hi} + \dot{y}_{hi}\sin\theta_{hi} = v_{hi}$) of the car-like mobile robots.

Next, to ensure that each robot safely steers past an obstacle, we adopt the nomenclature of [12] and construct circular regions that protect the robot. With reference to Fig. 1, given the *clearance parameters* $\epsilon_1 > 0$ and $\epsilon_2 > 0$, we enclose the each vehicle by a protective circular region centered at (x_{hi}, y_{hi}) with radius $r_{hi} = \frac{1}{2}\sqrt{(L_{hi} + 2\epsilon_1)^2 + (l_{hi} + 2\epsilon_2)^2}$ for $h = 1, \ldots, M$, and $i = 1, \ldots, n$.

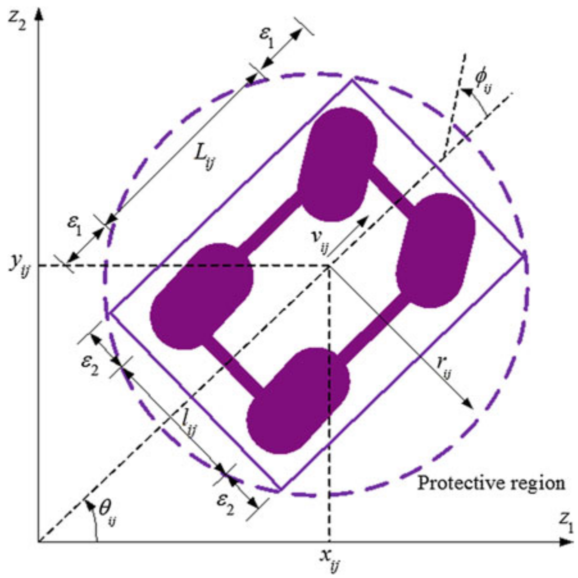


Fig. 1 Kinematic model of the agent $\mathcal{A}_{hi}$

These generate the *nonholonomic constraints* on the system. The kinodynamic model of the system, adopted from [11] is

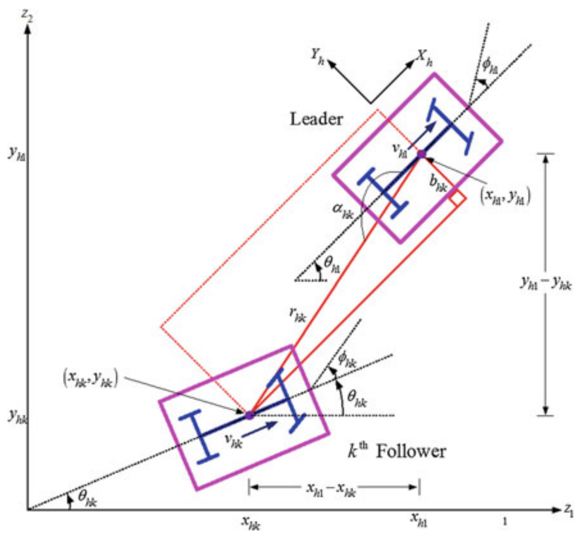
$$\left.\begin{array}{l}\dot{x}_{hi} = v_{hi}\cos\theta_{hi} - \frac{L_{hi}}{2}\omega_{hi}\sin\theta_{hi}, \quad \dot{y}_{hi} = v_{hi}\sin\theta_{hi} + \frac{L_{hi}}{2}\omega_{hi}\cos\theta_{hi}, \\ \dot{\theta}_{hi} = \frac{v_{hi}}{L_{hi}}\tan\phi_{hi} =: \omega_{hi}, \quad \dot{v}_{hi} := \sigma_{hi1}, \quad \dot{\omega}_{hi} := \sigma_{hi2},\end{array}\right\} \tag{1}$$

for $h = 1, \ldots, M$, and $i = 2, \ldots, n$. Here, v_{hi} and ω_{hi} are, respectively, the instantaneous translational and rotational velocities of $\mathcal{A}_{hi}$, while σ_{hi1} and σ_{hi2} are the instantaneous translational and rotational accelerations of $\mathcal{A}_{hi}$. Without any loss of generality, we assume that $\phi_{hi} = \theta_{hi}$. Now, system (1) is a description of the instantaneous velocities and accelerations of $\mathcal{A}_{hi}$.

2.1 Leader-Follower Based Formation Scheme

To desire a substantial degree of rigidness in our formation of $\mathcal{A}_h$, we define two reference frames: the body frame which is fixed with the rotating body of each leader $\mathcal{A}_{h1}$ in $\mathcal{A}_h$, and a space frame, the inertial frame similar to one proposed in [13]. We assign a Cartesian coordinate system $(X_h - Y_h)$ fixed on the leaders body of the $\mathcal{A}_h$, as shown in Fig. 2, based on the concept of an instantaneous co-rotating frame of reference. Thus, when the leader, $\mathcal{A}_{h1}$ rotates, we have a rotation of the $X_h - Y_h$ axes. To define the co-rotating frame of reference, first an origin is selected on the leader robot, $\mathcal{A}_{h1}$ in $\mathcal{A}_h$ at (x_{h1}, y_{h1}). An axis of rotation is then setup that is perpendicular to the plane of motion of the leader, $\mathcal{A}_{h1}$. Thus, at any selected moment t, the chosen rotating frame of reference rotates at an angular rate equal to

Fig. 2 The proposed scheme utilizing a rotation of axes with axis fixed at the leader



the rate of rotation of the leader, $\mathcal{A}_{h1}$ about (x_{h1}, y_{h1}). Thus, given the leader's position and its orientation, as long as (r_{hk}, α_{hk}) for h = 1, …, M, and k = 1, …, n is fixed, the kth follower robot's position in $\mathcal{A}_h$ will be unique. We define the shape of $\mathcal{A}_h$ as $\zeta_h = [\zeta_{h2}, \zeta_{h3}, \ldots, \zeta_{hn}]^T$, where for $\zeta_{hk} = [r_{hk}, \alpha_{hk}]^T$ for h = 1, …, M, and k = 2, …, n.

Definition 1 Let $2r_{hk} < r_{hk}^d < 2r_{hk} + \xi_h$ where for $\xi_h > 0$, for h = 1, …, M, and k = 2, …, n, the group of mobile robots make a $\zeta_h^d = [\zeta_{h2}^d, \zeta_{h3}^d, \ldots, \zeta_{hn}^d]^T$ formation, if $\exists T > 0, \forall t > T$:

$$r_{hk} = r_{hk}^d \quad and \quad \alpha_{hk} = \alpha_{hk}^d.$$

This gives then the polar coordinate representation of the follower's position relative to that of the leader in a given formation. However, such representations using polar coordinates lead to certain singularities in the controllers. To eliminate such singular points, we consider the position of the kth follower in $\mathcal{A}_h$ by considering the relative distances of $\mathcal{A}_{hk}$ in $\mathcal{A}_h$, from its leader, $\mathcal{A}_{h1}$ along the given X_h and Y_h directions. Thus, we have:

$$\begin{aligned} A_{hk} &= -(x_{h1} - x_{hk})\cos\theta_{h1} - (y_{h1} - y_{hk})\sin\theta_{h1}, \\ B_{hk} &= (x_{h1} - x_{hk})\sin\theta_{h1} - (y_{h1} - y_{hk})\cos\theta_{h1}, \end{aligned} \tag{2}$$

for h = 1, …, M, and k = 2, …, n and A_{hk} and B_{hk} are the relative positions with respect to the X_h-Y_h coordinate system of the kth followers in A_h. If A_{hk} and B_{hk} are known and fixed, the follower's position will be distinctive. Thus, to obtain a desired formation, one needs to know distances a_{hk} and b_{hk}, the desired relative positions along the X_h-Y_h directions, such that the control objective would be to achieve $A_{hk} \to a_{hk}$ and $B_{hk} \to b_{hk}$. i.e., $r_{hk} \to r_{hk}^d$ and $\alpha_{hk} \to \alpha_{hk}^d$, where $r_{hk}^d = \sqrt{a_{hk}^2 + b_{hk}^2}$ and $\alpha_{hk}^d = \tan\left(\frac{a_{hk}}{b_{hk}}\right)$ for $h = 1, \ldots, M$, and $k = 2, \ldots, n$.

3 Artificial Potential Field Function

This section formulates collision free trajectories of the robot system under kinodynamic constraints in a given workspace. We want to design the acceleration controllers, σ_{hi1} and σ_{hi2}, so that the group of robots moves safely towards the leaders target while maintaining a desired low-degree formation.

3.1 Attractive Potential Field Functions

3.1.1 Attraction to Target

For the establishment and advancement of the group of M formations having n mobile robots, we incorporate the leader-follower scheme within the framework of the Lyapunov-based control scheme (LbCS) [11]. A target is assigned to the leader of each formation. The leader, $\mathcal{A}_{h1}$ for $h = 1, \ldots, M$ will move towards its defined target with center (p_{h11}, p_{h12}), while the follower robots, $\mathcal{A}_{hk}$ move with their leader while maintaining a desired relative position to their leader. For the attraction of the leader, $\mathcal{A}_{h1}$ to its designated target, we consider an attractive potential function for $h = 1, \ldots, M$

$$V_{h1}(\mathbf{x}) = \frac{1}{2}\left[(x_{h1} - p_{h11})^2 + (y_{h1} - p_{h12})^2 + v_{h1}^2 + \omega_{h1}^2\right]. \tag{3}$$

For $\mathcal{A}_{hi}$ to maintain its desired relative position with respect to the leader, $\mathcal{A}_{h1}$, we utilize the following potential function for $h = 1, \ldots, M$ and $i = 2, \ldots, n$

$$V_{hi}(\mathbf{x}) = \frac{1}{2}\left[(A_{hi} - a_{hi})^2 + (B_{hi} - b_{hi})^2 + + v_{hi}^2 + \omega_{hi}^2\right]. \tag{4}$$

3.1.2 Auxiliary Function

To guarantee the convergence of the mobile robots to their designated targets, we design an auxiliary function as

$$G_{h1}(\mathbf{x}) = \frac{1}{2}\left[(x_{h1} - p_{h11})^2 + (y_{h1} - p_{h12})^2 + \rho_{h1}(\theta_{h1} - p_{h13})^2\right], \tag{5}$$

where p_{h13} is the prescribed final orientation of the leader robot, $\mathcal{A}_{h1}$ and

$$G_{hi}(\mathbf{x}) = \frac{1}{2}\left[(A_{hi} - a_{hi})^2 + (B_{hi} - b_{hi})^2 + \rho_{hi}(\theta_{hi} - \theta_{h1})^2\right], \tag{6}$$

for $i = 2, \ldots, n$ and $h = 1, \ldots, M$. The constant ρ_{hi} is a binary constant which we shall call the *angle-gain parameter* and will take a value of one only if a final pre-defined orientation is warranted, else it takes the default value of zero [11]. This auxiliary function is then multiplied to the repulsive potential field functions to be designed in the following subsections.

3.2 Repulsive Potential Field Functions

We desire the leader, $\mathcal{A}_{h1}$ and its followers, $\mathcal{A}_{hk}$ avoid all fixed and moving obstacles intersecting their paths.

3.2.1 Fixed Obstacles in the Workspace

Let us fix $q \in \mathbb{N}$ solid obstacles within the boundaries of the workspace. We assume that the lth obstacle is a circular disk with center (o_{l1}, o_{l2}) and radius ro_l. For $\mathcal{A}_{hi}$ to avoid the lth obstacle, we consider

$$FO_{hil}(\mathbf{x}) = \frac{1}{2}\left[(x_{hi} - o_{l1})^2 + (y_{hi} - o_{l2})^2 - (ro_l + r_{hi})^2\right], \tag{7}$$

as an avoidance function, where $h = 1, \ldots, M$, $i = 1, \ldots, n$, and $l = 1, \ldots, q$. These obstacle avoidance functions will be combined with appropriate tuning parameters to generate repulsive potential field functions in the workspace.

3.2.2 Moving Obstacles

To generate feasible trajectories, we consider moving obstacles of which the system has prior knowledge. Here, each mobile robot, has to be treated as a moving obstacle for all other mobile robots in the workspace.

3.2.3 Minimum Inter-robot Distance

We desire to maintain a minimum inter robot separation distance between the robots. This prevents $\mathcal{A}_{hi}$ from getting very close to (or colliding with) $\mathcal{A}_{hj}$ [11], especially during the re-establishment of the prescribed formation when the system is distorted. We can consider the following obstacle avoidance function

$$MO_{hij}(\mathbf{x}) = \frac{1}{2}\left[\left(x_{hi} - x_{hj}\right)^2 + \left(y_{hi} - y_{hj}\right)^2 - (r_{hi} + r_{hj})^2\right], \tag{8}$$

for $h = 1, \ldots, M$, and $i, j = 1, \ldots, n$, with $j \neq i$.

3.2.4 Inter Formation Avoidance

We also desire for each formation structure in the system to avoid any other formation structure in the workspace. For ith body of $\mathcal{A}_h$ to evade the uth body of $\mathcal{A}_m$, we adopt

$$DO_{himu}(\mathbf{x}) = \frac{1}{2}\left[(x_{hi} - x_{mu})^2 + (y_{hi} - y_{mu})^2 - (r_{hi} + r_{mu})^2\right], \tag{9}$$

for $i, u = 1, \ldots, n$ and $h, m = 1, \ldots, M$ with $m \neq h$.

3.2.5 Dynamic Constraints

Practically, the steering angles of the mobile robots are limited due to mechanical singularities while the translational speed is restricted due to safety reasons. Subsequently, we have $|v_{hi}| < v_{\max}$, where $v_{\max}$ is the *maximal achievable speed* of the $\mathcal{A}_{hi}$ and $|\omega_{hi}| < \frac{v_{\max}}{|\rho_{\min}|}$, where $\rho_{\min} := \frac{L_{hi}}{\tan(\phi_{\max})}$. This condition arises due to the boundness of the steering angle ϕ_{hi}. That is, $|\phi_{hi}| \leq \phi_{\max} < \pi/2$, where $\phi_{\max}$ is the *maximal steering angle*. Hence, we consider the following avoidance functions:

$$U_{hi1}(\mathbf{x}) = \frac{1}{2}(v_{\max} - v_{hi})(v_{\max} + v_{hi}), \tag{10}$$

$$U_{hi2}(\mathbf{x}) = \frac{1}{2}\left(\frac{v_{\max}}{|\rho_{\min}|} - \omega_{hi}\right)\left(\frac{v_{\max}}{|\rho_{\min}|} + \omega_{hi}\right), \tag{11}$$

for $h = 1, \ldots, M$ and $i = 1, \ldots, n$. These positive functions would guarantee the adherence to the limitations imposed upon the steering angle and the velocities of $\mathcal{A}_{hi}$ when encoded appropriately into the Lyapunov function.

4 Design of the Acceleration Controllers

The nonlinear acceleration control laws for system (1), will be designed using LBCS as proposed in [11].

4.1 Lyapunov Function

We now construct the total potentials, that is, a Lyapunov function for system (1).

$$
\begin{aligned}
L_{(1)}(\mathbf{x}) = & \sum_{h=1}^{M}\sum_{i=1}^{n}\left\{V_{hi}(\mathbf{x}) + G_{hi}(\mathbf{x})\left[\sum_{l=1}^{q}\frac{\alpha_{hil}}{FO_{hil}(\mathbf{x})} + \sum_{s=1}^{2}\frac{\beta_{his}}{U_{his}(\mathbf{x})}\right]\right\} \\
& + \sum_{h=1}^{M}\sum_{i=1}^{n}G_{hi}(\mathbf{x})\left\{\sum_{\substack{j=1\\ j\neq 1}}^{n}\frac{\eta_{hij}}{MO_{hij}(\mathbf{x})} + \sum_{u=1}^{n}\sum_{\substack{m=1\\ m\neq h}}^{M}\frac{\gamma_{himu}}{DO_{himu}(\mathbf{x})}\right\}
\end{aligned} \tag{12}
$$

4.2 Nonlinear Acceleration Controllers

The design of the feedback controllers begins by noting that the functions f_{hik} to g_{hij} for $h = 1, \ldots, M,\ i = 1, \ldots, n,\ j = 1,\ 2$ and $k = 1, \ldots, 3$ are defined as (on suppressing $\mathbf{x}$):

$$
\begin{aligned}
f_{h11} = & \left[1 + \sum_{l=1}^{q}\frac{\alpha_{h1l}}{FO_{h1l}} + \sum_{s=1}^{2}\frac{\beta_{h1s}}{U_{h1s}}\right](x_{h1} - p_{h11}) - G_{h1}\sum_{l=1}^{q}\frac{\alpha_{h1l}}{FO_{h1l}^2}(x_{h1} - o_{hl1}) \\
& + \left[\sum_{\substack{j=1\\ j\neq 1}}^{n}\frac{\eta_{h1j}}{MO_{h1j}} + \sum_{u=1}^{n}\sum_{\substack{m=1\\ m\neq 1}}^{M}\frac{\gamma_{h1mu}}{DO_{h1mu}}\right](x_{h1} - p_{h11}) \\
& - \sum_{i=2}^{n}\left[\sum_{\substack{j=1\\ j\neq 1}}^{n}\frac{\eta_{h1j}}{MO_{h1j}} + \sum_{u=1}^{n}\sum_{\substack{m=1\\ m\neq 1}}^{M}\frac{\gamma_{h1mu}}{DO_{h1mu}}\right](A_{hi} - a_{hi})\cos\theta_{h1} \\
& + \sum_{i=2}^{n}\left[\sum_{\substack{j=1\\ j\neq 1}}^{n}\frac{\eta_{h1j}}{MO_{h1j}} + \sum_{u=1}^{n}\sum_{\substack{m=1\\ m\neq 1}}^{M}\frac{\gamma_{h1mu}}{DO_{h1mu}}\right](B_{hi} - b_{hi})\sin\theta_{h1} \\
& + \sum_{i=2}^{n}\left[1 + \sum_{l=1}^{q}\frac{\alpha_{h1l}}{FO_{h1l}} + \sum_{s=1}^{2}\frac{\beta_{h1s}}{U_{h1s}}\right][-(A_{hi} - a_{hi})\cos\theta_{h1} + (B_{hi} - b_{hi})\sin\theta_{h1}] \\
& - 2G_{h1}\sum_{\substack{j=1\\ j\neq 1}}^{n}\frac{\eta_{h1j}}{MO_{h1j}^2}(x_{h1} - x_{hj}) - 2G_{h1}\sum_{u=1}^{n}\sum_{\substack{m=1\\ m\neq 1}}^{M}\frac{\gamma_{h1mu}}{DO_{h1mu}^2}(x_{h1} - x_{mu}),
\end{aligned}
$$

$$
\begin{aligned}
f_{h12} = {} & \left[1 + \sum_{l=1}^{q} \frac{\alpha_{h1l}}{FO_{h1l}} + \sum_{s=1}^{2} \frac{\beta_{h1s}}{U_{h1s}}\right](y_{h1} - p_{h12}) - G_{h1} \sum_{l=1}^{q} \frac{\alpha_{h1l}}{FO_{h1l}^2}(y_{h1} - o_{hl1}) \\
& + \left[\sum_{\substack{j=1\\ j\neq 1}}^{n} \frac{\eta_{h1j}}{MO_{h1j}} + \sum_{u=1}^{n} \sum_{\substack{m=1\\ m\neq 1}}^{M} \frac{\gamma_{h1mu}}{DO_{h1mu}}\right](y_{h1} - p_{h12}) \\
& - \sum_{i=2}^{n} \left[\sum_{\substack{j=1\\ j\neq 1}}^{n} \frac{\eta_{h1j}}{MO_{h1j}} + \sum_{u=1}^{n} \sum_{\substack{m=1\\ m\neq 1}}^{M} \frac{\gamma_{h1mu}}{DO_{h1mu}}\right](A_{hi} - a_{hi}) \sin\theta_{h1} \\
& - \sum_{i=2}^{n} \left[\sum_{\substack{j=1\\ j\neq 1}}^{n} \frac{\eta_{h1j}}{MO_{h1j}} + \sum_{u=1}^{n} \sum_{\substack{m=1\\ m\neq 1}}^{M} \frac{\gamma_{h1mu}}{DO_{h1mu}}\right](B_{hi} - b_{hi}) \cos\theta_{h1} \\
& - \sum_{i=2}^{n} \left[1 + \sum_{l=1}^{q} \frac{\alpha_{h1l}}{FO_{h1l}} + \sum_{s=1}^{2} \frac{\beta_{h1s}}{U_{h1s}}\right][(A_{hi} - a_{hi}) \sin\theta_{h1} + (B_{hi} - b_{hi}) \cos\theta_{h1}] \\
& - 2G_{h1} \sum_{\substack{j=1\\ j\neq 1}}^{n} \frac{\eta_{h1j}}{MO_{h1j}^2}(y_{h1} - y_{hj}) - 2G_{h1} \sum_{u=1}^{n} \sum_{\substack{m=1\\ m\neq 1}}^{M} \frac{\gamma_{h1mu}}{DO_{h1mu}^2}(y_{h1} - y_{mu}),
\end{aligned}
$$

$$
\begin{aligned}
f_{h13} = {} & \left[\sum_{l=1}^{q} \frac{\alpha_{h1l}}{FO_{h1l}} + \sum_{s=1}^{2} \frac{\beta_{h1s}}{U_{h1s}} + \sum_{\substack{j=1\\ j\neq 1}}^{n} \frac{\eta_{h1j}}{MO_{h1j}} + \sum_{u=1}^{n} \sum_{\substack{m=1\\ m\neq 1}}^{M} \frac{\gamma_{h1mu}}{DO_{h1mu}}\right] \rho_{h1}(\theta_{h1} - p_{h13}) \\
& - \sum_{i=2}^{n} \left[\sum_{l=1}^{q} \frac{\alpha_{h1l}}{FO_{h1l}} + \sum_{s=1}^{2} \frac{\beta_{h1s}}{U_{h1s}} + \sum_{\substack{j=1\\ j\neq 1}}^{n} \frac{\eta_{h1j}}{MO_{h1j}} + \sum_{u=1}^{n} \sum_{\substack{m=1\\ m\neq 1}}^{M} \frac{\gamma_{h1mu}}{DO_{h1mu}}\right] \rho_{hi}(\theta_{hi} - \theta_{h1}),
\end{aligned}
$$

$$
g_{h11} = 1 + G_{h1} \frac{\beta_{h11}}{U_{h11}^2}, \quad g_{h12} = 1 + G_{h1} \frac{\beta_{h12}}{U_{h12}^2},
$$

and for $i = 2, \ldots, n$

$$
\begin{aligned}
f_{hi1} = {} & \left[1 + \sum_{l=1}^{q} \frac{\alpha_{hil}}{FO_{hil}} + \sum_{s=1}^{2} \frac{\beta_{his}}{U_{his}}\right] [(A_{hi} - a_{hi}) \cos\theta_{h1} - (B_{hi} - b_{hi}) \sin\theta_{h1}] \\
& + \left[\sum_{\substack{j=1 \\ j \neq 1}}^{n} \frac{\eta_{hij}}{MO_{hij}} + \sum_{u=1}^{n} \sum_{\substack{m=1 \\ m \neq 1}}^{M} \frac{\gamma_{himu}}{DO_{himu}}\right] (A_{hi} - a_{hi}) \cos\theta_{h1} \\
& - \left[\sum_{\substack{j=1 \\ j \neq 1}}^{n} \frac{\eta_{hij}}{MO_{hij}} + \sum_{u=1}^{n} \sum_{\substack{m=1 \\ m \neq 1}}^{M} \frac{\gamma_{himu}}{DO_{himu}}\right] (B_{hi} - b_{hi}) \sin\theta_{h1} \\
& - 2 \sum_{\substack{j=1 \\ j \neq 1}}^{n} \frac{\eta_{hij} G_{hi}}{MO_{hij}^2} (x_{hi} - x_{hj}) - 2 \sum_{u=1}^{n} \sum_{\substack{m=1 \\ m \neq 1}}^{M} \frac{\gamma_{himu} G_{hi}}{DO_{himu}^2} (x_{hi} - x_{mu}) \\
& - \sum_{l=1}^{q} \frac{\alpha_{hil} G_{hi}}{FO_{hil}^2} (x_{hi} - o_{hl1}),
\end{aligned}
$$

$$
\begin{aligned}
f_{hi2} = {} & \left[1 + \sum_{l=1}^{q} \frac{\alpha_{h1l}}{FO_{h1l}} + \sum_{s=1}^{2} \frac{\beta_{his}}{U_{his}}\right] [(A_{hi} - a_{hi}) \sin\theta_{h1} + (B_{hi} - b_{hi}) \cos\theta_{h1}] \\
& + \left[\sum_{\substack{j=1 \\ j \neq 1}}^{n} \frac{\eta_{hij}}{MO_{hij}} + \sum_{u=1}^{n} \sum_{\substack{m=1 \\ m \neq 1}}^{M} \frac{\gamma_{himu}}{DO_{himu}}\right] (A_{hi} - a_{hi}) \sin\theta_{h1} \\
& + \left[\sum_{\substack{j=1 \\ j \neq 1}}^{n} \frac{\eta_{hij}}{MO_{hij}} + \sum_{u=1}^{n} \sum_{\substack{m=1 \\ m \neq 1}}^{M} \frac{\gamma_{himu}}{DO_{himu}}\right] (B_{hi} - b_{hi}) \cos\theta_{h1} \\
& - 2 \sum_{\substack{j=1 \\ j \neq 1}}^{n} \frac{\eta_{hij} G_{hi}}{MO_{hij}^2} (y_{hi} - y_{hj}) - 2 \sum_{u=1}^{n} \sum_{\substack{m=1 \\ m \neq 1}}^{M} \frac{\gamma_{himu} G_{hi}}{DO_{himu}^2} (y_{hi} - y_{mu}) \\
& - \sum_{l=1}^{q} \frac{\alpha_{hil} G_{hi}}{FO_{hil}^2} (y_{hi} - o_{hl1}),
\end{aligned}
$$

$$
f_{hi3} = \left[\sum_{l=1}^{q} \frac{\alpha_{hil}}{FO_{hil}} + \sum_{s=1}^{2} \frac{\beta_{his}}{U_{his}} + \sum_{\substack{j=1 \\ j \neq 1}}^{n} \frac{\eta_{hij}}{MO_{hij}} + \sum_{u=1}^{n} \sum_{\substack{m=1 \\ m \neq 1}}^{M} \frac{\gamma_{himu}}{DO_{himu}}\right] \rho_{hi} (\theta_{hi} - \theta_{h1}),
$$

$$
g_{hi1} = 1 + G_{hi} \frac{\beta_{hi1}}{U_{hi1}^2}, \quad g_{hi2} = 1 + G_{hi} \frac{\beta_{hi2}}{U_{hi2}^2}.
$$

So, we design the following theorem:

Theorem 1 *Consider a team of M formation structures each comprising of n car-like mobile robots whose motion is governed by the ODE's described in system* (1). *The principal goal is to establish and control each follower robot in each formation structure to track its designated leader in a cooperative manner and facilitate maneuvers within a constrained environment and reach the target configuration. The subtasks include; restrictions placed on the workspace, convergence to predefined targets, and consideration of kinodynamic constraints. Utilizing the attractive and repulsive potential field functions, the following continuous time-invariant acceleration control laws can be generated, that intrinsically guarantees stability, in the sense of Lyapunov, of system* (1) *as well*:

$$\left.\begin{aligned} \sigma_{hi1} &= -[\delta_{hi1} v_{hi} + f_{hi1}\cos\theta_{hi} + f_{hi2}\sin\theta_{hi}]/g_{hi1}, \\ \sigma_{hi2} &= -\left[\delta_{hi2}\omega_{hi} + \tfrac{L_{hi}}{2}(f_{hi2}\cos\theta_{hi} - f_{hi1}\sin\theta_{hi}) + f_{hi3}\right]/g_{hi2}, \end{aligned}\right\} \tag{13}$$

for $h = 1, \ldots, M$ and $i = 1, \ldots, n$, *where* $\delta_{hi1} > 0$, *and* $\delta_{hi2} > 0$ *are constants commonly known as convergence parameters*.

5 Stability Analysis

Theorem 2 *Let* (p_{h11}, p_{h12}) *be the position of the target of the leader in* $\mathcal{A}_h$, *and* p_{hi3} *for* $i = 1, \ldots, n$, *be the desired final orientations of the robots in each* $\mathcal{A}_h$. *Given* a_{hi} *and* b_{hi} *in* $\mathcal{A}_h$, *let* p_{hi1} *and* p_{hi2} *satisfy*

$$\begin{aligned} a_{hi} &= -(p_{h11} - p_{hi1})\cos\theta_{h1} - (p_{h12} - p_{hi2})\sin\theta_{h1}, \\ b_{hi} &= (p_{h11} - p_{hi1})\sin\theta_{h1} - (p_{h12} - p_{hi2})\cos\theta_{h1}, \end{aligned}$$

for $i = 2, \ldots, n$ *and* $h = 1, \ldots, M$.

Given $\mathbf{x}^*_{hi} := (p_{hi1}, p_{hi2}, p_{hi3}, 0, 0) \in \mathbb{R}^5$, *if* $\mathbf{x}_e := (\mathbf{x}^*_{11}, \mathbf{x}^*_{12}, \ldots, \mathbf{x}^*_{hn}) \in \mathbb{R}^{5\times n\times M}$ *is an equilibrium point for* (1), *then* $\mathbf{x}_e \in D(L_{(1)}(\mathbf{x}))$ *is a stable equilibrium point of system* (1).

Proof One can easily verify the following, for $i = 1, \ldots, n$ and $h = 1, \ldots, M$:

1. $L_{(1)}(\mathbf{x})$ is defined, continuous and positive over the domain $D(L_{(1)}(\mathbf{x})) = \{\mathbf{x} \in \mathbb{R}^{5\times M\times n} : FO_{hil}(\mathbf{x}) > 0, \quad l = 1, \ldots, q; \quad DO_{himu}(\mathbf{x}) > 0, \quad m = 1, \ldots, M, u = 1, \ldots, n, m \neq h; \; MO_{hij}(\mathbf{x}) > 0, \quad j = 1, \ldots, n, j \neq i; \; U_{his}(\mathbf{x}) > 0, s = 1, 2\}$;
2. $L_{(1)}(\mathbf{x}^*) = 0$;
3. $L_{(1)}(\mathbf{x}) > 0 \;\; \forall \mathbf{x} \in D(L_{(1)}(\mathbf{x}))/\mathbf{x}_e$.

Next, consider the time derivative of the candidate Lyapunov function along a particular trajectory of system (1):

$$\dot{L}_{(1)}(\mathbf{x}) = \sum_{i=1}^{n} \left[f_{hi1}\dot{x}_{hi} + f_{i2}\dot{y}_{hi} + f_{hi3}\dot{\theta}_{hi} + g_{hi1}v_{hi}\dot{v}_{hi} + g_{hi2}\omega_{hi}\dot{\omega}_{hi} \right].$$

Substituting the controllers given in (13) and the governing ODEs for system (1), we obtain the following semi-negative definite function

$$\dot{L}_{(1)}(\mathbf{x}) = -\sum_{i=1}^{n} \left(\delta_{hi1}v_{hi}^2 + \delta_{hi2}\omega_{hi}^2 \right) \leq 0.$$

Thus, $\dot{L}_{(1)}(\mathbf{x}) \leq 0 \; \forall \mathbf{x} \in D(L_{(1)}(\mathbf{x}))$ and $\dot{L}_{(1)}(\mathbf{x}_e) = 0$. Finally, it can be easily verified that $L_{(1)}(\mathbf{x}) \in C^1\left(D(L_{(1)}(\mathbf{x}))\right)$, which makes up the fifth and final criterion of a Lyapunov function. Hence, $L_{(1)}(\mathbf{x})$ is classified as a Lyapunov function for system (1) and $\mathbf{x}_e$ is a stable equilibrium point in the sense of Lyapunov.

Remark 1 This result is in no contradiction with Brockett's Theorem [14] as we have not proven asymptotic stability.

6 Simulation Results

In this section, we illustrate the effectiveness of the proposed continuous time-invariant controllers within the framework of the Lyapunov-based control scheme by simulating a virtual scenario.

We consider the motion of a pair of 4 cars in a double platoon formation in a two dimensional space with static obstacles in its path. Each follower robot in each formation structure is assigned a unique position relative to its leader as seen in Fig. 3. This is achieved by assigning appropriate values to (a_{hk}, b_{hk}) to obtain a geometric formation structure. While a leaders $\mathcal{A}_{11}$ and $\mathcal{A}_{21}$ moves towards its intended target, the followers, $\mathcal{A}_{hk}$ for h = 1, 2 and $k = 2, \ldots, 4$ maintains the desired relative position to its designated leader, thus maintaining a desired formation structure. Assuming that the appropriate units have been accounted for, Table 1 provides the corresponding initial and final configurations of the two robots and other necessary parameters required to simulate the scenario. Figure 4 shows the evolution of the Lyapunov function and its time derivative along the system trajectory while Fig. 5 depicts the time evolution of the nonlinear controllers of the leaders $\mathcal{A}_{11}$ and $\mathcal{A}_{21}$. Figures 4 and 5 show the boundness and convergence of the variables at the final state, respectively, implying the effectiveness of the controllers.

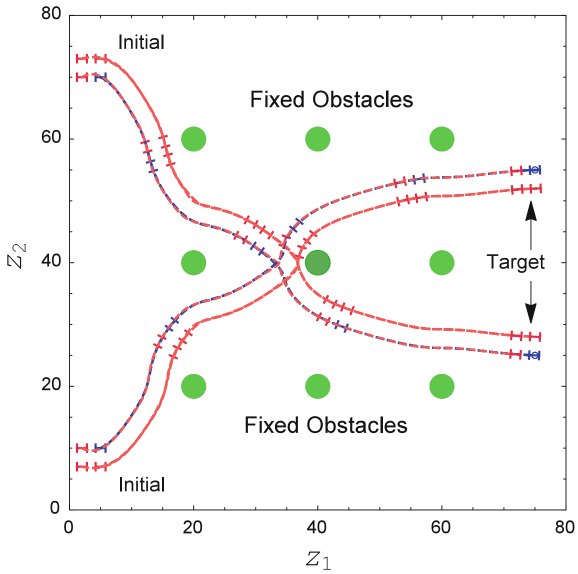


Fig. 3 The proposed scheme utilizing a rotation of axes with axis fixed at the leader

Table 1 Numerical values of initial and final states, constraints and parameters

Initial configuration	
Rectangular positions of leaders	$(x_{11}, y_{11}) = (5, 70), (x_{21}, y_{21}) = (5, 10)$
Desired relative distances of followers	$a_{12} = a_{23} = b_{14} = b_{22} = 0, b_{12} = b_{13} = -3$ $a_{13} = a_{14} = a_{22} = a_{24} = b_{23} = b_{24} = 3$
Translational velocity	$v_{hi} = 0.5$ for $i = 1, \ldots, 4$, $h = 1, 2$
Rotational velocities	$\omega_{hi} = 0$, for $i = 1, \ldots, 4$, $h = 1, 2$
Angular positions	$\theta_{hi} = 0$, for $i = 1, \ldots, 4$, $h = 1, 2$
Constraints and parameters	
Dimensions of robots	$L_{hi} = 1.6$, $l_{hi} = 1.2$ for $i = 1, \ldots, 4$, $h = 1, 2$
$\mathcal{A}_1$ leader's target	$(p_{11}, p_{12}) = (79, 24.5)$, $rt_{11} = 0.5$,
$\mathcal{A}_2$ leader's target	$(p_{21}, p_{22}) = (79, 56)$, $rt_{21} = 0.5$
Fixed obstacles	$(o_{11}, o_{12}) = (20, 20), (o_{21}, o_{22}) = (20, 40)$
	$(o_{31}, o_{32}) = (20, 60), (o_{41}, o_{42}) = (40, 20)$
	$(o_{51}, o_{52}) = (40, 40), (o_{61}, o_{62}) = (40, 60)$
	$(o_{71}, o_{72}) = (60, 20)$, $(o_{81}, o_{82}) = (60, 40)$
	$(o_{91}, o_{92}) = (60, 60)$, $ro_l = 2$ for $l = 1, \ldots, 9$
Max. translational velocity	$v_{\max} = 5$
Max. steering angle	$\phi_{\max} = \pi/2$
Clearance parameters	$\varepsilon_1 = 0.1$, $\varepsilon_2 = 0.05$
Control and convergence parameters	
Collision avoidance	$\alpha_{hil} = 1$, for $i = 1, \ldots, 4$, $h = 1, 2$, $l = 1, \ldots, 9$
	$\eta_{hij} = 0.01$ for $h = 1, 2$, $i, j = 1, \ldots, 4$, $j \neq i$,
	$\gamma_{himu} = 0.01$, for $h, m = 1, 2$, $i, j = 1, \ldots, 4$, $h \neq m$
Dynamics constraints	$\beta_{his} = 1$, for $i = 1, \ldots, 4$, $h, s = 1, 2$,
Convergence	$\delta_{11j} = 12{,}000$, $\delta_{21j} = 8000$, for $j = 1, 2$
	$\delta_{hij} = 50$, for $i = 1, \ldots, 4$, $h, j = 1, 2$

Fig. 4 Evolution of $L_1(\mathbf{x})$ and its time derivative $\dot{L}_{(1)}(\mathbf{x})$

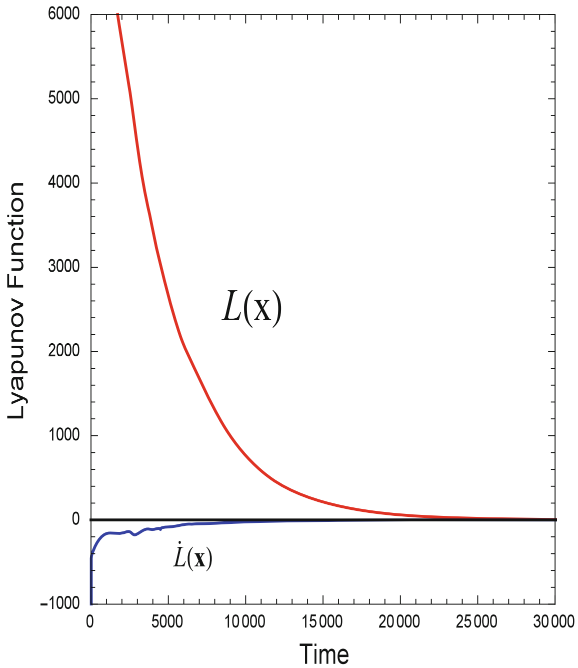
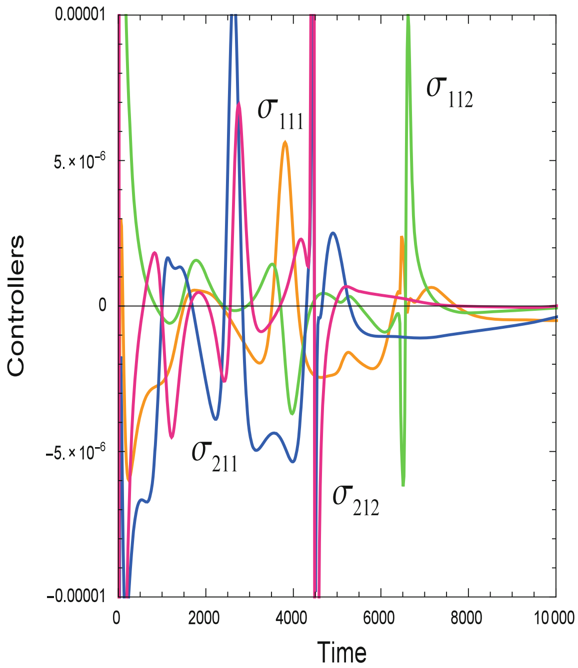


Fig. 5 Translational and rotational accelerations of the leaders $\mathcal{A}_{11}$ and $\mathcal{A}_{21}$



7 Conclusion

In this paper we have proposed a leader follower scheme for the coordination of MRSs in an environment populated with obstacles. The contribution of this paper is the capability of the robots in a desired geometric formation to move from some initial configuration to some final state while maintaining a desired low degree formation. The advantage of the proposed scheme is that we can have multiple formations structures having different geometric shape. The approach also considers inter-robot and inter-formation collision avoidance. Collision free maneuvers with fixed obstacles are imbedded in the overall framework. The derived controllers produced feasible trajectories and ensured a nice convergence of the system to its equilibrium state while satisfying the necessary kinematic and dynamic constraints. The effectiveness of the proposed control laws were demonstrated via a computer simulation. The control system presented is also easily scalable for any number of agents in any geometric shape. Future research will address more general formation applications in 3D.

References

1. Kanjanawanishkul, M., Hofmeister, K., Zell, A.: Smooth reference tracking of a mobile robot using nonlinear model predictive control. In: Proceedings of the 4th European Conference on Mobile Robots, ECMR'09, pp. 161–166, Croatia (2009)
2. Ze-su, C., Jie, Z., Jian, C.: Formation control and obstacle avoidance for multiple robots subject to wheel slip. Int. J. Adv. Robot. Syst. **9**, 1–15 (2012)
3. van den Broek, T.H.A., van de Wouw, N., Nijmeijer, H.: Formation control of unicycle mobile robots: a virtual structure approach. In: Proceedings of Joint 48th IEEE International Conference on Decision and Control and 28th Chinese Control Conference, pp. 8328–8333, Shanghai, PR China (2009)
4. Toibero, J., Roberti, F., Fiorini, P., Carelli, R.: Hybrid formation control for non-holonomic wheeled robots. In: Recent Progress in Robotics: Lecture Notes in Control and Information Systems, pp. 21–34. Springer, Berlin (2008)
5. Mas, I., Kitts, C.: Object manipulation using cooperative mobile multi-robot systems. In: Proceedings of the World Congress on Engineering *and* Computer Science. San Francisco, USA (2012)
6. Hou, S.P., Cheah, C.C., Slotine, J.J.E.: Dynamic region following formation control for a swarm of robots. In: Proceedings of IEEE International Conference on Robotics and Automation, pp. 1929–1934, Kobe, Japan (2009)
7. Olfati-Saber, R., Dunbar, W.B., Murray, R.M.: Cooperative control of multi-vehicle systems using cost graphs and optimization. In: Proceedings of the American Control Conference, pp. 1–15 (2003)
8. Guo, J., Yan, G., Lin, Z.: Cooperative control synthesis for moving-target enclosing with changing topologies. In: Proceedings of IEEE International Conference on Robotics and Automation, Alaska, USA (2010)
9. Yamaguchi, H.: A distributed motion coordination strategy for multiple nonholonomic mobile robots in cooperative hunting operations. Robot. Auton. Syst. **43**(4), 257–282 (2003)

10. Chung, S.-J., Slotine, J.-J.: Cooperative robot control and concurrent synchronization of Lagrangian systems. In: Proceedings of 46th IEEE International Conference on Design and Control (2007)
11. Sharma, B.: New directions in the applications of the Lyapunov-based control scheme to the find path problem. Ph.D. thesis, University of the South Pacific, Suva, Fiji Islands (2008)
12. Raghuwaiya, K., Singh, S.: Formation types of multiple steerable 1-trailer mobile robots via split/rejoin maneuvers. New Zealand J. Math. **43**, 7–21 (2013)
13. Kang, W., Xi, N., Tan, J., Wang, Y.: Formation control of multiple autonomous robots: theory and experimentation. Intell. Autom. Soft Comput. **10**(2), 1–17 (2004)
14. Brockett, R.W.: Asymptotic stability and feedback stabilisation. In: Differential Geometry Control Theory, pp. 181–191. Springer, Berlin (1983)

Swarming Intelligence of 1-Trailer Systems

Jai Raj, Krishna Raghuwaiya, Shonal Singh, Bibhya Sharma and Jito Vanualailai

Abstract In this paper, we propose a new solution to motion planning and control problem for a flock of 1-trailer systems. A set of artificial potential field functions is proposed for the flock of 1-trailer robots via the Lyapunov-based control scheme for the avoidance of swarm of boids and attraction to their designated targets. The dynamic environment for the first time includes a swarm of boids, which is governed separately by a system of ODE's. The swarm exhibits collective emergent behaviors in the vicinity of the workspace while the flock of 1-trailer systems safely maneuver from their initial configuration to designated targets. The effectiveness of the control laws is demonstrated via computer simulations. The novelty of the paper lies in the simplicity of the controllers and the ease in the treatment of the dynamic environment.

Keywords Swarm · Obstacle avoidance · 1-trailer system · Emergent · Stability

1 Introduction

Devising motion planning algorithms for multi-agents sharing a common work-space is inherently difficult. This is a result of the environment being no longer static but dynamic. Static environments have provided excellent breeding grounds for high-powered algorithms so far [1]. However, more recently there has been a shift of emphasis to include dynamic environments due to its applications in the real world. The dynamic obstacles can incorporate the mobile robots themselves as well as other moving solid objects or obstacles in the environment.

Fundamental to the motion planning problem of multi-agents is the need to control and plan the motions of the agents that would yield inter-agent and agent to

J. Raj (✉) · K. Raghuwaiya · S. Singh · B. Sharma · J. Vanualailai
School of Computing, Information and Mathematical Sciences,
The University of the South Pacific, Suva, Fiji
e-mail: raj_j@usp.ac.fj

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_21

obstacle collision avoidances. Numerous papers have discussed this problem, some of which includes methods such as discretization of the configuration time-space using sequential space slicing [2], plan-merging [3], negotiations [4], online artificial potential fields strategy [5, 6], decomposition of the problem into path planning and velocity planning sub-problems [7] and a Lyapunov-based control scheme for various nonholonomic multi-agents [8], to name a few.

Biologically inspired algorithms that mimic the flocking behavior are essential in accomplishing the control objective of a group while ensuring collision-free flight path [8]. Common objectives nowadays include formation flight control, satellite clustering, exploration, surveillance, foraging and cooperate manipulation [8, 9]. Basically, robots working towards a common objective can satisfy stringent time, manpower and monetary demands, enhance performance and robustness, and harness desired multi-behaviors, each of which is extremely difficult if not entirely impossible to obtain from single agents [9].

2 System Modeling

2.1 A Model of the 1-Trailer Robot

Two different trailer systems can be distinguished from literature; standard and the general trailer systems, grouped into two different categories based upon their different hooking schemes [10]. The standard 1-trailer system embodies a car-like robot and an on-axle hitched two wheeled passive trailer. The authors will consider n standard 1-trailer system, in Euclidean plane. The connections between the two bodies give rise to the following holonomic constraints on the system:

$$x_{i2} = x_{i1} - \frac{L_1}{2}\cos\theta_{i1} - \frac{L_2+2d}{2}\cos\theta_{i2}; \quad y_{i2} = y_{i1} - \frac{L_1}{2}\sin\theta_{i1} - \frac{L_2+2d}{2}\sin\theta_{i2}$$

We define $d := \varepsilon_1 + c$, where c is a small offset (see Fig. 1). The model of the ith tractor-trailer, adopted from [10], is

$$\left.\begin{array}{l} \dot{x}_{i1} = v_i\cos\theta_{i1} - \frac{L_1}{2}\omega_i\sin\theta_{i1}, \quad \dot{y}_{i1} = v_i\sin\theta_{i1} + \frac{L_1}{2}\omega_i\cos\theta_{i1}, \\ \dot{\theta}_{i1} = \omega_i, \ \dot{\theta}_{i2} = \frac{v_i}{L_2}\sin(\theta_{i1}-\theta_{i2}), \quad \dot{v}_i = \sigma_{i1}, \ \dot{\omega}_i = \sigma_{i2}, \end{array}\right\} \tag{1}$$

where v_i and ω_i are, respectively, the instantaneous translational and rotational velocities, while σ_{i1} and σ_{i2} are the instantaneous translational and rotational accelerations of the ith tractor.

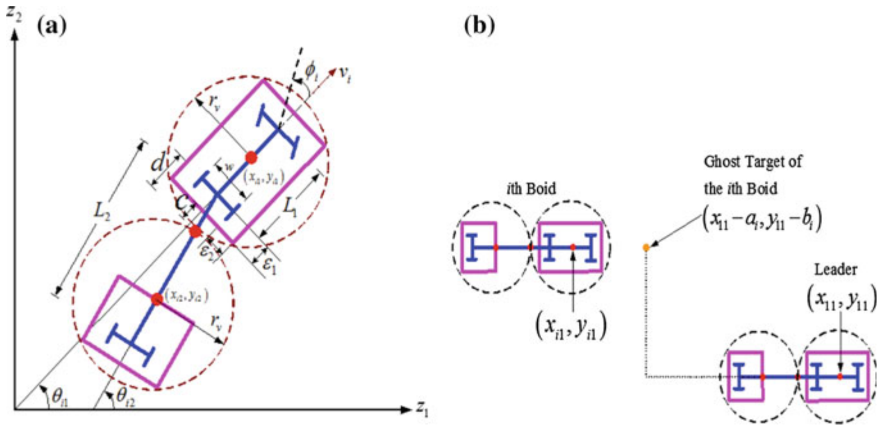


Fig. 1 The kinematic model and the positioning of the 1—trailer robots. **a** Kinematic model of the *i*th 1-trailer boid ($t = 0$). **b** Positioning of a mobile target relative to the position of the leader

2.2 A Swarm Model

A general swarm model, formulated by Mogilner et al. [11] will be utilized, moving with the velocity of the swarm's centroid. Following the nomenclature of Reynolds [12], each member of the flock is denoted as a boid. At time $t \geq 0$, let $(xb_k(t), yb_k(t)), k = 1, \ldots, h$, be the planar position of the kth individual, which we shall define as a boid residing in a disk of radius $rb_k \geq 0$,

$$B_k = \left\{ (z_1, z_2) \in \mathbb{R}^2 : (z_1 - xb_k)^2 + (z_2 - yb_k)^2 \leq rb_k^2 \right\}. \tag{2}$$

Let us define the *centroid of the swarm* as

$$(xb_k, yb_k) = \left(\frac{1}{h} \sum_{k=1}^{h} xb_k, \frac{1}{h} \sum_{k=1}^{h} yb_k \right). \tag{3}$$

At time $t \geq 0$, let $(\upsilon b_k(t), \omega b_k(t) := (\dot{x}b_k(t), \dot{y}b_k(t)))$ be the instantaneous velocity of the kth boid. Using the above notations, we obtain a system of first order ODE's for the kth boid, assuming the initial conditions at $t = t_0 \geq 0$:

$$\dot{x}b_k = \upsilon b_k(t), \; \dot{y}b_k = \omega b_k(t) \; \} \tag{4}$$

for $k = 1, \ldots, h$.

2.3 Model of the Dynamic Team

The dynamic model of the 1-trailer system and the swarm of boids is

$$\left.\begin{array}{l}\dot{x}_{i1} = v_i \cos\theta_{i1} - \frac{L_1}{2}\omega_i \sin\theta_{i1},\\ \dot{y}_{i1} = v_i \sin\theta_{i1} + \frac{L_1}{2}\omega_i \cos\theta_{i1},\\ \dot{\theta}_{i1} = \omega_i,\\ \dot{\theta}_{i2} = \frac{v_i}{L_2}\sin(\theta_{i1} - \theta_{i2}),\\ \dot{v}_i = \sigma_{i1},\\ \dot{\omega}_i = \sigma_{i2}, i = 1, \ldots, n,\\ \dot{x}b_k = \upsilon b_k,\\ \dot{y}b_k = \omega b_k, k = 1, \ldots, h.\end{array}\right\} \tag{5}$$

3 Deployment of the Lyapunov-Based Control Scheme

3.1 Target of the 1-Trailer Mobile Robots

To initiate movement, we propose to have a target. The leader moves towards its target (t_{11}, t_{12}). For the followers, the mobile ghost targets are positioned relative to the position of the leader whose center is given by $(t_{i1}, t_{i2}) = (x_{11} - a_i, y_{11} - b_i)$, for $i = 2, \ldots, n$. For the attraction to these targets, we consider an attractive potential function

$$V_i(\mathbf{x}) = \frac{1}{2}\left[(x_{i1} - t_{i1})^2 + (y_{i1} - t_{i2})^2 + v_i^2 + \omega_i^2\right] \tag{6}$$

for $i = 1, \ldots, n$.

3.2 Auxiliary Function

To guarantee the convergence of the tractor-trailer mobile robot to its designated target, we design an auxiliary function defined as

$$G_i(\mathbf{x}) = \frac{1}{2}\left[(x_{i1} - t_{i1})^2 + (y_{i1} - t_{i2})^2 + (\theta_{i1} - t_{i3})^2 + (\theta_{i2} - t_{i4})^2\right] \tag{7}$$

for $i = 1, \ldots, n$ where t_{i3} is the desired final orientation of the ith tractor and t_{i4} is the desired final orientation of the ith trailer.

3.3 Collision Avoidance Amongst the 1-Trailer Robots

To generate feasible trajectories, we consider moving obstacles of which the system has *priori* knowledge. Here, each member of the flock becomes a moving obstacle for all the other members. Therefore, for the mth body of the ith tractor trailer to avoid the uth body of the jth tractor-trailer, we have

$$M_{muij}(\mathbf{x}) = \frac{1}{2}\left[\left(x_{im} - x_{ju}\right)^2 + \left(y_{im} - y_{ju}\right)^2 - (2 \times r_v)^2\right] \tag{8}$$

for $i, j = 1, \ldots, n$ with $i \neq j$ and $m, u = 1, 2$.

3.4 Dynamic Constraints

Practically, the steering and bending angles of an articulated robot is limited due to mechanical singularities while the translational speed is restricted due to safety reasons. Subsequently, we have; (i) $|v| \leq v_{max}$, where v_{max} is the maximal speed of the tractor; (ii) $|\phi_i| \leq \phi_{max} < \frac{\pi}{2}$, where ϕ_{max} is the maximal steering angle of the tractor; and (iii) $|\theta_{i1} - \theta_{i2}| \leq \theta_{max} < \frac{\pi}{2}$, where θ_{max} is the maximum bending angle of the trailer with respect to the orientation of the tractor. This prevents a jack knife situation. Thus, the trailer is free to rotate within $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. Considering these constraints as artificial obstacles, we have the following potential field functions:

$$DC_{i1}(\mathbf{x}) = \frac{1}{2}\left[(v_{\max} - v_i)(v_{\max} + v_i)\right] \tag{9}$$

$$DC_{i2}(\mathbf{x}) = \frac{1}{2}\left[\left(\frac{v_{\max}}{|\rho_{\min}|} - \omega_i\right)\left(\frac{v_{\max}}{|\rho_{\min}|} + \omega_i\right)\right] \tag{10}$$

$$DC_{i3}(\mathbf{x}) = \frac{1}{2}\left[(\theta_{\max} - (\theta_{i2} - \theta_{i1}))(\theta_{\max} + (\theta_{i2} - \theta_{i1}))\right] \tag{11}$$

for $i = 1, \ldots, n$.

3.5 Avoidance of the Swarm of Boids by the 1-Trailer Robots

For the boids to avoid the 1-trailer mobile robots, we design repulsive potential field functions of the form

$$S_{ik} = \frac{1}{2}\left[(x_{im} - xb_k)^2 + (y_{im} - xb_k)^2 - (r_i + rb_k)^2\right], \tag{12}$$

where $i = 1, \ldots, n, m = 1, 2$ and $k = 1, \ldots, h$.

For the attraction of the swarm to the centroid and for the inter-individual avoidance of the swarm, the functions are:

3.6 Attraction to the Centroid

To ensure that the individuals of the swarm are attracted towards each other and also form a cohesive group by having a measurement of the distance from the kth individual to the swarm centroid, we use the following attraction function:

$$R_k(\mathbf{x}) = \frac{1}{2}\left[\left(xb_k - \frac{1}{h}\sum_{j=1}^{h} xb_j\right)^2 + \left(yb_k - \frac{1}{h}\sum_{j=1}^{h} yb_j\right)^2\right]. \tag{13}$$

The role is to ensure that the ith individual is attracted to the swarm centroid.

3.7 Avoidance of the Boundaries of the Workspace

For the avoidance of the left, upper, right and lower boundaries, the following position obstacle avoidance functions are utilized, respectively:

$$\begin{aligned} WB_{k1} &= xb_k - rb_k, \\ WB_{k2} &= yb_k - rb_k, \\ WB_{k3} &= b_1 - (xb_k - rb_k), \\ WB_{k4} &= b_2 - (yb_k - rb_k). \end{aligned} \tag{14}$$

3.8 Intra Swarm Avoidance

For the boids to avoid each other, we design obstacle avoidance function of the form

$$Q_{kj}(\mathbf{x}) = \frac{1}{2}\left[(xb_k - xb_j)^2 + (yb_k - yb_j)^2 - (rb_k + rb_j)^2\right], \tag{15}$$

for $k, j = 1, \ldots, h, j \neq k$. The function is an Euclidean measure of the distance between the individual boids.

4 Design of Nonlinear Controllers

This section will represent a Lyapunov function and the nonlinear control laws for system (5) will be designed using the LbCS.

4.1 Lyapunov Function

As per the LbCS, we combine all the attractive and repulsive potential field functions, and introducing *tuning parameters (or control parameters)*, $\psi_k > 0$, $\eta_{ks} > 0$, $\beta_{kj} > 0$, $\sigma_{ik} > 0$, $\xi_{ku} > 0$, $\zeta_{kr} > 0$ and $\lambda_{kr} > 0$ for $i, j, k, l, m, n, r, s, u \in \mathbb{N}$, we define a Lyapunov function for system (5)

$$L(\mathbf{x}) = \sum_{i=1}^{n}\left[V_i(\mathbf{x}) + G_i(\mathbf{x})\left(\sum_{s=1}^{3}\frac{\gamma_{is}}{DC_{is}(\mathbf{x})} + \sum_{\substack{j=1\\ i\neq j}}^{n}\sum_{m=1}^{2}\sum_{u=1}^{2}\frac{\beta_{muij}}{M_{muij}(\mathbf{x})} + \sum_{k=1}^{h}\frac{\sigma_{ik}}{S_{ik}(\mathbf{x})}\right)\right] + \sum_{i=1}^{n}\sum_{k=1}^{h}G_i(\mathbf{x})\left[\psi_k R_k(\mathbf{x}) + R_k(\mathbf{x})\left(\sum_{s=1}^{4}\frac{\eta_{ks}}{WB_{ks}(\mathbf{x})} + \sum_{\substack{j=1\\ j\neq k}}^{h}\frac{\beta_{kj}}{Q_{kj}(\mathbf{x})}\right)\right] \quad (16)$$

4.2 Controller Design

To extract the control laws for the kinodynamic system, we differentiate the various components of $L(\mathbf{x})$ separately with respect to t along a solution of system (5), carry out the necessary substitutions and upon suppressing $\mathbf{x}$, we have the controllers for the dynamic model.

1-Trailer Agents The process of designing the feedback controllers begins by noting that the functions f_{ik} for $i = 1, \ldots, 4, k = 1, \ldots, n$ and $k = 1, \ldots, 3$ are defined as (on suppressing $\mathbf{x}$):

$$f_{11} = \begin{bmatrix} 1 + \sum_{s=1}^{3} \frac{\gamma_{1s}}{DC_{1s}} + \sum_{j=2}^{n} \sum_{m=1}^{2} \sum_{u=1}^{2} \frac{\beta_{mu1j}}{M_{mu1j}} \\ + \sum_{k=1}^{h} \frac{\sigma_{1k}}{S_{1k}} + \psi_k R_k + R_k \sum_{s=1}^{4} \frac{\eta_{ks}}{WB_{ks}} + R_k \sum_{\substack{j=1 \\ j \neq h}}^{h} \frac{\beta_{kj}}{Q_{kj}} \end{bmatrix} (x_{11} - t_{11})$$

$$+ \sum_{\substack{j=1 \\ i \neq j}}^{n} \sum_{m=1}^{2} \sum_{u=1}^{2} G_j \frac{\beta_{muj1}}{M_{muj1}^2} \left(x_{jm} - x_{1u}\right)$$

$$- \sum_{i=2}^{n} \left[1 + \sum_{s=1}^{3} \frac{\gamma_{is}}{DC_{is}} + \sum_{\substack{j=1 \\ i \neq j}}^{n} \sum_{m=1}^{2} \sum_{u=1}^{2} \frac{\beta_{muij}}{M_{muij}} \right] (x_{i1} - t_{i1})$$

$$- G_1 \sum_{\substack{j=1 \\ i \neq j}}^{n} \sum_{m=1}^{2} \sum_{u=1}^{2} \frac{\beta_{mu1j}}{M_{mu1j}^2} \left(x_{1m} - x_{ju}\right) - G_1 \sum_{i=1}^{n} \frac{\sigma_{1k}}{S_{1k}^2} (x_{11} - xb_k)$$

$$f_{21} = \begin{bmatrix} 1 + \sum_{s=1}^{3} \frac{\gamma_{1s}}{DC_{1s}} + \sum_{j=2}^{n} \sum_{m=1}^{2} \sum_{u=1}^{2} \frac{\beta_{mu1j}}{M_{mu1j}} + \sum_{k=1}^{h} \frac{\sigma_{1k}}{S_{1k}} \\ + \psi_k R_k + R_k \sum_{s=1}^{4} \frac{\eta_{ks}}{WB_{ks}} + R_k \sum_{\substack{j=1 \\ j \neq h}}^{h} \frac{\beta_{kj}}{Q_{kj}(\mathbf{x})} \end{bmatrix} (y_{11} - t_{12})$$

$$+ \sum_{\substack{j=1 \\ i \neq j}}^{n} \sum_{m=1}^{2} \sum_{u=1}^{2} G_j \frac{\beta_{muj1}}{M_{muj1}^2} \left(y_{jm} - y_{1u}\right)$$

$$- \sum_{i=2}^{n} \left[1 + \sum_{s=1}^{3} \frac{\gamma_{is}}{DC_{is}} + \sum_{\substack{j=1 \\ i \neq j}}^{n} \sum_{m=1}^{2} \sum_{u=1}^{2} \frac{\beta_{muij}}{M_{muij}} \right] (y_{i1} - t_{i2})$$

$$- G_1 \sum_{\substack{j=1 \\ i \neq j}}^{n} \sum_{m=1}^{2} \sum_{u=1}^{2} \frac{\beta_{mu1j}}{M_{mu1j}^2} \left(y_{1m} - y_{ju}\right) - G_1 \sum_{i=1}^{n} \frac{\sigma_{1k}}{S_{1k}^2} (y_{11} - yb_k)$$

$$f_{1i} = \begin{bmatrix} 1 + \sum_{s=1}^{3} \frac{\gamma_{is}}{DC_{is}} + \sum_{j=2}^{n} \sum_{m=1}^{2} \sum_{u=1}^{2} \frac{\beta_{muij}}{M_{muij}} + \sum_{k=1}^{h} \frac{\sigma_{ik}}{S_{ik}} \\ + \psi_k R_k + R_k \sum_{s=1}^{4} \frac{\eta_{ks}}{WB_{ks}} + R_k \sum_{\substack{j=1 \\ j \neq k}}^{h} \frac{\beta_{kj}}{Q_{kj}} \end{bmatrix} (x_{i1} - t_{i1})$$

$$+ \sum_{\substack{j=1 \\ i \neq j}}^{n} \sum_{m=1}^{2} \sum_{u=1}^{2} G_j \frac{\beta_{muji}}{M_{muji}^2} \left(x_{jm} - y_{iu}\right)$$

$$- G_i \sum_{\substack{j=1 \\ i \neq j}}^{n} \sum_{m=1}^{2} \sum_{u=1}^{2} \frac{\beta_{muij}}{M_{muij}^2} \left(x_{im} - x_{ju}\right) - G_i \sum_{i=1}^{n} \frac{\sigma_{1k}}{S_{1k}^2} (x_{i1} - xb_k)$$

$$f_{i2} = \left[\begin{array}{l} 1+\sum\limits_{s=1}^{3}\frac{\gamma_{is}}{DC_{is}}+\sum\limits_{j=2}^{n}\sum\limits_{m=1}^{2}\sum\limits_{u=1}^{2}\frac{\beta_{muij}}{M_{muij}}+\sum\limits_{k=1}^{h}\frac{\sigma_{ik}}{S_{ik}} \\ +\psi_k R_k + R_k\sum\limits_{s=1}^{4}\frac{\eta_{ks}}{WB_{ks}}+R_k\sum\limits_{\substack{j=1\\ j\neq k}}^{h}\frac{\beta_{kj}}{Q_{kj}} \end{array}\right](y_{i1}-t_{i2})$$

$$+\sum_{\substack{j=1\\ i\neq j}}^{n}\sum_{m=1}^{2}\sum_{u=1}^{2}G_j\frac{\beta_{muji}}{M_{muji}^2}\left(y_{jm}-y_{iu}\right)$$

$$-G_i\sum_{\substack{j=1\\ i\neq j}}^{n}\sum_{m=1}^{2}\sum_{u=1}^{2}\frac{\beta_{muij}}{M_{muij}^2}\left(y_{im}-y_{ju}\right)-G_i\sum_{i=1}^{n}\frac{\sigma_{1k}}{S_{1k}^2}(y_{i1}-yb_k)$$

$$f_{i3} = \left[\begin{array}{l} \sum\limits_{s=1}^{3}\frac{\gamma_{is}}{DC_{is}}+\sum\limits_{\substack{j=1\\ i\neq j}}^{n}\sum\limits_{m=1}^{2}\sum\limits_{u=1}^{2}\frac{\beta_{muij}}{M_{muij}}+\sum\limits_{k=1}^{h}\frac{\sigma_{1k}}{S_{1k}} \\ +\psi_k R_k + R_k\sum\limits_{s=1}^{4}\frac{\eta_{ks}}{WB_{ks}}+R_k(\mathbf{x})\sum\limits_{\substack{j=1\\ j\neq k}}^{h}\frac{\beta_{kj}}{Q_{kj}} \end{array}\right](\theta_{i1}-t_{i3})$$

$$-G_i\frac{\gamma_{i3}}{DC_{i3}^2}(\theta_{i2}-\theta_{i1})-G_i\frac{L_1}{2}\sum_{\substack{j=1\\ i\neq j}}^{n}\sum_{u=1}^{2}\frac{\beta_{2uij}}{M_{2uij}^2}\left(x_{i2}-x_{ju}\right)\sin\theta_{i1}$$

$$+G_i\frac{L_1}{2}\sum_{\substack{j=1\\ i\neq j}}^{n}\sum_{u=1}^{2}\frac{\beta_{2uij}}{M_{2uij}^2}\left(y_{i2}-y_{ju}\right)\cos\theta_{i1}$$

$$+\frac{L_1}{2}\sum_{\substack{j=1\\ i\neq j}}^{n}\sum_{m=1}^{2}G_j\frac{\beta_{m2ji}}{M_{m2ji}^2}\left[\left(x_{jm}-x_{i2}\right)\sin\theta_{i1}-\left(y_{jm}-y_{i2}\right)\cos\theta_{i1}\right]$$

$$f_{i4} = \left[\begin{array}{l} \sum\limits_{s=1}^{3}\frac{\gamma_{is}}{DC_{is}}+\sum\limits_{\substack{j=1\\ i\neq j}}^{n}\sum\limits_{m=1}^{2}\sum\limits_{u=1}^{2}\frac{\beta_{muij}}{M_{muij}}+\sum\limits_{k=1}^{h}\frac{\sigma_{1k}}{S_{1k}} \\ +\psi_k R_k + R_k\sum\limits_{s=1}^{4}\frac{\eta_{ks}}{WB_{ks}}+R_k(\mathbf{x})\sum\limits_{\substack{j=1\\ j\neq k}}^{h}\frac{\beta_{kj}}{Q_{kj}} \end{array}\right](\theta_{i2}-t_{i4})$$

$$+G_i\frac{\gamma_{i3}}{DC_{i3}^2}(\theta_{i2}-\theta_{i1})-G_i\frac{L_2+2d}{2}\sum_{j=1,j\neq i}^{n}\sum_{u=1}^{2}\frac{\beta_{2uij}}{M_{2uij}^2}\left(x_{i2}-x_{ju}\right)\sin\theta_{i2}$$

$$+G_i\frac{L_2+2d}{2}\sum_{j=1,j\neq i}^{n}\sum_{u=1}^{2}\frac{\beta_{2uij}}{M_{2uij}^2}\left(y_{i2}-y_{ju}\right)\cos\theta_{i2}$$

$$+\frac{L_2+2d}{2}\sum_{\substack{j=1\\ i\neq j}}^{n}\sum_{m=1}^{2}G_j\frac{\beta_{m2ji}}{M_{m2ji}^2}\left[\left(x_{jm}-x_{i2}\right)\sin\theta_{i2}-\left(y_{jm}-y_{i2}\right)\cos\theta_{i2}\right]$$

for $i=2,\ldots,n,\;\; k=1,\ldots,h$ and

$$f_{i5} = 1 + G_i \frac{\gamma_{i1}}{DC_{i1}^2}, f_{i6} = 1 + G_i \frac{\gamma_{i2}}{DC_{i2}^2} \tag{17}$$

for $i = 1, \ldots, n$.

Swarm of boids The nonlinear velocity controllers for the swarm of boids are:

$$\begin{aligned} Lx_i = {} & G_k \left(\psi_i + \sum_{s=1}^{4} \frac{\eta_{is}}{WB_{is}(\mathbf{x})} + \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{\beta_{ij}}{Q_{ij}(\mathbf{x})} \right) \left(xb_i - \frac{1}{n} \sum_{j=1}^{n} xb_j \right) \\ & + R_i G_k \left(\frac{\eta_{i3}}{WB_{i3}^2} - \frac{\eta_{i1}}{WB_{i1}^2} \right) - 2R_i G_k \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{\beta_{ij}}{Q_{ij}^2} (xb_i - xb_j), \end{aligned}$$

$$\begin{aligned} Ly_i = {} & G_k \left(\psi_i + \sum_{s=1}^{4} \frac{\eta_{is}}{WB_{is}(\mathbf{x})} + \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{\beta_{ij}}{Q_{ij}(\mathbf{x})} \right) \left(yb_i - \frac{1}{n} \sum_{j=1}^{n} yb_j \right) \\ & + R_i G_k \left(\frac{\eta_{i4}}{WB_{i4}^2} - \frac{\eta_{i2}}{WB_{i2}^2} \right) - 2R_i G_k \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{\beta_{ij}}{Q_{ij}^2} (yb_i - yb_j). \end{aligned}$$

5 Stability Analysis

So, we design the following theorem:

Theorem 1 *Consider a flock of nonholonomic 1-trailer mobile robots and a swarm of boids whose motion is governed by the ODEs described in system (5). The principal goal is to establish a prescribed formation, facilitate split/rejoin maneuvers of the robots within a constrained and dynamic environment and reach the target configuration with the original formation. The subtasks include; restrictions placed on the workspace, convergence to predefined targets, and consideration of kinodynamic constraints. Utilizing the attractive and repulsive potential field functions, the following continuous control laws are*:

$$\left. \begin{aligned} \sigma_{i1} &= -[\delta_{i1} v_i + f_{i1} \cos\theta_{i1} + f_{i2} \sin\theta_{i1} + f_{i4} \tfrac{1}{L_2} \sin(\theta_{i1} - \theta_{i2})]/f_{i5}, \\ \sigma_{i2} &= -\left[\delta_{i2}\omega_i + \tfrac{L_1}{2}(f_{i2}\cos\theta_{i1} - f_{i1}\sin\theta_{i1}) + f_{i3}\right]/f_{i6}, \\ vb_i &= -\alpha_{i1} Lx_i, \\ \omega b_i &= -\alpha_{i2} Ly_i, \end{aligned} \right\} \tag{18}$$

for $i = 1, \ldots, n$ and $k = 1, \ldots, m$, where $\alpha_{i1}, \alpha_{i2}, \delta_{i1}, \delta_{i2} > 0$ are constants commonly known as convergence parameters.

6 Stability

Theorem 2 *If a fixed point* $\mathbf{x}_i^* = (t_{i1}, t_{i2}, t_{i3}, t_{i4}, 0, 0,) \in \mathbb{R}^6$ *is an equilibrium point of* $\mathcal{A}_i$, $i \in \{1, \ldots, n\}$, *then* $\mathbf{x}^* = (\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_n^*) \in D(L_{(5)}(\mathbf{x}))$ *is a stable equilibrium point of system* (5).

Proof One can easily verify the following, for $i \in \{1, \ldots, n\}$, $k = 1, \ldots, h$ and $m \in \{1, 2\}$:

1. $L_{(5)}(\mathbf{x})$ is defined, continuous and positive over the domain $D(L_{(5)}(\mathbf{x})) = \{\mathbf{x} \in \mathbb{R}^{6n} : MO_{muij}(\mathbf{x}) > 0, j = 1, \ldots, n, j \neq i; \quad DC_{is}(\mathbf{x}) > 0, s = 1, \ldots, 3; WV_{is}(\mathbf{x}) > 0, s = 1, \ldots, 4; \; S_{ik}(\mathbf{x}) > 0, k = 1, \ldots, h; \; WB_{ks}(\mathbf{x}) > 0, s = 1, \ldots, 4; Q_{kj}(\mathbf{x}) > 0, j = 1, \ldots, h, j \neq k; \; W_{kl}(\mathbf{x}) > 0, l = 1, \ldots, q\}$;
2. $L_{(5)}(\mathbf{x}^*) = 0$;
3. $L_{(5)}(\mathbf{x}) > 0 \; \forall \mathbf{x} \in D(L_{(5)}(\mathbf{x}))/\mathbf{x}^*$.

Next, consider the time derivative of the candidate Lyapunov function along a particular trajectory of system (5):

$$\dot{L}_{(5)}(\mathbf{x}) = \sum_{i=1}^{n} \left[f_{i1}\dot{x}_{i1} + f_{i2}\dot{y}_{i1} + f_{i3}\dot{x}_{i2} + f_{i4}\dot{y}_{i2} + g_{i1}\dot{\theta}_{i1} + g_{i2}\dot{\theta}_{i2} + g_{i3}v_i\dot{v}_i + g_{i4}\omega_i\dot{\omega}_i \right] + \sum_{i=1}^{h} [Lx_i\dot{x}b_i + Ly_i\dot{y}b_i].$$

Substituting the controllers given in (18) and the governing ODEs for system (5), we obtain the following semi-negative definite function

$$\dot{L}_{(5)}(\mathbf{x}) = -\sum_{i=1}^{n} \left(\delta_{i1} v_i^2 + \delta_{i2} \omega_i^2 \right) - \sum_{i=1}^{h} \left(\alpha_{i1} Lx_i^2 + \alpha_{i2} Ly_i^2 \right) \leq 0.$$

Thus, $\dot{L}_{(5)}(\mathbf{x}) \leq 0 \; \forall \mathbf{x} \in D(L_{(5)}(\mathbf{x}))$ and $\dot{L}_{(5)}(\mathbf{x}^*) = 0$. Finally, it can be easily verified that $L_{(5)}(\mathbf{x}) \in C^1\big(D(L_{(5)}(\mathbf{x}))\big)$, which makes up the fifth and final criterion of a Lyapunov function. Hence, $L_{(5)}(\mathbf{x})$ is classified as a Lyapunov function for system (5) and $\mathbf{x}^*$ is a stable equilibrium point in the sense of Lyapunov.

7 Simulation

This section demonstrates the effectiveness of the nonlinear control laws with simulation. The setup is where 1—trailer mobile robots have to navigate from an initial to a final configuration, whilst avoiding moving obstacles en route their target. The moving obstacles are the swarm of boids in each of the cases. The scenario mimics a circular behavior exhibited by the swarm of boids. Initially, the swarm are randomly placed within the workspace and the 1—trailer mobile robots are given its initial conditions. The swarm of boids then come to their center and start exhibiting a circular behavior. The behavior that is exhibited varies from animals to animals. The swarm circle and in the midst, the 1—trailer mobile robots considers the swarm as obstacles and avoids it reaching its designated target. Table 1 provides all the values of the initial conditions, constraints and different parameters utilized in the simulation (Fig. 2).

Table 1 Numerical values of initial state, constraints, and control and convergence parameters

Initial State of the 1-trailer robots	
Rectangular position	$(x_{11}, y_{11}) = (10, 75)$, $(x_{21}, y_{21}) = (10, 80)$
	$(x_{31}, y_{31}) = (10, 70)$
Angular positions	$\theta_{i1} = \theta_{i2} = 0$, for $i = 1, \ldots, 3$.
Translational velocities	$v_1 = 0.5$, $v_2 = v_3 = 3$
Rotational velocities	$\omega_1 = \omega_2 = \omega_3 = 0.3$
Final orientations	$t_{i3} = t_{i4} = 0$ for $i = 1, \ldots, 3$
Constraints	
Dimension of robot	$L_1 = 2$, $L_2 = 1.2$, $w = 0.5$
Leader target	$(t_{11}, t_{12}) = (140, 75)$
Position of ghost targets	$(a_2, b_2) = (0, -5)$, $(a_3, b_3) = (0, 5)$
Max. translational velocity	$v_{\max} = 10$
Max. steering angle	$\phi_{\max} = 7\pi/18$
Min. turning radius	$\rho_{min} = 0.75$
Clearance parameters	$\epsilon_1 = 0.1$, $\epsilon_2 = 0.2$
Workspace boundaries	$b_1 = b_2 = 150$
Parameters for the boids	
Boundary limitations	$\eta_{is} = 0.0001$ for $i = 1, \ldots, 15$ and $s = 1, \ldots, 4$
Cohesion parameter	$\psi_i = 25$, for $i = 1, \ldots, 15$
Inter-individual obstacle avoidance	$\beta_{ij} = \text{random}[100, 500]$, for $i, j = 1, \ldots, 15$, $i \neq j$
Avoidance of boids by agents	$\sigma_{ik} = 1$, for $i = 1, \ldots, 15$ and $k = 1, 2, 3$
Convergence	$\alpha_{i1} = \alpha_{i2} = 0.01$, for $i = 1, \ldots, 15$
Parameters for the 1-trailer robots	
Boid avoidance	$\beta_{muij} = 0.0001$, for $i, j = 1, \ldots, 3$, $i \neq j$ and $m = u = 1, 2$
Dynamic constraints	$\gamma_{ku} = 0.1$, for $i = 1, 2, 3$ and $s = 1, \ldots, 4$
Convergence	$\sigma_{i1} = \sigma_{i2} = 500$, for $i = 1, \ldots, 3$

There are 15 boids and 3 car-like mobile robots

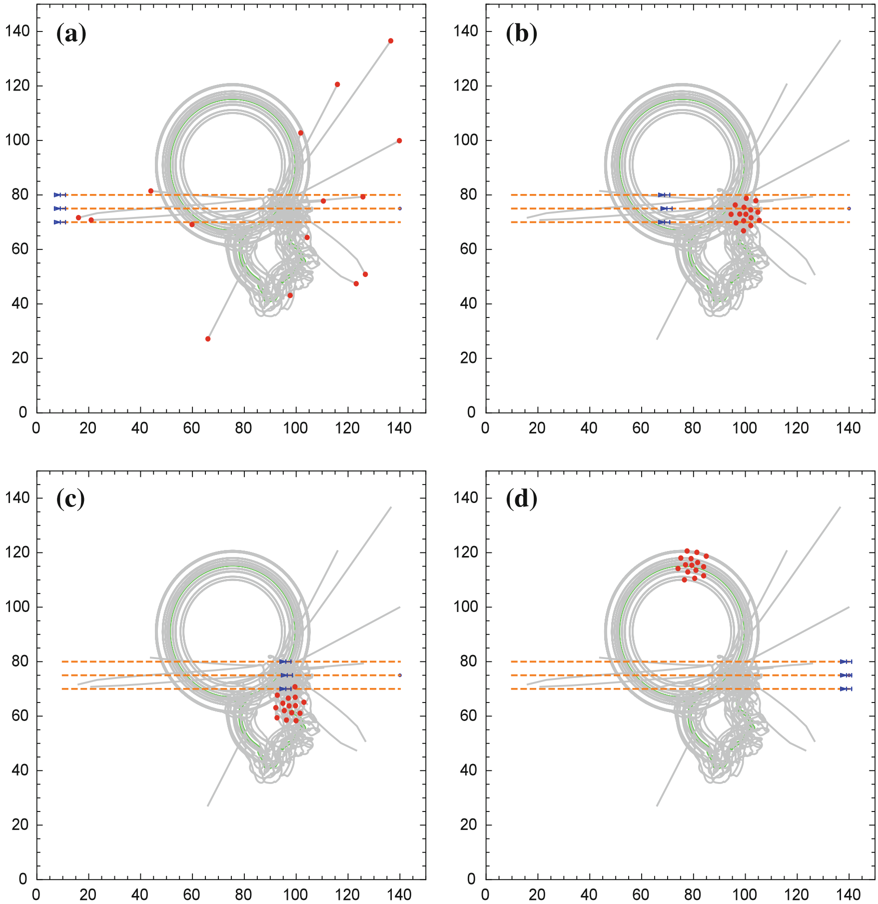


Fig. 2 There are $n = 15$ boids (shown in *red*), each with bin size 1, randomly positioned at the initial time $t = 0$. The *horizontal* and *vertical axes* give the coordinates $x_i(t)$ and $y_i(t)$ as $z_1(t)$ and $z_2(t)$, respectively, for the 1-trailer mobile robot and the swarm of boids at time $t \geq 0$. The *grey lines* show the trajectories of the individuals in the swarm. The path of the centroid for the swarm of boids is given by the Greek line. The 1-trailer mobile robots are shown in *blue color* with its trajectories in orange color. **a** The initial position of the 1-trailer mobile robots and the swarm of boids. **b** The 1-trailer mobile robots avoiding the swarm of boids at $t = 9,500$ units. **c** The 1-trailer mobile robots avoiding the swarm of boids at $t = 30,100$ units. **d** The final posture of the 1-trailer mobile robots and the position of the swarm at $t = 50,000$ units

8 Concluding Remarks

The design and control of a motion planner for multi-tasking of a robotic system is a complex, computer intensive yet an interesting problem. This paper presents a set of continuous acceleration control laws that successfully tackle the problem of formation control of 1-trailer robots moving in a dynamic environment which includes a swarm of boids pertaining to certain emergent behaviors. The controllers produced feasible trajectories and convergence of the system while satisfying the kinematic and dynamic constraints. To the author's knowledge, this is the first time in literature whereby moving 1-trailer robots are modelled together with swarm of emergent behaviors and successfully maintained within the Lyapunov-based control scheme via Lyapunov functions.

References

1. Berenson, D., Srinivasa, S.S., Ferguson, D., Collet, A., Kuffner, J.J.: Manipulation planning with workspace goal regions. In: IEEE International Conference on Robotics and Automation, 2009. ICRA'09, pp. 618–624. IEEE, 2009
2. Erdmann, M., Lozano-Perez, T.: On multiple moving objects. In: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 1419–1424, 1986
3. Alami, R., Fleury, S., Herrb, M., Ingrand, F., Robert, F.: Multi-robot cooperation in the martha project. IEEE Robot. Autom. Mag. **5**, 36–47 (1998)
4. Gerkey, B.P., Mataric, M.J.: Auction methods for multirobot coordination. IEEE Trans. Robot. Autom. **18**, 758–768 (2002)
5. Kostic, D., Adinandra, S., Caarls, J., Nijmeijer, H.: Collision-free motion coordination of unicycle multi-agent systems. In: 2010 American Control Conference, America, 2010
6. Latombe, J.C.: Robot motion planning. Kluwer Academic Publishers, USA (1991)
7. Kant, K., Zucker, S.W.: Toward efficiency trajectory planning: the path-velocity decomposition. Int. J. Robot. Res. **5**(3), 72–89 (1986)
8. Prasad, A., Sharma, A., Vanualailai, J.: A solution to the motion planning and control problem of a car—like robot via a single—layer perceptron. *Robotica* **32**(6), 935–952. ISSN 0263–5747, 2014
9. Prasad, A., Sharma, A., Vanualailai, J.: A new stabilizing solution for motion planning and control of multiple robots. Robotica, **First V**, 1–19. ISSN 0263–5747, 2014
10. Sharma, B., Vanualailai, J., Raghuwaiya, K., Prasad, A.: New potential field functions for motion planning and posture control of 1-trailer systems. Int. J. Math. Comput. Sci. **3**(1), 45–71 (2008)
11. Mogilner, A., Edelstein-Keshet, L., Bent, L., Spiros, A.: Mutual interactions, potentials, and individual distance in a social aggregation. J. Math. Biol. **47**, 352–389 (2003)
12. Reynolds, C.W.: Flocks, herds, and schools: a distributed behavioral model, in computer graphics. In Proceedings of the 14th annual conference on Computer graphics and interactive techniques, pp 25–34. New York, USA (1987)

Performance Evaluation of Face Classification Systems

Giovanni Betta, Domenico Capriglione, Mariella Corvino, Michele Gasparetto, Consolatina Liguori, Alfredo Paolillo and Emanuele Zappa

Abstract In this paper face classification systems based on 3D images are compared in terms of classification and metrological performance in presence of image uncertainty. In previous papers the authors proposed a new approach to classification and recognition problems. It is based on the evaluation of the image uncertainty and on the exploitation of such information to provide the confidence level of classification results. Such approach is here adopted for comparing several 3D architectures, different for camera specifications and geometrical positioning, with the aims of quantifying their performance from a metrological point of view and of identifying the configuration able to optimize the result reliability.

1 Introduction

Systems based on Face Recognition (FR) have roused interest for the last decade, and in particular the field of security research is going to spread the employment of such technology in many contexts [1, 2]. In a face recognition system, an unknown subject is recognized by comparing some quantities, extracted from an image of an unknown subject to be classified, with the corresponding ones extracted from a preexisting database [1, 3, 4]. Afterwards, a maximum likelihood approach is generally adopted to classify and recognize the unknown subject.

G. Betta · D. Capriglione · M. Corvino
DIEI, University of Cassino and of Southern Lazio, Cassino, FR, Italy
e-mail: betta@unicas.it

M. Gasparetto · E. Zappa
Department of Mechanical Engineering, Politecnico di Milano, Milano, Italy
e-mail: emanuele.zappa@polimi.it

C. Liguori (✉) · A. Paolillo
DIIn, University of Salerno, Fisciano, SA, Italy
e-mail: tliguori@unisa.it

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_22

However, the quantities measured for the classification are generally affected by uncertainty which propagates through all the processing stages, thus generating a risk in accepting the final decision [5, 6]. Some techniques are proposed to reduce this risk. They take into account the measurement uncertainty in the comparison stage [7–9], or adopt fuzzy logic [10–12], belief function theory [13], Bayesian networks.

These approaches try to handle the uncertainty by means of suitable representations, but whatever solution is adopted, the uncertainty is structurally included in the model and consequently the input uncertainty is considered always the same. Therefore, this solution is not optimal since each new case can be characterized by a different uncertainty, and consequently a fixed value of the considered input uncertainty can give rise to its overestimation or underestimation. In other words, these approaches do not consider the propagation of the measurement uncertainty according to ISO-GUM [14] that can be derived starting from input measured data. The state of art about the characterization of recognition system is quite rich [15–17], but the research is mainly focused on the evaluation of recognition reliability indexes that express the probability of a false positive and of a false negative. Vice versa, proper studies are still required for the analysis of the existing relationship between the overall uncertainty of the final results and the input influence quantities.

In this field, the authors tackled the problem of the metrological characterization of a face recognition classification system [7, 8]. It has been shown that the performance of these kinds of systems depends on several aspects from image acquisition to the classification procedure through the biometrical algorithm. In [18–20] the authors have proposed an original method for the evaluation of the measurement uncertainty in face recognition systems and for exploiting such a quantity in the classification phase. In particular, starting from the analysis of the acquired image, the proposed model on-line estimates the measurement uncertainty, that is used to provide the confidence level of system outputs. In this way, the proposed method allows improving the classification performance with respect to a traditional approach [21]. As a case study, it has been applied to a 2D face classification algorithm based on Linear Discrimination Analysis and to a 3D system employing the Active Appearance Models (AAM) algorithm for detecting the 2D significant points (landmarks), triangulating and extracting the features of interest for the classification.

As for 3D systems, from a designer point of view, particular aspects have to be generally considered, since several architectures have been proposed in order to solve the problem of the 3D reconstruction, using two or more cameras. From a general point of view, the 3D reconstruction problem is mathematically determined only if enough information about the sensing sub-system is available in terms of optical/intrinsic parameters (e.g. focal length, lens distortion, etc.) and geometrical/extrinsic parameters (e.g. the position and orientation of the cameras in space). Different approaches have been presented in the literature, for optimizing the performance of vision systems [22]. The variety of the proposals reported in literature does not help a system designer to choose an architecture for a specific problem. Although efforts have been made in order to set up wide-ranging strategies for the design planning of sensors and illuminators for computer vision applications [23], it is still difficult to give suggestions unless specific cases are considered. In this framework the authors propose to extend the use of the classification procedure developed in previous papers [20, 21] for comparing, from a metrological point of view, possible architectures of a 3D system for face recognition. To this aim a suitable experimental setup involving several couples of cameras characterized by different features and geometrical positioning, has been developed.

In the following, after a brief description of the classification procedure and of the experimental setup, the comparison of the metrological performance of different experimental configurations is shown.

2 The Used Biometric Approach

2.1 Biometric Algorithm

The facial images obtained by means of a couple of stereoscopic cameras are analyzed with the AAM-API software to automatically detect the 2D coordinates of a set of landmarks. A 3D mask is obtained triangulating the two 2D masks (see Fig. 1). The 3D mask of a person to be identified is finally compared with each of the 3D masks included in a database and a score is computed for each comparison.

To automatically detect the facial features in the images, the Active Appearance Model (AAM) technique is used. The AAM algorithm is based on the combined use of two different statistical models: the Shape Model and the Appearance Model [16]. 58 landmarks demarcating seven areas of the face have been used in this work: jaw, mouth, nose, eyes, and eyebrows, as shown in Fig. 1b, where an example of 3D mask is also shown. The choice of the 58 points is made according to the processing method carried out in previous studies [17, 18]. The 3D coordinates of each facial feature have been estimated by means of optical triangulation made using the Zhang's method.

The output of the biometric algorithm is a Score for each subject in the database; the Score basically represents the sum of squared discrepancy between the 3D coordinates of the mask to be recognized and the corresponding coordinates of each mask in the database. Prior to that evaluation of point to point distances, a roto-translation is computed in order for the coordinate frame of one mask to be moved onto the coordinate frame of the other mask with a rigid motion. The rototranslation allows to compensate for differences in position and orientation of the subject with respect to the stereoscopic system in acquisitions.

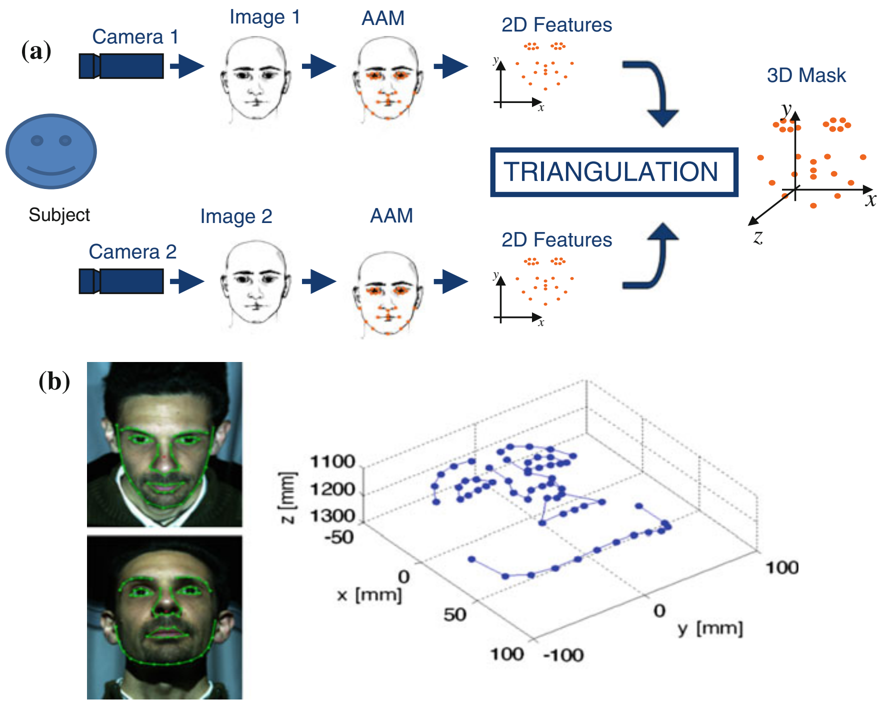


Fig. 1 **a** The schematic of biometric algorithm. **b** Stereoscopic images with the detected features and the obtained 3D mask

2.2 *Discussion of Bias and Uncertainty Sources*

The biometric algorithm can be schematized in four main steps, each one introducing uncertainty and/or bias, able to affect the whole system performance:

1. 2D image acquisition,
2. 2D image processing by means of AAM algorithm,
3. triangulation,
4. roto-translation.

The image acquisition is the prevalent measurement uncertainty cause, since the acquisition conditions like non-optimal focus, luminance or the presence of motion determine an uncertainty on the image itself. The image uncertainty propagates in the image processing algorithm, and generally further aspects can affect the final performance.

With reference to the considered biometric algorithm, an AAM approach was used, then the uncertainty in the building of the Shape Model and the Appearance Model [18] determine a better or worse feature localization. The triangulation is realized by means of the method proposed by Zhang and a residual error is present due mainly to the non-ideality of the calibration phase. Finally, the roto-translation is made by means of an iterative procedure and the residual error is mainly linked to the noise on the 3D features of the masks. The effects of the accuracy of AAM, triangulation, and roto-translation can be modeled by means of residual systematic Score greater than zero also for the correct class, whilst the measurement uncertainty introduces variability on the Score of the correct class that strictly depends on the acquisition condition.

In the following, all the systematic effects present are directly included in the measured Score, whilst only the measurement uncertainty (due to the acquisition process) is considered.

3 Influence of the Measurement Uncertainty

The uncertainty on the Score, u_S, depends mainly on the uncertainty of the 2D coordinates of the face features that in their turns depend on the characteristics of the processed images. As evidenced in previous papers [15–18] the main quantities of influence in face recognition problems can be related to luminance, defocus and motion blur. In order to quantify this uncertainty according to the ISO-GUM, a simple model was used in associating each quantity of influence to u_S. Denoting with u_i the contribute due to the ith quantity of influence on the Score uncertainty, we have posed:

$$u_i = f_i(\Delta_i) \tag{1}$$

where Δ_i is the value of the ith quantity of influence.

All of the quantities of influence are considered uncorrelated with the other ones, then the combined uncertainty on the score is evaluated as:

$$u_s = \sqrt{\sum_{i=1}^{N} u_i^2} \tag{2}$$

where N is the number of the considered quantities of influence. Applying these models, the uncertainty of the Scores is evaluated.

The modelled influence quantities are the luminosity, the defocus and the motion blur.

4 The Classification Procedure

With the aim to manage the measurement uncertainty for obtain a more reliable classification a new scheme was proposed in [17, 18] (Fig. 2):

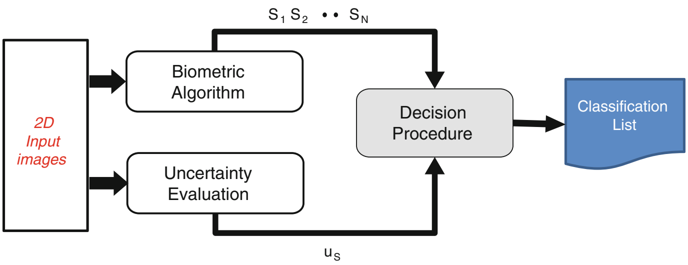


Fig. 2 The proposed classification scheme

– At first, the two 2D images of the unknown subject are processed by the biometric algorithm and the Score, S_i (i = 1,…,N), of each of the N subjects of the database is evaluated.
– The two images are processed also by the *uncertainty estimation* algorithm that determines the uncertainty on the obtainable scores (u_S).
– Finally, starting from all the evaluated scores and the estimated uncertainty, the decision procedure provides a classification list where all the subjects in which the unknown can be recognized are reported together with their Confidence Level, CL.

The goal of the decision procedure is the creation of the classification list composed by a record containing all the possible classes, each one characterized by a confidence level (CL). Thanks to this approach the decision is made on the basis of the estimated CL and consequently taking into account the actual acquisition conditions.

In Fig. 3, a diagram of the main steps forming the decision procedure is reported: (i) at first, the probability, P_j, that an input subject is the *j*-th subject of the database is evaluated; then, (ii) on the basis of the obtained probabilities a classification list is created with a selection of the probable subjects; finally, (iii) the confidence level for each subject (class) in the list is evaluated. In particular:

(i) P_j represents the probability that the unknown subject is the *j*-th subject of the database (i.e. the *j*-th class). Considering the score as a random variable, P_j represents the probability that the Score of the *j*-th class is equal to zero given

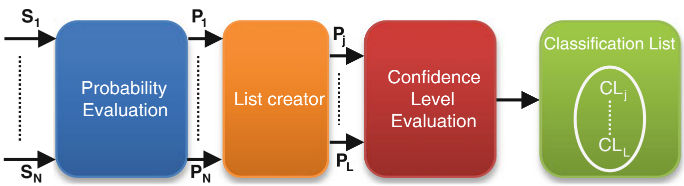


Fig. 3 Simplified block diagram of the decision procedure

a measured value S_j^*. This probability is evaluated by the score probability density function, *p*(*s*), with the following relationship:

$$P_j = \mathrm{P}(\mathrm{S_j} = 0|\mathrm{S}_j^*) = \begin{cases} 1 & \text{if } S_j^* \leq th \\ \int_{S_j^*}^{zc} \mathrm{p(s - th)ds} & \text{if } S_j^* > th \end{cases} \tag{3}$$

This function is applied to all the considered classes of the database, evaluating the probability of each one of them.

The Score of the correct class for each subject of the database are used in order to define *p*(*s*).

(ii) Starting from the so calculated probabilities, the classification list is composed of all the classes which show a probability greater than a second threshold, ***TH***. The value of TH was posed equal to 0.50.

(iii) The probabilities of all the classes included in the classification list are used in evaluating a normalization factor, *K,* defined as follows:

$$K = \sum_{\mathrm{j}} \mathrm{P}_j \text{ for all } j \text{ with } P_j > TH. \tag{4}$$

Then, the confidence level (CL) of each class belonging to the list is evaluated as the probability P_j divided by *K*:

$$\mathrm{CL_j} = \frac{\mathrm{p_j}}{\mathrm{K}} \tag{5}$$

5 Analyzed Architectures

In order to evaluate the performance of the classification procedure a multi-camera vision system has been developed, allowing to simultaneously acquiring images of the face from different points of view. The system is provided of 6 cameras; all the cameras are driven by a unique trigger to perform synchronous acquisitions of the same individual from different points of view. In this way, data referred to images of the face in frontal and rotated positions are perfectly comparable because they are acquired in the same instant and therefore the differences among them are due only to the camera position, while no change in facial expression occurs. The image acquisition system is shown in Fig. 4. Cameras are arranged in pairs: each pair of cameras is vertically aligned and placed at 0°, 5° and 10° referring to the axis of the face. With this arrangement, all the face characteristics are visible from all the cameras. Cameras 1 and 2 are AVT—Pike F-145B (Sony 2/3″ 1388 × 1038 pixel CCD progressive scan sensor) cameras, while cameras 3 to 6 are AVT Marlin

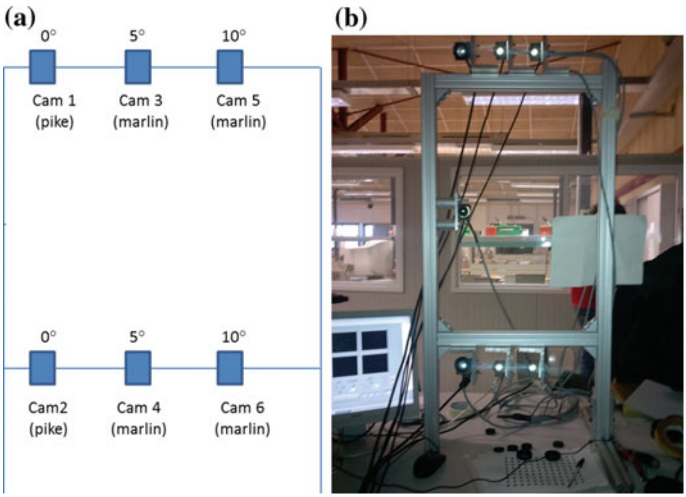


Fig. 4 The proposed multi-vision system **a** development scheme, and **b** picture

F-131B (equipped with 2/3″ 1280 × 1024 pixel CMOS sensor) cameras. All cameras are equipped with 25 mm focal length lenses and connected to a computer via three Firewire IEEE1394 acquisition cards (for cameras 1–6). The number of pixels of the cameras are different, but the different pixel size is such that it compensates for this difference, giving for each camera exactly the same millimeter resolution. The Stereoscopic vision can be obtained using any couple of camera, however coupling one of the upper cameras with the corresponding bottom one allows minimizing the nose undercut issue. An angle of 45° between upper and lower cameras optical axes has been used, which results to be the best compromise [19] because it maximizes the angle (and therefore allows to reduce the uncertainty in the distance estimation) and avoids occlusion phenomenon in the two views. The field of view of the system is approximately 300 × 400 mm, large enough to contain the subject's face placed 1000 mm away from the system. The illumination source is designed in order to obtain a uniform illumination in all the views. The stereoscopic system has been calibrated using the Zhang's method.

The acquisition system described above has been used to collect several facial images, in order to test the reliability of the recognition judgment as a function of the acquisition angle and considering natural variability of the expression and head position. The database includes facial images of 117 volunteers acquired as shown in Fig. 5. It should be underlined that, in the image acquisition, the environmental conditions (in particular the lighting) are controlled in order to ensure that the images are acquired without appreciable shadows, light reflections or motion blur; moreover the facial images are acquired with the person in frontal position and with the best possible focusing.

Fig. 5 Example of images collected in the database

6 Experimental Results

In this section, the results related to three different configurations of the analyzed architecture are reported. In particular, the couples of cameras 1–2, 3–4 and 5–6 have been considered. For each couple of cameras the uncertainty on the score is evaluated, then the metrological and classification performance of the related configurations are evaluated and compared.

6.1 Uncertainty on the Score

6.1.1 Luminosity

The average grey levels of the original images, considered with the optimal luminosity, were modified of ±3, ±10, ±30 (codes for an 8 bit representation), with respect to the reference image. In Fig. 6a the results are reported for the three configurations. As you can see, the uncertainty related to cameras 1–2 is significantly lower than the one achieved for the cameras 3–4 and 5–6. For each couple, the uncertainty is almost constant for each kind of luminosity variation but since the values are all very low a constant overestimated value of 0.05 can be considered for all the configurations.

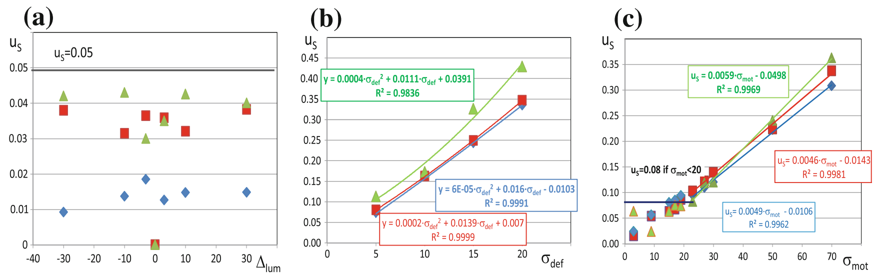


Fig. 6 u_S versus the quantities of influence **a** luminosity, **b** defocus, and **c** motion blur—(*blue curve* cameras *1–2*, *red curve* cameras *3–4*, *green curve* cameras *5–6*)

6.1.2 Defocus

Following the same procedure used for the evaluation of the luminosity influence, the effects of the lens defocus were analyzed. Also in this case different levels of defocus have been considered. The modified images are obtained by filtering the image with a Gaussian linear kernel filter characterized by a standard deviation of 3 to 20 pixels with respect to the reference image. Figure 6b shows the obtained results. A second order polynomial model well fits the observed data for all the configurations. In this case the camera couples 1–2 and 3–4 show a very similar behavior and a little worsening is observed for the couple 5–6.

6.1.3 Motion Blur

Generally, the motion blur is due to the movement between the subject and the camera. In particular, a motion between the subject and the camera pair along the horizontal direction has been considered. This effect has been simulated by filtering the image with a directional linear low-pass filter characterized by a given standard deviation with respect to the reference image. Figure 6c shows the obtained results. A linear model can be used for all the configurations, and the uncertainty values are similar.

The output of the biometrical algorithm, of the so created datasets (see Fig. 7 for the distributions of the datasets uncertainty), is further processed by the classification procedure. In order to quantify the overall classification performance we have considered the following indexes:

- *Correct classification* (*CC*): percentage of cases in which either the classification list includes only the right class with CL = 1 or the Classification list has more subjects where the right class has the highest value.
- *Long classification list* (*LCL*): percentage of cases in which the classification list includes the right class but there are some other classes with the same *CL.*

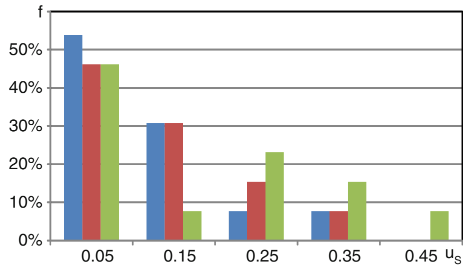


Fig. 7 Comparison of u_S histograms (*blue bar* is for cameras *1–2*, *red curve* is for cameras *3–4*, *green curve* is for cameras *5–6*)

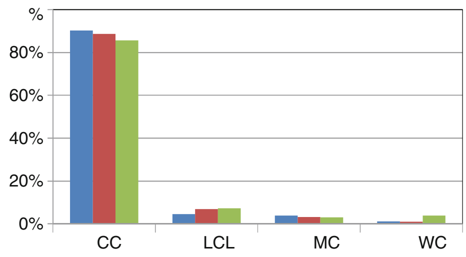


Fig. 8 Performance indexes of the three architectures (*blue* cameras *1–2*, *red* cameras *3–4*, *green* cameras *5–6*)

– *Missed classification* (*MC*): percentage of cases in which the subject is in the training database but the classification list is empty.
– *Wrong classification* (*WC*): percentage of cases in which the classification list has more subjects without the right class or the right class has not the highest value of CL.

Figure 8 reports the comparison of such indexes for the three considered architectures. It shows that whatever the couple of cameras, the correct decision is reached in more than the 85 % of cases. These good results are mainly imputable to the proposed strategy for the decision procedure which allows a good correct decision rate to be achieved even in presence of uncertainty on the acquired images [18].

Moreover, Fig. 8 shows that the couple of cameras 1–2 shows the best overall classification performance while the couple of cameras 5–6 is characterized by the worst one. In particular, comparing the results achieved by couples 1–2 and 3–4 we obtain a weak classification performance decay which should be mainly imputable to a little increase of the distance between subject and the cameras and of the horizontal angle with respect to the vertical axis of the face (see Fig. 5) [15].

Similar considerations could be made for explain the performance decay achieved for the pair 5–6 with respect to the pair 3–4.

7 Conclusions

The paper has compared the metrological performance of three different architectures for face detection systems based on 3D features. The study has been conducted by considering a popular algorithm based on 3D features (the AAM) and main causes of uncertainty generally affecting the performance of face recognition algorithms.

The method proposed by the authors in previous papers has been here adopted to evaluate the uncertainty on the score and suitable figures of merit associated to the classification performance. The achieved results quantify the performance of the considered system configurations highlighting that the distance between the subject (to be recognized) and the pair of cameras, as well as the horizontal angle with respect to the vertical axis of the face, generally affect the metrological performance of the system with a direct impact on the reliability of the final decision in the classification stage. The quantification of such kind information could be very useful also for system designers which, in practical application, have often to select the best tradeoff between cameras arrangement and system performance. The followed approach can be used also for exploiting other configurations including more than two cameras.

References

1. Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Jaesik, M., Worek, W.: Overview of the face recognition grand challenge. In Proceedings of IEEE Computer Vision and Pattern Recognition Conference, pp. 947–954 (2005)
2. Zhao, W., Chellapa, R., Phillips, P.J., Rosefeld, A.: Face recognition: a literature survey. J. ACM Comput. Surv., 399–458 (2003)
3. Gross, R., Shi, J., Cohn, J.: Quo vadis face recognition? Report of Robotics Institute of Carnegie Mellon University (2001)
4. Hea, Y., Zhao, L., Zou, C.: Face recognition using common faces method. Pattern Recogn., 2218–2222 (2006)
5. Betta, G., Capriglione, D., Crenna, F., Gasparetto, M., Liguori, C., Paolillo, A., Rossi, G.B., Zappa, E.: Face-based recognition techniques: proposals for the metrological characterization of global and feature-based approaches. Meas. Sci. Technol. **22**(12) (2011)
6. Jain, A.K., Duin, R.P.W., Mao, J.: Statistical pattern recognition: a review. IEEE Trans. Pattern Anal. Mach. Intell. **22**(1), 4–37 (2000)
7. Gianfelici, F., Turchetti, C., Crippa, P.: A non-probabilistic recognizer of stochastic signals based on KLT. Sig. Process. **89**(4), 422–437 (2009)
8. Zappa, E., Testa, R., Barbesta, M., Gasparetto, M.: Uncertainty of 3D facial features measurements and its effects on personal identification. Measurement (2014)
9. Crenna, F., Zappa, E., Bovio, L., Testa, R., Gasparetto, M., Rossi, G.B.: Implementation of perceptual aspects in a face recognition algorithm. J. Phys. Conf. Ser. **459**, 012031 (2013). doi:10.1088/1742-6596/459/1/012031
10. Zhang, G.: Face recognition based on fuzzy linear discriminant analysis. IERI Procedia **2**, 873–879 (2012)

11. Aisjah, A.S., Arifin, S.: Maritime weather prediction using fuzzy logic in Java sea for shipping feasibility. Int. J. Artif Intell. **10**(S13), 112–122 (2013)
12. Zhang, D., Wang, Q-G., Yu, L., Song, H.: Fuzzy-model based fault detection for a class of nonlinear systems with networked measurements. IEEE Trans. Instrum. Meas. **62**(12), 3148–3159 (2013)
13. Yager, R.R.: On the fusion of imprecise uncertainty measures using belief structures. Inf. Sci. **181**(15), 3199–3209 (2011)
14. JCGM 100.: Evaluation of measurement data—guide to the expression of uncertainty in measurement. ISO-GUM (2008)
15. Betta, G., Capriglione, D., Corvino, M., Liguori, C., Paolillo, A.: Face based recognition algorithms: a first step toward a metrological characterization. IEEE Trans. Instrum. Meas. **62** (5), 1008–1016, art. no. 6493520 (2013)
16. Betta, G., Capriglione, D., Corvino, M., Liguori, C., Paolillo, A.: Face based recognition algorithms: the use of uncertainty in the classification. IEEE Proc. I^2MTC13, 1098–1103 (2013)
17. Betta et alii, G.: Managing the uncertainty for face classification with 3D features. IEEE Proc. I^2MTC14, 412–417 (2014)
18. Betta, G., Capriglione, D., Corvino, M., Liguori, C., Paolillo, A.: A proposal for the management of the measurement uncertainty in classification and recognition problems. IEEE Trans. Instrum Meas. (2014). doi:10.1109/TIM.2014.2347218
19. Cootes, T.F., Gareth, J.E., Christopher, J.T.: Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. **23**(6), 681–685 (2001)
20. Zappa, E., Mazzoleni, P.: Reliability of personal identification base on optical 3D measurement of a few facial landmarks. Proc. Comput. Sci., 2769–2777 (2010)
21. Zappa, E., Mazzoleni, P., Hai, Y.: Stereoscopy based 3D face recognition system. Procedia Comput. Sci. **1**(1), 2521–2528 (2010)
22. Liguori, C., Paolillo, A., Pietrosanto, A.: A discussion about stereo vision techniques for industrial image-based measurement systems. In: Proceedings of the 20th IEEE Instrumentation and Measurement Technology Conference IMTC '03, vol. 1, pp. 77–82 (2003)
23. Di Leo, G., Liguori, C., Pietrosanto, A., Paciello, V., Paolillo, A.: Illumination design in vision-based measurement systems. In: Proceedings of the IEEE International Instrumentation and Measurement Technology Conference (I2MTC), pp. 1491–1495 (2013)

Reviewing and Classification of Software Model Checking Tools

Mogeeb A.A. Mosleh, Musaed A. Alhussein, Mohd Sapiyan Baba, Sorayya Malek and Sitihafizah ab Hamid

Abstract In this study, we provide historical accounts with an overview of essential research on model-checking development tools. This study has two main objectives; first, it is intended to investigate whether model checking still an active area; second, to classify existing model-checking tools by providing an illustration of each dimension scope, an analysis of similarities and differences among them, and a prediction of the future direction of typical model-checking tools. We found that existing model-checking tools show significant effects in automated system testing and verification. We also found that system testing and verification are still active areas of research. Current model-checking tools work efficiently on limited environment, and a lot of work need to perform for verifying the functional and nonfunctional attributes of complex systems. Despite the limitations of existing model-checking tools, universal model-checking tools can probably be developed if a good framework is established to fulfill the requirements of fully automated tools.

Keywords Model checkers · Software testing · Verification · Model-checking tools · Automated software testing · Classification model checking tools

M.A.A. Mosleh (✉) · M.S. Baba · S. Malek · S. ab Hamid
Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia
e-mail: MogeebMosleh@um.edu.my

M.S. Baba
e-mail: pian@um.edu.my

S. Malek
e-mail: Sorayya@um.edu.my

S. ab Hamid
e-mail: sitihafizah@um.edu.my

M.A. Alhussein
Faculty of Computer and Information Science, King Saud University, Riyadh 51178, Saudia Arabia
e-mail: musaed@ksu.edu.sa

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_23

1 Introduction

Software quality is an important element in evaluating the successful development of a software system, in which safety and freedom from errors are strict requirements. However, because of the complex nature of software systems, achieving of these requirements satisfactorily is relatively difficult. Research on manual testing strategies has been established for this purpose; however, such research remains time consuming, expensive, and difficult to achieve completely. An automated testing process is becoming necessary because manual testing tasks require much effort and are subject to human errors. Model checking is one of automatic testing strategy methods that have been proposed to generate test cases for software verification purposes [1]. Model checking is a model based testing technique involving in creation abstract model to generate system test cases automatically. A model checker is a tool designed primarily to verify system correctness. It uses two input arguments, the automaton-based model and a temporal logic property, to determine whether the model violates the system property. The result output of these tools is a counterexample. A counterexample provides feedback about error types and paths. In this research, we investigated the model checking research area to highlight the current progress, and issues. Then, we tried to categorize current model checking tools to help in shaping a future of this area.

2 Historical Overview

Early model checking research has focused on manual reasoning of program verification, as well as on the development of semantics and logic reasoning to verify programs as logical objects. As software systems size increased dramatically, the manual testing process became too inefficient to be trusted [2]. This resulted in a trend toward automation of testing tools. Research on model-checking extended the scope of automated testing to handle the programs, and properties. Pioneering work in software model checking using temporal logic formulas was established by Clarke and Emerson in the early 1980s [3].

A decade of research on using model checkers for testing lead to various techniques with diverse model constructions, different forms for representing models, and several approaches in generating test cases. Consequently, research on software model checking has increased, exploring new dimensions such as designing model-checking algorithms [4], developing model-checking tools, and applying model checking for static analysis and system testing [5]. Software model checking development has been impacted by three distinct areas: (1) program logic, (2) associated decision procedures, (3) Promela language [6]. Second, the development of automated model-checking approaches for solving state exploration issues and other existing problems [7]. The last, the development of compiler analysis and formal verification software using abstract interpretation to connect between the logical and the actual representations of finite state space [8].

Research on the application of model checking tools has also been extended in various areas such as in generating system test cases [9], verifying and validating software [10], security applications [11], Web testing [12], compiler analyses [13], networking [14], embedded systems [15], memory management [16], distributed systems [17], and programming language [18]. Recent research has focused more on delivering robust model-checker tools that can be used in the automatic testing process for verifying the safety and liveness of a computer system [19], which can be applied in the static analysis and test generation of programming languages [20, 21].

2.1 Current Model Checking Tools

Several model-checking tools developed for automated verification and software application testing. There are two types of model checking tools, enumerative based, and symbolic based.

2.2 Enumerative-Based Model-Checking Tools

Enumerative tools were designed mainly for use in execution-based model checkers led by VeriSoft. This approach implements enumerative state space exploration using the runtime system of a programming language. State search and stateless search are the two main approaches used in implementing the execution-based techniques. It has been used to effectively adopt executions for finding bugs during runtime.

VeriSoft is an execution-based stateless tool. Several UNIX processes are used as input to communicate with message queues and shared variables using the VeriSoft scheduler. The scheduler is designed to find the interleaving processes by selecting executed process at each point [22].

Java Pathfinder is designed to implement a systematic search over diverse threads for Java programs [23]. This tool has some advantages, such as the ability to use the Java virtual machine to store visited states, the ability to use different algorithms for a heuristic search without the limitations of a stateless search, the ability to use other model-checking techniques including symbolic and abstraction approaches, and available as open-source tools.

C model checker (CMC) is an execution-based tool designed to explore different executions for C programs by controlling the scheduler at operating system level. CMC stores the visited states in a hash table to identify different states of allocated structures [24].

Mace model checker (MaceMC) is an execution-based tool implemented as a domain-specific language of C++ for distributed systems. MaceMC has two techniques in solving the exploration issue. First, it exploits higher-level constructs to explore event transitions of shared operations. Second, it combines an exhaustive random search to identify all possible violations [25].

CHESS is designed for windows multithreaded programs. It intercepts system operations to explore the space of the scheduler. CHESS employs a new search technique called iterative context bounding to explore executions counter at k context-switches. It has been used widely in Microsoft testing framework [26].

Enumerative model checking was initially applied in testing and simulation, such as in testing network protocols as comprehensive approaches to ensure correctness. Later on, enumerative tool was used to generalize techniques in temporal logic specifications [27]. Several enumerative model checking of finite state programs developed such as Murphi, and Spin, which both show significant effects, particularly in domain verification [24]. However, state explosion is the main problem of enumerative tools, wherein the state space increases rapidly to be larger than the program properties itself. Thus, research in the context of enumerative tool has focused more in controlling the state explosion problem [28]. Two techniques have been used in handling the state explosion issue, reduction-based techniques and compositional techniques [29, 30].

Reduction-based techniques compute the equivalence relations of program behavior and explore one candidate state from each class as implemented in partial-order reduction [29], symmetry reduction [7], and minimization-based techniques [31]. Compositional techniques reduce safety verification problems with the original program and prove program properties to realize the safety of the original source. Research has used heuristics search to implement the compositional techniques, including iterative deepening and best-first search or A* search [32]. However, when the search space becomes too large, other techniques, such as bitstate hashing, are applied to explore each small portion of the system state space.

2.3 Symbolic-Based Model-Checking Tools

Symbolic model or implicit model checkers represent a set of states logically, such as binary decision diagrams (BDDs). Symbolic algorithms manipulate the logical representation of sates rather than the individual state. Thus, symbolic tools able to control larger state space. Research has provided an effective symbolic representation to improve the performance of constraint solvers such as binary decision, satisfiability solvers [33], and more recently, first-order theories [29], where BDDs were used to represent symbolic model checkers in various tools, such as the symbolic model verifier (SMV), to enhance exploration of a large state space [34].

Bounded model checking (BMC) is one of the common model checking approaches; it uses a symbolic representation to control the flow graph of state spaces. BMC techniques allow searching of overall program using backtracking search algorithms within a constrained solver [35]. Several tools were developed using BMC techniques [36], such as C-bounded, F-SOFT, and Calysto [37]. Other types of BMC tools were designed to generate model constraints using first-order theory and decision procedures. Abstract model checking was designed recently to improve the efficiency of symbolic algorithms. Abstract model checking employs semantic abstract of programs to perform reachability analysis on a specific abstract domain [38].

The SLAM model checker was the first symbolic representation tools, which used to represent C program. The C2BP tool in SLAM was used to implement abstraction of C programs, wherein the input to C2BP is the C program and the output is the Boolean program. BEBOP is a tool developed to implement symbolic model for the output of C2BP [9]. A SLAM tool which was developed mainly for microsoft device drive verification has become a commercial product, and is now called the Static Driver Verifier [39]. The SLAM introduced numerous concepts in model checking research, such as predicate abstraction [40], and BDD-based model checking [34].

The BLAST model checker, also called the “lazy abstract,” optimizes the program abstract for refinement. Abstraction and refinement are generated by constructing the abstract model and refining it automatically. BLAST is designed to build model checkers for recursive programs [41].

The MAGIC model checker was designed mainly to apply verification module for concurrent messages passing in C programs. It allows users to specify some transition systems, and verifies that the set is generated as program properties. MAGIC combines numerous approaches to optimize the amount of predicates by saving multiple counter paths [38].

The F-SOFT model checker was developed to check standard runtime errors in C programs by predicating the program abstraction for refinement with abstract domains. F-SOFT is implemented as a symbolic tool and combines both BDDs and polyhedral techniques for iterative refinement [42].

3 Systematic Review Finding

According to previous overview, the model-checking executing process involves with main four steps: designing the model, identifying properties, running the model checker, and analyzing output results as described in more details below.

Design system model: A special numeric or symbolic language is used to implement model-checking tools to define the system as a model such as, a finite state machine with a specific description of system properties. A description model language, such as Promela, is a modeling language used mostly to convert system components into specific forms of finite state machines that will specify system behavior in terms of variables, initial values, environmental assumptions, and system requirement.

Define system properties: System properties are identified using logic notations such as linear temporal logic (LTL) or computation tree logic (CTL) [43]. Model-checking tools designed to support different kinds of property-specific language and modern tools are supported by both temporal logic properties (LTL, and CTL) [44].

Run model checker: a finite-state description of system operations are usually the inputs of model checker that can be analyzed with system properties that expected to be achieved, and which are expressed as formulas of temporal logic. A model-checker tool takes two inputs to interrupt the system into modeling languages, and is executed to check the validity of the defined properties against the system model.

Analyze Results: Finally, the results are analyzed to either confirm that the properties hold or violated. If the results for the current property of the system model are satisfactory, then the model checker continues to check the next property. However, if the property of the system model is violated, then a counterexample or error path is reported to refine the system model. In the latter case, the model checker provides a counterexample, which is considered as a report for violated properties or system bugs [25].

Model-checking tools generate a huge state during the model-checking process, particularly with a large and complex system which may cause an "out of memory" error. The state space explosion problem is a critical issue for most model-checking tools [28], wherein the possible state space of a large and complex system cannot be handled using available resources (processor and memory). Research shows that model checkers can handle 108 to 109 states with explicit state space and approximately 10,000 lines of codes [2]. In such cases, research has developed several sophisticated techniques employing efficient algorithms to reduce the generated state, such as system model reduction techniques [45].

To achieve first objective of this research which aimed to investigate current trend, and progress of model checking tools, a systematic review to search for existing research on model checking was initiated on the Thomson Reuters Web of Science on April 24, 2013. Extensive literature on model-checking research is available on the Thomson Reuters (formerly ISI) Web of Science because model-checking research began in the early 1980s. We found that approximately 72 papers present at least one model-checking tool and specific algorithms useful in model-checking research. Moreover, 67 papers present an application of model checking in the field of computers for testing and verifying software applications, whereas the remaining 26 present the concept of using model checking in checking system behaviors in other areas not relevant to computer science. Figure 1 represents journal and conference articles on model-checking research published

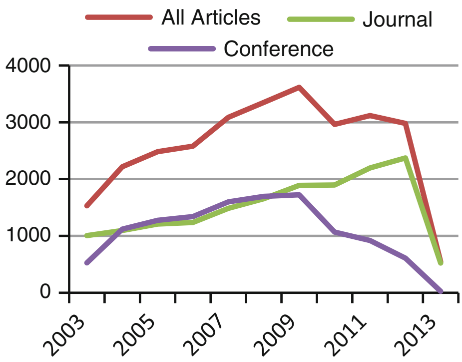


Fig. 1 Articles on model checking research between (2003–2013)

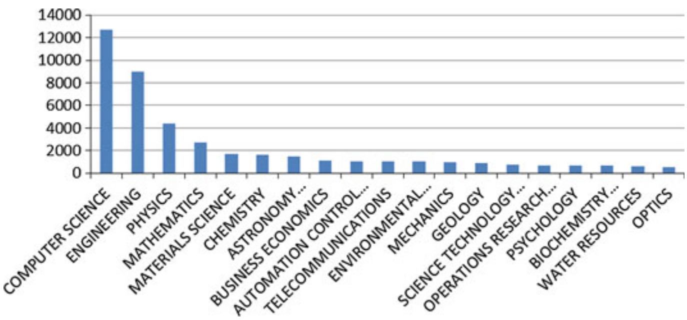


Fig. 2 Model checking research publications common area

between 2003 and 2013. The increasing similarity between conference and journal papers can be attributed to the circumstance that conference papers were indexed in the Thomson Reuters Web of Science database from the early stages of this research field. Model checking research has been increasingly applied after year 2000, when developed tools became more efficient in handling the system verification and testing issues, where the research before 2000 was focused mainly on theoretical approaches of model checking.

Model checking has been applied in different areas including computer science, engineering, physics, mathematics, materials science, chemistry, and so on. Figure 2 shows the top 19 areas where model checking has been applied. Research on model checking in the fields of computer science and engineering has resulted in the largest number of publications in terms of different checking-model techniques delivered and the resulting diversity of model-checking applications for system verification and validation as shown in Fig. 2. Finally, we used the systematic review to draw the historical publication chart of model-checking research publications per year (Fig. 3).

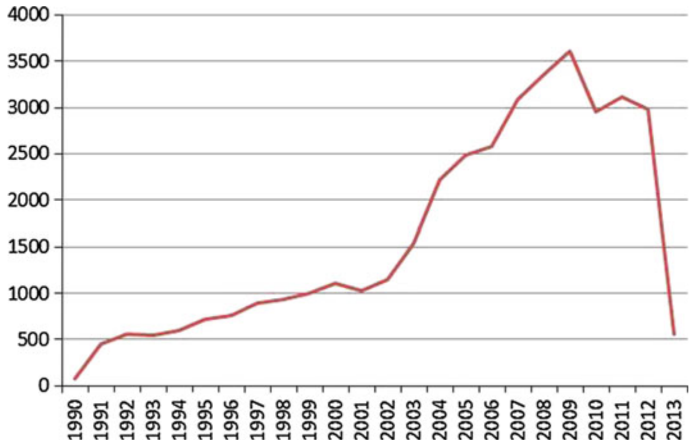


Fig. 3 Total publications on model-checking research (1990–2013)

3.1 Taxonomy of Model Checking Tools

According to the review, several model checking tools was developed by various approaches using different programming language for different purposes, platforms, and target system. Classifying model-checking tools into specific categories is difficult because an extensive variety of tools has been developed using various techniques. A good beginning for taxonomy model checking tools can be established based on their target design and testing strategy. Then, those two categorizes were further classified according to some specific dimensions.

3.1.1 Model Design

Model design involves methods and techniques used to build model-checking tools. The model design of model-checking tools is used to classify existing tools based on their techniques, which were developed to interrupt a system into specific models to build state space and system properties including, type of structures, notations, representation, modeling language, and characterization features.

I. Model Structures

Model-checking tools are designed with various types of finite states and temporal logic to express the requirements and properties of a system as logical formulas for verification. Most model-checking tools use either linear- or branching-time logic, or both.

Linear time tools: LTL, also known as linear-time temporal logic, is a modal of temporal logic with modalities that refer to time. In LTL tools, system states are represented as a sequence of operation.

Branch time tools: Branch time, also called CTL, is branching-time logic. Its model of time is structured like a tree in which the future is not determined. In CTL tools, system states are represented as a tree wherein the process can be run in parallel mode. A number of tools have been designed recently with hybrid temporal logic to combine both LTL and CTL approaches [43].

II. Model Notation

The notation dimension describes model-checking notation. Different modeling notations are used to model system behavior for generating test cases. We have identified the most common notation approaches, as adapted from [5].

State-based notation: This model uses a collection of variables to represent system states with certain operations to modify these variables. The operations are usually defined by a precondition and post-condition rather than by the coding.

Transition-based notation: This notation is used to describe the transition states of a system. It is similar to finite state machines which represent system components as nodes and arcs, wherein nodes represent system states and arcs represent system actions or operations. This type of notation usually costs more than computational types because of the addition of data variables, and because it represents system states and operations within the hierarchies of machines.

History-based notation: This notation is used to describe the ability of tracing the behavior of a system over time. Different notations, such as continuous, discrete, linear, or branching, are used. This notation also uses textual representation to specify the sequence of interactions among system components.

Functional notation: This notation type is used to describe a system as a collection of mathematical functions such as first order or higher order. More difficulties in representation are encountered when using this notation compared with other type.

Operational notation: This type is used to describe a system as a collection of executable processes running in parallel. Such notation is more suitable for representing distributed systems and communication protocols.

Stochastic notation: This notation is used to describe a system as a probabilistic model of events, input parameters, and environments. It is used to generate expected testing cases, such as the Markov chain model [46].

Data-flow notation: This notation is used to represent system data rather than control flow. It is used extensively to simulate the modeling of continuous systems [47].

III. Model Representation

Model-checking tools have two types of representations: explicit (enumerative) and implicit (symbolic). Additionally, Abstract representation has been used recently to represent the abstract of systems.

Explicit model checkers: Explicit model or enumerative model checking essentially traverses the graph of program states and transitions using various graph search techniques. The tool constructs a searchable representation of the design model and stores a representation of each visited state.

Implicit model checkers: Implicit model or symbolic model checkers use logical representations with a set of states, such as binary decision diagrams (BDDs), to describe the regions of a model state space.

IV. Model Design Language

Several model checkers tools developed to allow suitable representation for model checking in different types of systems. Most model-checking tools were built with their own modeling language, which can translate the system code into a specific modeling language for interpreting system requirements and properties into state machines and temporal logic, respectively. Several examples of these modeling languages are Promela, SMV, Spin, and visual modeling language (VML) [48, 49].

V. Model Design Characteristics

Model-checking tools also have varied characteristics which can be used to classify them, such as deterministic or non-deterministic, timed or untimed, and continuous or discrete.

3.1.2 Testing Strategies

Testing strategy is about the testing approaches used on model-checking tools. The testing strategy of model-checking tools is used to classify existing tools based on their testing criteria, techniques, target language, and case generation as described in the succeeding subsections.

I. Test Criteria

Test-selection criteria are common approaches used to control the generation of test cases. Model-checking tools can be classified according to their support for which kind of test-selection criteria. Test-selection criteria are employed to exploit the structure of model-checking tools, such as the nodes and arcs of a transition-based model or the conditional statements in a system model. Structural model coverage [50]; data coverage; requirement coverage; test case specifications; and random, stochastic, and fault-based are the common test criteria found in most model-checking tools.

II. Test Generation

The main concept behind the development of model checking is to generate test cases by formulating test criteria. Model-checking tools have used either manual or automatic approaches to generate system test cases. Test generation in model-checking tools is usually designed to generate test cases automatically using various methods such as random generation, graph search algorithms, symbolic execution, and theorem proving [51].

III. Target System and Languages

Model-checking tools were developed to work in different environments, such as UNIX and Windows platforms, and to be used for specific programming languages, including Java, C and C++, Web application, and embedded systems.

IV. Test Execution

Model-checking tools can also be categorized into two types, namely, online or offline, according to execution time. The methods used in model-checking tools for various levels of abstraction test selection include path, data-flow, event-flow, and stat-based testing. Online tools or real-time model-checking tools were designed to

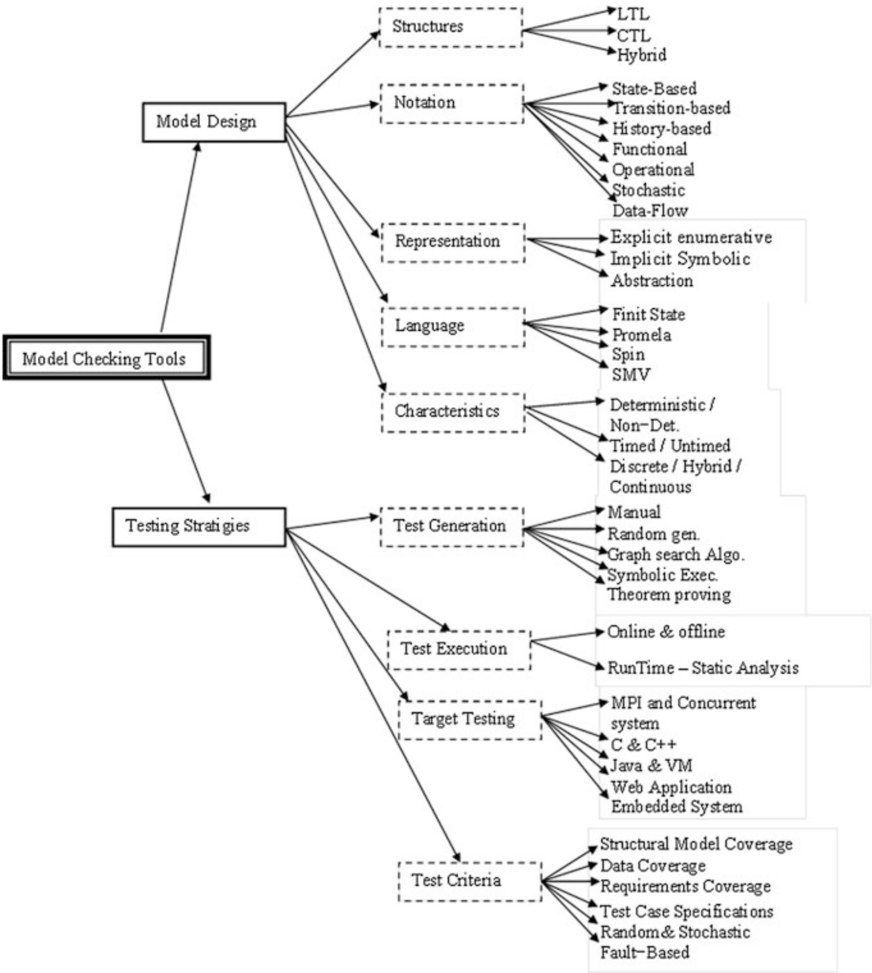


Fig. 4 Research classification of model-checking tools

generate test cases to predicate errors during system runtime, whereas offline-based tools were designed to analyze system structures and to perform system verification during compilation times. Model checking and static analysis are the two main approaches for automated verification capabilities to ensure limited correctness properties and to find bugs in software.

As a result, we categorized model-checking tools into two main categories based on their design target and testing strategy. Then, those two categorizes were further classified into nine dimensions, as shown in Fig. 4. Model design includes five dimensions, namely, structures, notations, representation, modeling language, and characterization. Testing strategy includes test generation, test execution, target system and programming language, and testing-criteria selection.

4 Discussion

Evidently, in the last three decades, extensive research has been performed in the area of model checking for testing and verifying computer systems. Early research was focused on testing systems using finite state machine theory. Afterwards, research gave special consideration to developing automated testing tools for different programming languages and to solving common issues associated with those tools. Research on model checking has moved from theoretical concepts to developing particular tools which have been applied extensively in the industry. Model checking has dealt with different algorithms, theories, and technology on developing model-based testing tools. Thus, system testing and verification are still active areas of research because of several issues that need to be resolved. Meanwhile, the huge number of available articles on model-checking theories and applications indicates the maturity of this field. Model-checking research has clearly received widespread attention across various scientific communities.

The main issue is state explosion problem, which limits the performance of system computation and storage spaces. Extensive research has been conducted to solve this problem to make the performance of developed tools acceptable; however, more research is required to increase the performance of model-checking tools for fully automated testing processes. In addition, studies on test case-generation tools do not cover testing-selection criteria of model-checking tools; an unsatisfactory performance can be exhibited by numerous developed tools. Several development tools still require special treatment. For example, nondeterministic models require a nonlinear counterexample to be generated.

Model-checking tools have been used primarily to verify system requirements based on system properties. They have been applied to verify the functional requirements of a system and the two non-functional requirements of safety and liveness. However, other nonfunctional requirements, such as a system-friendly interface, system security, and system integrability, have not been considered. Software model checking typically operates on the semantic level of a programming language by identifying finite state abstraction and by checking the model for system abstraction. We agree with previous research that without user interaction, software model checking approaches are still not mature enough to cope efficiently with real-life codes.

In addition, identifying the best model-checking tools is difficult because each tool has its own criteria, such as modeling language, environments, and programming languages. Most model-checking tool was designed for only one programming language, with the exception of PRISM, TAPAAL, and UPPAAL which were designed for both C and Java. Furthermore, most existing tools generate a counterexample report, which is important for refinement. Several modeling languages were developed to represent system properties for verification. This variety in modeling languages is acquired because of different algorithms applied in solving state explosion and test generation issues. We also noticed that most current tools are free to be used and developed. Taxonomy of existing model checking tools will be helped researcher to investigate each dimension well, and to design the appropriate solution for constrain, limitation, and issues of each dimension. It can also help to minimize the scope problems of current tools, and create a research opportunity to fulfill gaps in each dimension.

5 Conclusion

Model-checking research has expanded in different areas to close the gap between the intent of the programmer and the actual source code of the developed system. However, current model-checking tools work efficiently on bounded system properties, and we are still far from verifying the functional and nonfunctional properties of complex software systems. Several issues concerning algorithms and scaling techniques of large and modern software systems exist. More research needs to be conducted to improve the performance of model-checking tools, such as improving the test case-generation process, the accuracy and feasibility of the results of the test case-generation process, and the quality issues for coverage-based testing techniques.

Universal model-checking tools will probably be developed if a good framework is established to fulfill the requirements of fully automated tools; however, no research has yet been conducted on this issue. In this study, we classified current model-checking tools by identifying the main components of the unifying conceptual framework for this field. Thus, taxonomy is used to explore the design and testing dimension of the development of model-checking tools.

Despite the limitations of existing model-checking tools, we believe that research in the field has made excellent progress in the development of verification and algorithmic techniques for model-checking tools. The development of efficient model-checking tools is expected from future research. Special attention should also be given to developing an automatic testing tool into agile software which will help increase the productivity of programmers by giving them more time to settle other issues in the software-development process.

Acknowledgments The authors wish to thank University of Malaya, Taiz University, and King Saud University for their support. We also would like to thank and appreciate the help and support from the Research Center and College of Computer Sciences at King Saud University in providing the necessary facilities to accomplish this work.

References

1. Baier, C., Katoen, J.-P.: Principles of Model Checking. MIT press, Cambridge (2008)
2. Chen, H., Dean, D., Wagner, D.: Model checking one million lines of C code. In: NDSS, pp. 171–185 (2004)
3. Clarke, E.M., Emerson, E.A., Sistla, A.P.: Automatic verification of finite-state concurrent systems using temporal logic specifications. ACM Trans. Program. Lang. Syst. (TOPLAS) **8**, 244–263 (1986)
4. Visser, W., Havelund, K., Brat, G., Park, S., Lerda, F.: Model checking programs. Autom. Softw. Eng. **10**, 203–232 (2003)
5. Jhala, R., Majumdar, R.: Software model checking. ACM Comput. Surv. **41** (2009)
6. Lee, D., Yannakakis, M.: Principles and methods of testing finite state machines-a survey. Proc. IEEE **84**, 1090–1123 (1996)
7. Emerson, E.A., Sistla, A.P.: Symmetry and model checking. Formal Methods Syst. Des. **9**, 105–131 (1996)
8. Vardi, M.Y., Wolper, P.: Reasoning about infinite computations. Inf. Comput. **115**, 1–37 (1994)
9. Ball, T., Rajamani, S.K.: Generating Abstract Explanations of Spurious Counterexamples in C programs. Technical Report MSR-TR-2002-09, Microsoft Research (2002)
10. Cimatti, A., Micheli, A., Narasamdya, I., Roveri, M.: Verifying system C: a software model checking approach. In: Formal Methods in Computer-Aided Design IEEE, pp. 51–59 (2010)
11. Schwarz, B., Chen, H., Wagner, D., Morrison, G., West, J., Lin, J., Tu, W.: Model checking an entire Linux distribution for security violations. In: 21st Annual Computer Security Applications Conference, pp. 10–22 (2005)
12. Huang, H., Tsai, W.-T., Paul, R.: Automated model checking and testing for composite web services. In: Eighth IEEE International Symposium on: Object-Oriented Real-Time Distributed Computing, pp. 300–307 (2005)
13. Merz, F., Falke, S., Sinz, C.: LLBMC: bounded model checking of C and C++ programs using a compiler IR. In: Verified Software: Theories, Tools, Experiments, pp. 146–161. Springer (2012)
14. Petridou, S., Basagiannis, S., Roumeliotis, M.: Survivability analysis using probabilistic model checking: a study on wireless sensor networks (2013)
15. Gadkari, A.A., Yeolekar, A., Suresh, J., Ramesh, S., Mohalik, S., Shashidhar, K.: AutoMOTGen: Automatic model oriented test generator for embedded control systems. In: Computer Aided Verification, pp. 204–208. Springer (2008)
16. Kim, K.H., Yavuz-Kahveci, T., Sanders, B.A.: JRF-E: using model checking to give advice on eliminating memory model-related bugs. Autom. Softw. Eng. **19**, 491–530 (2012)
17. Basu, A., Bensalem, S., Peled, D., Sifakis, J.: Priority scheduling of distributed systems based on model checking. Formal Methods Syst. Des. **39**, 229–245 (2011)
18. Clarke, E., Kroening, D., Lerda, F.: A tool for checking ANSI-C programs. Tools and algorithms for the construction and analysis of systems, pp. 168–176. Springer (2004)
19. Sistla, A.P., Gyuris, V., Emerson, E.A.: SMC: a symmetry-based model checker for verification of safety and liveness properties. ACM Trans. Softw. Eng. Methodol. (TOSEM) **9**, 133–166 (2000)
20. Ball, T., Rajamani, S.K.: The S LAM project: debugging system software via static analysis. In: ACM SIGPLAN Notices, pp. 1–3. ACM (2002)

21. Paradkar, A.: Case studies on fault detection effectiveness of model based test generation techniques. In: ACM SIGSOFT Software Engineering Notes, pp. 1–7. ACM (2005)
22. Godefroid, P.: Software model checking: the VeriSoft approach. Formal Methods Syst. Des. **26**, 77–101 (2005)
23. Havelund, K., Pressburger, T.: Model checking java programs using java pathfinder. Int. J. Softw. Tools Technol. Transf. **2**, 366–381 (2000)
24. Edelkamp, S., Leue, S., Lluch-Lafuente, A.: Directed explicit-state model checking in the validation of communication protocols. Int. J. Softw. Tools Technol. Transf. **5**, 247–267 (2004)
25. Killian, C., Anderson, J.W., Jhala, R., Vahdat, A.: Life, death, and the critical transition: finding liveness bugs in systems code. Networked Sys. Des. Imp., 243–256 (2007)
26. Zheng, Y., Zhou, J., Krause, P.: A model checking based test case generation framework forweb services. In: 4th International Conference on Information Technology, pp. 715–722 (2007)
27. Emerson, E.A.: Temporal and modal logic. In: Handbook of Theoretical Computer Science, Formal Models and Semantics (B), vol. B, pp. 995–1072 (1990)
28. Godefroid, P., van Leeuwen, J., Hartmanis, J., Goos, G., Wolper, P.: Partial-order Methods for the Verification of Concurrent Systems: An Approach to the State-explosion Problem. Springer, Heidelberg (1996)
29. Flanagan, C., Godefroid, P.: Dynamic partial-order reduction for model checking software. In: ACM Sigplan Notices, pp. 110–121. ACM (2005)
30. Liu, J., Basu, S., Lutz, R.R.: Compositional model checking of software product lines using variation point obligations. Autom. Softw. Eng. **18**, 39–76 (2011)
31. Bustan, D., Grumberg, O.: Simulation-based minimization. ACM Trans. Comput. Logic (TOCL) **4**, 181–206 (2003)
32. Korf, R.E.: Depth-first iterative-deepening: an optimal admissible tree search. Artif. Intell. **27**, 97–109 (1985)
33. Silva, J.P.M., Sakallah, K.A.: GRASP—a new search algorithm for satisfiability. In: Proceedings IEEE/ACM on Computer-Aided Design, pp. 220–227. IEEE Computer Society (1996)
34. Biere, A., Cimatti, A., Clarke, E.M., Fujita, M., Zhu, Y.: Symbolic model checking using SAT procedures instead of BDDs. In: Proceedings of the 36th Annual ACM/IEEE Design Automation Conference, pp. 317–320. ACM (1999)
35. Biere, A., Cimatti, A., Clarke, E.M., Strichman, O., Zhu, Y.: Bounded model checking. Adv. comput. **58**, 117–148 (2003)
36. Armando, A., Mantovani, J., Platania, L.: Bounded model checking of software using SMT solvers instead of SAT solvers. Int. J. Softw. Tools Technol. Transf. **11**, 69–83 (2009)
37. Babic, D., Hu, A.J.: Calysto. In: Software Engineering. ICSE'08. ACM/IEEE 30th International Conference on IEEE, pp. 211–220 (2008)
38. Clarke, E., Grumberg, O., Jha, S., Lu, Y., Veith, H.: Counterexample-guided abstraction refinement for symbolic model checking. J. ACM (JACM) **50**, 752–794 (2003)
39. Ball, T., Bounimova, E., Cook, B., Levin, V., Lichtenberg, J., McGarvey, C., Ondrusek, B., Rajamani, S.K., Ustuner, A.: Thorough static analysis of device drivers. ACM SIGOPS Operating Syst. Rev. **40**, 73–85 (2006)
40. Jain, H., Ivančić, F., Gupta, A., Shlyakhter, I., Wang, C.: Using statically computed invariants inside the predicate abstraction and refinement loop. In: Computer Aided Verification, pp. 137–151. Springer (2006)
41. Beyer, D., Henzinger, T.A., Jhala, R., Majumdar, R.: The software model checker Blast. Int. J. Softw. Tools Technol. Transf. **9**, 505–525 (2007)
42. Ivančić, F., Yang, Z., Ganai, M.K., Gupta, A., Shlyakhter, I., Ashar, P.: F-Soft: software verification platform. In: Computer Aided Verification, pp. 301–306. Springer (2005)
43. Heljanko, K., Niemelä, I.: Bounded LTL model checking with stable models. In: Logic Programming and Nonmonotonic Reasoning, pp. 200–212. Springer (2001)

44. Chutinan, A., Krogh, B.H.: Computational techniques for hybrid system verification. IEEE Trans. Autom. Control **48**, 64–75 (2003)
45. Zhu, L.-X.: Model checking of dimension-reduction type for regression. Statistica Sin. **13**, 283–296 (2003)
46. Baier, C., Haverkort, B., Hermanns, H., Katoen, J.-P.: Model-checking algorithms for continuous-time Markov chains. IEEE Trans. Softw. Eng. **29**, 524–541 (2003)
47. Fischer, J., Jhala, R., Majumdar, R.: Joining dataflow with predicates. In: ACM SIGSOFT Software Engineering Notes, pp. 227–236. ACM (2005)
48. Schmidt, A., Varró, D.: CheckVML: a tool for model checking visual modeling languages. In: «UML». Modeling Languages and Applications, pp. 92–95. Springer (2003)
49. Holzmann, G.J.: The model checker SPIN. IEEE Trans. Softw. Eng. **23**, 279–295 (1997)
50. Ntafos, S.C.: A comparison of some structural testing strategies. IEEE Trans. Softw. Eng. **14**, 868–874 (1988)
51. Reps, T., Horwitz, S., Sagiv, M.: Precise interprocedural dataflow analysis via graph reachability. In: Proceedings of the 22nd ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages, pp. 49–61. ACM (1995)

Comparative Analysis of GUI Reverse Engineering Techniques

Ibrahim Anka Salihu and Rosziati Ibrahim

Abstract With the increasing number of mobile devices, the demand for mobile applications is ever increasing. Mobile applications are recently moving to more business-critical areas, becoming more and more complex, hence making it difficult to understand their behavior. Reverse engineering has been embraced by the software engineering community to improve their ability to understand a given system quickly by creating models of a system. The aim of this paper is to investigate the state-of-art in reverse engineering of GUI applications. The focus is on mobile applications reverse engineering techniques based on the point of GUI models generation, where the generated models can be used for program comprehension and testing. Firstly, we performed an exhaustive literature review on GUI reverse engineering approaches for model generation, followed by an assessment of capabilities of the reverse engineering techniques/tool for mobile applications based on the approaches. A discussion is presented on the result of the comparative assessment. Based on the results, limitations of the techniques and approaches in GUI model generation were identified and the areas that require further improvements were identified.

1 Introduction

Reverse engineering (RE) is gaining more popularity from the research community as an approach for reconstructing original specifications or design from software. It is used as an act of analyzing a software system, either in whole or in part, to extract

I.A. Salihu (✉) · R. Ibrahim
Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat
Johor, Malaysia
e-mail: hi130015@siswa.uthm.edu.my

R. Ibrahim
e-mail: rosziati@uthm.edu.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_24

design and implementation information [1]. The practices of reverse engineering improve the ability to understand a given system quickly and efficiently [2]. RE started in software engineering as a result of the need for program understanding in order to improve maintenance and evolution of software systems.

Nowadays, Reverse engineering is used for various purposes other than software comprehension such as software testing, checking software reliability, software reusability, updating user interfaces, migration and porting user interfaces to new platforms [3, 4]. Most developers interact with Graphical User Interface (GUI) to understand software applications by creating mental models of the application [5]. Graphical User Interface is defined as a hierarchical, graphical front-end to a software system that accepts, as input user-generated and system-generated events, from a fixed set of events and produces deterministic graphical output [6]. Most of the recent software applications feature a graphical user interface front-end which serves as the main UI for interacting with the software. The study in [7] shows that about 60 % of the software code is made up of the GUIs.

Mobile applications are a subset of a more general class of event-driven systems and specifically event-driven GUI applications [8]. According to [9], the amount of GUI related code is higher in mobile applications than in traditional desktop applications. In view of this, GUI analysis can aid the reverse engineer of mobile apps. Several techniques and tools have been proposed to reverse engineer mobile applications. However, a fully automatic analysis of mobile apps GUI remains challenging [10]. There is inadequate coverage of applications by the reverse engineering techniques and the scalability that is associated with large amounts of data collected at run time [11]. This paper presents an assessment of GUI reverse engineering techniques. The aim is to assess the effectiveness of the techniques in generating quality models from mobile applications. The paper is organized as follows. Section 2 discussed GUI reverse engineering based on static and dynamic approaches. Section 3 provides a framework for the comparative assessment. In Sect. 4, results of the assessment of five mobile apps reverse engineering tools is presented and Sect. 5 presents the conclusion and future work.

2 Reverse Engineering Approaches

There are two approaches to reverse engineering GUI applications; the static approach and dynamics approach. The static approach performs an analysis on the application's source code or binary code without executing the application [12, 13]. On the other hand, dynamic approach extracts information from an application by executing and analyzing its external behavior [14, 15]. The software engineering community has embraced similar approaches to reverse engineer mobile applications for various purposes such as program comprehension, testing and security analysis. This section presents a review of reverse engineer techniques/tools based on these approaches.

2.1 Static Approach

The static approach is not widely used in reverse engineering GUI apps. Nonetheless, few researchers have explored the static approach to reverse engineer GUI applications. Reference [16] introduced a knowledge-based reverse engineering technique for migration of the user interface of a system. Using static analysis on the source code of user interface of an original system, the structure of the GUI can be reverse engineered, which can later be recompiled to a target system's GUI implementation by applying a rule-based inference.

The emergence of mobile devices has brought the need to present the same web content on multiple devices with different layout sizes. The authors in [17] used static approach for reverse engineering to build models representing the hierarchical structure of user interface elements contained in HTML pages and their layout relationships which could be used for porting the contents to wide range of platforms automatically. However, Stroulia et al. [18] stated that, using static analysis for the purposes of GUI migration have a high risk of violating the internal coherence of an application. As such it is not widely adopted.

Staiger [19] proposed a technique to reverse engineer desktop programs written in C or C++ which use a GUI library, using static analysis to generate models representing the architecture of the application. Their analysis technique is integrated into Bauhaus tool suite [20]. The generated models called window graph and inter-procedural static single assignment (ISSA) graph represents the program's windows and their structure and the attributes of the widgets with their values. The models can be used for program comprehension to support maintenance tasks, GUI testing and GUI architecture recovery. However, few sources of inaccuracy in the widgets hierarchy (e.g. in the instantiation of these roles, were points-to based ISSA is used) were identified, which affects the accuracy of the GUI analysis [19].

The GUISurfer by Silva et al. [21] automatically extracts GUI behavioral models from the applications source code of java applications. The tool first uses a parser to generate Abstract Syntax Tree (AST) from the application. Secondly, code slicing is used to extract user interface relevant data from the AST and finally used strategic programming which traverse the AST and generate the GUI tree. This enables extraction of different behavioral models from application's source code. The goal is to use the models to reason about the quality of the system both from usability and an implementation perspective, as well as being used to help systems' maintenance, evolution, and redesign.

2.2 Dynamic Approach

The dynamic approach to reverse engineering does not require access to applications' source code and it could be easier to implement. Several reverse engineering techniques have been proposed to improve the quality of GUI model exploration in dynamic analysis.

One of the earliest tools is GUITAR proposed by Memon et al. [6] to reverse engineer desktop application's GUI directly from the executing application. The GUI is represented as models called GUI forest and event flow graph representing the structure and execution behavior of the GUI. The technique called GUI ripping is described as dynamic process in which the software's GUI is automatically analyzed by opening all its windows and extracting all their widgets (GUI objects) properties, and values. However, the models generated by the tool misses some important parts of the GUI, which indicates its inability to model dynamically changing GUIs particularly when the GUI components' visibility depends on the state of another component [22]. Consequently, some parts of the information extracted may be incomplete. An extension of the tool has been proposed for the android known as Android GUITAR [23].

Mesbah et al. [24, 25] proposed a crawling technique that uses dynamic analysis of user interface state changes in Web browsers to reverse engineer AJAX web-based applications. It uses state machine representation to capture the models due the state-full nature of Ajax user interface. Their technique is implemented a tool called CRAWLJAX which performs a scan on the DOM tree to identify elements that are capable of changing the state. It fires the elements to and incrementally deduce a state machine that models the navigational paths and states of the application. The technique can be used for testing and maintenance of web applications.

The techniques by Amalfitano et al. [11] reverse engineer the GUI of Rich Internet Applications (RIAs) to model the behavior of the application. The tool called CReRIA starts by dynamically interacting with the RIA firing an event on its current user interface, to allow a phased construction of a Finite State Machine from the GUIs. A heuristic clustering criterion is used by the tool for reducing the extracted data which contain false event sequences. The models can be used for program comprehension. However, the degree of automation is inadequate. Therefore, the process requires user interaction with an application during the analysis.

The GUI Driver tool by Aho et al. [26] dynamically reverse engineer Java GUI applications by simulating user actions to automatically execute, interact and analyze the Java GUI applications at runtime. The tool automatically generates structural models of each GUI state of the application and saved into XML files, and an FSM based internal GUI model describing the behavior of the GUI application as transitions between the GUI states. The structural model represents the hierarchical tree structure of all the objects of the focused GUI window including the relevant properties of the objects and values of the properties. The models created can be used for checking the conformance of the system against the requirements. The tool uses advanced GUI automation to deal with the challenge with dynamically changing GUI, thus interacts more with the GUIs [27]. Therefore, it is able to explore and model more states of a GUI application.

Similarly, Amalfitano et al. [28] proposed an extension of the crawling technique in [25] for the reverse engineering of android apps. The technique is implemented in a tool called Android Automatic Testing Tool (A^2T^2). A^2T^2 reverse engineer android apps by automatically building models representing the structure and behavior of the application's GUIs. The structure and behavior of the GUIs is represented with GUI tree and state machine models respectively. The generated models are used for model-based testing.

Shah and Tilevich [29] presented a technique to reverse engineer user interfaces to and across mobile devices and platforms. The approach examines an interface's runtime representation by means of aspect-oriented programming (AOP) and reflection. The program's control flow is intercepted by AOP at the point when all the components of an interface are laid out on the screen, while the examination of object structures is aided by reflection at runtime. The approach is unable to extract information about components that are visible only when other components are explored.

Amalfitano et al. [10] presents an extension of the ripping technique in their tool the AndroidRipper for automated testing of Android apps through their Graphical User Interface. According to them, a fully automatic analysis to generate models from executing GUI remains challenging for the Android GUIs. Considering the challenge, AndroidRipper mainly designed to automatically traverse the application's UI to generate and execute test not to develop re-usable models of the app. However, it maintains a state machine model of the GUI.

Joorabchi and Mesbah [9] presents ICRAWLER tool that automatically reverse engineer iOS mobile apps by dynamically running and analyzing the application to cover the interaction state space to extract information about the runtime behavior. It generates a state model of an application representing the user interface states and transitions between them. However, the tool has no support for advanced gestures such as swiping pages and pinching.

2.3 Hybrid Approach

The hybrid approach is utilized in order to improve the efficiency of reverse engineering techniques, which takes advantages and best features of both static and dynamic analysis. Not many researchers have utilized this approach to GUI reverse engineering. Some examples of GUIs reverse engineering techniques based on the hybrid approach are discussed as follows.

Systä [30] describes a hybrid approach that uses information from static analysis for the generation of dynamic information from java applications. The approach starts with static analysis to extract information about only the class files of interest using JDK debugger and generating static dependency graph which is later used in the dynamic analysis to generate models of a specific behavior of the software.

Such information slicing can greatly reduce the size of generated events while containing the relevant information. The generated model can be used for program comprehension.

Silva and Campos [31] proposed a hybrid approach for the reverse engineering of web applications through their GUIs. In order to have models that capture the behavior of system comprehensively, their approach starts with a dynamic analysis of the applications at runtime to obtain the first models and identify the event handlers associated with each UI control. Static analysis is used on the source code of the event handlers found from the initial analysis, which enable them to add more information to the dynamic exploration and support completion of the first models generated. The technique can be used to gain understanding of web applications.

The ORBIT tool by Yang et al. [8] was designed based on the concept of hybrid approach to reverse engineer android applications. Their approach uses static analysis to extract the set of actions supported by the Graphical User Interface of the application and further use dynamic crawling technique to reverse engineer a model from an application by exercising these events automatically on the running application. However, the tool does not model the structure of the GUIs which is essential in understanding applications.

3 Framework for Tool Assessment

Reverse engineering tools deal with information retrieval, storage, organization and representation. They typically used different approaches to retrieve information which is represented in various graphical models such as GUI tree and state machine for the structure and behavior of application respective. This section describes the framework used for assessment of tools. Three important aspects of the tools were covered. These comprise of the approach used by the techniques, the criteria for comparative assessment and capabilities of the tool.

3.1 Approach

The main objective of this comparison is to identify which reverse engineering approach has large number of research works and why? Based on the retrieved research papers, few research works were found based on the static approach [16, 19, 21]. This shows that it is not widely used for the reverse engineering of GUI applications. This is due to the dynamic object-oriented nature of the GUI apps which can sometimes complicate the analysis and makes it difficult or even impossible to cover the behavioral aspect of the GUIs comprehensively [14, 30].

For the dynamic approach, several works were retrieved for GUI apps [6, 9–11, 24, 26, 28, 29]. This concludes that the dynamic approach is widely used for the reverse engineering of GUI apps. In the case of hybrid approach, there are few research works on the reverse engineering of mobile applications for automated model generation [8, 30, 31]. It requires access to the source or bytecode code of an application which is rarely available for mobile apps.

3.2 Criteria for Comparative Assessment

In order to assess the effectiveness of technique/tool in reverse engineering GUI application, it is essential to define the relevant criteria used for the assessment. Therefore, the framework includes the following features. Coverage of the tool which indicates the number of functions covered with their states and the efficiency in terms of time taken for the exploration.

3.3 Tool Capabilities

As with most reverse engineering processes, the aim is to generate models to support program comprehension, either to support maintenance and evolution of an application or for model-based testing. The analysis performed in this section is to assess the capabilities of the tools. This is by identifying the type of information extracted from an app by the tool and how the information is being represented for easy understanding by the users.

4 Assessment of Five Mobile Apps Reverse Engineering Tools

As the study is focused on mobile apps, the literature search identifies reverse engineering tools that are designed for mobile apps which are capable of reverse engineering mobile apps to various models. Appropriate discussion and conclusions were drawn based on the results (Tables 1 and 2).

With reference to the coverage as shown in Table 1, Android GUITAR has low coverage of app compared to others. A^2T^2, ICRAWLER and Android Ripper have medium coverage while ORBIT has a high coverage of app. This indicates the

Table 1 Result of comparative assessment of reverse engineering tools

Tool	Coverage	Efficiency
Android GUITAR [23]	Low	Longer time
A^2T^2 [28]	Medium	Average
ICRAWLER [9]	Medium	Average
Android Ripper [10]	Medium	Average
ORBIT [8]	High	Less time

Table 2 Assessment of the capabilities of reverse engineering tools

Tool	Technique	Approach	Advantage	Limitation
Android GUITAR [23]	GUI ripping	Dynamic analysis	Facilitates model generation for MBT	Information retrieve is incomplete with severa false states
A^2T^2 [28]	GUI crawling	Dynamic analysis	Ability to recognize when two interface are equivalent	Manages only a subset of the possible widgets
ICRAWLER [9]	GUI crawling	Dynamic analysis	Emphasis on identifying new UI state from a visited UI state	No support for some UI elements, such as advanced gestures e.g. swiping and pinching
Android Ripper [10]	GUI ripping	Dynamic analysis	Emphasis on reducing false event sequence in data extracted	It designed to detect bugs based on event traces, does not create re-usable models
ORBIT [8]	GUI crawling	Static and dynamic	Enhance GUI exploration using static analysis	The order of event sequence is not controlled. It does not model the GUI structure and some gesture events are not covered

weakness of dynamic approach for the generation of quality models from GUI apps. For the efficiency, the results in Table 1 indicated that ORBIT is faster than the other tools. A^2T^2, ICRAWLER and Android Ripper takes more time for exploration while Android GUITAR takes longer time compared to all the tools.

With regards to the information extracted by the tools, there are some similarities. Most of the techniques extract structural and behavioral information from the GUIs, such as GUI windows and their hierarchy, the widgets (including their properties and values), and events with their resulting states. The techniques in [10, 28] extract structural models of the application while [8, 9, 11, 21] extract behavioral models of an application. Moreover, some of the techniques extract both the structural and behavioral models of the application [6, 32].

Another aspect of interest in the study is how the extracted information is being represented. All the techniques reviewed modeled the GUI structure in the form of GUI tree showing windows hierarchy, the widgets with properties and values of each. Regarding the GUIs behavioral information, there are some variations in methods of their representation. The technique in [6] uses event flow graph (EFG) and integration tree (IT) to model the GUI behavior. However, many false events are included in the EFG and the resulting Integration Tree which may need to be filtered later. Therefore, most tools modeled the GUI behavior as a finite states machine (FSM). Table 2 presents the assessment of tools capabilities highlighting their advantages and limitations. As indicated in Table 2, all the tools that are based on dynamic approach have limitations with respect to the information extracted. ORBIT tools which is based on hybrid static/dynamic analysis has better results compared to the others, however, it is also not optimal.

4.1 Discussion

In summary, both static and dynamic approaches for the reverse engineering of GUI applications have a number of advantages and drawbacks. The static approach is capable of extracting more accurate and complete information from an application, but the dynamic nature of GUI applications makes it difficult to extract the behavior of GUI applications by pure static analysis such as the runtime aspects, like timing and layout [14, 30, 31, 33]. The dynamic approach has shown better results in reverse engineering the behavior of GUI applications. However, the information extracted by dynamic approaches is incomplete [8, 14, 31, 32]. This affects the quality of the models generated. The most challenging issue with any GUI dynamic analysis technique includes the way and order in which GUI events are found and fired, the criteria used to stop the exploration of a given GUI and the scalability issue that is associated with large amounts of data collected at run-time [14]. This tasks still requires human intervention.

5 Conclusion and Future Work

Reverse engineering techniques are widely used to aid the complex task of understating GUI apps. Graphical User Interface reverse engineering is promising in supporting testing, security analysis, porting GUIs to other platform and program comprehension in general. This paper has presented a comparative assessment of reverse engineering techniques for GUI applications based on a framework that consist of the approach used by the tools, the criteria for the assessment and capabilities of the tools. Based on the assessment results, we found that dynamic approach is widely used for the reverse engineering of GUI apps for testing purpose due to its ability to deal with the dynamic behavior of GUI apps.

With regards to the models generated, we found that several aspects of app's behavior are missed by the reverse engineering techniques which indicated that the techniques are not optimal. Therefore, there is the need for further contributions in improving the efficiency of the techniques in covering the unexplored aspects of apps. We are currently furthering our research towards this direction by combining static and dynamic approaches to increase the effectiveness of the reverse engineering techniques for generation of high quality models. We aim to develop a tool based on the proposed approach that can reverse engineer mobile apps for model generation considering the limitation highlighted above.

Acknowledgment This research is supported under the Graduate Research Incentive Grants (GIPS), vote 1256, Universiti Tun Hussein Onn Malaysia.

References

1. Cipresso, T., Stamp, M.: Software reverse engineering. In: Handbook of Information and Communication Security, pp. 659–696. Springer, Berlin (2010)
2. Ali, M.R.: Why teach reverse engineering? SIGSOFT Softw. Eng. Notes **30**(4), 1–4 (2005)
3. Canfora, G., Penta, M.D., Cerulo, L.: Achievements and challenges in software reverse engineering. Commun. ACM **54**(4), 142–151 (2011)
4. Krijnen, J.: Software reverse engineering (2013)
5. Roehm, T., et al.: How do professional developers comprehend software? In: Proceedings of the 34th International Conference on Software Engineering, pp. 255–265. IEEE Press, Zurich (2012)
6. Memon, A., Banerjee, I., Nagarajan, A.: GUI ripping: Reverse engineering of graphical user interfaces for testing. In: 10th Working Conference on Reverse Engineering (WCRE 2003). IEEE Computer Society, Washington DC (2003)
7. Brooks, P., Robinson, B., Memon, A.M.: An initial characterization of industrial graphical user interface systems. In: International Conference on Software Testing Verification and Validation, ICST'09 (2009)
8. Yang, W., Prasad, M.R., Xie, T.: A grey-box approach for automated GUI-model generation of mobile applications. In: Fundamental Approaches to Software Engineering, pp 250–265. Springer, Berlin (2013)
9. Joorabchi, M.E., Mesbah, A.: Reverse engineering iOS mobile applications. In: 19th Working Conference on Reverse Engineering (WCRE). IEEE, New Jersey (2012)
10. Amalfitano, D., et al.: Using GUI ripping for automated testing of Android applications. In: Proceedings of the 27th IEEE/ACM International Conference on Automated Software Engineering, pp. 258–261. ACM, Essen (2012)
11. Amalfitano, D., Fasolino, A.R., Tramontana, P.: An iterative approach for the reverse engineering of rich internet application user interfaces. In: 2010 Fifth International Conference on Internet and Web Applications and Services (ICIW) (2010)
12. Grilo, A.P., Paiva, A.R., Faria, J.P.: Reverse engineering of GUI models for testing. In: 5th Iberian Conference on Information Systems and Technologies (CISTI) (2010)
13. Campos, J.C., et al.: GUIsurfer: A reverse engineering framework for user interface software. In: Reverse Engineering-Recent Advances and Applications, pp. 31–54. INTECH Open Access Publisher, Croatia (2012)
14. Kull, A.: Automatic GUI model generation: State of the art. In: IEEE 23rd International Symposium on Software Reliability Engineering Workshops (ISSREW) (2012)
15. Moore, M.M.: Rule-based detection for reverse engineering user interfaces. In: Proceedings of the Third Working Conference on Reverse Engineering (1996)
16. Moore, M., Rugaber, S., Seaver, P.: Knowledge-based user interface migration. In: Proceedings of International Conference on Software Maintenance. IEEE, New Jersey (1994)
17. Bouillon, L., Vanderdonckt, J., Souchon, N.: Recovering alternative presentation models of a web page with VAQUITA. In: Computer-Aided Design of User Interfaces III, pp. 311–322. Springer, Berlin (2002)
18. Stroulia, E., El-Ramly, M., Sorenson, P.: From legacy to web through interaction modeling. In: Proceedings of International Conference on Software Maintenance. IEEE, New Jersey (2002)
19. Staiger, S.: Reverse engineering of graphical user interfaces using static analyses. In: Proceedings of the 14th Working Conference on Reverse Engineering, pp. 189–198. IEEE Computer Society, Washington DC (2007)
20. Raza, A., Vogel, G., Plödereder, E.: Bauhaus—A tool suite for program analysis and reverse engineering. In: Reliable Software Technologies—Ada-Europe 2006, pp. 71–82. Springer, Berlin (2006)

21. Silva, J.C., et al.: The GUISurfer tool: Towards a language independent approach to reverse engineering GUI code. In: Proceedings of the 2nd ACM SIGCHI Symposium on Engineering Interactive Computing Systems. ACM, Berlin (2010)
22. Yuan, M., Xuebing. Y.: An FSM based GUI test automation model. In: 11th International Conference on Control Automation Robotics and Vision (ICARCV) (2010)
23. Android GUITAR: http://sourceforge.net/apps/mediawiki/guitar/index.php?title=Android_GUITAR. Accessed 02 Jan 2015
24. Mesbah, A., van Deursen, A., Lenselink, S.: Crawling AJAX-based web applications through dynamic analysis of user interface state changes. ACM Trans. Web **6**(1), 1–30 (2012)
25. Mesbah, A., Bozdag, E., van Deursen, A.: Crawling AJAX by inferring user interface state changes. In: Eighth International Conference on Web Engineering ICWE'08 (2008)
26. Aho, P., et al.: Automated java GUI modeling for model-based testing purposes. In: Eighth International Conference on Information Technology: New Generations (ITNG) (2011)
27. Aho, P., Raty, T., Menz, N.: Dynamic reverse engineering of GUI models for testing. In: International Conference on Control, Decision and Information Technologies (CoDIT) (2013)
28. Amalfitano, D., Fasolino, A.R., Tramontana, P.: A GUI crawling-based technique for android mobile application testing. In: IEEE Fourth International Conference on Software Testing, Verification and Validation Workshops (ICSTW). IEEE, New Jersey (2011)
29. Shah, E., Tilevich, E.: Reverse-engineering user interfaces to facilitate porting to and across mobile devices and platforms. In: Proceedings of the Compilation of the Co-located Workshops on DSM'11, TMC'11, AGERE!'11, AOOPES'11, NEAT'11, & VMIL'11. 2011. ACM, Washington, DC (2011)
30. Systä, T.: Static and Dynamic Reverse Engineering Techniques for Java Software Systems. Citeseer (2000)
31. Silva, C.E., Campos, J.C.: Combining static and dynamic analysis for the reverse engineering of web applications. In: Proceedings of the 5th ACM SIGCHI Symposium on Engineering Interactive Computing Systems, pp. 107–112. ACM, London (2013)
32. Coimbra Morgado, I., Paiva, A.C., Pascoal Faria, J.: Dynamic reverse engineering of graphical user interfaces. Int. J. Adv. Softw. **5**(3 and 4), 224–236 (2012)
33. Coimbra Morgado, I., Paiva, A., Pascoal Faria, J.: Reverse engineering of graphical user interfaces. In: The Sixth International Conference on Software Engineering Advances ICSEA (2011)

Artificial Bee Colony for Vertical-Handover in Heterogeneous Wireless Networks

Shidrokh Goudarzi, Wan Haslina Hassan, Seyed Ahmad Soleymani, Omar Zakaria and Lalitha Bhavani Jivanadham

Abstract Heterogeneous wireless networks are converging towards an all-IP network as part of the so-called next-generation network. In this paradigm, different access technologies need to be interconnected; thus, vertical handovers are necessary for seamless mobility. In this paper, an artificial bee colony (ABC) algorithm for real-time vertical handover using different objective function has been presented to find the optimal network to connect. It can select an optimal set of weights for specified values, and find the optimal network selection solution. Simulation results illustrate that the proposed ABC algorithm has better performances than the existing methods in many evaluating parameters, and the computational time is also minimized.

S. Goudarzi (✉) · W.H. Hassan · O. Zakaria · L.B. Jivanadham
Communication System and Network (IKohza) Research Group,
Malaysia Japan International Institute of Technology (MJIIT),
Universiti Teknologi Malaysia, Johor Bahru, Malaysia
e-mail: shidrokhgoudarzi@gmail.com

W.H. Hassan
e-mail: whh2@hotmail.com

O. Zakaria
e-mail: omar_zak_82@yahoo.com

L.B. Jivanadham
e-mail: bjlalitha@gmail.com

S.A. Soleymani
Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru, Malaysia
e-mail: ahmad.soleymani@gmail.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_25

1 Introduction

Seamless mobility over heterogeneous networks presents significant challenges in future mobile IP networks. This is due to the presence of varying factors that have significant impact in facilitating optimized handovers between different access network technologies (i.e. vertical handovers) [1].

In heterogeneous wireless networks, a mobile user may have to perform vertical handovers across different network domains to maintain its data connection and Quality-of-Service (QoS) parameters. This process consists of three phases: discover, decision and execution. Eventually, the available networks links are introduced using the accumulated information on factors including the throughput, handover rate, cost, location, packet loss ratio, received signal strength (RSS), carrier interference ratio, signal-to- interference and noise ratio (SINR), bit error rate and QoS parameters of the discovery stage. The collected data can furthermore provide the mobile device status, indicating the service class, resources, speed and even the battery charge status. Other attributes and user preferences such as the required services and the users budget can be derived from the information gathered. Recent works have shown some degree of success in providing vertical handovers but most, if not all, consider only partial environmental factors. Furthermore, the use of adaptive behavior has not been fully investigated [2].

Every Vertical Handover (VHO) mechanism should go through a decision-making process to evaluate the available wireless access networks. The outcome of the process is the identification of the network that is most suitable for the handing over of a mobile terminal based on the criteria gathered and evaluated during the discovery stage. The decision-making phase is the essential core stage of the VHO process because it selects the most suitable network for fulfilling both the user and system demands. Many proposed VHO algorithms can be found in the interoperability literature. In basic terms, three steps constitute the decision-making phase: (1) selecting the parameters, (2) processing the parameters, and (3) aggregating the parameters [3].

Rakovic and Gavrilovska [4] proposed an innovative method to select the radio access technology (RAT), namely, the Hopfield neural network RAT selection Mechanism (HRM), that utilized the Hopfield neural networks as a monitoring decision making method. They employed the packet success rate parameter as a link estimator instead of the RSS for making the handover decision. The proposed method is considered for use on the terminal side, but it is more beneficial to modify the mechanism for operation on the network side due to the fact that user side based mechanisms are limited through managing capacity and energy utilization, subsequent in easier but less efficient implementation. Existing algorithm by ÇAlhan and ÇEken [5] consider service fee, received signal strength information (RSSI), and user preference, etc. The proposed algorithm in comparison with the traditional RSSI based algorithm, improved results significantly for user and network as an outcome of the proposed handover systems. Moreover, a fuzzy based algorithm importantly decreases the number of handovers in comparison with a Simple

Additive Weighting (SAW) based algorithm. SAW is commonly used multi-attribute decision technique, which is based on the average of weights. The weights control the importance of parameters from access points (APs). The coefficients, i.e. weights of parameters, are nominated to highlight the effect of contributions on choosing candidate AP. However, this algorithm possesses some shortcomings in terms of high execution time, which could lead to high handover latency. Also, the study did not include the impact of other environmental factors, nor did their results investigate mobile QoS parameters such as handover delay and packet loss. In Chandralekha and Praffula [6], the decision problem is expressed as multiple objective optimization problems and genetic algorithm is selected in order to simulation. The simulation result showed that the number of handovers could be reduced if optimized network factor values are taken. Also, this approach has limitations on negative effect of selecting parameter configuration on the quality of the solution and the computation time. A new approach by ÇAlhan and ÇEken [7] using information about data rate, monetary cost and received signal strength as different parameters to make a handover decision has been reported. The vertical handover decision algorithm utilized in the training process was the Levenberg–Marquardt back propagation algorithm. This algorithm adjusts the weights in a classified manner, following the application of all the training vectors. It is the efficient training algorithm to feed forward neural networks. However, this algorithm also possesses some drawbacks. The main weaknesses are about the computation of the error function and Jacobian inversion to achieve a matrix in which the dimensions are equal to the total of all the weights in the neural network. Hence, the necessity for memory is extreme. A vertical handover decision function by Nasser et al. [8] is proposed, which enables devices to assign weights to different network parameters. The above mentioned network selection schemes have been widely studied. Many challenges are present in the VH decision phase during handover process. On the one hand, sometimes the terminals are moving quickly along their routes, so the related algorithms which provide the vertical handover decision stage need to be very fast and able to give a solution near to real-time in such dynamic sceneries. On the other hand, some decision algorithms handle many parameters that involve quite a lot off loading-point arithmetic calculation, and the computational effort increases with the required precision for the solutions, the number of QoS parameters or available networks discovered during the movement of the terminal. A high computational effort is in conflict with the low response time restriction, especially taking into account the low-performance processors embedded in many mobile devices. As such, there is a critical need to develop effective vertical handover decision (VHD) algorithms that not only consider significant factors, but are also able to dynamically adapt to varying conditions in a timely manner given the rapid change in the wireless environment. Therefore, to obtain an efficient algorithm for vertical handover is a challenging task. Therefore, to eliminate such problems, an artificial bee colony (ABC) algorithm using diverse objective function has been used for optimal network selection. Population initialization is a vital stage in evolutionary algorithms due to the fact that it can interrupt the convergence speed and in result it has negative effect on quality of last solution. In ABC algorithm, an improved solution search algorithm is used very first time for the network selection problem, which is based on assessing and searching of bees only over the greatest result of past iteration to develop exploitation. The objective of this study is to propose a novel, new network selection optimization algorithm that takes advantage of ABC optimization method to meet the better performance by employing (SEarch by Fixed Intervals) SEFI [9] to acquiring better solutions near to real time. SEFI is used to search solutions directly, which can determine the computation time. SEFI is a direct search scheme, which every part of the possible results for specified search accuracy are discovered.

The rest of the paper is arranged as follows: Sect. 2 presents the optimization problem for network selection during the VH process. Section 3 gives a brief introduction and concepts of ABC algorithm. Section 4 presents the proposed algorithm with explanation of its essential components, principles and procedures. Section 5 shows the details of the proposed algorithm with ABC. Section 6 gives concluding notes of this paper.

2 The Optimization Problem

The efficient adjusting the QoS weights to determine the best network among available ones is very significant for wireless networks. To discover the best network, we request to know the goodness of each accessible network. For this purpose, we design a structure to acquire the goodness of each available network. Firstly, a set of weights allocated to each QoS factors utilized to keep the degree of the quality of per network, and they are established on two main parts includes the user's preferences and network characteristics [10]. The basic form includes all of the QoS parameters, which can allocate limited weight range from 0 to 1. This measure is given by a function, which is called cost function. This function is evaluated in the VH decision-making stage. Thus, the optimization problem involves of looking the optimum result, which useful for per network then repay the smallest cost for the related network that would be chosen as the best solution for the VH decision stage.

3 Artificial Bee Colony Algorithm

ABC algorithm simulating seeking the manner of bees was designed by Karaboga [11]. The Artificial Bee Colony (ABC) algorithm is a famous swarm-based algorithm. ABC simulates the intelligent foraging behaviour of a bee swarm. In the ABC algorithm, each cycle contains three steps: At the beginning, a set of food source is chosen by bees in a random manner and their amounts are given. A new food source is produced, and the new nectar amount can be obtained from the new

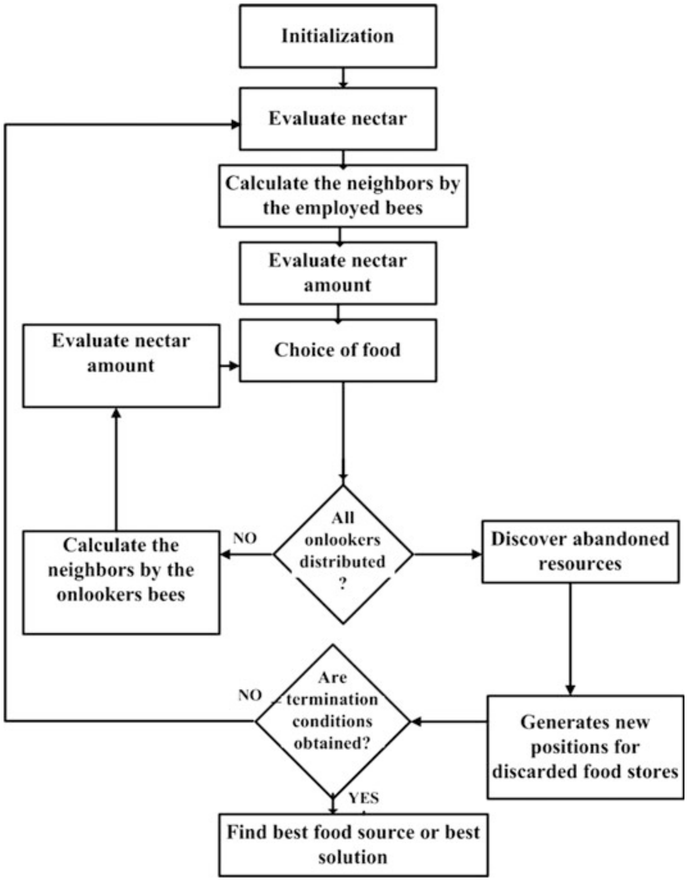


Fig. 1 Main steps of the ABC model

food source [12]. Figure 1 shows the ABC algorithm flowchart. The following steps are to be undertaken to discuss the basic ABC algorithm process.

3.1 Producing Initial Population

In the first stage, the scout bees are established for looking the food sources (solutions) and related control parameters must be set. In this step, a food source population is generated randomly by Eq. (1). The initial population of solutions is completed by SN number of randomly produced n-dimensional real-valued vectors (i.e., food sources). Let $X_i = \{x_i, 1, x_i, 2, \ldots, x_i, n\}$ corresponds to ith food source in the population which is followed by each food source, and can be expressed as follows:

$$x_{ij} = x_{min,j} + rand(0,1)\left(x_{max,j} - x_{min,j}\right) \quad (1)$$

where i = 1, 2, …, SN, j = 1, 2, …, n. SN is the number of food sources and n is the number of optimized parameters. $x_{min,j}$ and $x_{max,j}$ are the lower and upper limits for dimension j, correspondingly.

3.2 Employed Bees

In this step, employed bees have responsibilities to seek for new sources of foods. They can store all of the information of their neighborhoods in their memories. They discovery a food source in their neighborhood and then evaluate the fitness of all food sources. Then they create new food sources. After generating the new food source, fitness of all food sources must be computed and a greedy selection is used between each one and its parent. Then, employed bees should distribute their information of food source to onlooker bees by dancing in the hive. In this step, every employed bee X_i produces a new food source V_i in the nearby of its current location by using solution search equation as:

$$V_{i,j} = x_{i,j} + \Phi_{i,j}\left(x_{i,j} - x_{k,j}\right) \quad (2)$$

where k ∈ {1, 2, …, SN} and j ∈ {1, 2, …, n} are randomly chosen indexes; k has to be different from i; whereas $\Phi_{i,j}$ is a uniformly distributed real random number with range of [0, 1].

3.3 Calculating Probability Values Involved in Probabilistic Selection

In ABC algorithm, an onlooker bee calculates nectar amount received from the entire employed bees and selects a best source with a chance related to its nectar quantity. The probability based nectar amount selection is related to the fitness values of solutions in the population. In this process, a roulette wheel, ranking based, stochastic universal sampling, tournament selection or another selection structure can obtain the fitness criteria. The probability value in fluencies the behavior of onlooker bees, which select food sources based on probability, and is calculated as:

$$pi = \frac{fitnessi}{\sum_{i=1}^{SN} fitnessi} \quad (3)$$

3.4 Onlooker Bee Phase

In this step, the onlooker bees select the sources base on the information obtained of the employed bees. So, the roulette wheel selection method can use as a main method to assign the fitness. Whenever a food source is selected for an onlooker bee, a neighborhood source is established by computing the value of fitness function. Also, in the employed bees phase, a greedy selection is useful among two sources.

3.5 Abandonment Criteria: Scout Bee Phase

In this phase, employed bees, which cannot improve solutions within a predetermined number of trials, called "limit", then they become scouts and the solutions of should be abandoned. Then, the scouts search for new solutions, randomly. So, those sources that are poor should be abandoned and negative reaction behaviour increases to equalize the positive reaction. These stages are continued frequently till a termination a criterion is fulfilled, as a maximum round number or a peak CPU time.

3.6 Main Steps of the ABC Algorithm

It is clear from the above explanation that there are main control parameters in the basic ABC: The number of food sources, which are equal to the numbers of employed, or onlooker bees (SN), the value of limit and the maximum cycle number (MCN). In a strong search process, exploration and exploitation processes must be carried out together. In ABC algorithm that onlookers and employed bees continue the exploitation process in the search space, the scouts control the exploration route. Detailed pseudo-code of the ABC algorithm is given below:

Main steps:

Step 1: Generate initial population of solutions with the help of $x_{i,j}$, i = 1, 2… SN, j = 1, 2… n, trial $_i$ = 0, which is non improvement number of the solution X_i, utilized for abandonment
Step 2: Generate the population
Step 3: cycle = 1
Step 4: repeat
Step 5: {– Generate a new food resource population for employed bees –}
Step 6: for i = 1 to SN do
Step 7: Generate a new food resource Vi for the employed bee of the food resource Xi using Eq. (2) and calculate its amount.

Step 8: Employ a greedy choice scheme among Viand Xi and choose the better one
Step 9: If solution Xi does not modify $trial_i = trial_i + 1$, otherwise $trial_i = 0$
Step 10: end for
Step 11: Evaluate the probability values pi using Eq. (3) to get better solution by fitness criteria {– Generate a new food resource population for onlooker bees –}
Step 12: Now starts iteration t = 0 and i = 1
Step 13: repeat
Step 14: if random pi then
Step 15: Generate a new V_i food place using Eq. (2) for onlooker bees
Step 16: Exploit a greedy choice procedure among Vi and Xi and choose the superior one
Step 17: If solution Xi does not modify triali = triali + 1, otherwise triali = 0
Step 18: t = t + 1
Step 19: terminate if
Step 20: until (t = SN) {– Compute scout bees –}
Step 21: if max $(trial_i)$ > limit then
Step 22: Replace Xi by means of a new randomly created solution using Eq. (1).
Step 23: terminate if
Step 24: Remember the finest solution obtained so far
Step 25: cycle = cycle + 1
Step 26: until (cycle = max cycle number)

4 System Model

In this paper, proposed algorithm starts from an algorithm named SEFI, which Jaraz-Simon et al. [13] have designed to search solutions directly. This algorithm not only can determine the computation time, but also can integrate with heuristic proposal to acquiring better solutions near to real time.

We name SEFIABC the new heuristic proposal, which starting from SEFI [13], is based on (ABC) algorithm. ABC is a comparatively innovative optimization technique, which has been shown to be competitive to other population-based algorithms [9]. According to [13], SEFI explores the area of solutions and creates all the possible patterns for a given h, calculates the cost of results for the accessible networks and select the best network with the maximum cost function. SEFI was programmed in C language using specific loops to prepare the related structure of all the codes. In this paper, all of codes of SEFI has been successfully changed on MATLAB m-code programming language to integrate with ABC algorithm, also in order to show the results in simulation stage. The operation of SEFIABC is shown in the pseudo-code given in Algorithm 1 and in Fig. 1, and it is explained in algorithm description section as follows.

4.1 Algorithm Description

In the paper, the scenarios of vertical handover using the ABC algorithm are developed in the heterogeneous wireless networks. The operation of the SEFIABC optimization approach is shown in the pseudo-code given in Algorithm 1. The main purpose of the operations of this scheme is to increase the performance, especially in terms of computational time. As we have realized previously, the main action to look the best network is the high number of QoS parameters, but this means that to decrease the exploration precision to maintain the real-time control. The obvious disputation motivate us to plan a heuristic which could discover best choice in real time using low level of search accuracies [9].

The cost function of vertical handover is a quantity of the cost utilized by moving a network to a specific network. It is estimated for all networks n, which can cover the service region of a user. The network choice within the lowest computed cost can provide the most benefit to the users. As we know, different types of services require various combinations of reliability, latency and data rate. So, we consider service type as main metric. Also, monetary cost has a main attention for users, as different networks may employ different charging policies that may affect the user's choice to handover. In terms of network settings, network-related factors, such as available bandwidth, and network latency may need to be measured for operational network control. Usage of information of network in the selection of optimal to handover can be suitable in order to load balancing across diverse networks. Also, to warranty the system performance, a diversity of factors can be employed in the handover decision-making stage, such as the bit error rate (BER). Moreover, the mobile terminal circumstance includes active features for example velocity, moving pattern, moving histories and location information.

Consider C_n is the cost function calculated for network n. C_n is the cost of obtaining each of the user's demanded services from network n. In other words, each network is receiving user's requested services and each request is calculated (4) by own cost function.

$$\mathrm{c_s^n} = H^n_{s,i}\left(QoS^n_{s,j}\right) \tag{4}$$

Equation (4) is considered for each service. Where $H^n_{s,i}$ shows whether the lowest constraint i for service s can be met by network n. If an accessible network cannot be able to prepare the minimum requested QoS for a real time service, and would be removed from attention as a handover goal for the requested service rapidly. Also, $\mathrm{QoS^n_{s,j}}$ is the QoS offered by network n for parameter j for service s.

Since a low level of cost will result in a network being removed from attention, thus, $H^n_{s,i}$ (5) is calculated by:

$$H_{s,i}^{n} = \frac{1}{h_{s,i}^{n}} \tag{5}$$

where $h_{s,i}^{n}$ (6) is the following function:

$$h_{s,i}^{n} = \begin{cases} o & \textit{if i cannot be satisfied} \\ 1 & \textit{if i can be satisfied} \end{cases} \tag{6}$$

Also, we consider that each QoS factor has own weight and this weight shows the effect of factor on network or user. So each QoS factor can be calculated by Eq. (7):

$$Q_N = \sum W_{s,j}^{n} QoS_{s,j}^{n} \tag{7}$$

The cost function can be formulated as in (8):

$$\mathrm{C_{n=}} \sum \prod_{i} H_{s,j}^{n} \sum_{s,j}^{n} Q_N \tag{8}$$

Summary of Eq. (8) are explained in Table 1:

Initially the algorithm begins from SEFI wherever the search spaces of ranges P, which related to the QoS weights w_i. For a mobile node with N access networks, the best network can be selected for handover process as follows:

Step 1: Initialize population by SEFI, all individuals containing k-selected access networks which chosen randomly:

- (i) Set QoS parameters
- (ii) Run SEFI obtain optimal w_i

Step 2: Evaluate the cost function of per individual, where:

- (i) For per node n_i in the network:
  - (a) Calculate the distance (n_i, m_i) between terminal and access networks

Table 1 Description of components of the cost function

Formula	Descriptions
$H_{s,i}^{n}$	Network elimination of service *s* at network n
$\sum W_{s,j}^{n} QoS_{s,j}^{n}$	QoS factor represents the QoS qualified by selecting to receive from the network n
$W_{s,j}^{n}$	It is the weight representing the effect of QoS parameters to the user and network
$QoS_{s,j}^{n}$	It can normalize the QoS offered by network n in order to j to service *s*

(b) Assign n_i to m_i, where distance (n_i, m_i) = min $_{nm=1,2,\ \ldots nm}$ {distance ($n_{i,}$ m_i)}

(ii) Calculate the cost function

Step 3: Update the positions by the optimization algorithm.
Step 4: Repeat stages 2–4 till the highest number of cycle is obtained.

Each node can receive the advertisement messages from selected access networks at various signal strengths, and then calculates distances using Eq. (7) [14].

$$\mathrm{dij} = \mathrm{s} \cdot \left(\mathrm{P^r}\right)^{-1/2,} \quad \text{where } \mathrm{s} = \mathrm{c} \cdot \left(\mathrm{P^r}\right)^{-1/2} \Big/ 4\pi\mathrm{f} \tag{7}$$

In the calculation, dij is the distance among terminal i and access network j, f is the communication frequency, c is the speed of light and P^r is the received signal strength.

5 Simulation Results and Discussions

Based on MATLAB, this network selection scheme is structured with the ABC. As we can see, the solutions obtained by SEFIABC have good performances in each situation, thus showing the good implementation of our scheme. Clearly, in direct search like SEFI the computing time will be increased by high quantity of QoS factors, also a division element in order to the search spaces will be increased. In result, if we select high numbers of QoS factors, we cannot usage great division elements; also, if *NQoS* is small, we can execute analyses with all of the considered division elements. The networks categorized through the standards of the following QoS factors in Table 2.

We have measured the related contributions for SEFIABC, and we will show the related outcomes in Fig. 2. This paper uses four hexagonal honeycomb topologies with many different numbers and kinds of access networks. Based on simulation,

Table 2 QoS parameters

Net	Bandwidth (Mbps)	BER (dB)	Delay (ms)	Security level	Monetary cost (eur/MB)
UMTS	14.4	0.001	19	8	0.9
Wi-Fi (802.11n)	30	0.01	20	2	0.2
WiMax (802.16)	75	0.01	30	3	0.3
Wi-Fi (802.11ac)	72	0.01	15	3	0.2

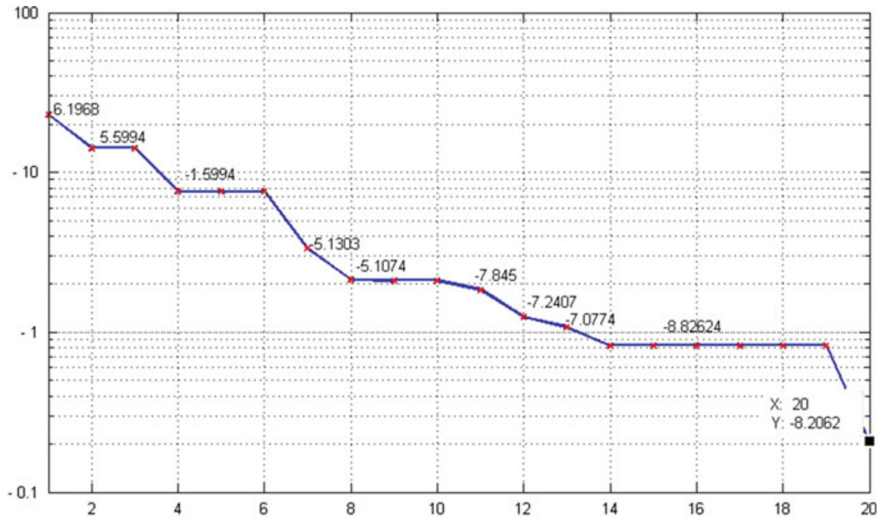


Fig. 2 Best cost found by SEFABC with

the solutions found by SEFIABC have low cost all the situations, even with small level of precision, thus showing the good implementation of proposed algorithm. The QoS parameters in this study are as follows: E = BER (dB), D = Delay (ms), B = Bandwidth (Mbps), C = Monetary cost (eur/MB), S = Security level [15]. The solutions given by SEFIABC for the aforementioned keys have suitable cost in all the cases, even when low precision. The Optimal solutions found by SEFIABC are presented for five QoS parameters, and a division factor of five and the number of population 4 and iteration 20 in Fig. 2. This figure represents an acceptable behavior of SEFABC (for the combinations or solutions the cost found by SEFIABC is low). As we can see in Fig. 2, the results of running our proposed algorithm are shown and in each iteration one network is selected as best network for handling off and we fixed number of iterations equal to 20, and in the last iteration (iteration = 20), the best network with lowest cost function is found with cost function equal to Best Cost = −8.2062.

Our heuristic shows better performance in comparison with SEFISA [16], for the same inputs. SEFIABC has better fitness because optimal solutions found by SEFISA up to four networks and five QoS parameters is Best solution = −6.36615 (Fig. 3).

Also, Fig. 4 shows results based on different QoS parameters, number of population equal 4 and iteration 20. As it's mentioned before, each QoS factor within own weight affect on network or user. The values of the QoS parameters characterize a given network and each parameter is combined with own weight, also each weight has a value between 0 and 1. Based on obtained results, we can see that in

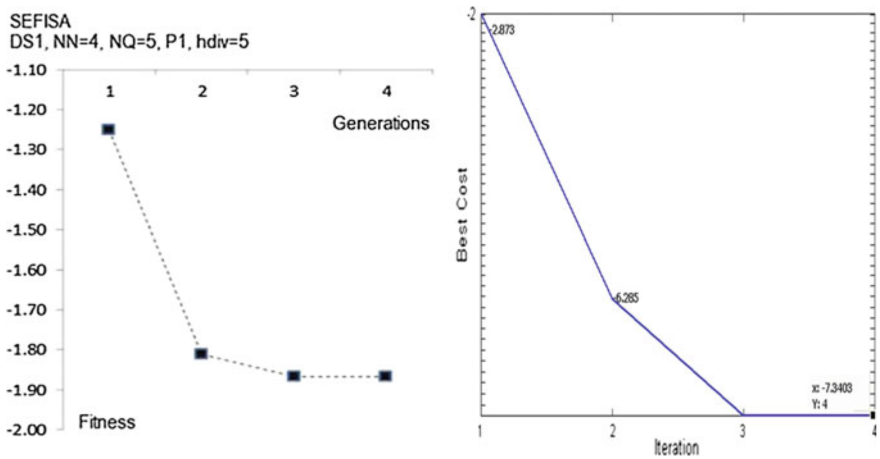


Fig. 3 Best solution = −6.36615 found by SEFISA [16]. Best solution = −7.3403 found by SEFIABC

first iterations such as iterations between 1 and 4, network 2 is selected as the best solution with available bandwidth = 10.06, and BER has acceptable rate at B = 0.01, D = 10.26, C = 0.2, and Security Level = 2. These values show a level of the network quality by means of cost function. It must consider that in per iteration the unavailable networks is eliminated based on elimination factor in cost function equation. Lastly, each solution is determined for the diverse networks, with changed cost values due to the fact that each network has specific QoS sets.

6 Computation Efficiency

Since, the real-time applications need less running time in addition to high performance, CPU time of this algorithm has been examined to examine the efficiency of algorithm. The results demonstrate the comparison of CPU timing between our heuristic with employing SEFI in initiation phase and without employing SEFI in initialization phase (generation of first population by randomly selection). As we can see SEFIABC with employing SEFI has less CPU time. In other words, It is evident that computation of the proposed SEFIABC algorithm based on the SEFI is faster (CPU time is less) than that using randomly selection in initialization phase. Corresponding graph is given in Fig. 5. This figure shows the results of CPU time in five different running times with Iteration = 500 and Number of Networks = 20.

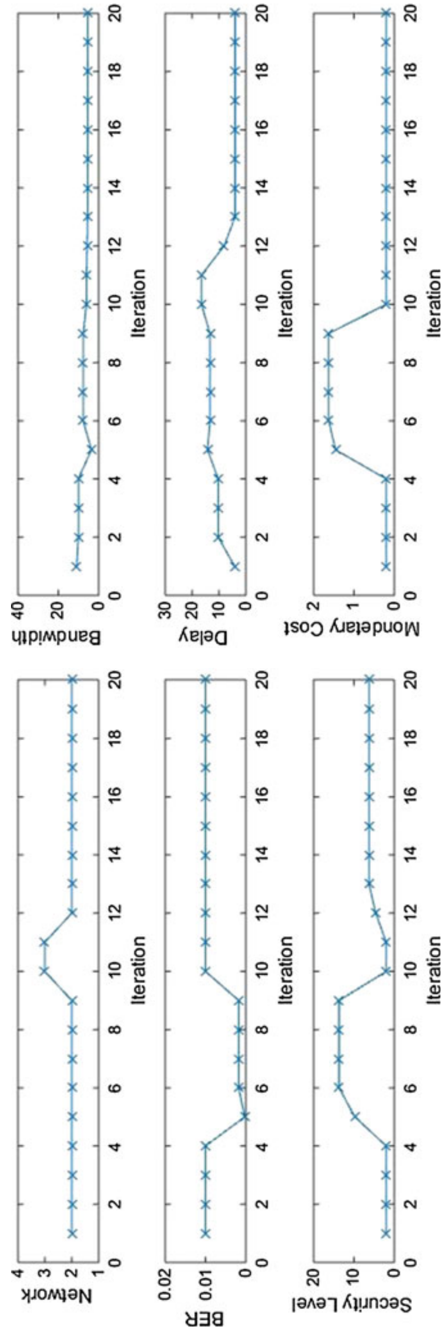


Fig. 4 Results based on different parameters found by SEFABC with iteration = 20 and network's number = 4

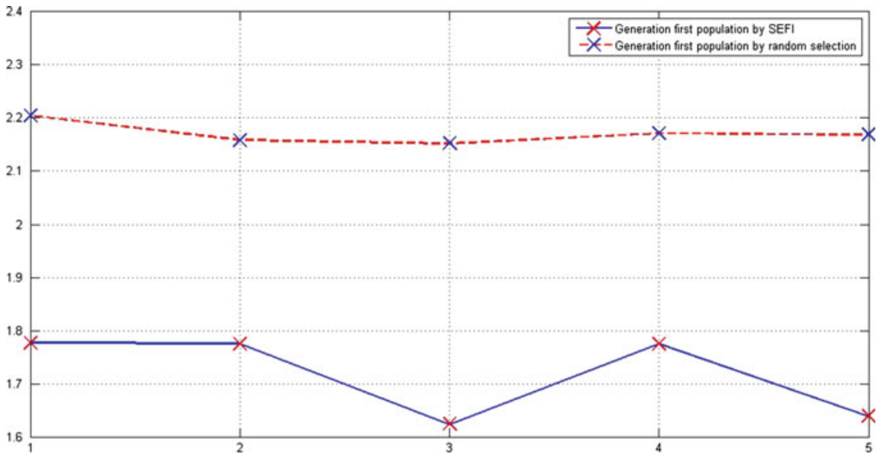


Fig. 5 Comparison of CPU Timing between SEFIABC with employing SEFI and without employing SEFI with iteration = 500 and number of networks = 20

7 Conclusions

We propose a SEFIABC algorithm to find the near-optimal solutions for the problem. In this study, new optimization for the vertical handover decision algorithm has been developed to raise the advantage of the handover for the user and network. The simulation results show that this network selection scheme has better performances than the existing schemes in the above mentioned factors, and it ensures the profits of both sides. With the help of the artificial bee colony algorithm, our network selection scheme can seek the optimal solution at the rational time.

References

1. Paul, S., Pan, J., Jain, R.: Architectures for the future networks and the next generation internet: a survey. Comput. Commun. **34**(1), 242 (2011)
2. Ahmed, A., et al.: Enabling vertical handover decisions in heterogeneous wireless networks: a state-of-the-art and a classification. Commun. Surv. Tutorials IEEE **16**, 776–811 (2014)
3. TalebiFard, P., Wong, T., Leung, V.C.M.: Access and service convergence over the mobile internet—a survey. Comput. Netw. **54**(4), 545557 (2010)
4. Rakovic, V., Gavrilovska, L.: Novel RAT selection mechanism based on Hopfield neural networks. In: IEEE International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT) (2010)
5. ÇAlhan, A., ÇEken, C.: Case study on handoff strategies for wireless overlay networks. Comput. Stand. Interfaces **35**(1), 170–178 (2013)
6. Chandralekha, P.B., Praffula, K.B.: Minimization of number of handoff using genetic algorithm in heterogeneous wireless networks. Int. J. Latest Trends Comput **1**(2), 24–28 (2010)

7. Çalhan, A., Çeken, C.: Artificial neural network based vertical handoff algorithm for reducing handoff latency. Wireless Pers. Commun. **71**(4), 2399–2415 (2013)
8. Nasser, N., Hasswa, A., Hassanein, H.: Handoffs in fourth generation heterogeneous networks. Commun. Mag. IEEE **44**(10), 96–103 (2006)
9. Jaraíz-Simon, M.D., et al.: Simulated annealing for real-time vertical-handoff in wireless networks. In: Advances in Computational Intelligence, pp 198–209. Springer, Berlin, Heidelberg (2013)
10. McNair, Janise, Zhu, Fang: Vertical handoffs in fourth-generation multinetwork environments. Wirel. Commun. IEEE **11**(3), 8–15 (2004)
11. Karaboga, D.: (2005) An idea based on honey bee swarm for numerical optimization. Technical Report, Computer Engineering Department, Engineering Faculty, Erciyes University
12. A comprehensive survey: artificial bee colony (ABC) algorithm and applications
13. Jaraz-Simon, M.D., Gomez-Pulido, J.A., Vega-Rodriguez, M.A., Sanchez-Perez, J.M.: Fast decision algorithms in low-power embedded processors for quality-of-service based connectivity of mobile sensors in heterogeneous wireless sensor networks. Sensors **12**(2), 1612–1624 (2012)
14. Karaboga, Dervis, Okdem, Selcuk, Ozturk, Celal: Cluster based wireless sensor network routing using artificial bee colony algorithm. Wirel. Netw. **18**(7), 847–860 (2012)
15. Zhu, F., McNair, J.: Optimizations for vertical handoff decision algorithms. In: IEEE Wireless Communications and Networking Conference, WCNC, vol. 2. IEEE (2004)
16. Jaraiz-Simon, M.D., Gomez-Pulido, J.A., Vega-Rodriguez, M.A.: Embedded intelligence for fast QoS-based vertical handoff in heterogeneous wireless access networks. Pervasive Mobile Comput (2014)

Dynamics Classifications of Extreme Doubly Stochastic Quadratic Operators on 2D Simplex

Rawad Abdulghafor, Farruh Shahidi, Akram Zeki and Sherzod Turaev

Abstract We study the dynamics of extreme doubly stochastic quadratic operators (d.s.q.o.) on two dimensional (2D) simplex. We provide some examples of d.s.q.o. which have infinitely many fixed points. We prove that the trajectory of extreme d.s.q.o., starting at some interior point of the simplex is convergent. Finally, we classify the dynamics of all extreme points of d.s.q.o. on 2D.

Keywords Doubly stochastic quadratic operators · Fixed point · Trajectory · Convergence and simplex

1 Introduction

It is well known that many biological processes can be considered as some nonlinear dynamical systems. From this point of view, the main problem the state of the process in a certain time frame which is the same as studying the limit behavior of the trajectories of corresponding dynamical systems. One of the important dynamical systems is the one generated by quadratic stochastic operators which arise in many problems of population genetics [1]. As of today, this area is

R. Abdulghafor (✉) · A. Zeki · S. Turaev
Kulliyyah of Information and Communication Technology, International Islamic University Malaysia, 53100 Kuala Lumpur, Malaysia
e-mail: raaac2004@yahoo.com

A. Zeki
e-mail: akramzeki@iium.edu.my

S. Turaev
e-mail: sherzod@iium.edu.my

F. Shahidi
Department of Mathematics, Pennsylvania State-University-UP, State College, PA 16801, USA
e-mail: Fus144@psu.edu

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_26

potentially developing [2, 3, 4]. Even so, there are many interesting problems left open. One of them is to study d.s.q.o. which are defined using majorization concept. The class of these operators is quite large [5], so one needs to consider some particular cases. In [5, 6], the problem of description of extreme points of the set d.s.q.o. was addressed. It was shown that, up to permutation, there are 37 extreme d.s.q.o. on two dimensional simplex. The present paper is devoted to classification of dynamics of extreme d.s.q.o. Namely, we will study the limit behavior of the trajectories of extreme d.s.q.o.

The present paper therefore aim to provide a couple of convergence result for general d.s.q.o. In addition, the paper classify the dynamics of extreme d.s.q.o. on 2D simplex. The motivation comes from paper [7] where general quadratic stochastic operators having coefficients 1 or 0 were classified on 2D simplex. As mentioned earlier, extreme d.s.q.o. have coefficients $1, \frac{1}{2}$ or 0. So our result will extend the results in Vallander [7] in the class of d.s.q.o. on 2D simplex. The paper is organized as follows. In the next section we give some preliminaries concerning majorization and d.s.q.o. In Sect. 3, we focus on dynamics classifications of extreme d.s.q.o. on 2D simplex. Section 4, we present a sub-classifications of extreme d.s.q.o. on 2D simplex. We note that a couple of examples of dynamics classifications of extreme d.s.q.o. on 2D simplex have been considered in Shahidi et al. [8].

2 Preliminaries

In this section we provide some definition based on majorization theory and defines doubly stochastic operator.

We define the $(m-1)$—dimensional simplex as follows.

$$S^{m-1} = \left\{ x = (x_1, x_2, \ldots, x_m) \in R^m : x_i \geq 0, \quad \forall_i = \overline{1,m}, \sum_{i=1}^{m} x_i = 1 \right\} \tag{1}$$

The set int$S^{m-1} = \{x \in S^{m-1} : x_i \geq 0\}$ is called the interior of the simplex. The points $e_k = \left(0, 0, \ldots, \underbrace{1}_{k}, \ldots, 0\right)$ are the vertices of the simplex and the scalar vector $\left(\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m}\right)$ is the center of the simplex.

A quadratic stochastic operator $V : S^{m-1} \to S^{m-1}$ is defined as:

$$(Vx)_k = \sum_{i=1}^{m} p_{ij,k} x_i x_j \tag{2}$$

where the coefficients $p_{ij,k}$ satisfy the following conditions

$$p_{ij,k} = p_{ij,k} \geq 0, \sum_{i=1}^{m} = 1 \tag{3}$$

If we let $A_k = \left(p_{ij,k}\right)_{i,j} = \overline{1,m}$, then the operator can be written in matrix form as follows

$$V = (A_1|A_2|\ldots|A_m) \tag{4}$$

where matrices A_i are non-negative and symmetric. For any $x = (x_1, x_2, \ldots, x_m) \in S^{m-1}$, we define $x_{\downarrow} = \left(x_{[1]}, x_{[2]}, \ldots, x_{[m]}\right)$ where $x_{[1]} \geq x_{[2]} \geq \ldots \geq x_{[m]}$—non-increasing rearrangement x. Recall in [9, 10] that for two elements x, y of the simplex S^{m-1} the element x is majorized by y and write $x \prec y$ or $y \succ x$ if the following condition holds

$$\sum_{i=1}^{m} x_{[i]} \leq \sum_{i=1}^{m} y_{[i]} \tag{5}$$

for any $k = \overline{1, m-1}$. In fact, this definition is referred to as weak majorization [10], the definition of majorization requires $\sum_{i=1}^{m} x_{[i]} = \sum_{i=1}^{m} y_{[i]}$. However, since we consider points only from the simplex, we may drop this condition.

A matrix $P = \left(p_{ij}\right)_{i,j=\overline{1,m}}$ is called stochastic (sometimes bistochastic), if

$$\begin{aligned} &p_{ij} \geq 0, \quad \forall i,j = \overline{1,m}, \\ &\sum_{i=1}^{m} = 1, \quad \forall i = \overline{1,m}, \quad \sum_{j=1}^{m} = 1, \quad \forall j = \overline{1,m}. \end{aligned} \tag{6}$$

For a doubly stochastic matrix $P = \left(p_{ij}\right)$, if its entries consist of only 0's and 1's, then the matrix is a permutation matrix.

A linear map $T : S^{m-1} \rightarrow S^{m-1}$ is said to be T-transform, if $T = (\lambda)I + (1-\lambda)P$ where I is an identity matrix, P is a permutation matrix which is obtained by swapping only two rows of I and $0 \leq \lambda \leq 1$.

Lemma 1 Olkin and Marshall [10] *for the concept of majorization and* $x, y \in S^{m-1}$, *the following assertions are equivalent.*

1. $x \prec y$ that is $\sum_{i=1}^{k} x_{[i]} \leq \sum_{i=1}^{k} y_{[i]}, k = \overline{1, m-1}$.
2. $x = Py$ for some doubly stochastic matrix P.
3. The vector x belongs to the convex hull of all $m!$ permutation vectors of y.

4. The vector x can be obtained by a finite compositions of T-transforms of the vector y, that is, there exist $T-$transforms $T_1, T_2, \ldots, T_k$ such that $x = T_1 T_2 \ldots T_k y$.
5. The inequality $\varphi(x) \leq \varphi(y)$ holds for any Schur-convex function.

From the above lemma, it follows that doubly stochasticity of a matrix P is equivalent to $Px \prec x$ for all $x \in S^{m-1}$. Motivated by this, in Olkin and Marshall [10], the definition of doubly stochastic operator is given as follows.

Definition A continuous stochastic operator $V : S^{m-1} \rightarrow S^{m-1}$ is called doubly stochastic, if $Vx \prec x$ for all $x \in S^{m-1}$. Identity operator, permutation operators (that is the linear operators with permutation matrix) and T-transforms are all doubly stochastic.

Let V be doubly stochastic operator and $x^0 \in S^{m-1}$. The sequence is infinite $\{x^0, V(x^0), V^2(x^0), \ldots, V^n(x^0), \ldots\}$ and it is called the trajectory starting at x^0. Here, $V^0(x^0) = x^0$ and $V^n(x^0) = V(V^{n-1}(x^0))$. We denote by $\omega(x^0)$ the set of limit points of the trajectory starting at x^0 and it is said to be the ω-limit set of the trajectory starting at x^0.

The point x^0 is called p-periodic, if there is a positive integer p such that $V^p(x^0) = x^0$ and $V^i(x^0) \neq x^0 \quad \forall i = \overline{1, p-1}$ if $p = 1$, we say that the point is fixed. We say periodic if the period is irrelevant and we usually refer to the point as periodic if the period is greater than one.

3 The Classification of Extreme Points of d.s.q.o. on S^2

Let us start with some examples of extreme d.s.q.o. on S^2. As we have mentioned, up to the level permutations of the matrices there are 222 extreme points of d.s.q.o. on 2D simplex. Our main result analyses all extreme points of d.q.s.o. on S^2. But, due to time limitation, we are not able to study each of them. Therefore, we provide some examples of extreme d.s.q.o., study them and then provide general result for all extreme points d.s.q.o. on S^2.

Example 1 The operator $V_1 : S^2 \rightarrow S^2$.

$$\begin{aligned} V_1(x) &= z^2 + xy + yz, \\ V_1(y) &= x^2 + xz + yz, \\ V_1(z) &= y^2 + xy + xz. \end{aligned} \tag{7}$$

is d.s.q.o. [5, 4].

The matrices of V_1 are

$$\begin{pmatrix} 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & 1 \end{pmatrix}.$$

Table 1 shows the iteration of the operator V_1 with some initial point from the interior of the simplex. One can see that the trajectory of this operator tends towards the center.

By solving the system of non-linear equation we can show that V_1 has a unique fixed point on the boundary of S^2

$$\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$$

Combining Theorems 3.3 and 3.4 we find that the trajectory of any initial point tends to the center of the simplex.

Figure 1 shows that the trajectory of initial values of x, y and z for V_1 tends to the same number $\left(\frac{1}{3}\right)$. In Fig. 2, it is shown that the trajectory on the 2D simplex converges to the center.

Table 1 The numerical solutions for operator V_1

Iterations	x	y	z
x^0	0.1	0.7	0.2
$V(x^0)$	0.25	0.17	0.58
$V^2(x^0)$	0.4775	0.3061	0.2164
$V^3(x^0)$	0.25923175	0.39757729	0.34319096
$V^4(x^0)$	0.357289624	0.292612025	0.350098351
$V^5(x^0)$	0.329559083	0.355185371	0.315255546
$V^6(x^0)$	0.328414782	0.324478676	0.347106542
$V^7(x^0)$	0.335803365	0.331255601	0.332941034
$V^8(x^0)$	0.33237506	0.334855202	0.332769738
$V^9(x^0)$	0.333522133	0.333651423	0.332826444
$V^{10}(x^0)$	0.333101593	0.333290015	0.333608392
$V^{11}(x^0)$	0.33350234	0.333270504	0.333227156
$V^{12}(x^0)$	0.333241613	0.333410629	0.333347758
$V^{13}(x^0)$	0.333328523	0.333363903	0.333307573
$V^{14}(x^0)$	0.33332635	0.333321539	0.333352111
$V^{15}(x^0)$	0.333341921	0.333334936	0.333323143
$V^{16}(x^0)$	0.33333673	0.333330471	0.333332799
$V^{17}(x^0)$	0.333333511	0.333332201	0.333334287
$V^{18}(x^0)$	0.333333565	0.333333247	0.333333187
$V^{19}(x^0)$	0.333333333	0.333333333	0.333333333
$V^{20}(x^0)$	0.333333333	0.333333333	0.333333333
$V^{21}(x^0)$	0.333333333	0.333333333	0.333333333

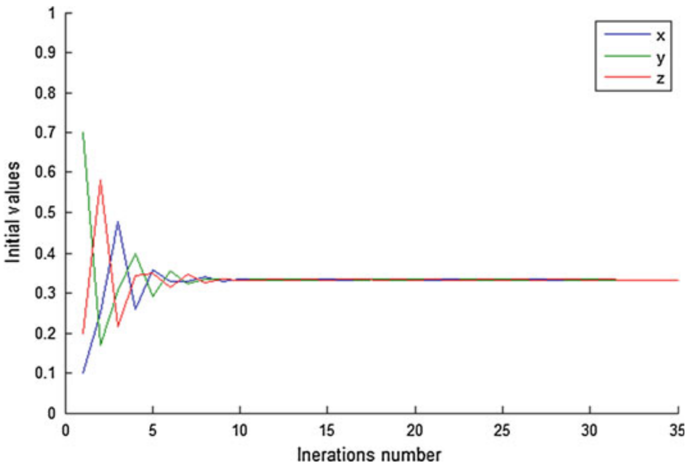


Fig. 1 The trajectory of operator V_1

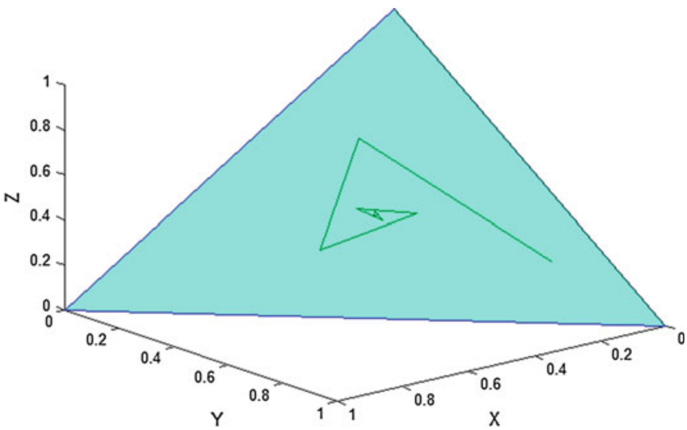


Fig. 2 The trajectory of operator V_1 on 2D simplex

Example 2 The operator $V_2 : S^2 \to S^2$.

$$\begin{aligned} V_2(x) &= xy + 2xz, \\ V_2(y) &= y^2 + xy + yz, \\ V_2(z) &= x^2 + z^2 + yz. \end{aligned} \tag{8}$$

is d.s.q.o. [4, 5].

The matrices of V_2 are

$$\begin{pmatrix} 0 & \frac{1}{2} & 1 \\ \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 1 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 1 \end{pmatrix}.$$

Note that there are many infinitely fixed points of the operator V_2. They are called

$$\{(x,y,z) \in S^2 | x = z\}, \{(x,y,z) \in S^2 | x = 0\}, \{(x,y,z) \in S^2 | y = 0\}$$

Let $y = c$, then the equation

$$\begin{aligned} V_2(x) &= xc + 2xz, \\ V_2(y) &= c, \\ V_2(z) &= x^2 + z^2 + cz. \end{aligned} \tag{9}$$

Table 2 shows that the trajectory does not converge to the center of the simplex for a particular fixed point.

Accordingly, we see that under any iteration y does not change (see Fig. 3). So $\{(x,y,z) \in S^2 | y = c,\ 0 < c < 1\}$ is the family of invariant lines (see Fig. 4). For a fixed, simple calculations show that the trajectory of any point from the line $y = c$ tends to $\left(\frac{1-c}{2}, c, \frac{1-c}{2}\right)$ (see Fig. 3).

Classification After studying operators V_1 and V_2 we come up with some observations. It is observed that the operator V_1 has a unique fixed point, which is the center of the simplex. The finding equally reveal that the trajectory of any initial point tends to the center of the simplex. The operator V_2 has been shown to behaves differently. It has infinitely many fixed points and the trajectories do not tend to the center. Rather, V_2 has infinitely many invariant lines of the form $y = c, 0 < c < 1$

Table 2 The numerical solutions for operator V_2

Iterations	x	y	z
x^0	0.8	0.1	0.1
$V(x^0)$	0.24	0.1	0.66
$V^2(x^0)$	0.3408	0.1	0.5592
$V^3(x^0)$	0.415231	0.1	0.484769
$V^4(x^0)$	0.444105	0.1	0.455895
$V^5(x^0)$	0.449341	0.1	0.450659
$V^6(x^0)$	0.449933	0.1	0.450067
$V^7(x^0)$	0.449993	0.1	0.450007
$V^8(x^0)$	0.449999	0.1	0.450001
$V^9(x^0)$	0.45	0.1	0.45
$V^{10}(x^0)$	0.45	0.1	0.45
$V^{11}(x^0)$	0.45	0.1	0.45

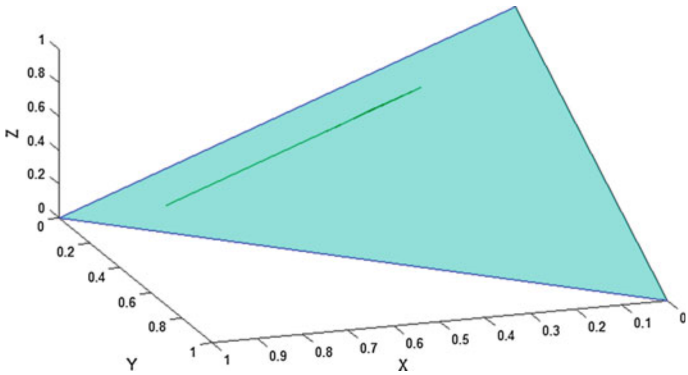


Fig. 3 The trajectory of operator V_1

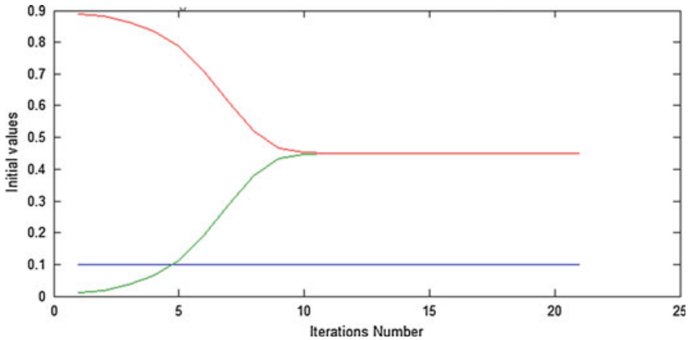


Fig. 4 The trajectory of operator V_1 on 2D simplex

and the trajectory of an initial point from this line tends to the point $\left(\frac{1-c}{2}, c, \frac{1-c}{2}\right)$. Consequently, of those 222 operators, the behaviors of 198 are similar to V_1, indicates that, the trajectory of any non-fixed point tends to the center of the simplex. 18 operators have behaviors similar to V_2, and this suggests that, they have infinitely many invariant lines and the trajectory is convergent along those lines. finally, there are 6 permutation operators, they are easy to study.

4 The Sub-classification of Extremes of d.s.q.o. on S^2

In this section we study the limit behavior of d.s.q.o. of the second group (18 extremes). Due to obvious reasons, we are not able to study each of them. Therefore, we provide some examples of extreme d.s.q.o., study them and then provide general result for all extreme of second group d.s.q.o. As we have seen that we can classify this group into two subgroups of 9 extremes. The trajectory of subgroup 1 converges monotonic increasing or decreasing, but the trajectory of subgroup 2 converges non-monotonically increasing or decreasing.

Now let us to examine examples of those subgroups.

Example 2.1 The following operator $V_3 : S^2 \to S^2$

$$\begin{aligned} V_3(x) &= x^2 + xy + yz, \\ V_3(y) &= xy + 2yz, \\ V_3(z) &= y^2 + z^2 + xz. \end{aligned} \tag{10}$$

is d.s.q.o. (see [4, 5]).

The matrices of V_3 are

$$\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & \frac{1}{2} \\ 0 & 1 & 0 \\ \frac{1}{2} & 0 & 1 \end{pmatrix}.$$

From the point that $x + y + z = 1$, we get the operator V_3 as follows

$$\begin{aligned} V_3(x) &= x, \\ V_3(y) &= xy + 2yz, \\ V_3(z) &= x^2 + z^2 + xz. \end{aligned} \tag{11}$$

Let us assume $x = \beta$, then $z = 1 - \beta - y$ or $y = 1 - \beta - z$, we find the operator V_3 as follows

$$\begin{aligned} V_3(x) &= \beta, \\ V_3(y) &= 2y - \beta y - 2y^2, \\ V_3(z) &= 2y^2 + \beta y - 2y - \beta + 1. \end{aligned} \tag{12}$$

The extreme of operator V_3 has an invariant curve of the form $x = c$, and for any point taken from this curve, the trajectory tends to $(c, (1-c)/2, (1-c)/2)$.

It has been shown in Table 3 that the trajectory of operator V_3 of sub-classifications of d.s.q.o. is convergent monotonic increasing or decreasing.

For a fixed, simple calculations show that the trajectory of any point from the line $x = c$ tends to $\left(c, \frac{1-c}{2}, \frac{1-c}{2}\right)$ (see Fig. 5).

Example 2.2 The following operator $V_4 : S^2 \to S^2$

Table 3 The numerical solutions for operator V_3

Iterations	x	y	z
x^0	0.1	0.01	0.89
$V(x^0)$	0.1	0.0188	0.8812
$V^2(x^0)$	0.1	0.035013	0.864987
$V^3(x^0)$	0.1	0.064073	0.835927
$V^4(x^0)$	0.1	0.113528	0.786472
$V^5(x^0)$	0.1	0.189926	0.710074
$V^6(x^0)$	0.1	0.288716	0.611284
$V^7(x^0)$	0.1	0.381846	0.518154
$V^8(x^0)$	0.1	0.433895	0.466105
$V^9(x^0)$	0.1	0.447871	0.452129
$V^{10}(x^0)$	0.1	0.449778	0.450222
$V^{11}(x^0)$	0.1	0.449978	0.450022
$V^{12}(x^0)$	0.1	0.449998	0.450002
$V^{13}(x^0)$	0.1	0.45	0.45
$V^{14}(x^0)$	0.1	0.45	0.45
$V^{15}(x^0)$	0.1	0.45	0.45

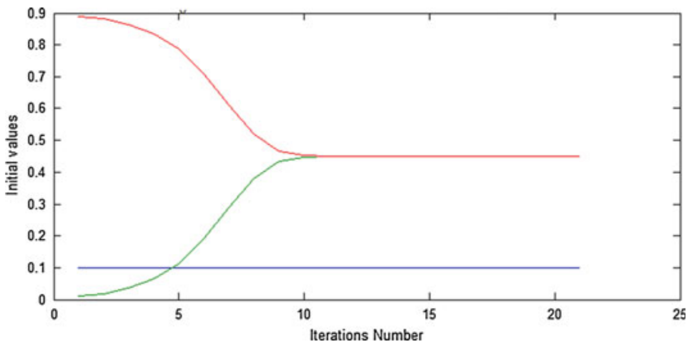


Fig. 5 The trajectory of operator V_3

$$\begin{aligned} V_3(x) &= x^2 + xy + yz, \\ V_3(y) &= xy + 2yz, \\ V_3(z) &= y^2 + z^2 + xz. \end{aligned} \tag{13}$$

is d.s.q.o. (see [4, 5]).

The matrices of V_3 are

$$\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & \frac{1}{2} \\ 0 & 1 & 0 \\ \frac{1}{2} & 0 & 1 \end{pmatrix}.$$

From the point that $x+y+z=1$, we get the operator V_4 as follows

$$\begin{aligned} V_3(x) &= x, \\ V_3(y) &= xy+2yz, \\ V_3(z) &= x^2+z^2+xz. \end{aligned} \tag{14}$$

Let us assume $x=\beta$, then $z=1-\beta-y$ or $y=1-\beta-z$, we find the operator V_4 as follows

$$\begin{aligned} V_3(x) &= \beta, \\ V_3(y) &= 2y-\beta y-2y^2, \\ V_3(z) &= 2y^2+\beta y-2y-\beta+1. \end{aligned} \tag{15}$$

The extreme of operator V_4 has an invariant curve of the form $x = c$, and for any point taken from this curve, the trajectory tends to $(c,\ (1-c)/2,\ (1-c)/2)$.

It has been shown in Table 4 that the trajectory of operator V_4 of sub-classifications of d.s.q.o. is convergent non-monotonically increasing or decreasing.

For a fixed, simple calculations show that the trajectory of any point from the line $x = c$ tends monotonically to $\left(c, \frac{1-c}{2}, \frac{1-c}{2}\right)$ (see Fig. 6).

5 Conclusions and Future Work

In this paper, we discussed an important class of quadratic stochastic operators d.s. q.o. on finite-dimensional simplex, which is developed based on majorization theory. Furthermore, we presented that the trajectory of many extremes of d.s.q.o. convergences to the center starting at some interior point of the simplex.

Table 4 The numerical solutions for operator V_4

Iterations	x	y	z
x^0	0.2	0.3	0.5
$V(x^0)$	0.2	0.44	0.362
$V^2(x^0)$	0.2	0.3952	0.4048
$V^3(x^0)$	0.2	0.401006	0.398994
$V^4(x^0)$	0.2	0.399801	0.400199
$V^5(x^0)$	0.2	0.40004	0.39996
$V^6(x^0)$	0.2	0.399992	0.400008
$V^7(x^0)$	0.2	0.400002	0.399998
$V^8(x^0)$	0.2	0.4	0.4
$V^9(x^0)$	0.2	0.4	0.4
$V^{10}(x^0)$	0.2	0.4	0.4

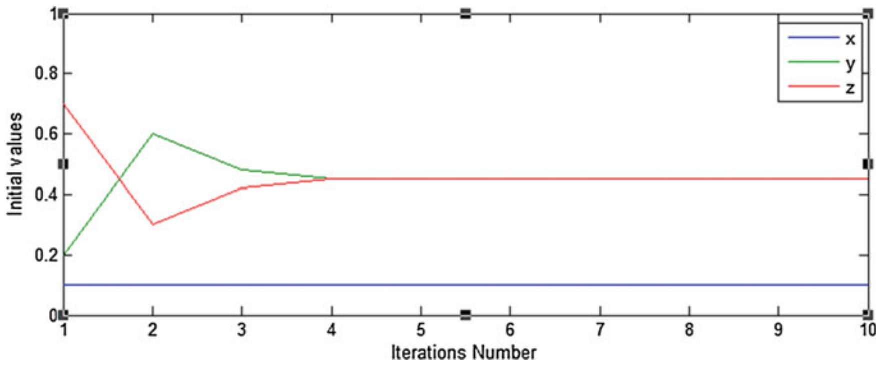


Fig. 6 The trajectory of operator V_4

Meanwhile, we showed that any extreme point of d.s.q.o. on 2D simplex has no periodic points on the interior of the simplex. Moreover, we classified the extreme points of d.s.q.o. on two dimensional 2D simplex. From the analysis, we found three different results of the dynamics classifications of 222 extreme points of d.s.q.o. on two dimensional 2D simplex. In the first classification, 198 extreme points converge to the center of the simplex. In the second classification, there are 18 extreme points, two of the three lines converge to the same value while the third line doesn't change its value from beginning (i.e., constant). In the third classification, 6 extreme points don't converge at any point, they are either never change or become periodic points. Finally, according to the second classifications, we found that there are two different sub-classifications, where the result observed that the trajectories are convergent regularly and monotonically. In our future work, we will study the limit behavior of the trajectories of DSQOs on infinitely dimensional simplex.

Acknowledgements We would like to thank the Kulliyyah of Information and Communication Technology and the Research Management Center in the International Islamic University Malaysia for their supports.

References

1. Lyubich, Y.I., Vulis, D., Karpov, A., Akin, E.: Mathematical structures in population genetics. Biomathematics(Berlin) (1992)
2. Ganikhodzhaev, R.: Quadratic stochastic operators, lyapunov functions, and tournaments. Russ. Acad. Sci. Sbornik Math. **76**(2), 489 (1993)
3. Shahidi, F., Abu Osman, M.: The limit behaviour of trajectories of dissipative quadratic stochastic operators on finite-dimensional simplex. J. Differ. Equ. Appl. **19**(3), 357–371 (2013)
4. Shahidi, F.: On dissipative quadratic stochastic operators. Appl. Math. Inf. Sci. **2**, 211–223 (2008)

5. Ganikhodzhaev, R., Shahidi, F.: Doubly stochastic quadratic operators and birkhoff's problem. Linear Algebra Appl. **432**(1), 24–35 (2010)
6. Shahidi, F.: On the extreme points of the set of bistochastic operators. Math. Notes **84**(3), 442–448 (2008)
7. Vallander, S.: On the limit behavior of iteration sequence of certain quadratic transformations. Sov. Math. Doklady. **13**, 123–126 (1972)
8. Shahidi, F., Ganikhodzhaev, R., Abdulghafor, R.: The dynamics of some extreme doubly stochastic quadratic operators. Middle-East J. Sci. Res. (Math. Appl. Eng.) **13**, 59–63 (2013)
9. Ando, T.: Majorization, doubly stochastic matrices, and comparison of eigenvalues. Linear Algebra Appl. **118**, 163–248 (1989)
10. Olkin, I., Marshall, A.W.: Inequalities: Theory of majorization and its applications. Academic, New York (1979)

Box-Counting Fractal Dimension Algorithm Variations on Retina Images

Mohd Zulfaezal Che Azemin, Fadilah Ab Hamid, Jie Jin Wang, Ryo Kawasaki and Dinesh Kant Kumar

Abstract This research work investigates the influences of FD algorithm variation on the measurement of retinal vasculature complexity. Forty retinal vasculature images from publicly available dataset were subjected to four variations of box-counting FD algorithm. Different positions of box-grid were found to significantly affect the measurement of FD ($p < 0.0001$, $d = 0.746$) due to non-identical vessels captured for measurement. By averaging multiple box-grid placements the FD mean shows no significant difference ($p = 0.12$, $d = 0.124$). Using different smoothing effect (big versus small) results in significantly different FD mean, the variation however was small ($d = 0.211$). The FD of skeletonized vasculature is significantly different than the segmentation ($p < 0.0001$) with a modest effect size ($d = 0.613$). More reliable FD measurement on retinal vasculature could be obtained by averaging the FD values using multiple positions of the grid.

M.Z. Che Azemin (✉) · F.A. Hamid
Kulliyyah of Allied Health Sciences, International Islamic University Malaysia, Kuantan, Malaysia
e-mail: zulfaezal@iium.edu.my

F.Ab. Hamid
e-mail: fadilah.hamid1905@gmail.com

J.J. Wang
Department of Ophthalmology and Westmead Millenium Institute, University of Sydney, Sydney, Australia
e-mail: jiejin.wang@sydney.edu.au

R. Kawasaki
Department of Public Health, Yamagata University, Yamagata, Japan
e-mail: rkawasaki@gmail.com

D.K. Kumar
School of Electrical and Computer Engineering, RMIT University, Melbourne, Australia
e-mail: dinesh@rmit.edu.au

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_27

1 Introduction

Diminishment in the multifaceted nature of the body structure makes physiological structure less ready to adapt to physiological pressure [1]. Aging is identified to be connected with the loss of complexity in the organ configuration of the human body [2, 3], which is associated with operational loss. Such deficiency of complexity has been detected in heart activity [4], magnetic resonance imaging data [5], electroencephalography (EEG) [6], and general physiological measures [7].

Complexity can be quantified based on fractal dimension (FD) analysis. Previous work has reported that FD of the retinal vasculature was inversely associated with eye refractive condition [8] and advancing age [9], and positively associated in diabetic retinopathy complication, a microvascular manifestation of diabetes [10]. Recent studies show associations between retinal vascular changes and either small infarcts detected on brain imaging, or with clinical stroke [11]. A loss of complexity of the cerebral microvasculature may lead to limited collateral formation at cerebral infarction sites [12].

However, it is shown that changes in focus, brightness, color and contrast influence the FD values ($\rho = 0.47 - 0.97$) [13]. Brightness has little effect to FD measurements as long as the median pixel value is kept above 90. Nevertheless, the study concluded that the parameters that control the quality of the retina image during the acquisition needs to be standardized to ensure consistent result of FD especially in comparative study.

In this study, we investigate further the effect of different algorithm used when quantifying the FD. Four variations of box-counting FD algorithm were used to estimate the FD values, namely (1) Different grid placement, (2) Averaging multiple box-grid placements, (3) Different smoothing effects, and (4) Reducing the vessel thickness to 1-pixel thickness (skeletonized vasculature). We used box-counting Fractal Dimension to estimate the FD of the vasculature. It involves two important steps, (1) Placement of box-counting grid and (2) Calculation of the slope *log* Count versus *log* Size. We aimed to investigate the impact of four algorithm variations occurring during these two important steps on FD calculation.

2 Materials and Methods

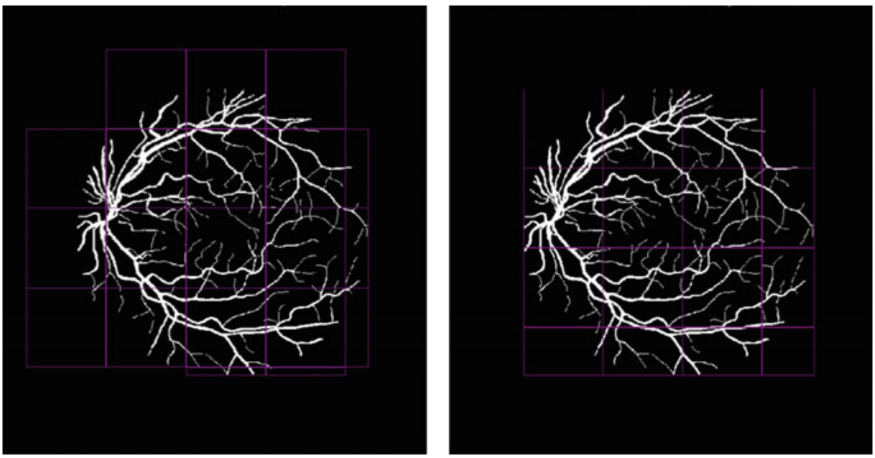
The dataset used to assess the outcome of algorithm variation on the retinal vasculature fractal dimension was a public dataset, Digital Retinal Images for Vessel Extraction (DRIVE) [14]. A set of forty digital images from the DRIVE dataset were downloaded from (http://www.isi.uu.nl/Research/Databases/DRIVE/). The dataset was originally from a screening program in The Netherlands. The retina images were photographed with a Canon CR5 fundus camera using a cropping region of 45°. The dataset also provides a set of manual tracing of a grader as instructed by an ophthalmologist. Using a public dataset has the advantage of making the current experiment repeatable by other researchers. In addition, errors introduced by vessel segmentation algorithm can be avoided and controlled as potential bias. All experiments were conducted using ImageJ plugin FracLac [15] for the FD calculation. The basic strategy for the FD measurement is by efficiently lay a progression of boxes of diminishing size, Box Size, over the retinal blood vessel image and enumerate the number of boxes, Count, that cover the blood vessels. The fractal dimension is then estimated from the gradient of *log* Count versus *log* Box Size.

2.1 Experiment 1: Different Placement of Box-Counting Grid

The positioning of the grid has an effect on the actual count of the vascular network and in turn gives difference FD values. Figure 1 illustrates the grids of same size with different initial position, the number of boxes required to cover the vascular network depends on where the grids are placed. It is shown that while the grid size and the white pixels are the same, the number of boxes needed to cover the white pixels slightly differs. Two grid positions as shown in Fig. 1 were used in this experiment, Grid Position 1 and Grid Position 2.

2.2 Experiment 2: Average of the Different Placement

The reliability of the algorithm can be improved by repeating the box-counting process several times with different initial grid positions. A series of box sizes was



Grid Position 1 Grid Position 2

Fig. 1 Different positioning of the box-counting grid results in variation in FD measurement applied to various positions, the more the positions used in the process would result in more reliable result. The final fractal dimension was taken from the average of fractal dimension calculated from four different grid orientations. This experiment was run twice (Multiple Position 1 versus Multiple Position 2) to compare the algorithm results when repeated in different time instance.

2.3 Experiment 3: The Effect of Smoothing

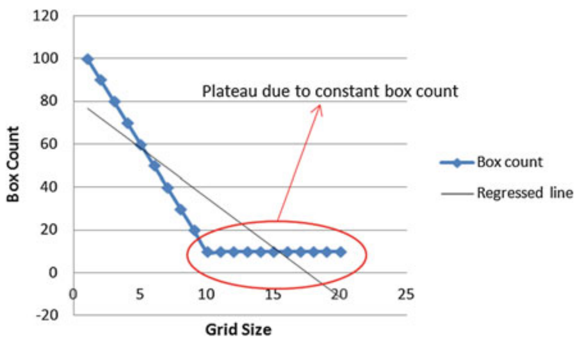
Box-counting fractal dimension algorithm employs a series of increasing box sizes and counts the number of boxes covering the vasculature. The number of boxes needed to cover the vasculature stays the same over an interval of change in box size. This can result in erroneous FD when the slope was taken from a regressed line affected by the data plateau. Figure 2 describes the phenomenon using a linear scale. This problem can be overcome by ignoring the data that contribute to the plateau. The data can be removed two algorithms:

1. Smoothed FD (big) method reduces the data by starting from the smallest box size and keeping only box-count larger than the previous box size.
2. Smoothed FD (small) method reduces the data by starting from the largest size and keeping only box-count smaller than the previous box size.

2.4 Experiment 4: The Effect of Skeletonization

In this experiment we used vessel skeleton to replace vessel segmentation for the FD calculation. Researchers have different definition on vessel skeleton, Azemin et al. [8] defined the vessel skeleton as tracing of the vasculature which is essentially the segmentation of the vessel, while Doubal et al. [16] used skeletal image of

Fig. 2 Illustration of data plateau during FD calculation due to the constant box count



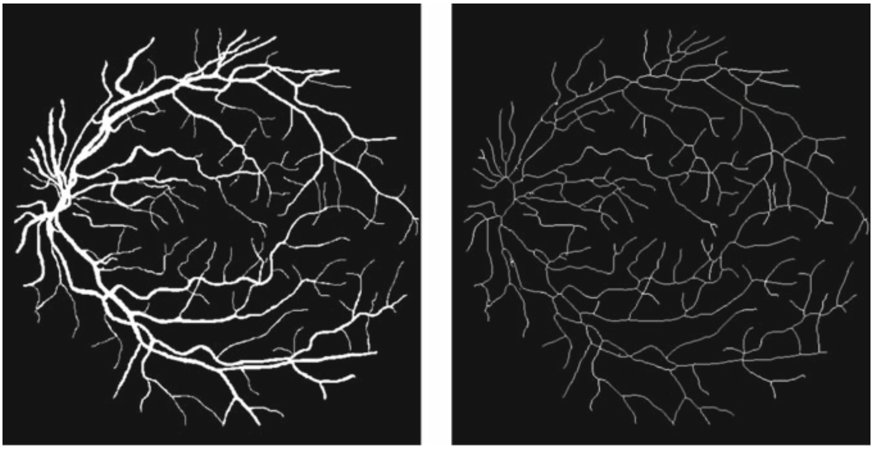


Fig. 3 Comparison between vessel segmentation and skeletonised vasculature

the vasculature when the width of the vessels were reduced to one-pixel thickness as shown in Fig. 3. We used the definition of the latter when describing skeletonization. This process preserves the branching characteristics of the vessels but not their width.

2.5 Statistical Analysis

The Student's t-test was employed to assess the mean difference between a pair of algorithms being evaluated with $p < 0.05$ was taken as statistically significant. The correlation between the algorithms was assessed using Pearson's correlation and effect size was graded with Cohen's method. Effect size was calculated to illustrate the degree of variation relative to the standard deviation of the measures from the reference images. The magnitude of effect size was considered to be small when at 0.2, moderate at 0.5 and large at 0.8 or more based on previous studies [13].

3 Results

Table 1 compares the results of experiment 1–4. Grid Position 1 (experiment 1) was shown to give different FD measurement ($p < 0.0001$, $d = 0.746$) when compared to Grid Position 2. By averaging multiple grid placements (experiment 2) the FD mean of Multiple Position 1 shows no significant difference (p = 0.120, d = 0.124)

Table 1 Effect of changes in box-counting algorithm on fractal dimension

Statistical test	Experiment 1: grid placement	Experiment 2: average grid placement	Experiment 3: smoothing effect	Experiment 4: skeleton effect
T-test (p-value)	<0.0001	0.120	0.009	<0.0001
Effect size (d)	0.746	0.124	0.211	0.613
Pearson's correlation (ρ)	0.887	0.938	0.999	0.930

compared to Multiple Position 2. In experiment 3, using Smoothed FD (big) method results in significantly different FD mean compared to Smoothed FD (small) method, the variation however was small (d = 0.211). Results from experiment 4 indicates that FD of skeletonized vasculature is significantly different compared to vessel segmentation (p < 0.0001, d = 0.613) with a modest effect size.

4 Discussion and Conclusion

Previous comparative analysis on the impact of image quality on FD has demonstrated the need for standardization of retinal image parameters. The current research work explored further how variations in algorithm could influence the FD values. The inconsistent FD measurements were revealed in experiment 1 and 4. The variation due to the grid placement can be compensated by using multiple grid placements and averaging the resulting FD values improves the reliability as shown in results from the experiment 2. FD measured from vessel segmentation [9] is significantly different than the values measured skeletonized vasculature [16], future work should clearly state the type of vascular network used in the FD analysis. While applying different smoothing filters (big versus small) on the *log* Count versus *log* Box Size produced significantly different results, the effect size was considered small, hence, negligible.

These experiments enable us to conclude that more reliable FD measurement on retinal vasculature could be obtained by averaging the FD values using multiple positions of the grid. The types of the vascular network either vessel segmentation or vessel skeleton on where the FD measurement is extracted should also be defined to improve comparability between studies. It should be noted, however, the FD variations will only be affected only when comparing studies using different techniques.

Acknowledgment This research was supported by Ministry of Education, Malaysia under Research Acculturation Grant Scheme RAGS13-029-0092.

References

1. Backx, P.H.: Complexity, confusion and controversy continue complicating the contribution of RyR2 channel phosphorylation to heart function. J. Physiol. **592**, 1911–1912 (2014)
2. Sleimen-Malkoun, R., Temprado, J.J., Hong, S.L.: Aging induced loss of complexity and dedifferentiation: consequences for coordination dynamics within and between brain, muscular and behavioral levels. Front. Aging Neurosci. **6,** 1–1 (2014)
3. Sandu, A.-L., Staff, R.T., McNeil, C.J., Mustafa, N., Ahearn, T., Whalley, L.J., Murray, A.D.: Structural brain complexity and cognitive decline in late life—a longitudinal study in the Aberdeen 1936 birth cohort. Neuroimage **100**, 558–563 (2014)
4. Naraghi, L., Peev, M.P., Esteve, R., Chang, Y., Berger, D.L., Thayer, S.P., Rattner, D.W., Lillemoe, K.D., Kaafarani, H., Yeh, D.D.: Others: the influence of anesthesia on heart rate complexity during elective and urgent surgery in 128 patients. J. Crit. Care **30**, 145–149 (2015)
5. Squarcina, L., De Luca, A., Bellani, M., Brambilla, P., Turkheimer, F.E., Bertoldo, A.: Fractal analysis of MRI data for the characterization of patients with schizophrenia and bipolar disorder. Phys. Med. Biol. **60**, 1697 (2015)
6. Zappasodi, F., Olejarczyk, E., Marzetti, L., Assenza, G., Pizzella, V., Tecchio, F.: Fractal dimension of EEG activity senses neuronal impairment in acute stroke. PLoS ONE **9**, e100199 (2014)
7. Di Ieva, A., Esteban, F.J., Grizzi, F., Klonowski, W., Martín-Landrove, M.: Fractals in the Neurosciences, part II clinical applications and future perspectives. Neurosci. **21**, 30–43 (2015)
8. Che Azemin, M.Z., Mohamad Daud, N., Ab Hamid, F., Zahari, I., Sapuan, A.H.: Influence of refractive condition on retinal vasculature complexity in younger subjects. Sci. World J. 2014 (2014)
9. Azemin, M.Z.C., Ab Hamid, F., Aminuddin, A., Wang, J.J., Kawasaki, R., Kumar, D.K.: Age-related rarefaction in retinal vasculature is not linear. Exp. Eye Res. **116**, 355–358 (2013)
10. Lee, J., Zee, B.C.Y., Li, Q.: Detection of neovascularization based on fractal and texture analysis with interaction effects in diabetic retinopathy. PLoS ONE **8**, e75699 (2013)
11. Aliahmad, B., Kumar, D.K., Hao, H., Unnikrishnan, P., Che Azemin, M.Z., Kawasaki, R., Mitchell, P.: Zone specific fractal dimension of retinal images as predictor of stroke incidence. Sci. World J. 2014 (2014)
12. Kawasaki, R., Azemin, M.Z.C., Kumar, D.K., Tan, A.G., Liew, G., Wong, T.Y., Mitchell, P., Wang, J.J.: Fractal dimension of the retinal vasculature and risk of stroke: a nested case-control study. Neurology **76**, 1766–1767 (2011)
13. Wainwright, A., Liew, G., Burlutsky, G., Rochtchina, E., Zhang, Y.P., Hsu, W., Lee, J.M., Wong, T.Y., Mitchell, P., Wang, J.J.: Effect of image quality, color, and format on the measurement of retinal vascular fractal dimension. Invest. Ophthalmol. Vis. Sci. **51**, 5525–5529 (2010)
14. Staal, J., Abràmoff, M.D., Niemeijer, M., Viergever, M.A., van Ginneken, B.: Ridge-based vessel segmentation in color images of the retina. Med. Imaging, IEEE Trans. **23**, 501–509 (2004)
15. Karperien, A.: FracLac for ImageJ (2013)
16. Doubal, F.N., Hokke, P.E., Wardlaw, J.M.: Retinal microvascular abnormalities and stroke: a systematic review. J. Neurol. Neurosurg. Psychiatry **80**, 158–165 (2009)

Performance of BER Channel Estimation and Tracking Based on DD-NLMS for Indoor and Outdoor Environment in MIMO OFDM

Suzi Seroja Sarnin and Siti Maisurah Sulong

Abstract This paper is focusing on the performance of the proposed extended MIMO channel model based on 3D wave scattering with NLMS channel estimator using simulations, in the context of communication system with MIMO antenna operating in outdoor and indoor environments. Different training rates and different Doppler frequencies are used to track time-variations of the channel. The adaptive algorithm is namely Least Means Square (LMS) algorithm and normalized LMS (NLMS). The performance is evaluated in system BER, for different Doppler frequencies (correspond to different mobility speeds). Simulation results have demonstrated that time-domain adaptive channel estimation and tracking in MIMO OFDM systems based on the DD-NLMS is very effective in slowly to moderate time-varying fading channels. This paper provides analysis, evaluation and computer simulations in MATLAB.

1 Introduction

Nowadays consumer aims at getting high data rate applications and services due to the increasing demand on wireless communication technology. But in wireless, we need to bear a few limiting factors occurs due to multipath propagation that can cause signal distortion. Previous researchers stated that multiple input multiple output orthogonal frequency division multiplexing (MIMO-OFDM) system can enhance the data rates in frequency selective fading channels and achieve high spectral efficiency by simultaneously exploiting the space, time and frequency domains [1].

The term multiple input multiple output refers to the use of an array of antennas for both transmitting and receiving. MIMO approaches show promise of enabling better wireless communications because they mitigate problems inherent in ground

S.S. Sarnin · S.M. Sulong (✉)
Department of Electrical Engineering, UiTM Shah Alam, 40000 Selangor, Malaysia
e-mail: sitimaisurah91@gmail.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_28

to ground links, which are the most common links used by wireless devices, including cell phones and WiFi while OFDM is an effective method in handling frequency-selective fading by converting a wideband frequency selective channel into parallel narrowband frequency subcarrier.

The communication medium is always unknown at the receiver and channel equalization to compensate for multipath shifts cannot be done. In [2], it is proved that channel estimate needs to be calculated before proper decoding can be done at the receiver. Channel estimation technique is introduced to improve accuracy of the received signal. Reference information symbols and interpolation technique can be used by receiver in order to estimate the channel [3].

In this research, adaptive algorithm will be used for channel estimation process. There are a few types of basic algorithm such as LMS, Recursive Least Square (RLS) and Constant Modulus Algorithm (CMA). The proposed Decision Directed Normalized Least Mean Square (DD-NLMS) is upgraded from basic LMS algorithm. Previously, many channel estimation technique with adaptive algorithm have been proposed for MIMO OFDM system. In their paper, researchers proposed LMS with step size ambiguity and evaluate the performance of symbol error rate [4]. Simulation in [5] proved that RLS channel estimation has better performance compare to LMS channel estimation. This is because RLS CE algorithm has better anti-noise as well as tracking capability but still it suffers from high complexity. For [6], they did a research to acquire better approximation of signal by studying behaviour of LMS and NLMS in interference cancellation of speech signal. The analysis conclude that new variable step size algorithms can be developed in order to optimize filter performance.

2 Methodology

Table 1 shows system parameter use for the simulation. The system will consider 2 × 2 MIMO OFDM. Optimum values of forgetting factor that used for indoor channel estimation is shown in Table 2. For outdoor environment, Table 3 provide us with multipath channel profile with total of 4 paths used in the simulation while Table 4 shows values of forgetting factor during outdoor simulation.

2.1 System Parameter

See Table 1.

Table 1 System parameter

Parameter	Value
Channel bandwidth (BW)	1 MHz
Sampling frequency (f_s)	1 MHz
RF frequency (f_c)	2.4 GHz
Channel length (L)	4 taps
FFT size (N_{FFT})	128
Number of subcarriers (K)	128
Subcarrier spacing (Δf)	7.813 kHz
Cyclic prefix length (N_{cp})	8 samples
Modulation type	QPSK
Useful symbol duration (T_B)	128 μs
OFDM symbol duration (T_s)	136 μs
Maximum delay spread (τ_{max})	3 μs

2.2 *Indoor Channel Estimation and Tracking*

See Table 2.

2.3 *Outdoor Channel Estimation and Tracking*

See Tables 3 and 4.

Table 2 Optimum values of forgetting factor

Training rate (Hz)	25 %	10 %	4 %
F_d = 10	0.995	0.995	0.997
F_d = 50	0.995	0.995	0.997
F_d = 100	0.995	0.995	0.997
F_d = 200	0.995	0.995	0.997

Table 3 Multipath channel profile

Path index, l	Path delay, $\tau_l\,(\mu s)$	Path power, P_l (dB)
1	0	0.0
2	1.0	−2.15
3	2.0	−3.058
4	3.0	−4.045

Table 4 Optimum values of the forgetting factor for outdoor

Training rate (Hz)	10 %	4 %	2 %
f_d = 40	0.992	0.995	0.997
f_d = 75	0.991	0.993	0.995

2.4 System Model

Figure 1 shows the MIMO OFDM system that operate with 2-transmit and 2-receive antennas for indoor and outdoor environment. The proposed NLMS channel estimator is employed to track the time variation of the channel estimates. The channel is assumed to be Rayleigh Fading, The individual channel impulse responses between transmit-receive antenna pairs are assumed to be uncorrelated. Channel estimation is performed in time-domain followed by zero-forcing equalization in frequency-domain. The proposed estimators are then extended to perform DD channel tracking in time-domain. For indoor, the mobile speed is 18 km/h while for outdoor it is roughly 34 km/h.

In the transmitter, the incoming bit stream is modulated by mapping to a constellation scheme (PSK or QAM), and subdivided into M_t parallel substreams that are transmitted from an independent OFDM transmitter. Each K data symbols are serial-to-parallel (S/P) converted, and then modulated by passing through a K-tone inverse fast Fourier transform (IFFT) block. The IFFT block converts the data symbols from frequency-domain to time-domain. Suppose that the input sequence into the IFFT block, corresponding to the pth transmit antenna at the transmission time of the mth OFDM symbol, is represented by $X_p\ (m, k)$, $(k = 0,1,\ldots, K - 1)$. Then, the OFDM symbols at the output of the IFFT are given by

$$x_p(m,s) = \frac{1-}{K}\sum_{k=0}^{K-1} X_p(m,k)e^{j\frac{2\pi ks}{K}} \tag{1}$$

Cyclic prefix is added to each OFDM symbol to eliminate intersymbol interference (ISI). The symbols are parallel-to-serial (P/S) converted and then the training sequences are periodically multiplexed with the OFDM data symbols in time-domain. At the receiver, noisy received signal assumed to be additive and

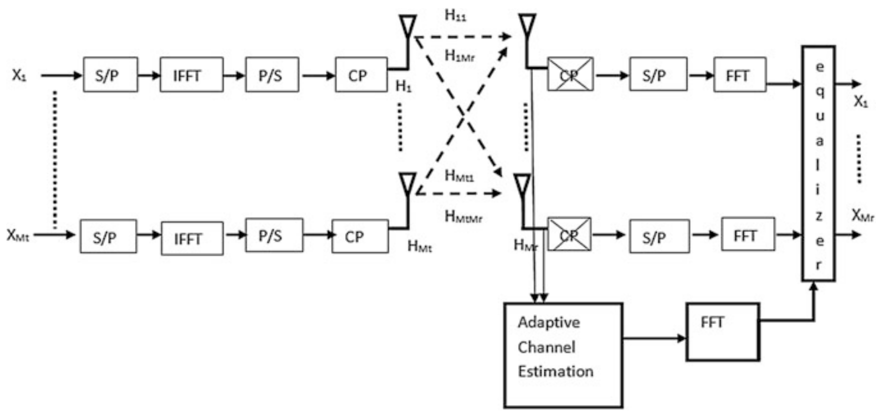


Fig. 1 Baseband MIMO OFDM system model

uncorrelated with input signal. Received signal during OFDM symbol transmission in matrix form can be written as

$$\tilde{\boldsymbol{y}}_q(m) = \tilde{\boldsymbol{\chi}}(m)\tilde{\boldsymbol{h}}_q(m) + \boldsymbol{v}_q(m) \tag{2}$$

where

$$\tilde{\boldsymbol{y}}_q(m) = \left[y_q(m,0), y_q(m,1), \ldots, y_q(m,K-1)\right]^T \tag{3}$$

$$\tilde{\boldsymbol{\chi}}(m) = \left[\tilde{\boldsymbol{X}}_1(m), \tilde{\boldsymbol{X}}_2(m), \ldots, \tilde{\boldsymbol{X}}_{M_t}(m)\right] \tag{4}$$

with

$$\tilde{\boldsymbol{X}}_p(m) = \left[\boldsymbol{x}_p^H(m,0), \boldsymbol{x}_p^H(m,1), \ldots, \boldsymbol{x}_p^H(m,N-1)\right]^H \tag{5}$$

is a $N \times L$ Toeplitz matrix contains delayed versions of the input data sent from the pth transmit antenna, where at the time interval i, the transmitted and delayed data symbols are arranged in a vector as

$$\boldsymbol{x}_p(m,i) = [x_p(m,i), x_p(m,i-1), \ldots, x_p(m,i-L+1)] \tag{6}$$

The $M_t\ L \times 1$ vector $\boldsymbol{h}_q(m)$ contains the stacked vectors corresponding to the individual channel impulse responses from the transmit antennas to the qth receive antenna, and it may be written as

$$\tilde{\boldsymbol{h}}_q(m) = \left[\boldsymbol{h}_{1q}^T(m), \boldsymbol{h}_{2q}^T(m), \ldots, \boldsymbol{h}_{M_t q}^T(m)\right]^T \tag{7}$$

where $\mathbf{h}_{pq}$ (m) is the channel impulse response between the pth transmit and the qth receive antennas;

$$\boldsymbol{v}_q(m) = [v(m,0), v(m,1), \ldots, v(m,K-1)]^T \tag{8}$$

is a $K \times 1$ vector representing additive white Gaussian noise (AWGN) at the qth receive antenna with complex elements. Then, the frequency-domain received signal at the qth antenna over the kth tone at the time m, Y_q (m, k), can be expressed as

$$Y_q(m,k) = \sum_{p=1}^{M_t} H_{pq}(m,k)X_p(m,k) + V_q(m,k) \tag{9}$$

where ***Xp** (m, k), **Hpq** (m, k), and **Vq** (m, k)* denote the kth data sample of the transmitted OFDM symbol from the pth antenna, the channel coefficient of the kth tone between the pth transmit and the qth receive antennas, and the noise at the qth receive antenna on the kth tone, respectively.

2.5 LMS Based Channel Estimation

The tap-weight vector update equation is given by

$$\hat{\boldsymbol{h}}(n+1) = \hat{\boldsymbol{h}}(n) + \mu x(n)e^{*}(n) \tag{10}$$

where μ is a positive constant called the step-size parameter,

$$\boldsymbol{x}(n) = [x(n), x(n-1),\ldots,x(n-p)]^{T} \tag{11}$$

and

$$\boldsymbol{h}_n = [h_n(0), h_n(1), \ldots, h_n(p)]^{T}] \tag{12}$$

The simplicity of LMS algorithm stems from the fact that the update for the lth coefficient is given by

$$h_l(n+1)(k) = h_l(n) + \mu x(n-l)e^{*}(n) \tag{13}$$

The optimum value of the step size parameter μ is chosen through extensive simulations.

2.6 NLMS Algorithm

The normalized LMS (NLMS) is a practical version of the LMS algorithm, which provides one way to automate this choice of varying step size. The NLMS is formulized as

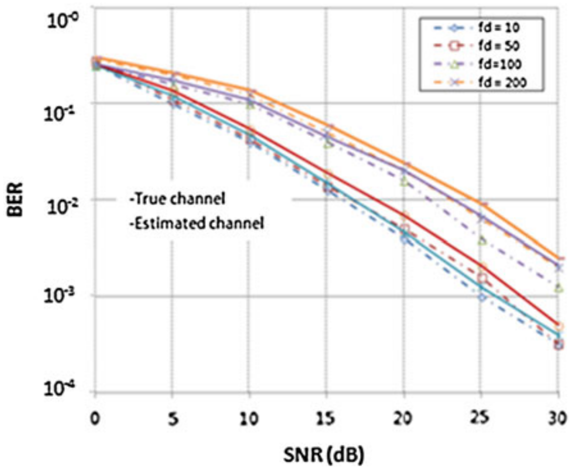
$$\boldsymbol{h}(n) = \boldsymbol{h}(n-1) + \frac{\tilde{\mu}}{\alpha + x^{H}(n)x(n)} e^{*}(n)\boldsymbol{x}(n) \tag{14}$$

where μ represents an adaptation constant, usually $0 < \mu < 2$, that guarantees the convergence of the normalized LMS algorithm in the mean square sense, and α is an arbitrary positive constant.

3 Simulation Results

The channel tracking performance of the DD-NLMS MIMO-OFDM system has been evaluated by simulation for indoor and outdoor environment. A focus is on the capabilities of the NLMS algorithms to track the time variations of the channel at different Doppler frequencies and training rate.

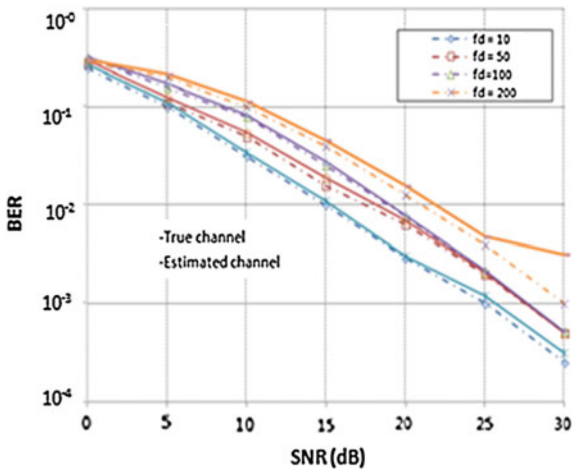
Fig. 2 The BER performance versus SNR, with DD-NLMS channel tracking and 10 % training data



3.1 Indoor Environment

The BER performance of DD-NLMS channel estimation and tracking for indoor environment with different doppler frequencies and training rates of 10 and 4 % respectively was presented in Figs. 2 and 3. It shows that the performance of this algorithm is very sensitive to the step size parameter and by properly choosen μ, which is optimized by simulation can make NLMS channel estimates converge to desired value. By increasing Doppler frequency and SNR, small increase in the BER is experienced. This degradation in the BER is referred to the slow convergence of the NLMS algorithm which negatively effects on its performance in tracking the variation of the channel as Doppler frequency increases. Its already

Fig. 3 The BER performance versus SNR, with DD-NLMS channel tracking and 4 % training data



known that small values of step size will lead to slow convergence rate while faster convergence can be obtain using suitable large value of step size. This is because too large value of step size can cause unstable of the estimator.

3.2 Outdoor Environment

Adaptive channel estimation method based on NLMS algorithm has been presented in Figs. 4, 5 and 6 for Doppler frequencies of 40 Hz, 75 Hz and training rates of 10, 4 and 2 % respectively for MIMO OFDM system for outdoor environment. The results show that the DD-NLMS has good tracking performance at low SNR and high training rates as both Doppler frequency and SNR increase, some degradation in the performance is experienced especially at reduced training rates. This degradation in the performance is attributed to the significant errors induced due to the inherent slow convergence of the NLMS and its maladjustments which increases as the SNR is increased. By fixing step size under different Doppler frequencies, it can reduce system performance.

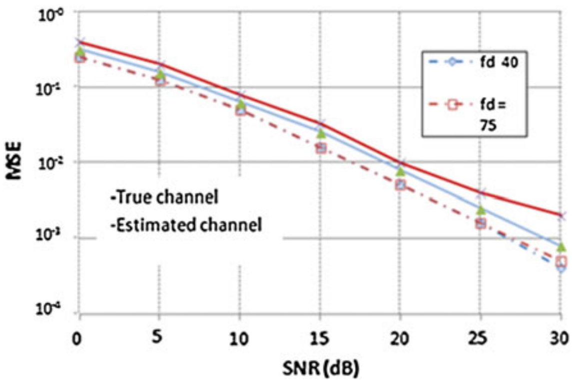


Fig. 4 The BER performance versus SNR with DD-NLMS channel tracking and 10 % training data

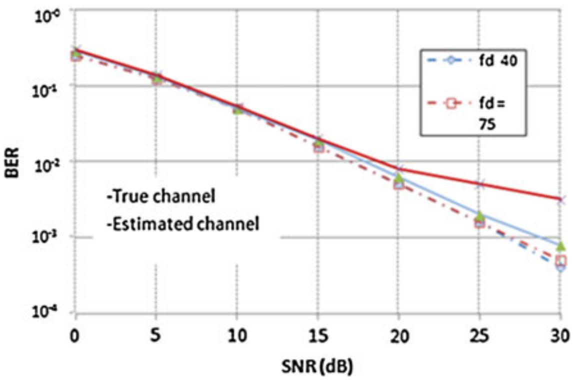


Fig. 5 The BER performance versus SNR with DD-NLMS channel tracking and 4 % training data

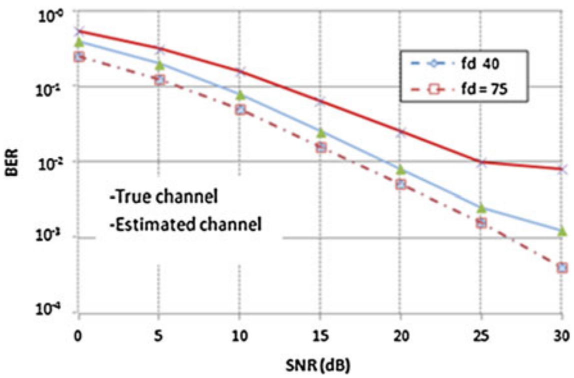


Fig. 6 The BER performance versus SNR with DD-NLMS channel tracking and 2 % training data

4 Conclusion

In indoor situation, the NLMS channel estimates converge to the desired values if the step size is properly chosen. For outdoor environment, a time adaptive channel estimation method based on NLMS algorithm has better tracking performance even in moderate time-varying channels at low SNR and low mobility. It is computationally simple, numerically robust but it has drawback of slow convergence which makes it unreliable in fast time-varying channels. All in all, the NLMS based channel estimation might be favored at low SNR and higher mobility. The proposed idea is better because time domain channel estimation and tracking have shown the accurate channel estimate, reliable tracking performance and significant reduction in the computational complexity.

Acknowledgments This work was supported in part by the Faculty of Electrical Engineering University Teknologi MARA, Shah Alam Selangor.

References

1. Van de Beek, J.J., Edfors, O., MSandell, M., Wilson, S.K., Borjesson, P.O.: On channel estimation in OFDM systems. In: Vehicular Technology Conference, vol. 2 (1995)
2. Uysal, M., Al-Dhahir, N.: A space time block coded OFDM for unknown frequency selective fading channels. IEEE Trans. Commun. **5**(10), 393–395 (2003)
3. Sklar, B.: Digital Communication Fundamentals and Application, 2nd edn. Pearson Education, Asia (2001)
4. Pandey, A., Malviya, L., Sharma, V.: Comparative study of LMS and NLMS algorithms in adaptive equalizer. Int. J. Eng. Res. Appl. **2**(3), 1584–1587 (2012)
5. Rana, M.: Performance comparison of LMS and RLS channel estimation algorithms for 4G MIMO OFDM systems. In: 14th International Conference on Computer and Information Technology, (ICCIT), pp. 22–24 (2011)
6. Kabir, A., Rahman, K., Hussain, I.: Performance study of LMS and NLMS adaptive algorithms in interference cancellation of speech signals. J. State Univ. Bangladesh. **1**:57–65 (2007)

Space Division Multiplexing in Multimode Fiber for Channel Diversity in Data Communications

Angela Amphawan, Yousef Fazea and Mohamed Elshaikh

Abstract Space division multiplexing (SDM) has recently gained eminence as a means to alleviate data traffic congestion to future-proof current network infrastructure. This paper reports on SDM of a new spiral-phased wavefront comprising modified Laguerre-Gaussian (LG) and Hermite-Gaussian (HG) modes on a wavelength of 1550.12 nm over a 2 km-long MMF. Power coupling coefficients, degenerate mode group delays and bit-error rates are analyzed for different vortex orders.

1 Introduction

The escalation of data traffic and machine-to-machine connections has catalyzed researchers around the globe to look for possible strategies to increase the capacity of optical fiber networks and optimize the limited optical spectrum [1–5]. While wavelength division multiplexing (WDM) has been the workhorse of large data networks for decades, recently, the notion of using distinct spatial mode profiles as signatures for individual channels—a technique termed space division multiplexing (SDM)—has gained significant attention. SDM is a revolutionary technology which increases the aggregate data rate by a factor of the number of modes that can be precisely generated and demultiplexed. SDM is of immense advantage to telecommunications companies

A. Amphawan (✉) · Y. Fazea
Integrated Optics Group, School of Computing, Universiti Utara Malaysia,
Sintok, Kedah, Malaysia
e-mail: angela.amphawan.dr@ieee.org

A. Amphawan
Research Laboratory of Electronics, Massachusetts Institute of Technology,
Cambridge, MA, USA

M. Elshaikh
School of Computer and Communication Engineering, Universiti Malaysia Perlis,
Arau, Perlis, Malaysia

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_29

as no extra spectrum is required. This is particularly important considering the scarcity of the optical spectrum and intense government regulation.

In SDM, signature spatial mode profiles are generated and launched into a few mode fiber (FMF) or multimode fiber (MMF) as independent data streams [6]. These signature mode profiles are generated by means of spatial light modulator encoding [7–10], adaptive optical optimization [11–14], mode converters/couplers [15–17] and add-drop multiplexers [18, 19].

Although Laguerre-Gaussian (LG) modes [20] are predominantly used, Hermite-Gaussian (HG) modes have recently been explored. A one-dimensional HG mode profile has been generated using a passive beam shaper on fused silica at the MMF input core for collimated Gaussian field [21, 22], exciting mid-order modes and achieving a data rate of 10 Gbps for a MMF length of 220 m. The experiment was then extended in [23, 24] to two-dimensional symmetrical profiles but with opposite phase distributions on both axes, achieving a data rate of 10 Gb/s over 250 m.

The main distinction of this paper, is in the selective excitation of a spiral-phased wavefront comprising modified LG and HG modes for different vortex orders ($m = 1$, $m = 2$, $m = 3$ $m = 4$). Also, in contrast, in this work, the predetermined mode groups are generated using a vertical-cavity surface-emitting laser (VCSEL) array instead of an etched silica mask.

This work focuses on SDM of a new combination of spiral-phased modes comprising modified LG and HG modes modeled using VCSEL array and a vortex lens. The spiral-phased LG and HG modes are coupled into a graded-index MMF. The spatial electric fields, degenerate mode group delays and power-coupling coefficients into individual fiber modes are analyzed for various vortex orders, m.

This paper proceeded as follows. Section 2 elucidates the modeling of a new MDM scheme for the spiral-phased wavefront comprising modified LG and HG modes. Section 3 presents the results and discussions and the conclusion of the paper presented in Sect. 4.

2 Simulation of the Spiral-Phased LG and HG Modes

SDM of spiral-phased LG and HG modes was simulated in Optsim 5.2 simulator [25] and Matlab [26], as depicted in Fig. 1. The transmitter consists of a pair of LG 20 and HG 20 modes on each of the two VCSELs operating on wavelengths 1550 nm and 1550.12 nm. Each VCSEL is driven by PRBS electrical signals and modulated to non-return-to-zero optical pulses. The power from each VCSEL array is assumed to be emitted uniformly into 6 μm beams. The LG modes are extensions of the simple Gaussian mode shape to higher order modes in a cylindrical coordinate system, and often used to describe the transverse mode shapes at the output of a VCSEL. They are characterized by azimuthal index l, radial index m, spot size w_0, and radius of curvature R_0. For fixed values of w_0 and R_0, the family of LG

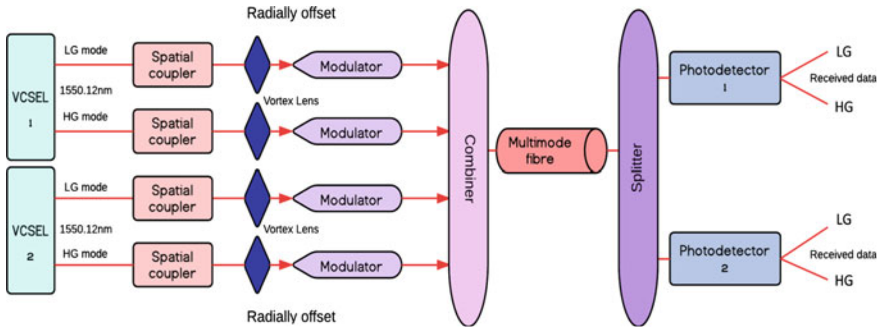


Fig. 1 MDM model for spiral-phased comprising modified LG and HG modes

modes for different azimuthal and radial indices are mutually orthogonal. The transverse electrical field profile of the launched LG_{ml} mode in the first VCSEL is expressed as:

$$\psi_{ml}(r,\phi) = \alpha \cdot \left(\frac{2r^2}{w_0^2}\right)^{L/2} \cdot L_m^l\left(\frac{2r^2}{w_0^2}\right) \cdot \exp\left(-\frac{r^2}{w_0^2}\right) \cdot \exp\left(j\frac{\pi r^2}{\lambda R_0}\right)\begin{cases} \cos(L\phi), & l \geq 0 \\ \sin(L\phi), & l < 0 \end{cases} \tag{1}$$

where α is normalization constant, $L = |l|$, λ is the field wavelength, and L_m^l is generalized Laguerre polynomial. At the beam waist, the inverse of R_0 is zero, indicating a flat phase front. At any distance to the left or right of the waist, the beam begins to diverge and R_0 become finite. Figure 2 shows the amplitude and phase distribution of the generated LG mode. The generated HG may be expressed as [27]:

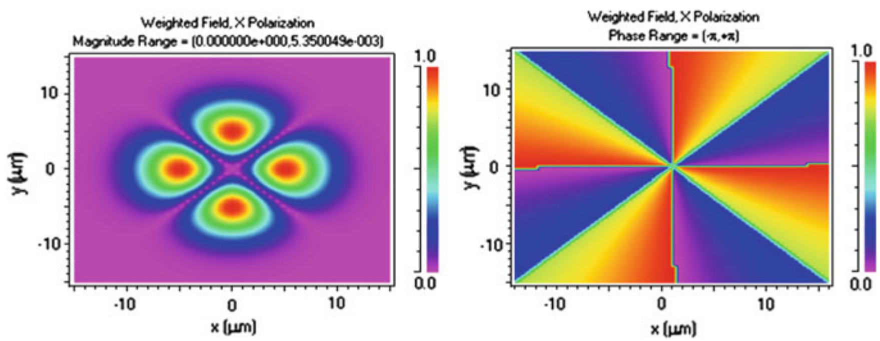


Fig. 2 Magnitude distribution (*left*) and phase distribution (*right*) of transverse electric field of LG_{20} mode from VCSEL array

$$\psi_{lm}(x,y) = \alpha \cdot H_l\left(\frac{\sqrt{2}(x-b)}{w_{o\,x}}\right) \cdot \exp\left(\frac{-(x-b)^2}{w_{0x}^2}\right) \cdot \exp\left(\frac{j\pi(x-b)^2}{\lambda R_{0x}}\right) \times H_m\left(\frac{\sqrt{2}y}{w_{0y}}\right) \exp\left(\frac{-y^2}{w_{0y}^2}\right) \cdot \exp\left(\frac{j\pi y^2}{\lambda R_{0y}}\right) \tag{2}$$

$w_{0x} = 2$ μm and $w_{0y} = 2$ μm are the x and y spot sizes respectively, b is the radial offset from the core center; $R_{0x} = 0$ and $R_{0y} = 0$ are the x and y radii of curvature respectively; H_l and H_m are Hermite polynomials.

Figure 3 shows the magnitude (left) and phase distribution (right) of the transverse electric field profile of HG mode from VCSEL array. The VCSEL is connected to a vortex lens used to transform the flat phase front to a spiral phase front. The focal length of lens is maintained to $f = 8.0$ mm and the vortex order is varied each run ($m = 1$, $m = 2$, $m = 3$ and $m = 4$). The applied phase transformation is expressed as [25]:

$$t(x,y) = \exp\left[-j\left(\frac{n\pi r^2}{2\lambda f} + m\theta\right)\right] \tag{3}$$

$$r = x^2 + y^2 \tag{4}$$

$$\theta = \tan^{-1}(y/x) \tag{5}$$

where x and y are transverse coordinates in the x-y plane, λ is the signal wavelength, m is the vortex order, n is the material index and f is the lens focal length.

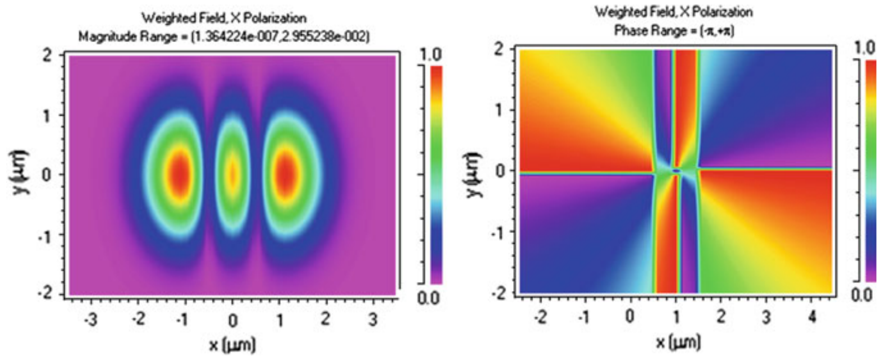


Fig. 3 Magnitude distribution (*left*) and phase distribution (*right*) of transverse electric field of HG_{20} mode from VCSEL array

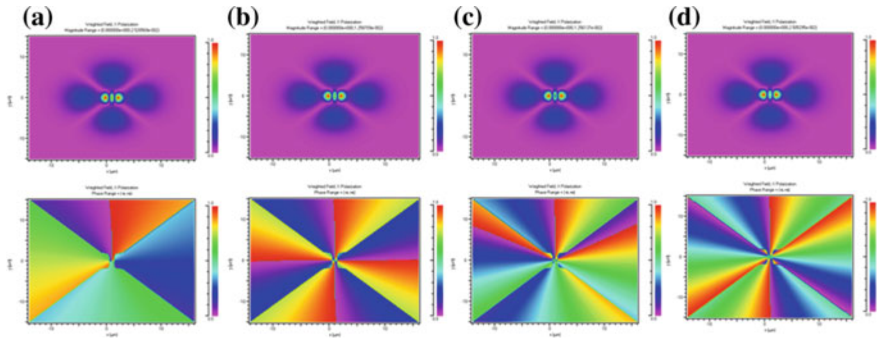


Fig. 4 Magnitude (*top*) and phase distributions (*bottom*) of the transverse electric field of both spiral-phased modified LG and HG modes; the focal length f is maintained to 8 mm and vortex order m is varied each run to **a** m = 1, **b** m = 2, **c** m = 3, **d** m = 4

Fig. 5 Measured refractive index profile of manufactured MMF in MDM model

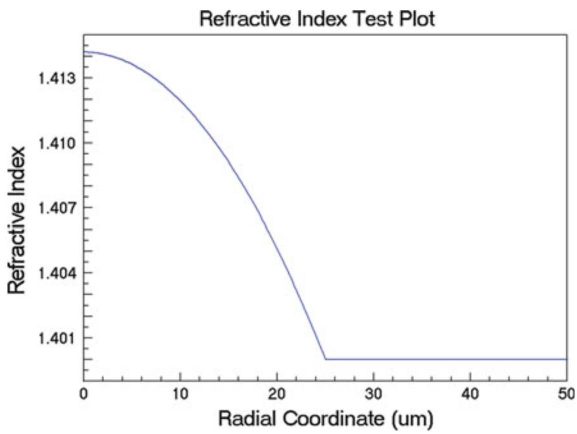


Figure 4 shows the magnitude (top) and phase distributions (bottom) of spiral-phased wavefront comprising modified LG and HG modes for each vortex order run. The modified LG and HG modes radial offset is set to 1 μm.

The SDM signals are then propagated through a 2 km-long manufactured MMF. The assumed value for attenuation is 1.5 dB/km with consideration of power modal coupling. The measured refractive index profile of the manufactured MMF is shown in Fig. 5.

Two photodetectors are used to retrieve the combined signals. The modes are then demultiplexed at the photodetector. The transverse electric field distributions, power-coupling coefficients, degenerate mode group delays and bit error rates are analyzed for different vortex order. The results and analysis are presented in Sect. 3.

3 Result and Discussion

Figure 6 qualitatively shows the transverse electric field after propagating through the MMF for various radial offsets at wavelength 1550.12 nm. The output transverse electric fields are indistinguishable from the input transverse electric fields due to power modal coupling. In Fig. 6b, for m = 2, the transverse electric field indicates a combination of more adjacent modes compared to previous cases.

For quantitative evaluation of the effects of different vortex orders, m. Figure 7 shows the power coupling coefficients versus modal delay after the photodetector for the wavelength 1550.12 nm based on noninterferometric modal decomposition [28]. In Fig. 7a, the power is coupled mostly into higher-order modes and lower-order modes. Hence the time delay between modes is very high, resulting in an erratic pulse for the case of m = 1. In Fig. 7b, the power is coupled mostly in higher-order modes producing the smallest pulse width for m = 2. In Fig. 7c, as observed, the power is coupled more dominantly in higher-order modes. Therefore the differential time delay between modes is minimized, leading to a narrow width pulse. In Fig. 7d, most of the power is relatively coupled to higher-order modes and medium order-modes. The time delay differences between modes are small.

Figure 8 shows the power coupling coefficients and the propagation delay of the degenerate mode groups. From the curves, it is evident that the best differential mode delay is achieved when the vortex order m = 2. Symmetric and anti-symmetric modes are observed in the power coupling coefficients of degenerate modes due to opposing propagation constants. A comparison of BERs of different vortex orders are examined in Table 1.

The lowest BER is attained for m = 2, followed by m = 3, m = 4, and m = 1. The BER indicates that acceptable BER is attained for vortex orders, m = 2 and m = 3.

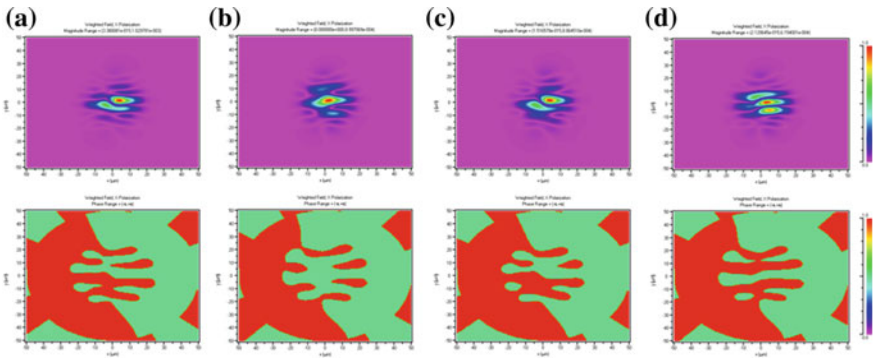


Fig. 6 Spatial electric field magnitude distribution (*top*) and phase distribution (*bottom*) for Channel 2 after the propagation through MMF whereby the focal length, f = 8 mm and vortex order, m is varied each run to **a** m = 1, **b** m = 2, **c** m = 3, **d** m = 4

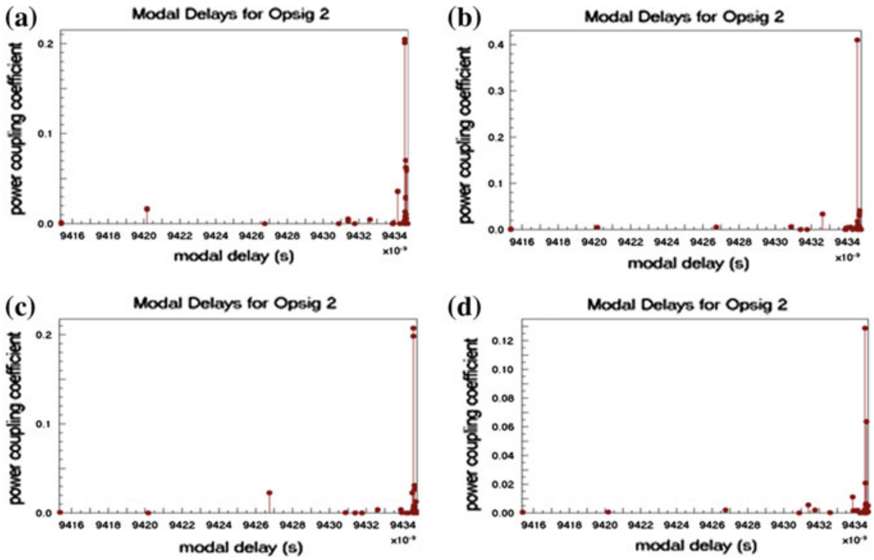


Fig. 7 Power coupling coefficient versus modal delay at the MMF output for Channel 2 for different vortex order, whereby the focal length, $f = 8$ mm and vortex order, m is varied: **a** $m = 1$, **b** $m = 2$, **c** $m = 3$, **d** $m = 4$

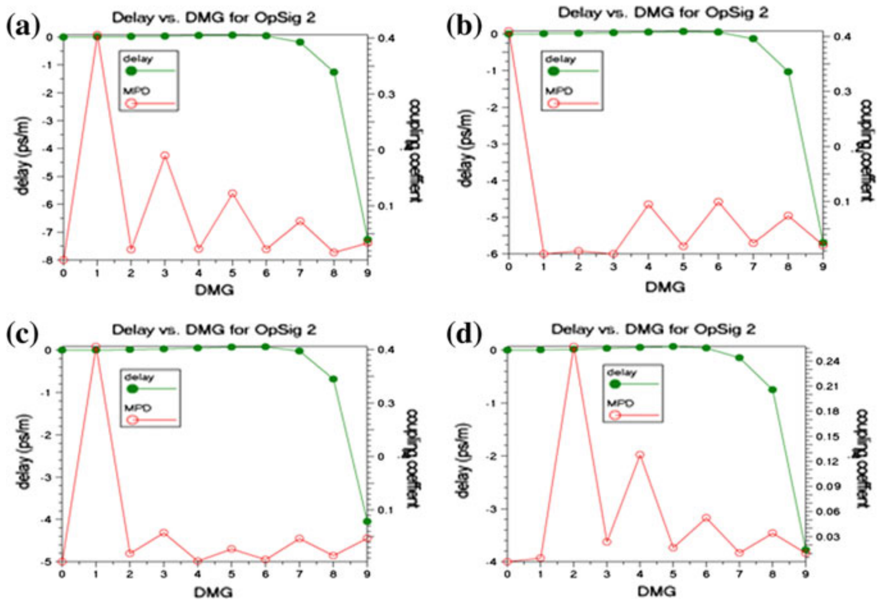


Fig. 8 Relative group delay of fiber for different vortex order for Channel 2: whereby the focal length, $f = 8$ mm and vortex order, m is varied each run to **a** $m = 1$, **b** $m = 2$, **c** $m = 3$, **d** $m = 4$

Table 1 The effects of BER for different vortex orders

Vortex order	BER
$m = 1$	1.39×10^{-5}
$m = 2$	4.51×10^{-15}
$m = 3$	1.14×10^{-10}
$m = 4$	1.07×10^{-8}

4 Conclusion

This paper reports on SDM of a new combination of spiral-phased LG 20 and HG 20 modes at 1550.12 nm. A data rate of 44 Gbps over a 2 km MMF was attained. The lowest BER is obtained for $m = 2$. The SDM model would be valuable in parallel optical interconnects for multiple-input-multiple-output (MIMO) data communications [29].

References

1. Talebi, S., et al.: Spectrum assignment in optical networks: A multiprocessor scheduling perspective. IEEE/OSA J. Opt. Commun. Network. **6**(8), 754–763 (2014)
2. Wang, R., Mukherjee, B.: Spectrum management in heterogeneous bandwidth optical networks. Opt. Switching Network. **11**(A), 83–91 (2014)
3. Cisco Visual Networking Index: Global mobile data traffic forecast update 2014–2019 white paper. Cisco, San Jose (2015)
4. Amphawan, A., Khair, M.A.M., Hasan, H.: Multimedia traffic routing in multilayer WDM networks. Network Complex Syst. **2**(3), 1–9 (2012)
5. Nisar, K., Amphawan, A., Hassan, S.B.: Comprehensive structure of novel voice priority queue scheduling system model for VoIP over WLANs. Int. J. Adv. Pervasive Ubiquitous Comput. (IJAPUC) **3**(4), 50–70 (2011)
6. van Uden, R.G.H., et al.: Ultra-high-density spatial division multiplexing with a few-mode multicore fibre. Nat. Photon. **8**(11), 865–870 (2014)
7. Amphawan, A.: Binary spatial amplitude modulation of continuous transverse modal electric field using a single lens for mode selectivity in multimode fiber. J. Mod. Opt. **59**(5), 460–469 (2012)
8. Amphawan, A., O'Brien, D.: Holographic mode field generation for a multimode fiber channel. In: IEEE International Conference on Photon (ICP2010). IEEE, Langkawi (2010)
9. Carpenter, J., Thomsen, B.C., Wilkinson, T.D.: Degenerate mode-group division multiplexing. J. Lightwave Technol. **30**(3946), 3946–3952 (2012)
10. Carpenter, J., Wilkinson, T.D.: All optical mode-multiplexing using holography and multimode fiber couplers. J. Lightwave Technol. **30**(12), 1978–1984 (2012)
11. Arik, S.O., Askarov, D., Kahn, J.M.: Adaptive frequency-domain equalization in mode-division multiplexing systems. J. Lightwave Technol. **32**(10), 1841–1852 (2014)
12. Arik, S.O., Kahn, J.M.: Adaptive MIMO signal processing in mode-division multiplexing. In: Photonics Society Summer Topical Meeting Series. IEEE (2014)
13. Amphawan, A.: Binary encoded computer generated holograms for temporal phase shifting. Optics Exp. **19**(23), 23085–23096 (2011)
14. Amphawan, A.: Backlighting for alignment of optics in first diffraction order path. In: Proceedings of SPIE International Conference on Applications of Optics and Photon (2011)

15. Amphawan, A., Nedniyom, B., Al Samman, N.M.: Selective excitation of LP01 mode in multimode fiber using solid-core photonic crystal fiber. J. Modern Optics **60**(20), 1675–1683 (2013)
16. Jung, Y., et al.: Dual mode fused optical fiber couplers suitable for mode division multiplexed transmission. Opt. Express **21**(20), 24326–24331 (2013)
17. Tsekrekos, C.P., Syvridis, D.: All-Fiber broadband LP02 mode converter for future wavelength and mode division multiplexing systems. IEEE Photon. Technol. Lett. **24**(18), 1638–1641 (2012)
18. Miller, D.A.B.: Reconfigurable add-drop multiplexer for spatial modes. Opt. Express **21**(17), 20220–20229 (2013)
19. Fang, L., Jia, H.: Mode add/drop multiplexers of LP02 and LP03 modes with two parallel combinative long-period fiber gratings. Opt. Express **22**(10), 11488–11497 (2014)
20. Amphawan, A., Wa'el Ali Alabdalleh, : Simulation of properties of the transverse modal electric field of an infinite parabolic multimode fiber. Microwave Optical Lett. **54**(6), 1362–1365 (2012)
21. Kwok, C., et al.: Novel passive launch scheme for ultimate bandwidth improvement of graded-index multimode fibers. In: Optical Fiber Communication Conference. Optical Society of America (2010)
22. Geng, L., et al.: Efficient line launch for bandwidth improvement of 10 Gbit/s multimode fibre links using elliptical Gaussian beam. ECOC, We, 6 (2010)
23. Geng, L., et al.: Symmetrical 2-D hermite-gaussian square launch for high bit rate transmission in multimode fiber links. In: Optical Fiber Communication Conference. Optical Society of America (2011)
24. Li, Y., et al. 20 Gb/s mode-group-division multiplexing employing Hermite-Gaussian launches over worst-case multimode fiber links. In: Optical Fiber Communication Conference. Optical Society of America (2014)
25. Rsoft Design Group, I: OptSim user guide (2010)
26. MATLAB. Natick, Massachusetts, United States, The MathWorks, Inc. Release (2013)
27. Xue, X., Kirk, A.G.: Transverse modal characterization of VCSELs based on intensity measurement. In: Optoelectronic Interconnects VII; Photon. Packaging and Integration II. SPIE, San Jose (2000)
28. Amphawan, A., O'Brien, D.: Modal decomposition of output field from holographic mode field generation in a multimode fiber channel. In: IEEE International Conference on Photon (ICP2010). IEEE, Langkawi (2010)
29. Amphawan, A., Review of optical multiple-input-multiple-output techniques in multimode fiber. Opt Eng 50:102001 (2011)

A Self-organizing Approach: Time Synchronization for the HeNodeBs in Heterogeneous Network

Mohammad Kamrul Hasan, A.F. Ismail, Aisha H. Abdalla, H.A. M. Ramli, Wahidah Hashim, Abdur Razzaque and M.H. Khairolanuar

Abstract In Heterogeneous Network (HetNet) for LTE/LTE-A system femtocells (HeNodeBs) are designed and implemented to extend coverage and capacity. The arbitrary usage of HeNodeBs supported for indoor coverage is linked through a third-party internet backhaul to the HeNodeB managing server, and this is a segment of the operators end. To sustain HeNodeBs to the control function IEEE1588, master slave strategy is put in effect. Nevertheless, caused by the absence of internet or unfortunate connectivity for a prolonged period HeNodeBs undergoes synchronization difficulties that leads to frequency mismatch With this paper, a kind of self-organizing method is suggested for time synchronization in Heterogeneous Network, and that is standardized using IEEE1588 master slave method along with Precision Timing Protocol (PTP). The proposed approach is two way message transmission system, adopting the Least-squares Linear Regression Algorithm (LSRA) to reduce offsets as well as frequency misalignments (drift).

Keywords Femtocell · Synchronization · Heterogeneous network · LTE/LTE-Advanced · IEEE 1588 · HeNodeB

M.K. Hasan (✉) · A.F. Ismail · A.H. Abdalla · H.A.M. Ramli · M.H. Khairolanuar
Department of Electrical and Computer Engineering, International Islamic University, 50728 Ombak, Malaysia
e-mail: hasankamrul@ieee.org

A.F. Ismail
e-mail: af_ismail@iium.edu.my

A.H. Abdalla
e-mail: aisha@iium.edu.my

W. Hashim
College of Information Technology, Universiti Tenaga Nasional, 47000 Kajang, Malaysia
e-mail: wahidah@uniten.edu.my

A. Razzaque
Department of Computer Science and Engineering, University of Dhaka, Dhaka, Bangladesh
e-mail: razzaque@ieee.org

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_30

1 Introduction

Heterogeneous Community of LTE/LTE-A method provided expanded exposure and capacity via HeNodeBs for in-house/apartment situations. The more implementation of plug and play HeNodeBs introduces frequency misalignments due to improper synchronization or limited bandwidth connectivity. The frequency alignments are necessary to control the HeNodeBs. In Fig. 1, a Heterogeneous network scenario is illustrated [1]. To make it possible for a dependable and effective distribution associated with the multilevel network services, the accurate efficiency assured synchronization is not just only important, but also the basic requirement of the network densification technique. Considering that all HeNodeBs usually are associated by IP backhaul on the operator's network, as a result incompatible delay may perhaps take place due to changing traffic blockage. The actual synchronization connected with HeNodeBs are needed to prepare the signal obtained for reducing interference due to a number of accesses at the as time ensuring the offset of the carrier within tolerable boundaries. HeNodeBs are expected to follow the standards for minimum time and frequency synchronization which are required by various cellular standards. The requirements which dictate time synchronization are mandatory in order to manage interference as well. Which is desirable to make sure perfect assignments in synchronized networks, exclusively in Time-Division Duplex (TDD) as well as LTE-A systems. As well, the criterions and provisions aimed at the frequency synchronization be present also spirited to retain the correct structure of frequency within macro-eNodeBs and HeNodeBs. Likewise, time synchronization cutting-edge multi-hop heterogeneous network also

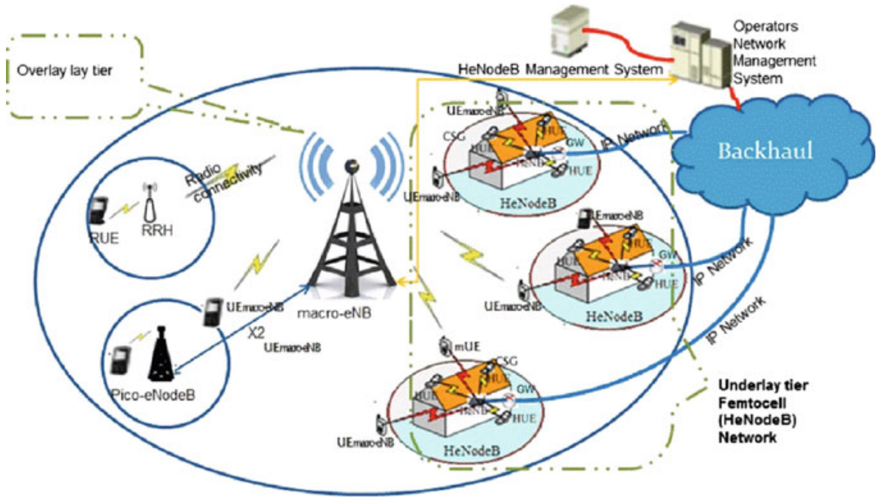


Fig. 1 Design of heterogeneous network consisting HeNodeBs, in multilevel apartment the scenario is more complex

Table 1 Heterogeneous network timing requirement

Classifications	Frequency alignments parts per billion (ppb)	Offset/phase (μs)
FDD	16	–
TDD	16	1.5
eICIC	16	1.32
LTE-A	16	0.5

requirements to circumvent the interference among the spectrum, precisely in place of Frequency-Division Duplex (FDD) [2, 3]. The heterogeneous network has more restrictive timing demands. Despite the fact of LTE FDD classifications prerequisite merely the rate of recurrence synchronization, while LTE TDD techniques be responsible for a further step requirement of 1.5 μs. Aimed at supplementary in advanced LTE-A systems, the constraint is even more unambiguous. In Table 1, the timing requirements of different access technologies [4, 5] are summarized.

All the way through orthodox 3G setups for pico-eNodeB and macro-eNodeB, synchronization was provided with TDM timing otherwise Global Navigation Satellite System (GNSS). Still, in 4G association such as HeNodeB is aligning the phases through IP backhaul comprising IEEE1588 [6], PTP [7] as well as enhanced IEEE 1588 [8]. Yet, in several cases, wherever deficiency of bandwidth or imperfect associations, the existing 1588 master-slave system has drawback. The IEEE 1588 uses pairwise messaging system via consistent medium. As a result, the delay increases and network becomes asymmetric. In Fig. 2 the IEEE1588 master/slave time synchronization method is depicted. Instead, operators pick GNSS for conveying synchronization and location credentials [2]. Usually, HeNodeBs inside apartment the GNSS signal is weaker, caused by the building

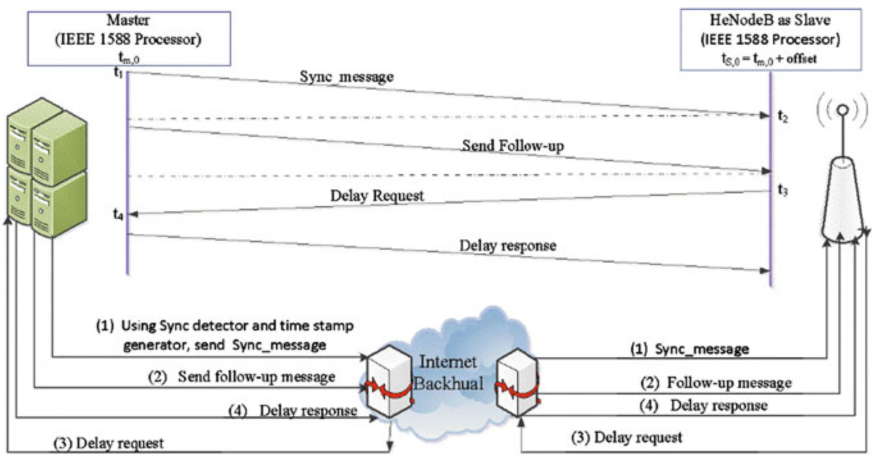


Fig. 2 IEEE1588 master/slave synchronization technique

wall width. Furthermore, some proposals are made which require extra hardware, for an example local Grandmaster. Nevertheless, extra hardware may increase cost.

This paper presents a self-organizing approach consuming Least Square Regression to synchronize the HeNodeBs in Heterogeneous Network. The proposed algorithm is novel and decentralized, thereby no need to have GNSS or hardware support anymore.

The rest of the paper is organized as: Sect. 2 deliveries the synchronization procedures and troubles to cope with the challenge, and recommended a proposed synchronization design, Sect. 3 contains encloses program arithmetic evaluation with the effect explanation, and Sect. 4 completes the paper with summary.

2 Synchronization Problems and Methods

To synchronize HeNodeBs, macro-eNodeB assistance has considered through broadcasting frame by using the Poisson clock [9]. The approach shows that each of the HeNodeBs needs to update their clocks with macro-eNodeB. The approach is able to synchronize the HeNodeBs. Yet, the tactic has succeeded up to 0.5 s offset which is beyond the requisition.

Microsecond phase aligning is speciously extreme to precise for LTE-A systems as the time duration of a single sub-frame in LTE-A is simply 1 µs [10]. In [10, 11], it is observed that by means of enhanced Inter-Cell Interference Cancellation (eICIC) even with flawed phase alignment between adjacent macro-eNodeBs for the performance tainted. Due to phase misalignment error between HeNodeBs and macro-eNodeB different interference suppression techniques performances can be disrupted. Thus stringent exactitude is obligatory for the HetNets association of LTE-A. Just in order to offset disturbance from a macro-eNodeB in order to Pico-eNodeB UEs, macro-eNodeBs every once in a while directs its own sub-frame with no data apart from some control signals and reference signals. Consequently, your moderate sub-frames usually are named Almost Blank Sub-frames (ABSs). On the other hand, macro-eNodeB must send vital system information along with Common Reference Signals (CRS) with regard to providing assistance to the resultant UEs. Using this type of tactic, HeNodeBs may plan the eICIC ABSs with regard to offering badly situated UEs. In [12], to synchronize the HeNodeBs listening supported scheme has proposed in TDD-LTE system. This scheme assumed an external synchronized clock in order to synchronize its neighbor networks. Each of the HeNodeB listens from the received signals from the neighbors and exchange message to synchronize. Moreover, in the proposed scheme considered Signal Interference to Noise Ratio (SINR) (γ_{ij}) with a threshold value because of interference barrier is shown in Eq. 1 [12].

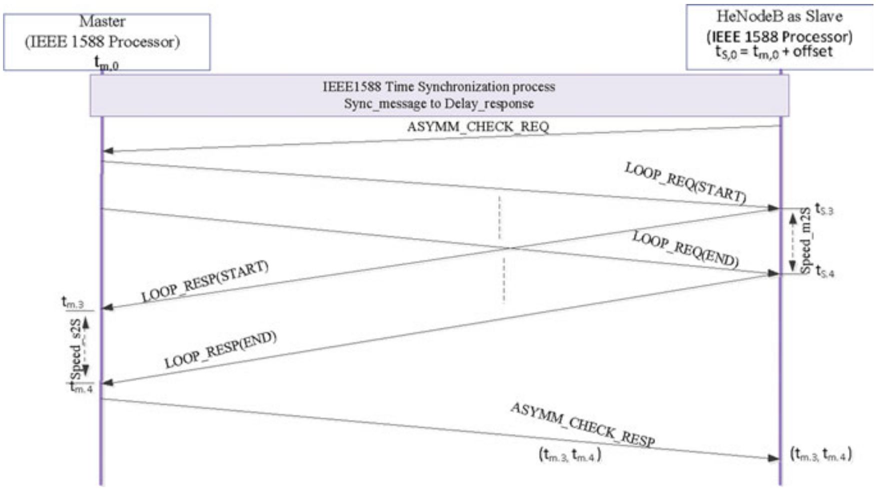


Fig. 3 Enhanced IEEE 1588 for the asymmetric ratio scheming

$$\gamma_{ij} = \frac{P_{ij}}{\sum_{k \in U,\, k \neq j} P_{ik} - \sum_{l \in T_j} P_{il} + N} \qquad (1)$$

where, transmitted power is denoted P_{ij}, i and j are the HeNodeBs for the transmission and reception, N is the noise at the receivers. The performance were estimated through SINR, synchronization time. However, this approach didn't consider about the required time for the synchronization, offsets as well as frequency errors.

An enhanced time synchronization algorithm is proposed over IEEE 1588 in paper [8] to minimize the offset (bias error) in the case of asymmetric IP based communication links, for example: Asymmetric Digital Subscriber Line (ADSL) and Very- High-Data Rate Digital Subscriber Line (VDSL) which uses IEEE 1588. IEEE 1588 synchronization algorithm is improved by the extra procedure which is called the block burst transmission (in Fig. 3). This procedure is utilized in order to calculate the asymmetric ratio for the communication link. However, the new and improved IEEE 1588 still introduces an offset error which cannot be reduced to zero. Thus, certain modifications are made so that it can effectively perform over the asymmetric backhaul link. Hence better accuracy is a necessity for the asymmetric communication link of ADSL or VDSL. Some more enhancements are required so that it can efficiently perform over asymmetric backhaul link.

Mobile station Assisted (MS) technique using receiver-receiver synchronization has presented in [13] which employs the least-squares linear regression algorithm to reduce offsets. Nevertheless, in receiver-receiver synchronization, clock offset achieved 2.5 μs and the accuracy in frequency reached is up to 250 parts per billion (ppb). However, for the LTE/LTE-A heterogeneous network, the requirement is not enough as per the data in Table 1. To synchronize HeNodeBs macro-eNodeB

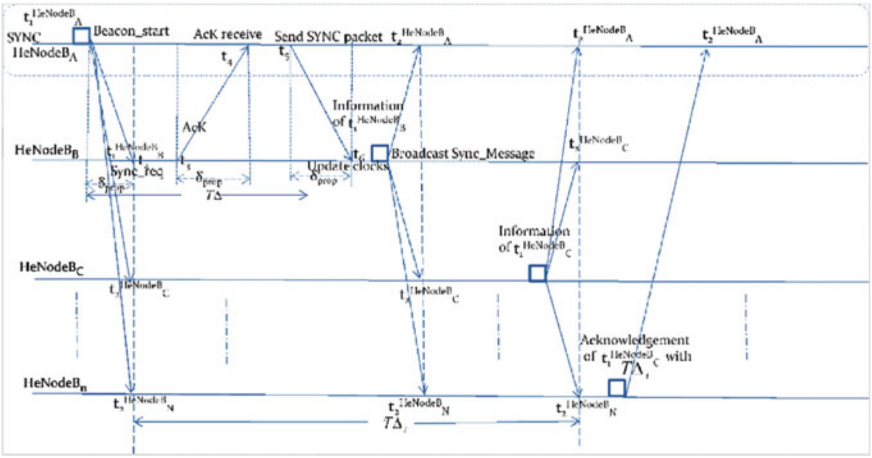


Fig. 4 Proposed synchronization approach for multi-hop synchronization scheme

assistance has considered through broadcasting frame using Reference Broadcast System (RBS) [13]. The approach shows that each of the HeNodeBs needs to update their clocks with macro-eNodeB or a reference node. The approach is able to synchronize the HeNodeBs. However, the approach has achieved larger offset than the requirement. In order to mitigate this problem, a self-organized approach is proposed which is inherited from RBS and enhanced IEEE 1855 [8] with the mechanism of LSRA to synchronize the multi-hop HeNodeBs. The proposed approach is illustrated in Fig. 4. The steps of the proposed approach is described as point wise below.

1. A coordinated synchronized HeNodeB will send *sync_beacon* to nearby *n*th *HeNodeBs*.
2. All neighboring HeNodeBs of *n*th *HeNodeBs* will receive Sync appeal and store arrival time which are $T_{n,i}^{HeNodeB_j}$ and the frequency alignment (drift) $\Upsilon_{\text{drift}}^{HeNodeB_{ij}}$, then send AcK to the reference *HeNodeB*.
3. Synchronized HeNodeB will send the correct clock reference.
4. HeNodeBs clocks will be swap as well as compare with up-to-date sync information from synchronized HeNodeB to regulate offset and frequency alignment. In every exchange of message the *n*th *HeNodeBs* directs acknowledgement to the synchronized *HeNodeB*.

3 Performance Analysis

To attain the goal of synchronization, the actual performance matric can be recognized as clock offset and clock drift. For the nth HeNodeB synchronization offset and drift is the key performance parameter. Therefore, the offset estimation can be represent in Eq. (2) applying LSRA headed for estimate and correct misalignment. Considering the propagation delay of $\mathrm{HeNodeB_{B \to N}}$. In the transmission reception there must have some receive delay. The delay difference can be form seeing that a Gaussian distributed random which is zero mean σ^2 variation. Due to wall attenuation in indoor environment propagation delay need to be considered. For the offsets and frequency error approximation Eq. (3) is applied.

$$\hat{\mathrm{T}}^{\mathrm{HeNodeB_{B\to N}}}_{\mathrm{Phase}} \cong \frac{1}{M}\sum_{i=1}^{M} \hat{\mathrm{T}}^{\mathrm{HeNodeB_{B\to N}}}_{\mathrm{Phase}} \tag{2}$$

$$\begin{aligned} \Upsilon_d = \Bigg[& \left(\Upsilon^{\mathrm{HeNodeB_{AN}}}_{prop} - \Upsilon^{\mathrm{HeNodeB_{AB}}}_{rec,i}\right) + \left(\Upsilon^{\mathrm{HeNodeB_{AN}}}_{rec,i} - \Upsilon^{\mathrm{HeNodeB_{AB}}}_{rec,i}\right) \\ & + \left(\Upsilon^{\mathrm{HeNodeB_{AN}}}_{rec,i} - \Upsilon^{\mathrm{HeNodeB_{AC}}}_{rec,i}\right)\Bigg] \end{aligned} \tag{3}$$

The relative frequency errors which is the drift between $HeNodeB_{A\to N}$ can be assess using Eq. (4).

$$\left(\mathrm{E}\left(\left|\emptyset_r^{\mathrm{A\to N}}\right|\right)\right) = \left(\emptyset_r^{\mathrm{HeNodeB_{AB}}} + \emptyset_r^{\mathrm{HeNodeB_{AC}}} + \emptyset_r^{\mathrm{HeNodeB_{AN}}}\right) \tag{4}$$

The performance exploration of the recommended arrangement, the simulation can be accomplished expending Monte Carlo simulation. The simulation strictures are in Table 2 [8, 12–15].

Table 2 Simulation parameters

Factors		Values
Network area		300 × 300 m^2
Number of HeNodeB		50
Distance(d)		20 m
Radius		30–40 m
Delay variation (σ)		10 µs, 15 µs
Synchronization process		15
HeNodeB radius		30
Number of samples for Monte Carlo iterations		1024
Carrier bandwidth		20 MHz
Time bandwidth factor		100 MHz
Path loss = {C*(d(a)}	Path loss exponent (a)	3
	Constant losses (C)	2

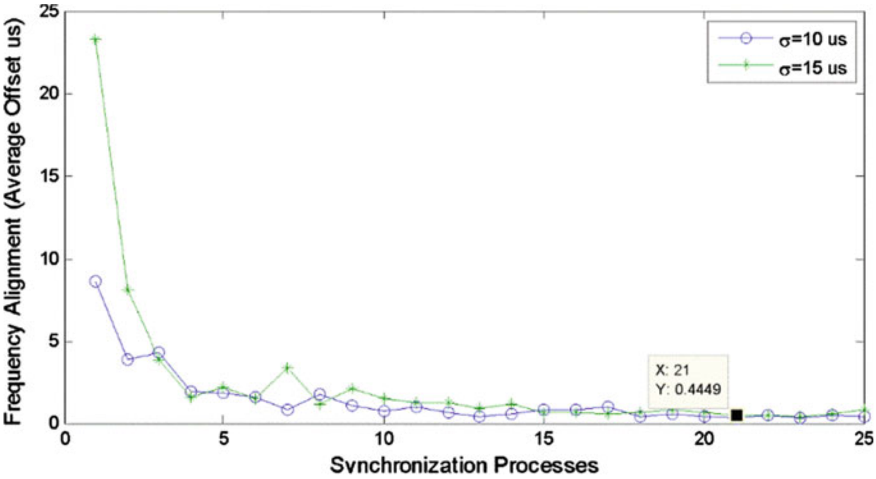


Fig. 5 Frequency alignment of offsets versus synchronization process for 15 HeNodeBs

The final result for the frequency precision of offset achieved for the LTE-A system surpasses the necessary timing as revealed in Table 1. The Fig. 5 shows the offset accuracy for the 15 HeNodeBs. The figure demonstrates that the marked data strips which is 0.44 μs in 21 synchronization process when σ is 10. However, if the number of HeNodeBs are increased during simulation (Fig. 6), the effect is a little bit different than the 15 HeNodeBs synchronization. It could be viewed in this

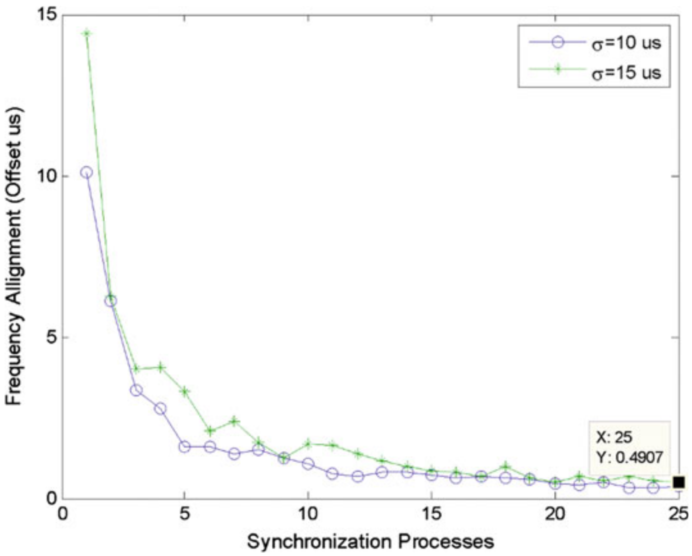


Fig. 6 Frequency alignment of offsets versus synchronization process for 20 HeNodeBs

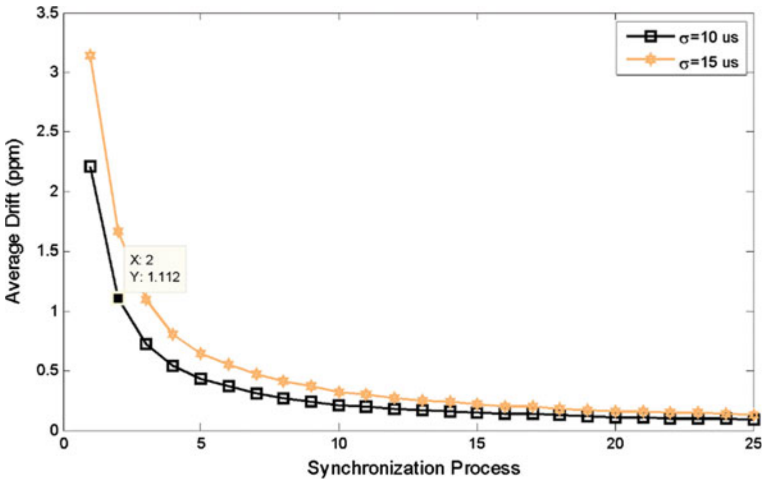


Fig. 7 Average drift versus synchronization process for 20 HeNodeBs

article that the synchronization gotten by the 25 processes, furthermore fulfills the actual LTE-A timing need.

In Fig. 7 the drift is evaluated through the 25 number of synchronization processes for σ = 10 µs, 15 µs. In 2 synchronization process, the drift accuracy (1.12 ppb) can be obtained the requirement of LTE-A.

4 Conclusion

This paper investigated IEEE1588 and RBS. It has been observed in the research that IEEE 1588 works extremely well inside Heterogeneous Network specifically for LTE/LTE-A methods; certainly where a pairwise communication is required and this depends on good wireless networks connectivity. Furthermore, for the depended media there is a big chance to enforce various delays intended for transmit and receive route, as a result asymmetry has been created, and this asymmetry cannot be predicted in pairwise synchronization processes. Thereby in the proposed scheme which is considered as prior issue. The proposed approach is able to avoid such typical delays the timing target requirement achieving (as shown in Figs. 5, 6 and 7) of the TDD, eICIC and LTE-A systems in Heterogeneous Network.

Acknowledgement A special thanks to Ministry of Education for the education sponsorship, and thanks to Research Management Center (RMC), International Islamic University Malaysia (IIUM).

References

1. Hasan, M.K., Ismail, A.F., Abdalla Khaizuran, A.H., Abdullah, K., Ramli, H.A.M.: Performance analysis of interference coordination techniques in heterogeneous network (2014)
2. Pesyna K.M., Wesson, K.D., Heath, R.W., Humphreys, T.E.: Extending the reach of GPS-assisted HeNodeB synchronization and localization through tightly-coupled opportunistic navigation. IEEE Int. Workshop HeNodeB Netw. 242–247 (2011). doi:10.1109/GLOCOMW.2011.6162445
3. Okada, M., Hara, T., Saitue, N., Wada, T.: Frame synchronization among base stations for TDD systems, In: 4th International Symposium on Communications, Control and Signal Processing (ISCCSP). IEEE, pp. 1–4 (2010). doi:10.1109/ISCCSP.2010.5463476
4. Hasan, M.K., Saeed, R.A., Hashim, A.H.A., Islam, S., Alsaqour, R.A., Alahdal, T.A.: Femtocell network time synchronization protocols and schemes. Res. J. Appl. Sci. Eng. Technol. **4**(23), 5136–5143 (2012)
5. HetNet Synchronization, Online [available]: http://www.qulsar.com/Applications/Telecommunications_and_Networks/HetNets.html
6. Andre, V., Dominik, S., Clock synchronization in telecommunications via PTP (IEEE 1588). In: IEEE International Frequency Control Symposium, pp. 334–341 (2007). doi:10.1109/FREQ.2007.4319093
7. Vallat, A., Schneuwly, D.: Clock synchronization in telecommunications via PTP (IEEE 1588). In: Frequency Control Symposium, 2007 Joint with the 21st European Frequency and Time Forum. IEEE International. IEEE, pp. 334–341 (2007)
8. Sungwon, L.: An enhanced IEEE 1588 time synchronization algorithm for asymmetric communication link using block burst transmission. IEEE Commun. Lett. **12**(9), 687–689 (2008). doi:10.1109/LCOMM.2008.080824
9. Shao-Yu, L., Lee, H., Sung-Yin, S., Chen, P., Kwang-Cheng. C.: Network synchronization among femtocells, IEEE GLOBECOM, pp. 248–253 (2011). doi:10.1109/GLOCOMW.2011.6162446
10. Wang, Y., Pedersen, K.I.: Performance analysis of enhanced inter-cell interference coordination in LTE-aAdvanced heterogeneous networks. IEEE VTC Spring, pp. 1–5 (2012)
11. Sahin, M.E., Guvenc, I., Moo-Ryong, J., Arslan, H.: Handling CCI and ICI in OFDMA femtocell networks through frequency scheduling. IEEE Trans. Consum. Electron. **55**(4), 1936–1944 (2009)
12. Kaushik, G., Prasad, V.: Network time synchronization in time division—LTE systems, MS thesis, School of Electrical Engineering Aalto University, Finland (2013)
13. Peng, J., Li, Z., Mclernon, D., Jibo, W.: Mobile station assisted receiver-receiver time synchronization scheme for HeNodeBs. In: IEEE Conference on Vehicular Technology. pp. 1–5 (2011). doi:10.1109/VETECS.2011.5956752
14. Kerö, N., Müller, T., Kernen, T., Deniaud, M.: Analysis of precision time protocol (PTP) locking time on non-PTP networks for generator locking over IP. SMPTE Motion Imaging J. **123**(2), 37–47 (2014)
15. Exel, R.: Mitigation of asymmetric link delays in IEEE 1588 clock synchronization systems. IEEE Commun. Lett. **18**(3), 507–510 (2014)

Assessment of Information Security Management on Indonesian Higher Education Institutions

Candiwan, Puspita Kencana Sari and Nadiailhaq Nurshabrina

Abstract Information is one of the valuable corporate's asset and must be protected, including for higher education institutions. Many security breach had happened on universities in Indonesia in many forms, such as penetration of the official website, website deface, and penetration to academic system to change the scores. The purpose of this research is to capture the implementation of information security management in some higher education institutions in Indonesia and not to generalize the condition of all institutions. This paper is used a descriptive approach to explore the maturity level of particular items in main clauses and controls of ISO27001:2013. The result showed that average score of maturity level is in level 2 (repeatable but initiative). It can be concluded that implementation of information security management is still limited on initiatives from its IT organization. Some controls are repeatable without adequate planning and documentations.

1 Introduction

Many cases occur in the information security organization from viruses, social engineering, DoS attack, hackers to data theft. Security breaches continue to increase, both in terms of the number of incidents as well as financial losses. Information security is a combination of systems, operations and internal controls to ensure the integrity and confidentiality of the data and operating procedures within an organization [1]. PWC Survey in 2014 said that the number of incidents of

Candiwan (✉) · P.K. Sari · N. Nurshabrina
School of Economic and Business, Telkom University, Jl. Telekomunikasi Terusan Buah Batu, Bandung, Indonesia
e-mail: candiwan@telkomuniversity.ac.id

P.K. Sari
e-mail: puspitakencana@telkomuniversity.ac.id

N. Nurshabrina
e-mail: nadiailhaq@gmail.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_31

security breaches rose to 48 % greater than in 2013, while the total financial losses ranging from organizations to large organizations increased by 34 % compared to the year 2013 [2].

Open Security Foundation released a survey in 2014 which stated that as many as 35 % of security breaches occur in Educational Institutions [3]. These institutions are targeted by people who are not responsible because they save many personal data. A survey to IT leaders held in 2014 showed the top challenges for the future of education institutions are to protect the personal data of students, organizations information and intellectual property [3]. According to data taken from the Indonesia Security Incident Response Team on Internet Infrastructure (Id-SIRTII) showed that the attack on the academy's website (ac.id) amounted to 18.98 %, this number is in the second position after the government website (go.id) amounted to 27.42 % [4].

Some examples of security incidents in higher education institution recently are:

(a) The official online library portal, specifically for students, owned by UGM was hacked. Hacker just left a short warning message to the web admin to enhance web security [5].
(b) A hacker team had penetrated an official website of Trisakti University. They warned that the website can be easily penetrated from outside, including sites across faculties. According to them from thousands of websites that have been audited, 80 % of them are still vulnerable [6].
(c) Official website of a faculty in UNPAD had been penetrated by a hacker that performed two actions. First, replacing the photo slides in the homepage with photos from hacker. Second, replacing the news content with a writing "Hello Admin" [7].

Security threats not only coming from cyberspace or external parties, but also come from environment and internal organization, such as sales of information by internal party, theft of documents, natural disasters, illegal data changes, and others. Considering those risks, this paper aims to assess the effectiveness of information security management in Indonesian higher education institutions, but not to generalize the condition of all institutions. The effectiveness of ISM will refer to the main clauses and controls in ISO/IEC 27001:2013 as the international standard for information security management systems. This is caused that the effectiveness of ISM depends on the implementation strength of security controls [8].

2 Literature Review

2.1 Information Security Management

Information security term is used to describe the protection of information assets, including both computer and non-computer equipments, facilities, and data to ensure confidentiality, integrity, and availability of information through the application policy, education and technology [9, 10]. The objective of information security is to ensure business continuity, minimize business loss, and maximize return on investment [11, 12]. Therefore, organization management is not only expected to keep secure information resources, but is also expected to keep the organization in order to continue functioning after a disaster security system.

Information security has three basic components that must be managed, i.e. confidentiality of sensitive information from unauthorized parties; integrity of information to ensure its accuracy and completeness; and availability of information and vital services to authorized users whenever needed [13]. Beside those three basic goal, information security also covers issues that can threaten accountability, reliability, nonrepudiation, privacy, authentication and trust of information [8, 14].

Information security management consists of four stages in basis form: (i) identify the threats that can attack corporate information resources, (ii) define the risks that may be caused by the threats, (iii) determine the information security policy, and (iv) implement the controls to address the risks that exists [9]. To implement the controls, organizations can refer to the ISM standard or best practice, e.g., ISO/IEC 27001:2013. ISM standard provides a model and guidelines to organization for implementing, operating, monitoring, reviewing, managing and improving an information security management by introducing a set of controls to an acceptable level of protection [12, 15].

2.2 ISO/IEC 27001:2013

In order to minimize the threats that continue to evolve, the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) published a standard guide for information security. ISO/IEC states to this guide as an Information Security Management System (ISMS) and describes the explanation in a series of ISO/IEC27000. In ISO 27001:2013, this information security management is based on Plan Do Check Action (PDCA) model that has purpose to improve continually the information security [16]. This standard consists of seven main clauses and some controls that is available in the ANNEX section. The items of the main clause consists of context of organisation, leadership, planning, support, operation, performation evaluation and improvement. There is a statement of applicability (SOA) to control the implementation of ISMS, consists of 114 controls from 14 domains and 35 objectives in ISO/IEC 27001:2013, as illustrated in Table 1.

2.3 Information Security in Higher Education

Some practitioners in information security also expressed that education institutions are vulnerable to security breaches. Like Rebecca Herold said "*Academic*

Table 1 Domains, objectives and number of controls in Annex A ISO27001:2013

NO annex	Domain ISO 27001:2013	Number of objectives	Number of controls
A.5	Information security policies	1	2
A.6	Organization of information security	2	7
A.7	Human resource security	3	6
A.8	Asset management	3	10
A.9	Access control	4	12
A.10	Cryptography	1	2
A.11	Physical and environmental security	2	15
A.12	Operations security	7	14
A.13	Communications security	2	7
A.14	System acquisition, development and maintenance	3	13
A.15	Supplier relationships	2	5
A.16	Information security incident management	1	7
A.17	Information security aspects of business continuity management	2	4
A.18	Compliance	2	8
Total		35	114

institutions offer a treasure trove of information of interest to cybercriminals. Data is like gold to cybercrooks, universities are like Fort Knox to them." [17]. Furthermore, universities are data-rich environments with multiple access points and a culture of collaboration and open sharing information [18]. That conditions make higher education institutions are vulnerable to security breaches. Therefore, it requires a commitment from management to achieving and maintaining a reasonable level of protection [17].

In Indonesia, there are some information requirements from higher education regulator (DIKTI) that should be fulfilled by institutions for each semester, such as students' transaction and personal information, faculty information, curriculum, facilities and capacity information, and so on. Therefore, higher education institutions should protect the integrity and availability of those information. If they cannot meet the data requirement, they will get administrative penalties.

3 Research Framework and Method

This paper is a preliminary research about information security management framework. The first step is assessment about information security implementation in some universities. For this research, we refer to the ISO/IEC 27001:2013, a standard used by many companies for information security management system (ISMS). Some structured questions are constructed from main clauses and some

Table 2 Maturity level criteria

Maturity level	Description
0 (Non-existence)	There is no recognition of the need for internal control
1 (Ad hoc)	There is some recognition of the need for internal control
2 (Repeatable but initiative)	Some controls are in place but they are not documented
3 (Defined)	Some controls are in place and are adequately documented
4 (Manageable and measurable)	There is an effective internal control and risk management environment
5 (Optimize)	An organization wide risk and control program provide continuous and effective control and risk mitigation

Table 3 The example of questions based on main clause and controls in ISO27001

No main clause/annex	Main clause/control	Question statement
7.1	Resources	Institution provide the necessary resources for the implementation of information security management
A.6.2.1	Mobile device policy	The use of mobile devices is governed by an information security policy

controls of 14 domains that exist in ANNEX or Statement of Aplicability (SOA) of ISO/IEC 27001:2013 (as showed in Table 1). From that questions, it will describe which controls are implemented completely, partially, or not implemented measured using criteria of maturity level as showed in Table 2 [19].

To describe clearly of main clauses or ANNEX in ISO/IEC 27001:2013 used in this research, some examples of question statement to explore how far ISM has been implemented, are given in Table 3. The answers from respondents then would be categorized into maturity level based on descriptions in Table 2.

This research used qualitative method with descriptive approach to explore the maturity level of each controls. In qualitative research, sampling can continue to do as long as for the breadth and depth of knowledge on the topic you are studying is still growing and this stops when you are not getting new knowledge and insights [20]. There are 117 higher education institutions in a city in Indonesia. But, not all of them have been utilizing ICT to support their business processes. Therefore, sample had been filtered from the availability of official websites. Only 66 institutions that fulfilled the criteria to become sample for this research. Data were collected from head of IT department from each institutions using about fifty structured questions. It was not easy for convincing them to participate in this research. To analyze the data, it used descriptive approach by calculating average score of maturity level per domain and per institutions. Interpretation of the score are also supported by observation to each institution. Finally, there were 12 institutions willing to share their information and the maturity level showed a convergent score. The result score and interpretation are discussed in the next part of this paper.

4 Result and Discussion

From twelve higher education institutions that participated for this research, 75 % of IT organizations are become a department, directly under head of university (as Chief Executive Officer or CEO), and the rest of IT organizations are merged with other unit under a department. So, most of them have a strategic position and wider span of service to all of business unit in their institutions. But, it doesn't relevant with the number of staff or employee in the institution as representation of organization size. From those institutions, 33 % of them only have employees less than 100 people, 25 % have 100–250 employees, 25 % have 250–500 employees, and 17 % have more than 500 employees. It means, only 42 % of them are considered as big institutions (have more than 250 employees) [21]. IT organization position are also not really relevant with adoption of IT standard in organization. Among those nine institution where IT organization are directly under rector CEO, only two of them have been use IT standard, but not specifically to information security.

According to the adoption of IT standard, we found that it is relevant with budgeting for information security in their organization. Those two institutions who are adopting IT standard have a fix budget for information security programs every year. While, the restten institutions only have conditional budgeting, so they will get different budget for their security programs based on their allocation every year. Even worst, two of them doesn't have any budget for their information security programs although they ever have security breach experience before. Totally, three of the institutions have security breach experiences. Unfortunately, only one of them then implement some controls to protect their information assets even though not adopting any IT standard.

Adoption of IT standard also isn't highly relevant with IT system usage in the institution. Institutions that have been implementing IT standard do not use so many IT system. One of them only use two kinds of IT systems, for academic and financial process. All of the institutions use IT system for supporting academic process, their main business process. The second famous system that is used in the institutions is library system; 9 of 12 institutions use it. It's not wondering because it is also related with their main business in education industry. Other systems used by those institutions are employment (8 institutions), financial (8 institutions), and other purposes (3 institutions).

Maturity Level. To get a description about effectiveness of information security implementation in the higher education institutions, some controls in ISO/IEC 27001:2013 had been selected to represent it. The measurement used criteria of the maturity level that consists of six levels, level 0 (not existent) to level 5 (optimize). Head of IT organization from each institution had defined their own maturity level, refer to the criteria that had been given, for each clauses (in main clauses and Annex) from ISO/IEC 27001:2013. Those answers then calculated to get average score for each domain and also average score for each institutions.

Regarding the main clauses of ISO/IEC 27001:2013, there are seven clauses evaluated. Based on the results, Performance Evaluation clause has good implementation especially for performance evaluation and effectiveness of information security management. Meanwhile, Leadership clause and Improvement clauses are also well implemented, after the performance evaluation. However, Planning clause is not implemented well especially for determination of information security objective for certained function and level. Overall, the main clauses are recognised to be implemented in those higher education institutions, but they are still done for a particular purpose and without a long-term planning.

Detail of score for each domain in Annex (A.5 to A.18) are shown in the Tables 4 and 5. From these tables, it can be seen that the highest score for maturity

Table 4 Detail of *maturity level* for each institutions (part. 1)

HE institutions	A.5	A.6	A.7	A.8	A.9	A.10	A.11
1	3	3	2.67	2.67	2.75	2	2.67
2	1.5	1.33	0.33	1.67	1.5	1.5	1.33
3	2	1.67	1.67	1.33	1.5	1	1.67
4	1	0.33	1	1	1.75	1	3.67
5	1	0.33	1	1.33	1.5	1	4
6	1	0.33	0.33	0.67	1.5	2	2
7	0.5	0.67	0	1.33	1.5	0	0.33
8	1.5	1.33	1.67	2.67	2.75	2	2.67
9	2.5	3	3.67	4.33	4	4	4
10	2	1.33	2	2.33	2.5	3	3
11	4	2.67	2.33	2	4.5	2	4
12	2.5	1.67	2	3.67	3	3	1.33
Avg. score per annex	1.88	1.47	1.56	2.08	2.40	1.88	2.56

Table 5 Detail of *maturity level* for each institutions (part. 2)

HE institutions	A.12	A.13	A.14	A.15	A.16	A.17	A.18
1	2.57	3	2.33	3	2.5	2	2
2	2.14	1.5	1.67	1	1	1.5	1
3	1.86	1	1	1	3	1	2
4	1.57	2	1.77	5	3	1.5	5
5	2.14	2	1.67	5	3.5	1.5	5
6	0.71	1	1	1	1	1	1
7	1.14	1.5	1	0	0	0.5	0.5
8	1.86	3	2	2.5	2	2	2.5
9	3.71	4	4	4	4	4	4
10	2.14	3	3	3	2.5	2.5	2.5
11	4.29	4	3.67	2	2	4	4
12	2.43	2	5	5	1.5	2	1.5
Avg. score per annex	2.21	2.33	2.34	2.71	2.17	1.96	2.58

level is refer to Annex A.15 (2.71), followed by Annex A.18 (2.58) and Annex A.11 (2.56). Annex A.15 is for Supplier Relationship, while Annex A.18 is about Compliance and Annex A.11 is for Physical and Environmental Security. It is a common thing in Indonesia that higher education institutions do not develop their IT system in-house, yet prefer using IT provider services. Therefore, most of those institutions have big concern to controls their contracts and agreement with their suppliers. Nevertheless, those controls are in place but not well documented. It might be related with maturity level of Compliance (Annex A.18) which also has a high enough scale (2.58). When they apply some controls in Supplier Relationship area, it can drive the compliance to contract and related regulation to monitor, review and audit the suppliers in delivering their services. Physical and Environmental Security area is also top-three maturity level. Protections of physical facilities, room, and office from unauthorized people or natural disasters are common things done by organization regardless information security risks. Therefore, those protections will also increase maturity level of this area.

The bottom-three area consists of Annex A.6 (Information Security Organization), A.7 (Human Resources security), A.5 (Information Security Policy) and A.10 (Cryptography). Information security organization area, as weakest area with maturity score 1.47, the weakest control is about mobile device policy. Whereas, most of employee (staff and faculty member) and students always bring their own mobile devices to access information through organization's network. This control should be consider as one of important policy to support information security in universities. Human resources security has maturity level 1.67 with the control that has the lowest score is term and condition of employment. Most of institution do not include information security as employees' responsibility in their contract. Other areas that have low level of maturity are Information Security Policy and Cryptography (with same score 1.85). Cryptography might has been implemented through firewall devices, but it wasn't supported with specific policy. Therefore, policy area is also has low level because some information security policy are not yet established.

Mapping of maturity level for all area in ISO 27001:2013 can be seen in Fig. 1. Overall, average score of maturity level for all area is 2.15 (repeatable but initiatives). It means that some controls for information security are implemented but not managed well. Those controls are done occasionally and might be repeatedly but only coming from initiatives of IT organization when they think it is needed for certain conditions. There is no adequate plan in implementing information security controls. It may be driven by no specific regulation from the government or industrial requirement for higher education organization to apply information security controls.

Maturity level for each institution can be seen in Fig. 2. Institution number 3 and 9 are institutions that have been adopting IT standard. But, they have different maturity level. Although institutions number 3 is older, it doesn't guarantee that it will implement information security better. It depends on the management

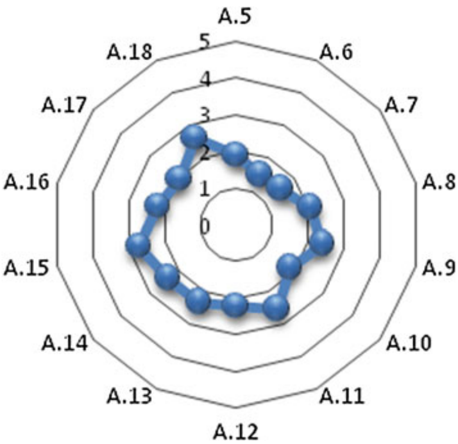


Fig. 1 *Maturity level* from ISO/IEC 27001:2013 supporting clauses (A.5–A.18)

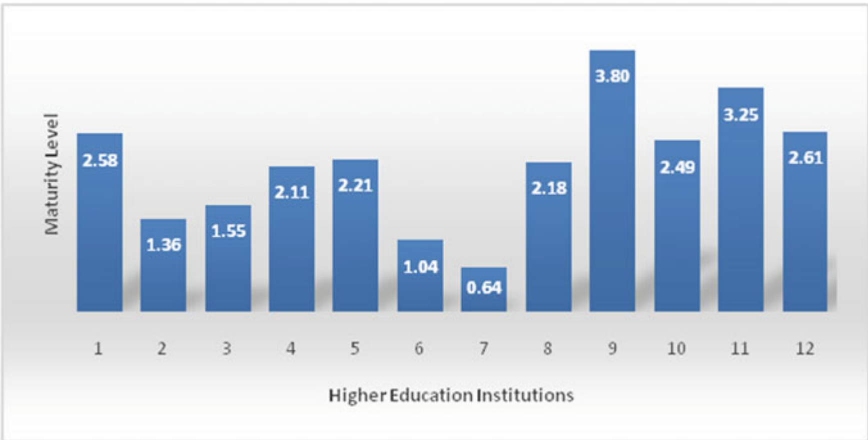


Fig. 2 *Maturity level* from twelve higher education institutions as respondents

awareness and commitment to evaluate and comply to the established policies. It can be seen in Table 4 that institution number 9 has higher compliance level than number 3. The lowest maturity level is 0.64 that refers to institution number 7. This institution doesn't adopt any IT standard and only has conditional budget for information security. For this institution, information security management is not a priority in their IT programs although they have experience in security breach.

5 Conclusion

From those results, it can be concluded that implementation of information security management, based on ISO/IEC 27001:2013, in Indonesian higher education institutions are still limited in initiatives from its IT organization. Some controls are repeatable without proper planning and documentation. Some of control areas are quite mature, such as for compliance, supplier relationship and physical and environmental security. However, some of controls need improvements such as: information security organization concerning mobile device policy and human resources security concerning information security as employees' responsibility in their contract. Some of institutions that became case study for this research have implemented IT standard but not specifically for information security management.

This paper is a preliminary for further research to build a framework for information security management in higher education institutions. From this research we tried to get a description of information security implementation in some institutions as a basis condition. Beside this assessment, we are also doing a study about influence factors in implementing information security and analysis of information security risks. Based on those study, we can arrange some priority steps or controls based on ISO/IEC 27001:2013 to address some risks in information security.

References

1. Hong, K.S., et al.: An integrated system theory of information security management. Inf. Manage. Comput. Secur. **11**(5), 243–248 (2003). www.emeraldinsight.com/0968-0277.htm. (Visited on 25 May 2012)
2. Price Waterhouse Coopers: Global information security survey. http://www.pwc.com/gx/en/consuling-services/information-secuirty-survey/index.html (2014). (Visited on November 2014)
3. Smith, D.F.: Educause 2014: Cyberattacks are a growing problem for higher education. http://www.edtechmagazine.com/higher/article/2014/10/educause-2014-cyberattacks-are-growing-problem-higher-education (2014). (Visited on 10 November 2014)
4. Sumantri, I.: Kebijakan Keamanan Informasi. http://folder.idsirtii.or.id/pdf/Kebijakan_Keamanan_Informasi.pdf (2013). (Retrieved on 1 November 2014)
5. Chandrataruna, M.: Perpustakaan online UGM dibobol hacker. http://teknologi.news.viva.co.id/news/read/166308-perpustakaan-online-ugm-dibobol-peretas (2010). (Visited on 20 January 2015)
6. Sarie.: Hacker: Situs Universitas Trisakti Mudah Dibobol. http://techno.okezone.com/read/2010/09/26/55/376069/hacker-situs-universitas-trisakti-mudah-dibobol (2010). (Visited on 20 January 2015)
7. FEB UNPAD.: Peristiwa hacking pada official website FEB unpad. http://www.fe.unpad.ac.id/id/arsip-fakultas-ekonomi-unpad/berita/239068-peristiwa-hacking-pada-official-website-feb-unpad (2013). (Visited on 20 Januari 2015)
8. Mukundan, N.R., Sai, L.P.: Perceived information security of internal users in Indian IT services industry. Inf. Technol. Manag. **15**, 1–8. Online: 9 April 2013, Springer Science+Business Media, New York (2013)

9. Mcleod Jr, R., Schell, G.P.: Management information system, 10th edn. Pearson Prentice Hall, USA (2007)
10. Whitman, M.E., Mattord, H.J.: Principles of information security, 4th edn. Course Technology, USA (2012)
11. Kruger, H., et al.: A vocabulary test to assess information security awareness. South African Information security multi-conference in Port Elizabeth, South Africa (2010)
12. Chang, S.E., Ho, C.B.: Organizational Factors to the effectiveness of implementing information security management. Ind. Manag. Data Syst. **106**(3), 345–361 (2006)
13. Mitchell, R.C., et al.: Corporate information security management. New Libr. World **100**, 213–227 (1999). (MCB University Press. London, ISSN 0307-4803)
14. Clark, D.L.: Enterprise security: The manager's defense guide, pp. 32–33. Pearson Education Inc, Boston (2003)
15. Gilies, A.: Improving the quality of information security management systems with ISO27000. The TQM J. **23**(4), 367–376 (2011). (Emerald Group Publishing Limited)
16. British Standard Institution: ISO/IEC 27001:2013 information tecnology-security techniques-information security management systems-requirements. BSI Standard Limited, Switzerland (2013)
17. Roman, J.: Universities breaches: A continuing trend. http://www.databreachtoday.com/university-breaches-continuing-trend-a-6660/p-2 (2014). (Visited on 5 Februry 2015)
18. Roman, J.: Universities: Prime breach target. http://www.databreachtoday.com/universities-prime-breach-targets-a-7865 (2015). (Visited on 5 Feb 2015)
19. Al-Mayahi, I.M.: SP: ISO 27001 GAP analysis-case study. In: International Conference on Security and Management (SAM' 12), Las Vegas (2012)
20. Cooper, C.R., Schindler, P.S.: Business research methods, 10th edn. McGraw-Hill, Boston (2008)
21. Jarvelainen, J.: Information security and business continuity management in interorganizational IT relationships. Inf. Manage. Comput. Secur. **20**(5), 332–349 (2012)

Optical Nerve Disc Segmentation Using Circual Integro Differencial Operator

Jan Kubicek, Juraj Timkovic, Martin Augustynek, Marek Penhaker and Martina Pokrývková

Abstract The paper deals with design of suitable methodology for the detection of the optic nerve disc from ophthalmological images. This issue was frequently solved in the past based on the Hough transform. The problem is that some the clinical images lack of sufficient contrast and so it is difficult to identify individual objects in the images. In our analysis, we process the image data, which are poorly distinguishable. Another significant drawback is the lower resolution of the analyzed data, which leads to poor identification of the optic nerve disc. For the detection of the optic nerve disc is used integro differential operator, which is looking for objects in the image that have a circular shape. The advantage of this proposed solution is sufficient sensitivity even in the weaker contrast environment.

Keywords Image processing · Optical nerve disc · Integro differencial operator · Image segmentation · Object detection · Retinopathy of prematurity · RetCam

J. Kubicek (✉) · M. Augustynek · M. Penhaker
VSB–Technical University of Ostrava FEI, K450, 17. Listopadu 15,
70833 Ostrava–Poruba, Czech Republic
e-mail: jan.kubicek@vsb.cz

M. Augustynek
e-mail: martin.augustynek@vsb.cz

M. Penhaker
e-mail: marek.penhaker@vsb.cz

J. Timkovic
Clinic of Ophthalmology, University Hospital Ostrava, Ostrava, Czech Republic
e-mail: timkovic.j@bluepoint.sk

J. Timkovic
Faculty of Medicine, Masaryk University, Brno, Czech Republic

J. Timkovic
Faculty of Medicine, University of Ostrava, Ostrava, Czech Republic

M. Pokrývková
Department of Neonatology, University Hospital Ostrava, Ostrava, Czech Republic
e-mail: martina.pokryvkova@seznam.cz

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_32

1 Introduction

Retinopathy of prematurity (ROP) is a vasoproliferative disease that affects especially prematurely-born infants with low birth weight. ROP is the second most frequent cause of blindness in children in developed countries [1, 2]. In the Czech Republic, it represents the second most frequent cause of severe vision impairment in children.

The aim of a ROP screening is an early detection of the first signs of the disease. That allows ophthalmologist to optimally schedule ophthalmological examinations, plan a possible treatment and hereby prevent a development of the severe vision impairment. For the purpose of the clinical rating of the size of retinal lesions and for the purpose of measuring distances of retinal changes and assessing their progression in time, we use the optic nerve disc diameter as a reference point.

RetCam 3 is a digital imaging system that represents a modern alternative of screening and diagnostics of retinal diseases in children. The main advantage of this system is the possibility of wide-angle view of the retina in comparison with the conventional method of the indirect ophthalmoscopy that is considered to be the standard method for the screening for ROP in prematurely-born infants. The possibility of making the picture or video during the examination allows ophthalmologists to assess the dynamics of the retinal changes or efficacy of a therapy over time.

The aim of the pilot project is formulation an algorithm for automatic image detection of the optic nerve disc that would enable the assessment of the size of retinal lesions in fully automated and standardized mode.

2 Problem Definition

Measurement of parameters and the detection of the optical nerve disc is in the field of ophthalmology significant problem. We do not need only detect the contour of the optic nerve disc, but it is very helpful to know the geometric parameters that represent optical nerve disc. In the past, this problem was solved on the basis of the modified Hough transform. This mathematical method is beneficial for detecting circular shapes that are sufficiently distinguishable visible. In case of noisy images, this method is difficult to apply because it does not achieve the required sensitivity.

For the purposes of our analysis, we used an integro differential operator, which has better sensitivity and contrast for less recognizable objects.

The advantage of the analyzed images is contrast imaging of blood vessels, their direction is easily detectable. We are focused on the problem of the optic nerve disc, which is not enough contrast distinguished. Another significant problem is the analysis of image edges. Very often it happens that the edges of the object are not fluent.

For this reason we decided to image preprocessing based on the contrast transformation with the aim of highlighting the contrast of the optic nerve disc [3–11].

3 Proposed Algorithm of Optical Nerve Disc Detection

Daugman's algorithm is based on applying an integro-differential operator to primarily find the iris and pupil contour. In our work, optimized Daugman's algorithm has been used for detection and localization of optical nerve disc. The equation of Daugman's operator is follows:

$$\max(r, x_0, y_0)|G(r) * \frac{d}{dr} \oint \frac{I(x, y)}{2\pi r} ds \tag{1}$$

where: x_0, y_0, r are the center and radius of direction detected circle, G(r) is Gaussian function and I(x, y) is the original retinal image. G(r) is a smoothing function, the smoothed image is then scanned for a circle that has a maximum gradient change, which indicates an edge. It worth mentioning here the problem is that the illumination inside the optical nerve disc is a perfect circle with very high intensity level. Therefore, we have a problem of sticking to the illumination as the max gradient circle. So a minimum nerve radius should be set. Another issue here is in determining the optical nerve disc boundary the maximum change should occur at the edge between the very dark nerve and the background, which is relatively darker than the bright spots of the illumination [12, 13].

Structure of the proposed segmentation algorithm is composed from several steps. Firstly, it is necessary to perform preprocessing of brightness transformations to improve the contrast differences between the optic nerve disc and the background, find the center of the optic nerve disc and after that a calculation of circular Daugman's operator. In the final step there is performed transformation the detected circle to ellipsoidal shape. The best approximation of the optic nerve disc is using the ellipse. Daugman's operator finds all circual shapes in the analyzed image which give circular features. For this reason, it is advantageous to restrict the area of the image where we are focused on (Figs. 1, 2).

Adjusting of brightness is performed on the basis of contrast stretching. Contrast stretching (often called normalization) is a simple image enhancement technique that attempts to improve the contrast in an image by "stretching" the range of intensity values it contains to span a desired range of values, e.g. the full range of pixel values that the image type concerned allows. It differs from the more sophisticated histogram equalization in that it can only apply a linear scaling function to the image pixel values. As a result the enhancement is less harsh. Before the stretching can be performed it is necessary to specify the upper and lower pixel value limits over which the image is to be normalized. Often these limits will just be the minimum and maximum pixel values that the image type concerned allows. For

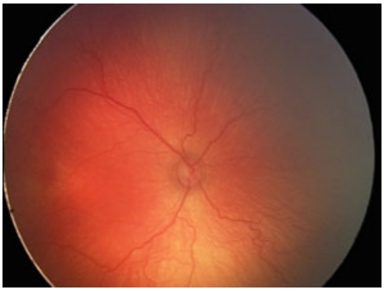


Fig. 1 The original analyzed image

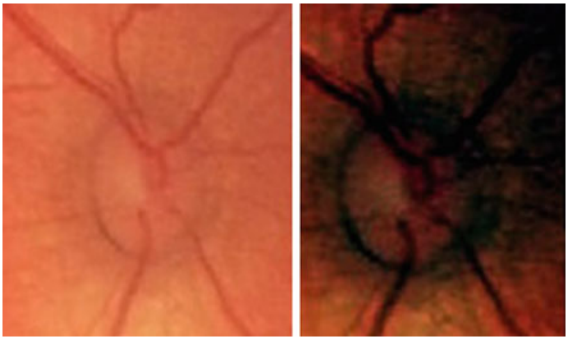


Fig. 2 Extraction of the region of interest (*left*) and contrast stretching (*right*)

example for 8-bit gray level images the lower and upper limits might be 0 and 255. Call the lower and the upper limits a and b respectively. [14–16].

The simplest sort of normalization then scans the image to find the lowest and highest pixel values currently present in the image. Call these c and d. Then each pixel P is scaled using the following function:

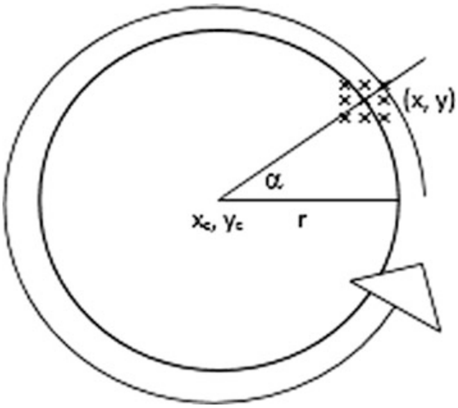
$$P_{out} = (P_{in} - C)\left(\frac{b-a}{d-c}\right) + c \tag{2}$$

Daugman's operator is calculated for each pixel with coordinates (x, y) and r values. An important step is to calculate the difference of the inner and outer pixel values of the contour edge the optical disc. There is Daugman's operator for circular disk with coordinates (x_c, y_c) on Fig. 3. Individual edge coordinates are given by the following system of equations:

$$x = x_c + r * \cos(\alpha) \tag{3}$$

$$y = y_c + r * \sin(\alpha) \tag{4}$$

Fig. 3 Calculation of Daugman´s operator



After taking differences, of all pixels in a circular contour are calculated individual, this process is in terms of successful detection insufficient. For this reason it is necessary to adjust the upper limit circle sample (CS). This procedure assesses how many points should be assessed on the contour of the optic nerve disc to calculate Daugman’s integral. Higher values of the parameter CS reduce the estimated error of calculation Daugman’s integral. After the calculation of individual difference values for each pixel, the detected contour can be calculated. These values are added together to calculate Daugman’s operator. The mathematical formulation for calculating the differences is given by the following relation:

$$diff(x, y) = I(x + \Delta_\alpha, y + \Delta_\alpha) - I(x - \Delta_\alpha, y - \Delta_\alpha) \quad (5)$$

There are only integer values for each axis (x, y). The problem is that we can´t make an accurate calculation of the difference for each angle α. For detection of the optic nerve disc, we used a modified version of Daugman’s operator which has higher sensitivity.

In this case four sensitive points on the diagonal are calculated. Structure of Daugman’s operator is determined as follows:

$$\max(x_c, y_c, r) \sum_{j=1}^{CS} diff\left(x_j, y_j\right) \quad (6)$$

Edge coordinates can be written in the generalized system of equations of the form:

$$x_j = x_c + r * \cos\left(\alpha_j\right), y_j = y_c + r * \cos\left(\alpha_j\right) \quad (7)$$

$$
\begin{aligned}
diff\left(x_j, y_j\right) &= \sum_{k=1}^{4} Ik \\
I_1 &= (I(x_j+1, y_j) - I(x_j-1, y_j)).\cos(\alpha_j) \\
I_2 &= (I(x_j, y_j+1) - I(x_j, y_j-1)).\sin(\alpha_j) \\
I_3 &= (I(x_j+1, y_j+1) - I(x_j-1, y_j-1)) \cdot \sin(\alpha_j+45) \\
I_4 &= (I(x_j+1, y_j-1) - I(x_j-1, y_j+1)) \cdot \cos(\alpha_j+45)
\end{aligned}
\tag{8}
$$

The last part of the algorithm is to transform the circular contours to the ellipsoidal shape that best approximates the shape of the optic nerve disc. Optimization process has been used for finding the minimum distance between the edge of the optic nerve disc and detected circular contour. Desired output is ellipsoidal contour that best approximates the characteristics of the optical nerve disc and geometric parameters of this shape [17–21].

We used the fact, that algebraic representation of the circle contour is follows

$$F(x) = ax^T x + b^T x + x = 0 \tag{9}$$

where $a \neq 0$ and $x, b \in R$. Ellipse is typically defined by quadratic equation:

$$F_1(x) = x_1^T A x_1 + b_1^T x_1 + c = 0 \tag{10}$$

On the base this facts, we solve optimization problem:

$$\min(a, b, x)|F(x) - F_1(x)| \tag{11}$$

[22–26]

4 Algorithm Testing

For the development and the assessment of the algorithm for the automated image detection of the optic nerve disc area was used a photo-documentation of the posterior segment of 47 prematurely born infants (26 boys, 21 girls) that had been examined by an ophthalmologist during the ROP-screening in neonatal intensive care in the University Hospital Ostrava from 14.8.2014 to 26.8.2014.

Their average gestational age at birth was 31 weeks (median: 31 weeks, SD ± 4.26, range from 24 to 40 weeks) and their average birth weight was 1635 g (median: 1460 g, SD ± 757.81, range from 600 to 3640 g). All the ophthalmological examinations were performed through dilated pupils (phenylephrine hydrochloride 2.5 % eye drops + tropicamide 0,5 % eye drops) and using an eyelid retractor (K1-5401/K1-5677, Katen Products Inc., Denville, NJ, USA) under the installation of local anaesthesia (oxybuprocaine hydrochloride 0.4 % eye drops).

The posterior segment of the eye was examined and photographed with the RetCam 3 digital system (Clarity Medical Systems Inc., Pleasanton, CA, USA) with D1300 lens that enables 130-grade-view of the retina (Fig. 4). The average post conceptual age (PCA) of these infants at the time of the photo-documentation was 39th week of PCA (median: 39th week of PCA, SD ± 5.20, range 31st–56th week of PCA). The images were taken with the best possible centring of the optic nerve disc. The image resolution was 640 × 480 pixels (96 dpi) with 24-bit colour depth.

The main goal is to make the detection of the optic nerve disc by automatic method even for cases where the optic nerve disc is accompanied by a weak contrast to their surroundings. The following example shows the effectiveness of the proposed algorithm (Fig. 5).

Fig. 4 Ophthalmological examination of the prematurely-born infant with the RetCam 3 digital imaging system

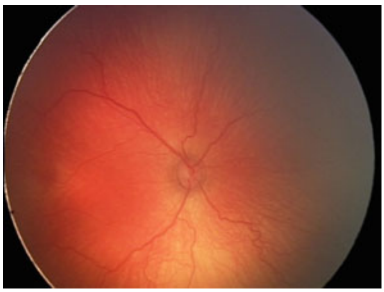


Fig. 5 Original database data

In the input image optical nerve disc is captured in the center of the image. As is obvious at first sight, identifiability is very weak, due to the weak contrast. In the second step, the region of interest is performed and the contrast transform of this area (Fig. 6).

After performing the contrast transformation, detected object is emphasized. On the other hand after transformation, there is noise which impairs the visual effect of the procedure.

Subsequently Daugman's operator is calculated that finds the maximum circular shape in the input image. This detection could be sufficient at first sight, but the output of operator does not correspond with some of the edges of the optical disc. To the extent possible eliminated these inaccuracies optimization procedure is used that searches for the minimum distance between the edge of the actual object and the detected contour. The final result is displayed on the Fig. 8, there is achieved the best approximation of the optic nerve disc.

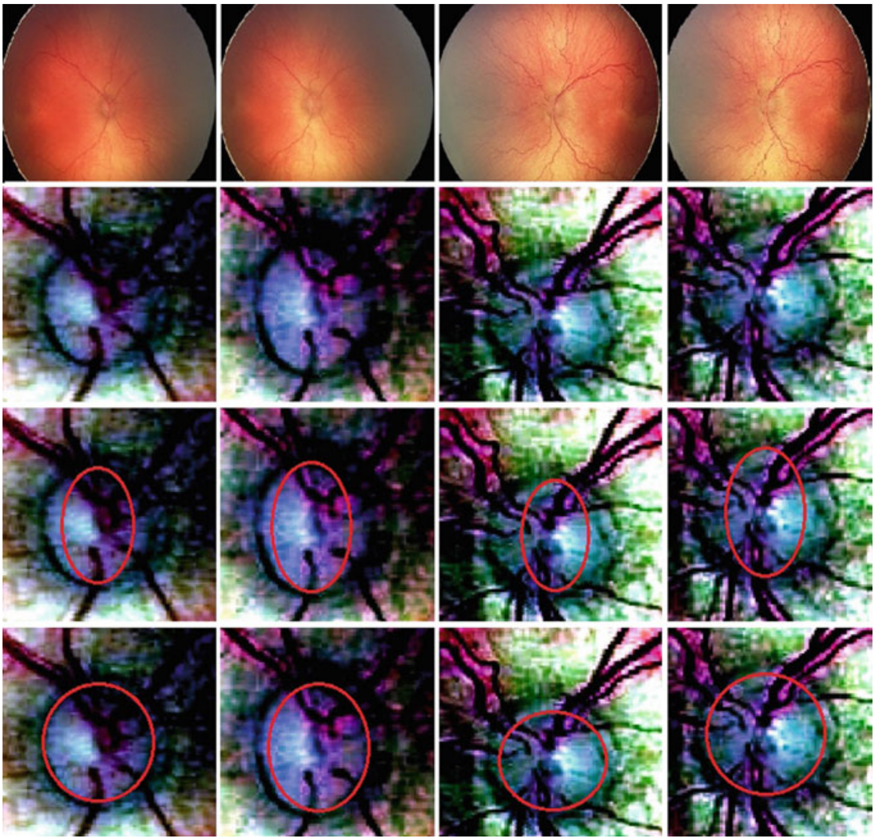


Fig. 6 Algorithm testing for database image data: *first row* original database images, *second row* optical nerve RoI, *third row* approximation region of interest by Daugman's operator, *fourth row* the final detection of optic nerve disc by minimizing areas

5 Conclusion

The proposed algorithm for automatic shape detection of the optic nerve disc appears to be a appropriate basis for further development of software applications that would help their functions in the clinical assessment of structural changes in retina of prematurely-born infants and thus improve the diagnosis and assessment of treatment efficacy of this severe retinal disease. The proposed algorithm can reliably localize the region of the optical disc. Optical disc is approximated by a symmetrical shape. In some cases this detection is less accurate due to asymmetry of the localized object. Consequently, we would like to focus on the implementation of algorithms that can automatically adapt to the shape of the optic nerve disc, for achieve better detection of objects of interest.

Acknowledgments The work and the contributions were supported by the project SP2015/179 'Biomedicínské inženýrské systémy XI', and this paper has been elaborated in the framework of the project "Support research and development in the Moravian-Silesian Region 2014 DT 1 - Research Teams" (RRC/07/2014). Financed from the budget of the Moravian-Silesian Region. This work is partially supported by the Science and Research Fund 2014 of the Moravia-Silesian Region, Czech Republic.

References

1. Kasik, V., Cerny, M., Penhaker, M., Snášel, V., Novak, V., Pustkova, R.: Advanced CT and MR image processing with FPGA. In: 13th International Conference on Intelligent Data Engineering and Automated Learning, IDEAL 2012. Natal2012, pp. 787–793
2. Steinkuller, P.G., Du, L., Gilbert, C., Foster, A., Collins, M.L., Coats, D.K.: Childhood blindness. J. Am. Assoc. Pediatr. Ophthalmol. Strabismus. **3**, 26–32 (1999)
3. Camus, T.A., Wildes, R.: Reliable and fast eye finding in closeup images. In: IEEE 16th International Conference on Pattern Recognition, Quebec, Canada, pp. 389–394 (2004)
4. Augustynek, M., Penhaker, M.: Non invasive measurement and visualizations of blood pressure. Elektronika ir Elektrotechnika. **16**(10):55–58 (2011)
5. Daugman, J.G.: How iris recognition works. In: Nicole, R. (ed.) Proceedings of 2002 International Conference on Image Processing, 2002, Title of paper with only first word capitalized, J. Name Stand. Abbrev., in press (2002)
6. Kasik, V., Penhaker, M., Novak, V., Pustkova, R., Kutalek, F.: Bio-inspired genetic algorithms on FPGA evolvable hardware. In: Intelligent Information and Database Systems (Aciids 2012), Pt Ii. 2012, vol. 7197, pp. 439–447
7. Wildes, R., Iris recognition: An emerging biometric technology. In: Proceedings of the IEEE, vol. 85(9) (1997)
8. Wildes, R., Asmuth, J., Green, G., Hsu, S.: A system for automated iris recognition. In: Proceedings IEEE Workshop on Applications of Computer Vision (2004)
9. Mahboubeh, S., Kenari, P.: Fast algorithm for iris localization using Daugman circular integro differential operator. In: Proceedings of the IEEE International Conference of Soft Computing and Pattern Recognition (2009)
10. Penhaker, M., Matejka, V. (eds.) Image Registration in Neurology Applications. In: 2010 International Conference on Networking and Information Technology (ICNIT), 11–12 June 2010

11. Park, M., Jin, J.S., Luo, S.: Locating the optic disc in retinal images. In: Proceedings of the International Conference on Computer Graphics, Imaging and Visualisation, page 5, Sydney, QLD, Australia. IEEE, July (2006)
12. Zhu, X., Rangayyan, R.M.: Detection of the optic disc in images of the retina using the Hough transform. In: Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Vancouver, BC, Canada, 20–24 August 2008. IEEE, pp. 3546–3549 (2008)
13. Hussain, A.R.: Optic nerve disc head segmentation using genetic active contours. In: Proceedings International Conference on Computer and Communication Engineering. Kuala Lumpur, Malaysia, IEEE, pp. 783–787 (2008)
14. Nathans, J.: The challenge of macular degeneration. Sci. Am. **285**(4), 68–75 (2001). doi:10.1038/scientificamerican1001-68
15. Ells, A., Holmes, J.M., Astle, W.F., Williams, G., Leske, D.A., Fielden, M., Uphill, B., Jennett, P., Hebert, M.: Telemedicine approach to screening for severe retinopathy of prematurity: A pilot study. Ophthalmology **110**(11), 2113–2117 (2003). doi:10.1016/S0161-6420(03)00831-5
16. Ma, L., Wang, Y., Tan, T.: Iris recognition using circular symmetric filters. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Young, M, The Technical Writer's Handbook. University Science, Mill Valley, CA, 1989 (2002)
17. Fred, L.: Bookstein fitting conic sections to scattered data. Comput. Graph. Image Process. **9**, 56–71 (1979)
18. Daugman, J.G.: High confidence visual recognition of persons by a test of statistical independence. IEEE Trans. Pattern Anal. Mach. Intell. **15**(11), 1148–1161 (1993)
19. Jacobs, I.S., Bean, C.P.: Fine particles, thin films and exchange anisotropy. In: Rado, G.T., Suhl, H. (eds.) Magnetism, vol. III, pp. 271–350. Academic, New York (1963)
20. Martin-Roche, D., Sanchez-Avila, C, Sanchez-Reillo, R.: Iris recognition for biometric identification using dyadic wavelet transform zero-crossing. IEEE Aerosp. Electron. Syst. Mag, pp. 3–6 (2002)
21. Proenca, H., Alexandre, L.A.: Iris segmentation methodology for non-cooperative recognition. Vis Image Signal Proc IEE Proc **153**(2), 199–205 (2006)
22. Osareh, A., Mirmehd, M., Thomas, B., Markham, R.: Comparison of colour spaces for optic disc localisation in retinal images. In: Proceedings 16th International Conference on Pattern Recognition. Quebec City, Quebec, Canada, pp. 743–746 (2002)
23. Hoover, A., Goldbaum, M.: Locating the optic nerve disc in a retinal image using the fuzzy convergence of the blood vessels. IEEE Trans. Med. Imag. **22**(8), 951–958 (2003). doi:10.1109/TMI.2003.815900
24. Cerny, M.,Penhaker, M., Ieee.: The HomeCare and circadian rhythm (2008). ISBN:978-1-4244-2254-8, doi:10.1109/ITAB.2008.4570546
25. Cerny, M., Penhaker, M.: Wireless body sensor network in health maintenance systems. Elektron Ir Elektrotechn, 113–116 (2011). ISSN: 1392-1215. doi:10.5755/j01.eee.115.9.762
26. Penhaker, M., Cerny, M., Rosulek, M.: Sensitivity analysis and application of transducers. Paper presented at the 5th international workshop on wearable and implantable body sensor networks, BSN 2008, in conjunction with the 5th International Summer School and Symposium on Medical Devices and Biosensors, ISSS-MDBS 2008, Hong Kong (2008)

Extraction of Blood Vessels Using Multilevel Thresholding with Color Coding

Jan Kubicek, Jan Valosek, Marek Penhaker, Iveta Bryjova and Jan Grepl

Abstract The proposed work deals with the design of algorithm for segmentation of blood vessels in order to extract the geometrical features of analyzed vessels. The algorithm was tested on a sample of 25 clinical data to identify blood vessels direction and suppress surrounding tissues. The proposed methodology is effective for practical purposes, because identifies minor visual differences on the surface of blood vessels. These changes often represent pathological lesions that cannot be identified from the native data. The algorithm is based on the identification of several significant levels to make ideal thresholding, according to object features. For even better diagnostic results image background is subsequently filtered out from the segmented image, background may have a disturbing effect in the diagnosis and the final segmented image contains only analyzed blood vessels.

Keywords Medical image processing · Multilevel thresholding · Blood vessels · Image segmentation · Shape detection · Angiography · Roi

---

J. Kubicek (✉) · J. Valosek · M. Penhaker · I. Bryjova · J. Grepl
VSB–Technical University of Ostrava, FEI, K450, 17. Listopadu 15, 70833 Ostrava–Poruba, Czech Republic
e-mail: jan.kubicek@vsb.cz

J. Valosek
e-mail: valosekjan@icloud.com

M. Penhaker
e-mail: marek.penhaker@vsb.cz

I. Bryjova
e-mail: iveta.bryjova@vsb.cz

J. Grepl
e-mail: jan.grepl@vsb.cz

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_33

1 Introduction

Blood vessels form an essential element of the vascular system. Vascular system consists arteries, veins, capillaries and lymph vessels. All these parts are essential for the good condition of the human organism. It is obvious, there is a strong requirement on contrast imaging of individual blood vessel for extraction geometrical features, which correspond with physiological condition of vascular system. Analysis of blood vessel from native data is quite difficult because we have to work with grey levels. In many cases we are not able to recognize proper direction and narrowing of analyzed vessels [1–3].

2 Medical Imaging Systems

Imaging systems allow better medical diagnosis of the examined parts of the human system, thus largely contributing to the prevention and treatment of disease. Examination of the blood vessels system is called angiography. Angiography can be performed using computed tomography or magnetic resonance imaging. Each of these modalities offers certain advantages and disadvantages. In some cases it is necessary the giving of contrast media to the patient and his subsequent exposure to X-rays in the case of using computed tomography angiography, as well as a longer examination time. In the case of examination on magnetic resonance, there are contraindications for the group of patients with metal implants or prostheses. Benefits in the form of visualization of blood stream in 3D mode and good resolution of the imagined vessels, however, are prevailing, and therefore these methods are widely used in clinical practice [4, 5].

3 Related Work

The main objective of work is design of suitable segmentation method for blood vessels extraction and elimination structures, which do not correspond with analyzed blood vessels. Significant benefit of proposed algorithm is color coding of individual blood vessels into separated sets. This feature is important especially in cases, when we need to perform mapping of vessels surface to identify pathological changes in early stage.

In the past more methods in this area have been published. Many of these method work with grey level intensities. For example detection of blood vessels has been performed on the base of eigenvalues of Hessian matrix. Other solutions based on active contours enhanced by the use of the interaction between the boundary of the object and expanding contours is presented in [6]. Thresholding for detection of objects in an image was presented in 1979 [7] and called Otsu method. This method is frequently used in many segmentation approaches for objects identification. We used modified Otsu method with multilevel thresholding for extraction of blood vessels. Significant benefit of proposed solution is color coding. In the color spectrum is much simpler to identify lesions. These lesions usually are not obviously visible from native MRI data [8, 9].

4 Proposed Multilevel Thresholding Algorithm

The main aim of this article is to present the design of an algorithm for segmentation of blood vessels images based on Otsu method. On angiographic images may not be vasculature system always well detectable and easily recognizable for medical staff. Main problem is the grey spectrum of native images. Individual structures are often badly recognizable because they are represented by similar grey levels. In many cases are difficult to identify pathological lesion from gray level intensities.

The proposed algorithm segments the object in image according to the number of selected thresholding levels, thus reduced the number of shades of gray in the image and on the base this fact direction of blood vessels is emphasized. Subsequently blood stream area is separated from image background. On the base of this procedure vascular system is highlighted. In the final part of algorithm individual detected vessel are transformed to color spectrum [7, 10–12].

The big advantage of segmentation algorithm is using for single image and for image series as well. This significant feature uses dynamic indexation in multi slice matrix. It is possible to load entire image series to multi slice matrix and perform segmentation for each slice. On the algorithm output we can compare original image and segmented result and arbitrarily move among individual slices. There is strong requirement for size of image. When we put images into matrix, all images must have same size. This requirement has been kept for our analyzed data.

4.1 Otsu Method

Otsu segmentation method is based on determining the most ideal level thresholding for the analyzed image. This thresholding level is found by calculating the value of either within-class variance or between class variance.

In terms of computing efficiency is more effective to set the threshold value based on the between class variance, where the threshold value between the highest-variance threshold is determined as an ideal. For the calculation of between class variance is necessary to first calculate the weight and the average intensity values.

The pixels of the image have different shades of gray, those are labeled as L interval [0,1, ..., L]. The number of pixels in given gray level i is signed as n_i. Whole number of pixels is given: $N = n_0 + n_1 + \cdots + n_L$. When we divide pixels into two groups 1 and 0 (0 represents the image background, 1 represents the image foreground), according to the level thresholding.

The weight W and the average intensity values μ are calculated by following equations:

$$W_0 = \sum_{i=1}^{k} \frac{n_i}{N} \tag{1}$$

$$W_1 = \sum_{i=k+1}^{L} \frac{n_i}{N} \tag{2}$$

$$\mu_0 = \sum_{i=1}^{k} \frac{n_i * i}{N_k} \tag{3}$$

$$\mu_1 = \sum_{i=k+1}^{L} \frac{n_i * i}{N - N_k} \tag{4}$$

Between class variance σ^2 is consequently given by equation:

$$\sigma^2 = W_0 * W_1 * (\mu_0 - \mu_1)^2 \tag{5}$$

Between class variance is calculated for every possible thresholding level and level with the highest value of between class variance is determined as the most ideal specified image is segmented according to it [7, 13–21].

4.2 Design of Proposed Multilevel Thresholding Algorithm

Segmentation method with only one thresholding is not suitable for processing of medical images. In this area, it is much better to use improved Otsu method with multiple level thresholding, which increases sensitivity of blood vessels detection.

Proposed Otsu thresholding method finds specific levels for a image based on the histogram distribution into equally large areas. Specific thresholding level is used for each area. The analyzed image is consequently segmented according to all thresholding levels.

Individual pixels with different shade levels are labeled as L with range: [0,1, …, L]. Number of thresholding levels is given as p. The size of one segmented area is given by equation:

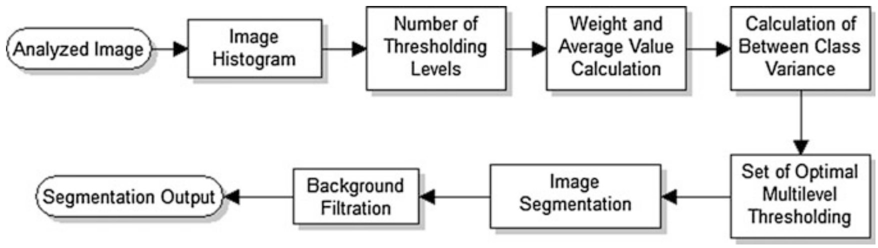


Fig. 1 Block diagram of proposed algorithm

$$a = \frac{L}{p} \tag{6}$$

Between class variance σ^2 is calculated similarly as Otsu method:

$$\sigma^2 = W_0 * W_1 * (\mu_0 - \mu_1)^2 \tag{7}$$

The number of divided regions of the histogram is equal to the number of thresholding levels p. Optimal thresholding levels for individual areas are given:

$$P_p = max_p\left(\sigma^2\right) \tag{8}$$

For the validity of the above formulas is necessary that the number of pixels in different shades of gray L is equal to 256 * j. Number of thresholding levels p must be from set: [2*j,4*j,8*j], where j belongs to set: [1,2,…,∞]. According to practical results, it is recommended to use set: [1, 2, …, 8].

The proposed algorithm for blood vessels segmentation is described by following diagram (Fig. 1).

4.3 Filtration of Image Background

After segmentation, resulting segmented image contains p + 1 gray shades. The segmented image still contains image background with extracted blood vessels. For achieving better diagnostic information is necessary to filter out image background and keep segmented blood stream only.

Filtration works on principle keeping of pixels with value which corresponds of blood vessels area. Other pixels should be suppressed. After segmentation process almost blood vessels are represented by set: [p − 1, p, p + 1], in some cases: [p − 2, p − 1, p, p + 1]. From these cases we can choose more appropriative alternative. Pixels with value less then p − 1 or p − 2 are filtered. Filtration is done by replacing zero values. With this procedure we get color map of blood vessel and background spectrum is suppressed to one color.

4.4 Algorithm Testing for Real Patient Data

The proposed algorithm has been tested on the sample of the 25 patients records with cooperation of Trinec Hospital. We were focused especially on data from MRI and CT angiography. The following image outputs show usability of proposed multilevel algorithm for cases of three patients. The first segmentation output is presented on the neck area which is acquired using magnetic resonance angiography. Second output is performed on neck area image, which is acquired from computer tomography angiography. These two outputs denotes on effective usability of segmentation algorithm for same area acquired from different modalities. The third case shows segmentation of pelvis area from MRI. In all cases is possible to perform segmentation either on the area of whole image or on extracted Region of Interest (RoI).

Algorithm testing for neck area acquired from MRI angiography (Figs. 2, 3 and 4).

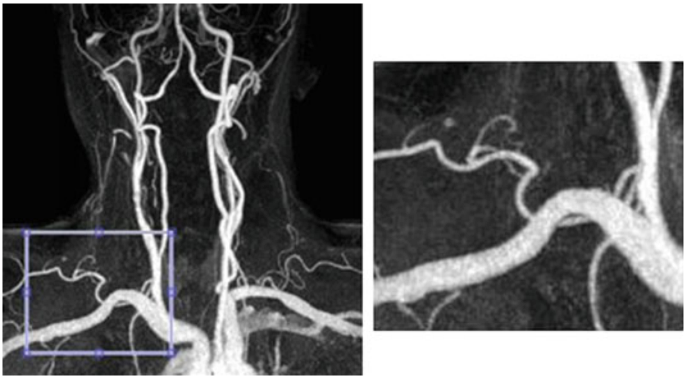


Fig. 2 Original image (*left*) and selected RoI (*right*)

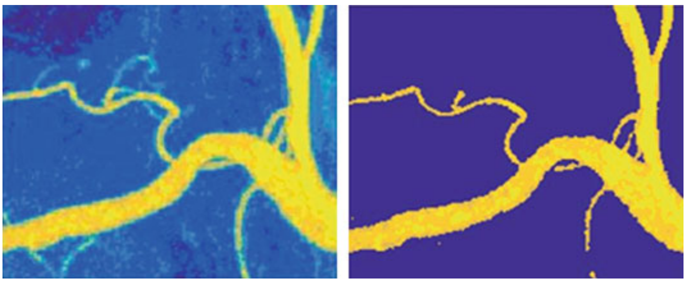


Fig. 3 Segmentation output (*left*) and background filtration (*right*)

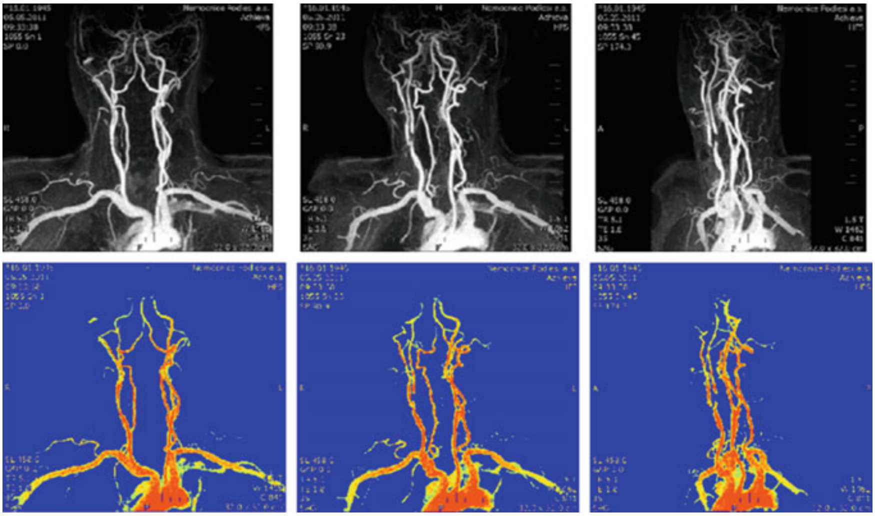


Fig. 4 Example of segmentation algorithm for neck image series. *Upper row*—original image data, *lower row*—segmentation output

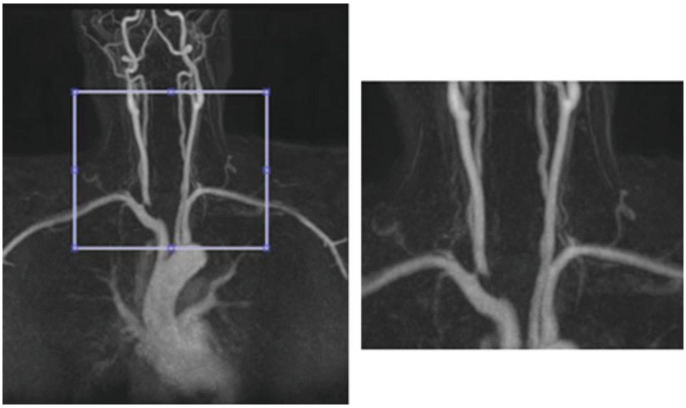


Fig. 5 Original image (*left*) and selected RoI (*right*)

Application of proposed multilevel thresholding method has been tested on the neck area from MRI angiography. For this case 256 gray levels have been substituted by 8 thresholding levels.

Algorithm testing for neck area acquired from CT angiography (Figs. 5, 6).

In this case 256 gray levels have been also substituted by 8 thresholding levels. Segmentation for bow cases gives satisfactory results with effective separation image background and vessels direction.

Algorithm testing for pelvis area acquired from MRI angiography (Figs. 7, 8).

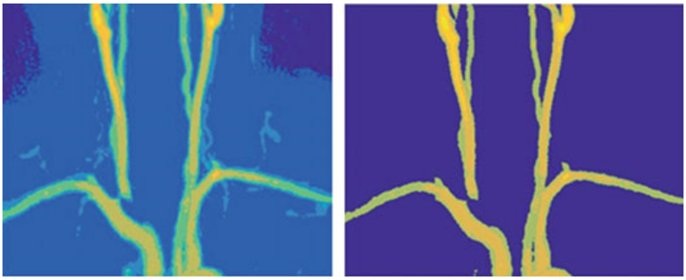


Fig. 6 Segmentation of neck area (*left*) and background filtration (*right*)

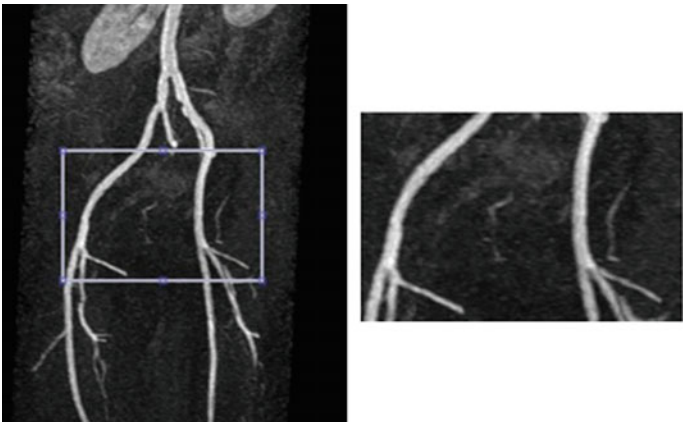


Fig. 7 Original image of pelvis area (*left*) and selected RoI (*right*)

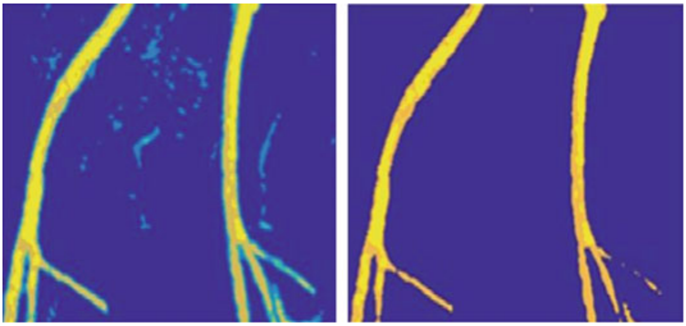


Fig. 8 Segmentation of pelvis blood vessels (*left*) and background filtration (*right*)

In this case proposed algorithm has been used for segmentation of pelvis blood vessels acquired from MRI angiography. For this image set number of thresholding levels could be reduced to 4 levels.

5 Conclusion

Segmentation and consecutive extraction objects of interest is frequently used task in the field of medical image processing. In this article we propose suitable method for extraction direction and shape of blood vessels. This segmentation procedure is necessary for extraction of geometrical parameters analyzed blood stream. On the base this method we are able to specify vessels diameter, blood filling and other parameters which asses condition of blood vessels system. Proposed segmentation algorithm gives considerable improvement for blood vessels extraction. We are able to extract individual blood vessels without adjacent structures. It is first step for measuring geometrical parameters of blood stream. In the coming time we want to focus on extraction of geometrical shape of individual blood vessels and consequently to calculation of tortuosity. It is key parameter for assessment of physiological condition blood vessels system.

Acknowledgments The work and the contributions were supported by the project SP2015/179 'Biomedicínské inženýrské systémy XI' and This work is partially supported by the Science and Research Fund 2014 of the Moravia-Silesian Region, Czech Republic and this paper has been elaborated in the framework of the project "Support research and development in the Moravian-Silesian Region 2014 DT 1—Research Teams" (RRC/07/2014). Financed from the budget of the Moravian-Silesian Region.

References

1. Holibková, Alžběta a Stanislav Laichman. Přehled anatomie člověka. 4. vyd. Olomouc: Univerzita Palackého v Olomouci, 140 s (2006). ISBN 80-244-1480-5
2. Javorka, K.: Lekárska fyziológia: učebnica pre lekárske fakulty. 1. vyd. Martin: Osveta, 678 s., obr. (2001). ISBN 80-806-3023-2
3. Stříteský, Jan. Patologie. 1. vyd. Olomouc: EPAVA, 2001, 388 s., obr. ISBN 80-86297-06-3
4. Penhaker, Marek a Martin Augustynek. Zdravotnické elektrické přístroje 1. 1. vyd. Ostrava: VŠB - Technická univerzita Ostrava, 2013, 411 s. ISBN 978-80-248-3107-7
5. Nekula, J.: Radiologie. 3. vyd. Univerzita Palackého, Olomouc (2005). ISBN 80-244-1011-7
6. Xiang, Y., Chung, A.C.S., Je, J.: An active contour model for image segmentation based on elastic interaction. J. Comput. Phys. **219**(1), 455–476 (2006). doi:10.1016/j.jcp.2006.03.026
7. A threshold selection method from gray-level histograms. IEEE Trans. Syst Man Cybernet. **9** (1), s. 62–66. doi:10.1109/TSMC.1979.4310076
8. Kubicek, J., Penhaker, M., Bryjova, I.: Articular cartilage defect detection based on image segmentation with colour mapping. In: 6th International Conference on Computational Collective Intelligence Technologies and Applications, 24th–26th September, 2014 Seoul, Korea, vol. 8733, pp. 214–222. Springer International Publishing, Print ISBN 978-3-319-11288-6, Online ISBN 978-3-319-11289-3, Series ISSN 0302-9743. doi:10.1007/978-3-319-11289-3_22
9. Kubicek, J., Penhaker, M.: Fuzzy algorithm for segmentation of images in extraction of objects from MRI. In: International Conference on Advances in Computing, Communications and Informatics (ICACCI, 2014, pp. 1422–1427, 24–27 Sept 2014. doi:10.1109/ICACCI.2014.6968264

10. Frangi, Alejandro Federico. Three-dimensional model-based analysis of vascular and cardiac images = Driedimensionale Modelgestuurde Analyse van Vaat- en Hartbeelden (met een samenvatting in het Nederlands)/Alejandro F. Frangi. Wageningen: Ponsen & Looijen, 2001. ISBN 9039326479
11. Lorenz, C., Carlsen, I.C., Buzug, T.M., Fassnacht, C., Weese, J.: CVRMed-MRCAS '97: First Joint Conference Computer Vision, Virtual Reality and Robotics in Medicine and Medical Robotics and Computer-Assisted Surgery, Grenoble, France, 19–22 March 1997. In: Proceedings Multi-scale Line Segmentation with Automatic Estimation of Width, Contrast and Tangential Direction in 2D and 3D Medical Images. č. 1205, s. 233–242 (1997). doi:978-3-540-62734-0
12. Zhang, F., Zhang, X., Liu, X., Cao, K., Haishun, D., Yanbin, C.: Blood vessel enhancement for DSA images based on adaptive multi-scale filtering. Optik-Int. J. Light Electron Opt. **125** (10), 2383–2388 (2014). doi:10.1016/j.ijleo.2013.10.111
13. Yao, C.A., Chen, H.J.: Automated retinal blood vessels segmentation based on simplified PCNN and fast 2D-Otsu algorithm. J. Cent. S. Univ. Technol. **16**(4), 640–646 (2009). doi:10.1007/s11771-009-0106-3
14. Rodríguez, R.: A strategy for blood vessels segmentation based on the threshold which combines statistical and scale space filter. Comput. Methods Programs Biomed. **82**(1), 1–9 (2006). doi:10.1016/j.cmpb.2005.10.008
15. Liu, S.: Detection of the optic disc on retinal fluorescein angiograms. J. Med. Biol. Eng. **31**(6), 405–412 (2011). doi:10.5405/jmbe.773
16. Guthrie, M.J., Osswald, C.R., Valio, N.L., William, F., Mieler, A., Kang-Mieler, J.J.: Objective area measurement technique for choroidal neovascularization from fluorescein angiography. Microvasc. Res. **91**, 1–7 (2014). doi:10.1016/j.mvr.2013.11.005
17. Augustynek, M., Pindor, J., Penhaker, M., Korpas, D., Society, I.C.: Detection of ECG significant waves for biventricular pacing treatment. In: Proceedings of Second International Conference on Computer Engineering and Applications: Iccea 2010, vol. 2, pp. 164–167. ISBN:978-0-7695-3982-9, doi:10.1109/ICCEA.2010.186 (2010)
18. Augustynek, M., Labza, Z., Penhaker, M., Korpas, D., Society, I.C.: verification of set up dual-chamber pacemaker electrical parameters. In: Proceedings Second International Conference on Computer Engineering and Applications: Iccea 2010, vol. 2, pp. 168–172. ISBN:978-0-7695-3982-9. doi:10.1109/ICCEA.2010.187 (2010)
19. Penhaker, M., Darebnikova, M., Cerny, M.: Sensor network for measurement and analysis on medical devices quality control. In: E-Technologies and Networks for Development, vol. 171, pp. 182–196 (2011)
20. Cerny, M., Pokorny, M.: Circadian rhythm evaluation using fuzzy logic. Adv. Methods Computat. Collective Intell. **457**, 289–298 (2013)
21. Cerny, M., Penhaker, M.: Wireless body sensor network in health maintenance systems. Elektron. Ir Elektrotechn. **9**, 113–116 (2011)

User Behavior Based Call Admission Control for Traffic Steering in Wi-Fi/Cellular Networks

Lim Yun Yi, Azizul Rahman Mohd Shariff and Khuzairi Mohd Zaini

Abstract The expanded growth in smartphones and applications, hence data traffic resulted in cellular networks being conspicuously overwhelmed with demand for resources. Service providers constantly looking for suitable technologies and mechanisms to reduce the network congestion by offloading the data traffic, where Wi-Fi is an excellent option. In the traffic steering process, network selection is the most vital process, as it affects the user's network experience. This paper focuses on enhancing the procedure of Call Admission Control (CAC). Two key functions were introduced into the call admission procedure, namely Channel Reserve Engine (CRE) and Traffic Steering Policy (TSP). Based on the simulation results, it is found the proposed CAC reduces the call blocking probability as compared with conventional CAC. Thus, it is suitable to be implemented in WLAN/Cellular integration.

1 Introduction

Over the past few years, the growth of cellular networks, smartphones and mobile applications has expanded the global information and communication technology (ICT) industry. This demanded very high resource requirements in already deployed cellular networks, consequently making the network overwhelmed with data traffic. In 2014, Cisco has released a new document, titled Cisco Visual Networking Index (VNI) Global Mobile Data Traffic Forecast for 2013–2018 [1]. According to this report, a sharp increase of 81 percent is shown in 2013 global

L.Y. Yi · A.R.M. Shariff (✉) · K.M. Zaini
School of Computer Science, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia
e-mail: azizulrahman@usm.my

L.Y. Yi
e-mail: lyy13_sk54@student.usm.my

K.M. Zaini
e-mail: khuzairi@uum.edu.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_34

mobile data traffic. The data traffic has reached 1.5 exabytes per month at the end of 2013 from 820 Petabytes per month at the end of 2012. This explosion in data growth pushed the service providers to constantly looking for suitable network mechanisms to offload data traffic and reduce network congestion, where Wi-Fi is an excellent option [2]. This is because Wi-Fi has enormous capacity and it supports dense access point deployments. In addition, newer Wi-Fi technologies have been proposed such as IEEE 802.11ac that not only provides much higher data capacity but also with some QoS guarantees suitable as secondary means to providing data resources. Based on previous studies, the implementation of Wi-Fi/Cellular network integration can be classified into two, which are loosely coupled integration [3] and tightly coupled integration [4]. This paper focuses on tightly coupled integration. The implementation of tight coupling Wi-Fi/Cellular network is by connecting the Wi-Fi access point to a Radio Network Controller (RNC). Tight coupling exhibit reduced architectural complexities in providing a seamless handover between both networks, improving the QoS hence much better user quality experience. In the tight coupling integration, there are three main components which responsible for regulating the process of handover. These components are policy provisioning module (PPM), Connection Admission Control module (CAC), and QoS Mobility Management Module (MMM). Connection Admission Control, also known as Call Admission Control (CAC) is responsible for managing incoming traffic flow and bandwidth allocation. Besides, it is also responsible for maintaining the QoS of existing connections in the network. When there is new traffic flow comes in, CAC will either accept or reject the traffic. The decision is made based on whether there is available bandwidth to allocate for the particular flow. In the conventional CAC, if there is available flow, it will accept the request. The opposite, it rejects the flow [5]. This work proposed an enhancement to the CAC by introducing two key functions, namely Channel Reserve Engine (CRE) and Traffic Selecting Policy (TSP). The rest of the paper is organized as follows. The related works are presented in Sect. 1. Section 2 provides the methodology of the proposed research work. Simulations results and related discussions are presented in Sect. 3. Lastly, the conclusion of this work is presented in Sect. 4.

1.1 Related Works

In traffic steering process, decision making is a very crucial step that it affects the changes in many aspects, such as latency, cost, security, power consumption, speed and etc. Fundamentally, traffic steering decision is dependent on a few criteria, which can be divided into 4 groups [6]; *Network* (*bandwidth, latency, signal strength, cost* etc.), *Terminal* (*velocity, battery, location*), *User* (*user profile, application usage, QoE*), *Service* (*service capacity, QoS*). Most work is focused on network criteria in making the handoff decision [7]. User preference is considered as one of the important criteria for traffic steering decision. Throughout the study in

[8], the authors figured that the user behavior affects the performance of cellular networks. Therefore, user behavior is considered as one of the important criteria. In [9], the authors stated that no existing work takes mobility criteria of users into consideration. However, in the study [10] shows that user mobility is the most important feature in the mobile wireless network. This is because user mobility has a profound effect on the QoS provisioning.

Besides mobility, the study of user behavior based criteria is also investigated. In [8], the work proposed three user behavior based criteria, which are call holding time, cell residency time, and channel holding time. In the research, the definition of call holding time is the duration of the requested call connection. Meanwhile, the cell residency time refers to the amount of time a mobile user spends in a cell. Lastly, channel holding time defines how long a call will use this channel before completion or change connection to another cell. In [11], the researcher stated that previous work has neglected some of the important network aspects during the decision making process, such as network congestion level, switching penalty, and pricing in the network. Thence, they proposed a group of criteria for traffic steering decision which consists of user behavior criteria and network criteria. The proposed criteria are user mobility; location, user, and time dependent Wi-Fi availabilities; network dependent switching time and switching cost for changing network connections; and usage based pricing. Literature review of existing works shows that, decision of network selection involves various types of criteria. Earlier works focuses on network based criteria or users preference or combination of both. This is because both network and user criteria are considered as principal factors. However, it is important as well to consider both network and users' condition, such as applications metrics (class of traffic, average download/uploads speeds), network congestion level (available resources, density of users, threshold settings), network architecture (offloading deployment plan, location of users, cell radius). Therefore, this work proposes a group of factors which comprises of users' behavior and network selection to be considered in selecting the network.

2 Proposed Research Work

In this section, there are three main focus of the proposed work to be discussed. Firstly, the proposed architecture of Wi-Fi/Cellular network. This work proposed a better definition and understanding of Wi-Fi/Cellular network architecture for future reference. Next, the proposed CAC included two proposed key functions, which are the Channel Reserve Engine (CRE) and Traffic Selecting Policy (TSP) to enhance CAC procedure. Finally, the proposed network selection criteria are the users' behavior context (user's network and mobility) and network selection, which has not been studied in previous work. The definition of Wi-Fi/Cellular network architecture is proposed, as shown in Fig. 1. As depicted, users are categorized into three categories. The first category of user is category 'W', where the users are located in Wi-Fi footprint and connected to the Wi-Fi. The second category is

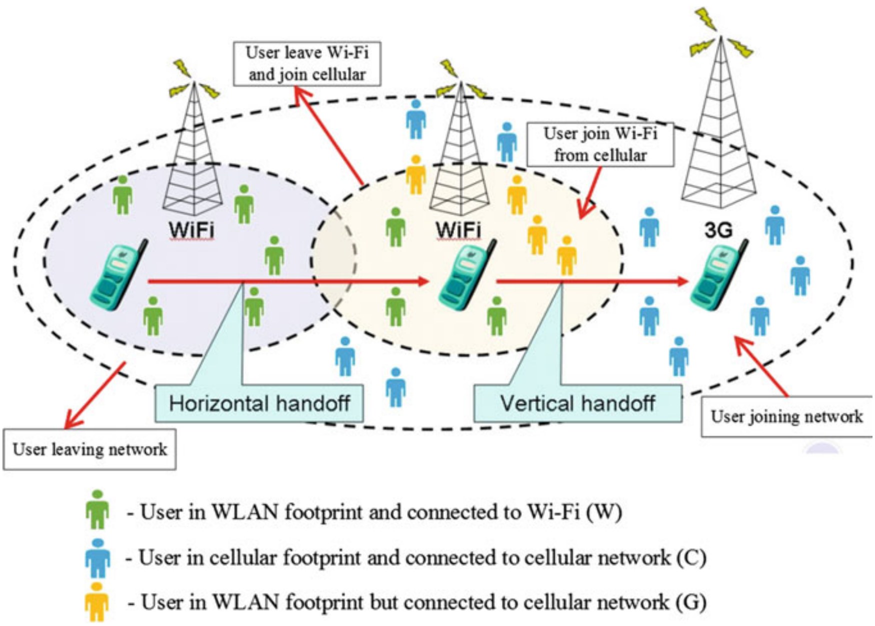


Fig. 1 Proposed Wi-Fi/Cellular architectural reference for traffic steering

category ‘C’, which refers to users in cellular footprint and connected to the cellular network. The last category is category ‘G’. Users from this category are located in Wi-Fi footprint, but connected to the cellular network. In the proposed architecture, only users under category ‘G’ are selected for the traffic steering process. This is because only users that are co-located in cellular and Wi-Fi networks can be steered i.e. a user connected to a cellular network, with an ongoing data session can be steered to a Wi-Fi network.

In conventional CAC, user’s request acceptance is made based on the whether there is available bandwidth to allocate. If there is no available bandwidth, a user’s request will be rejected. In order to enhance the CAC procedure for traffic steering, this work proposed two key functions to be included in call admission procedure. The two key functions are Channel Reserve Engine (CRE) and Traffic Selecting Policy (TSP). The responsibility of CRE is to compare a user’s requirement from CAC and check if there is sufficient bandwidth to allocate. If there is no sufficient bandwidth to allocate for the user, it triggers and passes the information to TSP to execute traffic offloading.

TSP is a set of policies for traffic offloading and bandwidth allocation. It is also responsible in selecting the appropriate network for handoff. In this step, TSP checks if there is a suitable user or users that can be offloaded based on proposed criteria, which are user behavior and network selection. For this, the TSP will first identify if a user is in category G, which is located within Wi-Fi footprint but connected to the cellular network. Subsequently, TSP identifies the user’s traffic profile in this case the downloading percentage or data progress. TSP only select user whose traffic is data services. This is because data services have less strict requirement on latency adhering commonly to the 'best-effort' service. As to maintain data services performance, TSP will only look into user whose data progress is less than or equal to 50 %. If there is no available user to offload, TSP reject the user's request. Figures 2 and 3 illustrates the general view of the proposed CAC procedure for Cellular/Wi-Fi interworking and the details of proposed call admission procedure for Cellular/Wi-Fi interworking respectively. The proposed criteria in this work focused on exploiting the combination of user behavior and network selection. In here, the user behavior context refers to user's traffic and mobility. For user's traffic, it is categorized into four categories, which are called

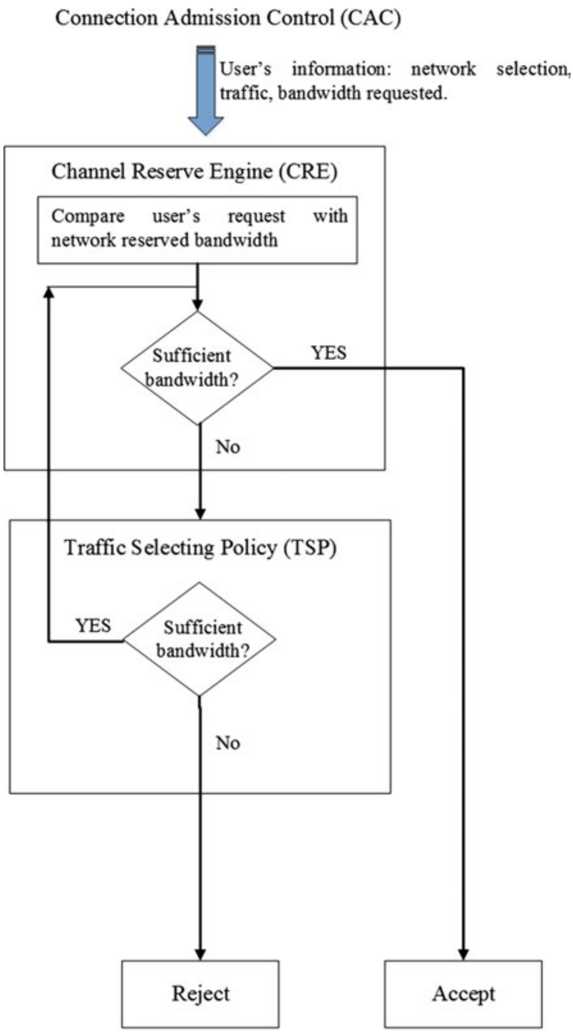


Fig. 2 Proposed enhancement to the conventional call admission control (CAC) procedure

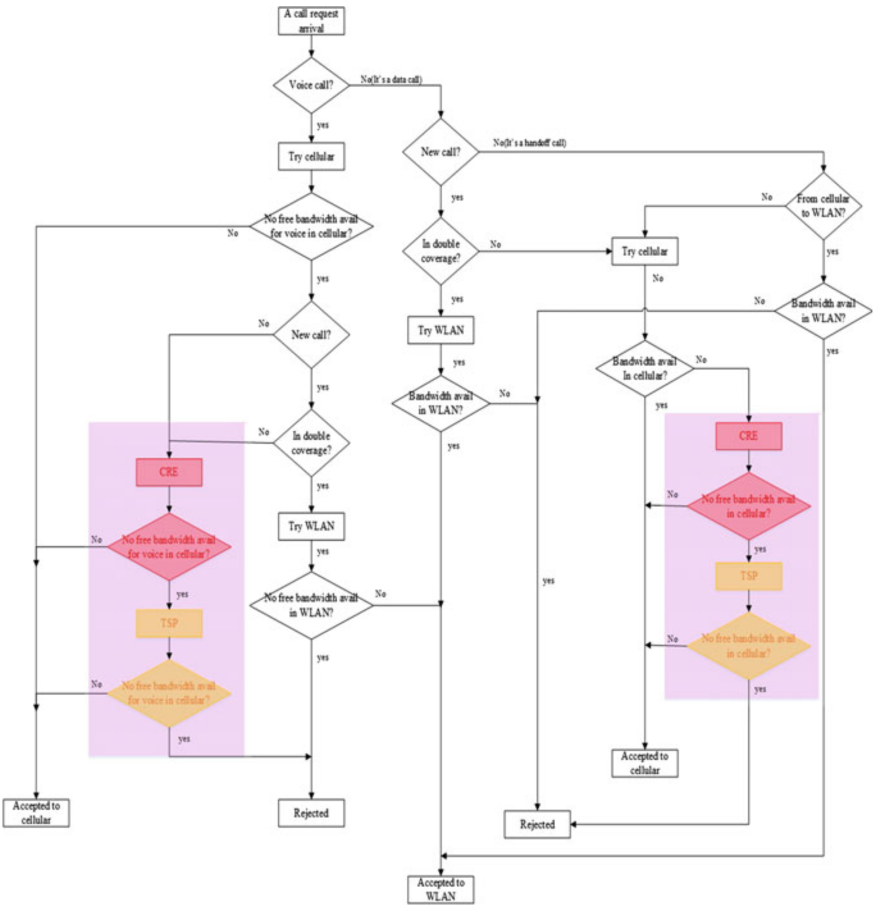


Fig. 3 Proposed call admission control (CAC) procedure for Wi-Fi/Cellular traffic steering

services, session services, data services, or no service. Each category is given different levels of priority, such that call services is categorized as high priority, medium priority for session services and low priority for data services. When traffic steering is triggered, only users with low priority traffic (data services) will be offloaded.

As for network selection, it refers to the user's connected network with respect to their located footprint area. As mentioned in previous paragraphs, the proposed architecture of Wi-Fi/Cellular network classified users into three: category 'C', 'W', and 'G'. When traffic steering is triggered, network will only select and offload user under category 'G' to Wi-Fi. This is because Wi-Fi only provides services within a limited area. Only users located within Wi-Fi coverage area able to make a connection. Moreover, by offloading user traffic from category 'G' to Wi-Fi reduces the pressure on the cellular network and enables new user's traffic to be accepted into the cell. Therefore, users under category ‘G’ are the most suitable to be offloaded compared with users under category ‘C’ who are not in Wi-Fi service coverage. The strength or capacity of how much steering can be performed is dependent on the number of users co-located in the cellular and Wi-Fi networks.

3 Simulation Results and Analysis

Call blocking is considered as one of the key performance metric [12]. Therefore, the evaluation of the proposed work focuses on the call blocking probability. As for the evaluation, four different parameters has been chosen to examine the call blocking probability. The four selected parameters are Wi-Fi thresholds, number of existing users in Wi-Fi and Cellular network, number of users joining Wi-Fi and Cellular network, and number of users leaving Wi-Fi and Cellular network.

3.1 Evaluation of Call Blocking Probability Based on the Varying Wi-Fi Threshold Value

In this, different value of Wi-Fi threshold is used. The Wi-Fi threshold value is preset to 600 kbps (60 %), then gradually increase to 700 kbps (70 %) and 800 kbps (80 %). It is to discover the best threshold value that can be set for Wi-Fi in order to optimize the network performance. The thresholds acts as the operational limits of Wi-Fi capacity, allowing spare capacity to absorb users that are offloaded or steered from the Cellular network.

As in Fig. 4, simulation results clearly shows that the proposed CAC achieved better blocking probability compared with conventional CAC. The call blocking

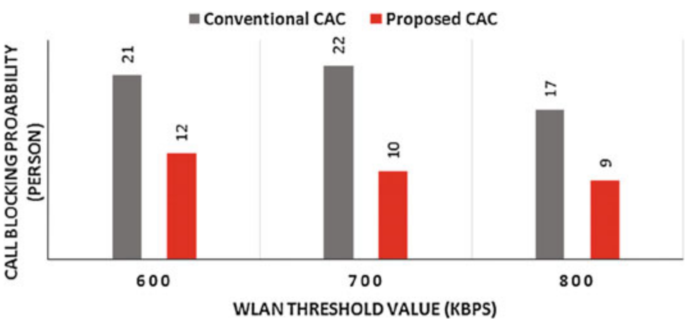


Fig. 4 Comparison of call blocking probability between conventional CAC and proposed CAC with varying thresholds

probability drops as the Wi-Fi threshold value increases. When Wi-Fi threshold is 600 kbps, the call blocking probability for conventional CAC and proposed CAC is 21 users and 12 users respectively. As Wi-Fi threshold is increased to 700 kbps, call blocking probability for conventional CAC is 22 users, and 10 users for proposed CAC. When Wi-Fi threshold value is increased to 800 kbps, the call blocking probability for both conventional CAC and proposed CAC show a slight decrement. Hence, it is understood that setting the appropriate thresholds determines the number of users that can be steered from a cellular to a Wi-Fi network. This threshold relates to the reserved capacity for the Wi-Fi network to accept users from the cellular network. From this results, it is understood that understanding and defining the optimum threshold is also key and they are highly dependent on local network and traffic dynamics.

3.2 *Evaluation of Call Blocking Probability Based on Varying Concurrent Number of Users in Wi-Fi and Cellular Network*

This simulation is to examine the call blocking probability with respect to varying number of concurrent users in Wi-Fi and Cellular network. From the simulation results in Fig. 5, call blocking probability increases as the number of existing users increases. A sharp increment is shown as the number of existing users is increased to 150, 200 and 150 users. When the number of existing users is 100 users, the call blocking probability for conventional and the proposed CAC almost achieved similar performance. As the number of existing users increase to 150 users, the call

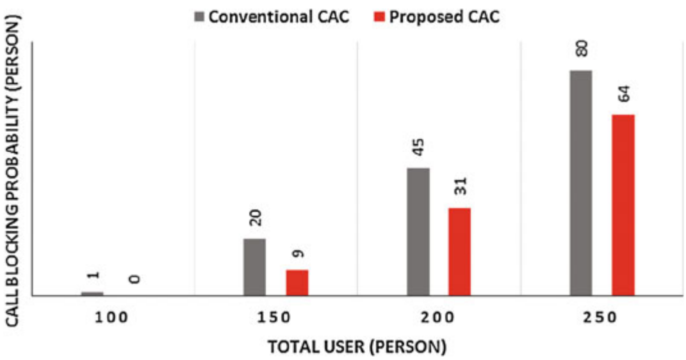


Fig. 5 Comparison of call blocking probability between conventional CAC and proposed CAC based on varying number of concurrent users in Wi-Fi and Cellular network

blocking probability for conventional CAC is 20 users and 9 users for proposed CAC. When total 200 of existing users are in the network, the call blocking probability for conventional CAC is 45 users and 31 users for proposed CAC. As the number of existing users increases to 250 users, the call blocking probability for conventional CAC is 80 users and 64 users for proposed CAC. The call blocking probability achieved by the proposed CAC shows a marked improvement as compared with conventional CAC.

3.3 *Evaluation of Call Blocking Probability Based on Varying Number of Users Joining Wi-Fi and Cellular Network*

This simulation is carried out by using different number of users joining the network. The objective is to study whether increasing the number of users joining, will affect the network usage and call blocking probability. The number of users joining the network is set to 20, 30 and 40 % of the number of users in Wi-Fi foot print.

From Fig. 6, blocking probability increases as the number of users joining increases. When the number of users joining is 20 % of the total users in Wi-Fi footprint, the call blocking probability for conventional CAC is twice compared to the proposed CAC. The call blocking probability shows a slight increase as the number of users joining increases to 30 and 40 % of the total users in Wi-Fi footprint. When 30 % of the users are joining, the call blocking probability for conventional CAC is 18 users and 9 users for proposed CAC. Similar trend is also seen with 40 % joining users. The call blocking probability achieved by the proposed CAC shows a better result compared with conventional CAC.

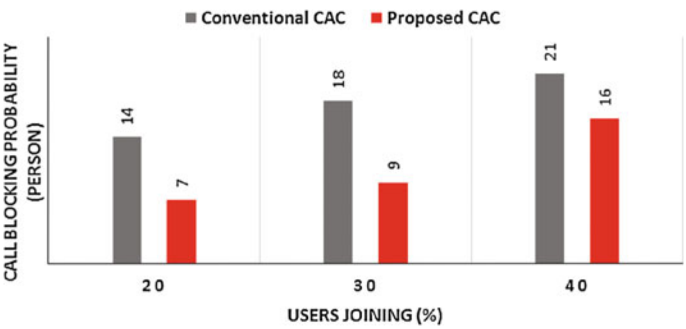


Fig. 6 Comparison of call blocking probability between conventional CAC and proposed CAC based on varying percentage of users joining Wi-Fi and Cellular network

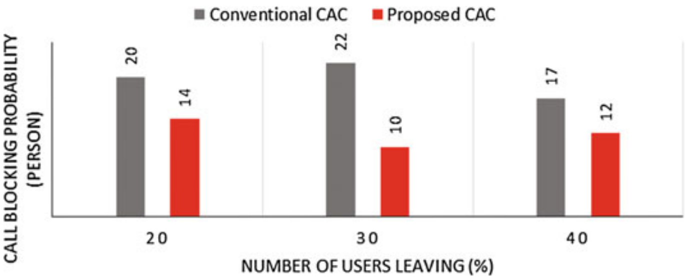


Fig. 7 Comparison of call blocking probability between conventional CAC and proposed CAC based on varying percentage of users leaving Wi-Fi and Cellular network

3.4 Evaluation of Call Blocking Probability Based on Varying Number of Users Leaving Wi-Fi and Cellular Network

This simulation studies the effect on call blocking probability with respect to the number of users leaving Wi-Fi and Cellular network. For this simulation, the total number of users leaving both networks is preset to 20 % of users in Wi-Fi footprint, and gradually increase to 30 and 40 %. For example, if there are 50 users in Wi-Fi footprint, then 20 % of the number of users leaving will be 10 users.

As depicted in Fig. 7, when 20 % of the users are leaving the network, the call blocking probability for conventional CAC and proposed CAC is 20 users and 14 users respectively. As the number of users leaving increased to 30 and 40 %, the call blocking probability for proposed CAC exhibits lower blocking on average by about 40 % lower than the conventional CAC. Overall, proposed CAC shows a lower pattern of call blocking, thus allowing better performance for cellular network.

4 Conclusions

From the proposed Wi-Fi/Cellular architecture and new CAC, the results proved that the proposed CAC achieved a lower overall call blocking probability as compared with conventional CAC. Offloading users on data service from a cellular network to a Wi-Fi network allows much better performance and reduces the resources demand pressure on cellular networks. This simple enhancements to the conventional CAC allows it to be feasibly implemented in Cellular/Wi-Fi network as a suitable CAC mechanism for traffic steering. As the integration of Wi-Fi/Cellular network is seen as a long term solution for the cellular network, further enhancements is by looking into fine-grained details of the CRE and TSP on the user behavior context, such as data service downloading/uploading ratio, signal strength, QoS levels and etc. Further studies are also needed to find the optimal threshold for cellular networks to accept new joining users i.e. category ‘C’ users, and optimal threshold for Wi-Fi to be able to accept category ‘G’ users steered from the cellular networks.

Acknowledgements This research is supported by the Universiti Sains Malaysia (USM) under the Fundamental Research Grant Scheme (FRGS).

References

1. Cisco Visual Networking Index.: Global mobile data traffic forecast update, 2013–2018. Retrieve on 27 Sept http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html (2014)
2. Ruckus.: Hotspot 2.0 making Wi-Fi as easy to use and secure as cellular. http://theruckusroom.typepad.com/files/hotspot_2.0_wp_2013.pdf (2013)
3. Song, W., Jiang, H., Zhuang, W., (Sherman) Shen, X.: Resource management for QoS support in cellular/Wi-Fi inter working, network, IEEE, vol. 19, Issue 5 (2005)
4. Wang, X., Mellor, J., Al-Begain, K.: Towards providing QoS for integrated cellular and Wi-Fi networks. http://www.cms.livjm.ac.uk/pgnet2003/submissions/Paper-23.pdf (2003)
5. Abaye, A., Lo, W., Sun, M.: U.S. Patent No. 7,260,060. U.S. Patent and Trademark Office, Washington, DC (2007)
6. Seth, A.: International journal of advanced research in IT and engineering. Vertical Handoff Decis. Algorithms Next Gener. Wireless Netw. **2**(8), 64–79 (2013)
7. Richa, S.K.: QOS parameters based vertical handoff decision in heterogeneous network: A practical approach. Int. J. Adv. Res. Comput. Commun. Eng. **3**(7), 4 (2014)
8. Ghaderi, M., Boutaba, R.: Call admission control in mobile cellular networks: a comprehensive survey. Wireless Commun. Mob. Comput. **6**(1), 69–93 (2006)
9. Hou, J., Fang, Y.: Mobility-based call admission control schemes for wireless mobile networks. Wireless Commun. Mob. Comput. **1**(3), 269–282 (2001)
10. Cheung, M., Mark, J.W.: Effect of mobility on QoS provisioning in wireless communication networks. In: IEEE Wireless Communications and Networking Conference, 1999. WCNC, pp. 306–310 (1999)
11. Cheung, M.H., Southwell, R., Huang, J.: Congestion-aware network selection and data offloading. In: 48th Annual Conference on Information Sciences and Systems (CISS), pp. 1–6. IEEE (2014)
12. Tetcos.: Global system for mobile communication (GSM). http://www.tetcos.com/GSM_White_Paper.pdf (2015)

Obtaining Porous Si Characteristic from SEM Images via Non-destructible Method; Image Segmentation

Siti Noraini Sulaiman, Muhammad Sailuddin Darus, Alhan Farhanah Abd Rahim and Fadzil Ahmad

Abstract This work aims to introduce an alternative method of obtaining porous density characteristics of porous silicon material by making use of images obtained from the Scanning Electron Microscope. The available and most commonly used method of obtaining the porous density characteristics of semiconductor materials is the gravimetric or quasi-gravimetric method. Using the gravimetric approach requires the sample material to go through various measurements during the multiple stages of processing and it would ultimately result in the destruction of the sample material. The gravimetric approach is flawed as the results it produces are questionable as it is less accurate. Also, it is refutable due to its destructive nature which is caused by the use of alkaline solution which dissolves the sample material at the final stage of the gravimetric process. Therefore, this research introduces an alternative image processing technique which would require only images of the sample material as an input which is obtained via the Scanning Electron Microscope. The image obtained is processed by segmenting the SEM images into two significant black and white regions which allow for the number of pores present on the image to be numerated. From the data obtained from the image, the porosity and porous density of the sample material can be calculated. While being a much simpler process than the commonly used gravimetric method, it is also non-destructive to the sample and is believed to produce a more precise and accurate result.

1 Introduction

Porous silicon (PS) is a popular and highly versatile material which finds applications in a wide variety of fields such as biomedical devices for drug delivery [1, 2] and photonic crystals [3]. One of the main reasons for which porous silicon is widely used in various applications is that many of the properties of porous silicon are dependent on its porosity [4]. Thus, the performance of a PS-based device is

S.N. Sulaiman (✉) · M.S. Darus · A.F. Abd Rahim · F. Ahmad
Faculty of Electrical Engineering, Universiti Teknologi MARA, 13500 Penang, Malaysia
e-mail: sitinoraini@ppinang.uitm.edu.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_35

dependent on the material's porosity. By quantifying this property-porosity relationship, the quality of a PS-based device can be determined and improved even before the production stages of the device. Porosity has emerged as an effective tool for controlling electronic and optical properties of the silicon quantum structure [5].

PS is derived from the chemical element silicon which microstructures are saturated with nanoporous holes. Typically, PS is produced using two techniques; through stain etching and anodization. Through the stain etching technique, PS is generated by etching crystalline silicon in aqueous ethanolic hydrofluoric acid (HF) electrolytes. Using anodization, possible anodization cells employ platinum cathode and silicon wafer anode immersed in the HF electrolyte. The surface of the silicon wafer anode that is immersed in the electrolyte is the process that forms PS. The corrosion of the anode is produced by running electrical current through the cell [6].

Conventionally, acquiring a PS film's porosity involves the gravimetric or quasi-gravimetric method [7] which is a tedious and time-consuming process. The gravimetric approach's accuracy is questionable and ultimately results in the destruction of the PS sample being tested. To remedy this situation, the image processing method is developed which could provide a faster and more accurate results despite it being non-destructive.

This technique requires digital images which depict the nanostructure of the sample PS. Nanostructure is a term used to define mostly semiconductor device structures which dimensions are only as small as a few nanometers to a few micrometers. Since normal cameras could not capture an image in the scale as small as a PS' nanostructure, a scanning electron microscope (SEM) is used.

This work estimates the porous density of the PS by the calculations of the porosity and the determination of the pore size distribution which allows the quality and the porous percentage of the material. The calculations of the porosity can be made by using the data extracted from the SEM images, particularly, the area of the SEM image and the sum of the total area of the pores present in the image.

Using image segmentation, the digital images are then processed to allow the pores and background of the PS sample to be distinguished. From there, the necessary parameters of the PS sample can be obtained to allow for the characteristics of the PS to be determined.

The remaining section of the paper is organized as such: Sect. 2 describes the techniques used to obtain PS specifically the Adaptive Fuzzy K-Mean (AFKM) segmentation and Black and White (BW) conversion techniques. Section 3 shows the methodology. Section 4 presents the results and discussion while Sect.5 concludes the experiment.

2 Image Processing Technique

This work attempts to introduce alternatives in determining the porous density of the porous silicon material by using the image processing technique on the image of the sample porous silicon material. The microscopically-enhanced image of the

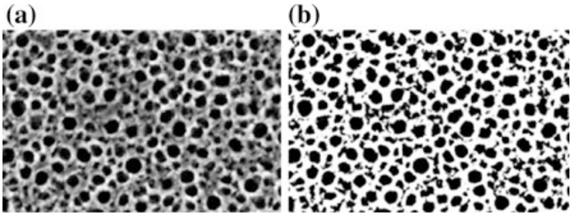


Fig. 1 SEM images of a PS sample converted to binary. **a** Original image before being converted to binary. **b** Resulting image after being converted to binary

porous silicon sample material would be obtained from a Scanning Electron Microscope (SEM) source. The processing of the image, however, is done using a computer application which is written and coded by using specific software on a Windows platform.

The SEM images will be processed by using either one of two techniques which are the Adaptive Fuzzy K-Mean (AFKM) [8] technique implemented in C-language or the BW conversion technique using MATLAB software. These two different techniques are used to convert the SEM images of the PS sample into binary images to allow the information embedded in the image to be processed. Figure 1 shows the SEM images of a PS sample converted to binary.

The Adaptive Fuzzy K-Mean (AFKM) [8] clustering algorithm converts the SEM image into binary. The AFKM clustering algorithm is the combination of two conventional methods of calculating distance from each data i.e. the K-Means (KM) and the Moving K-Means (MKM) clustering algorithm with the Fuzzy concept i.e. the Fuzzy C-Means (FCM). The KM method is to determine which set of data belongs to a cluster as opposed to another cluster. The weakness of this method is that it ignores small clusters and only considers their local convergence. The FCM method has allowed all data to belong to two or more clusters at different degrees of membership but its disadvantage is that the significant boundary is unclear. Another way introduced to overcome the clustering problem is MKM clustering which flaw lies in that it is sensitive to noise [9]. The latest version to solve the clustering problem is AFKM which implements the fuzzy concept and always updates the distance between the membership and centers.

On the other hand, BW conversion technique is implemented using the MATLAB software which has many applicable image processing tools and functions that prove to be very useful in implementing this work. The BW conversion refers to the MATLAB function ‘im2bw’ which is able to convert an image into the binary form. It converts each pixel in an image using a luminance threshold that is user-specified. Any pixels below the threshold will be set to ‘0’ or black while the pixels above the threshold will be set to ‘1’ or white.

The disadvantage with the BW method is that it requires different luminance threshold values depending on the image's brightness. Each SEM images uses different luminance thresholds depending on the brightness to yield the best possible result which is most similar to the original SEM image.

3 Methodology

The main goal of this work is to obtain the porous density characteristic of the porous silicon material from SEM images using the image processing technique. To that end, the image segmentation technique is employed on the input SEM images in order to separate the solid regions from the pores present in the sample material. Figure 2 shows the process to obtain the PS sample characteristic from the SEM image.

Six original PS images obtained from Nano-optoelectronics Research and Technology Laboratory (NOR Lab) USM are chosen to analyze the characteristics of porosity. The original images are shown in Fig. 3a–f namely SEM_01, SEM_02, SEM_03, SEM_04, SEM_05 and SEM_06, respectively.

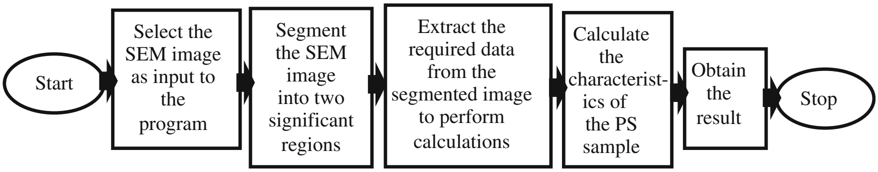


Fig. 2 Flowchart of the process to obtain the PS sample characteristics from the SEM image

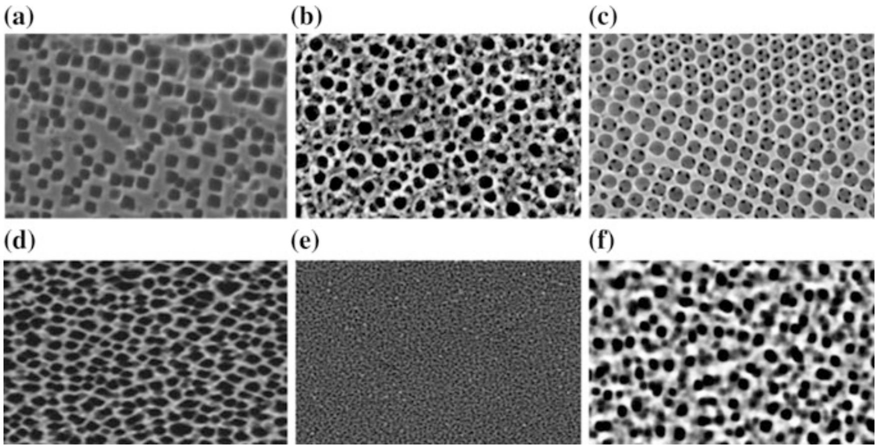


Fig. 3 The original SEM images used in this work. **a**SEM_01. **b**SEM_02. **c** SEM_03. **d** SEM_04. **e** SEM_05. **f** SEM_06

4 Result and Discussion

The details on how the image processing method manipulates raw data from SEM images to obtain the characteristics of the PS sample are presented in Figs. 4, 5 and Tables 1, 2 and 3. Without involving the physical PS sample and only by using the image data, this method could prove to be more efficient in obtaining the PS characteristics. The solid region and the pores in the image of the PS sample could be separated and isolated from each other, allowing for the pores to be enumerated using image manipulation.

The PS sample of various qualities and physical properties as shown in Fig. 3 is used to test and validate the performance of this method. Each of the SEM images is put through the program to obtain the PS parameters required to calculate the porous density of the sample. Figure 4 shows the results.

From the comparison between the AFKM and the BW technique, it can be seen that both of the methods used produce an almost similar binary image with subtle differences depending on the sample image used. While serving the same purpose, both of the methods employ different techniques to convert the SEM image data into the binary image which yield results with minor, but real differences.

A higher percentage of porosity is beneficial for porous silicon as the porous characteristics can positively affect the sensitivity of the material. A better percentage of porosity is preferable as it would allow the PS to be highly compatible for sensor and other suitable applications which would require a higher sensitivity.

While the porosity parameter can be obtained using this method, the pore size distribution would also need to be studied to show how the pattern of the pore sizes distribution is affected by the production condition of the PS material. Figure 5 below demonstrates how the pore size distribution is obtained using the BW technique with image SEM_02 serving as an input.

The distribution of the pore sizes that is divided over specific ranges can be seen from the graphs and the results are tabulated in Table 1. While it seems erratic and irregular in pattern, the distribution of pores is largely contributed by the smaller ones while the larger pores are few and far in between. The distribution pattern seems to favor smaller pores more than larger ones.

Other SEM images of the PS samples are also tested to see the pattern of the pore size distribution over specific ranges of pore diameters. The results are tabulated in Tables 2 and 3.

As the results obtained have shown, the different physical properties of the PS sample have vastly different nanostructures as evidenced by the SEM images. The porous density and the pore size distributions are obtained using the AFKM method and the BW method. Both of the methods used, when compared, would yield results with minor differences.

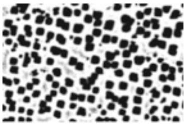
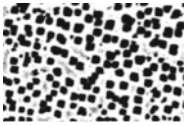
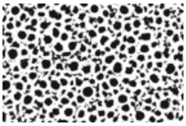
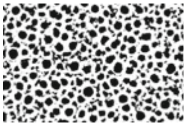
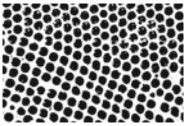
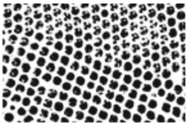
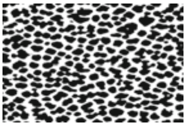
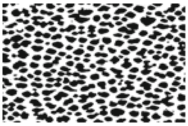
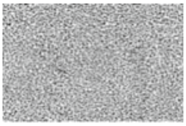
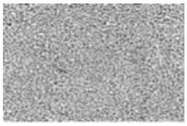
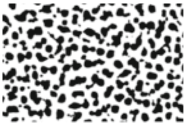
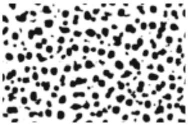
	AFKM		BW	
	Processed Image	Porous Density	Processed Image	Porous Density
SEM_01		43.21%		44.06%
SEM_02		41.36%		41.36%
SEM_03		55.87%		46.57%
SEM_04		54.54%		47.12%
SEM_05		39.43%		42.38
SEM_06		39.22%		29.42%

Fig. 4 Comparison between the porosity calculated from the binary converted image using the AFKM and BW

The different distributions of pore sizes across the different PS samples are due to the effects of the multiple criteria that are involved during the making of the porous silicon which could be the currents used in the etching, the density or the combinations of the chemicals used to produce it. Since the process of obtaining the porosity and porous density of the PS material is made easier using this alternative method, proper adjustments can be made during the manufacturing process to improve the quality of the porous silicon for future applications.

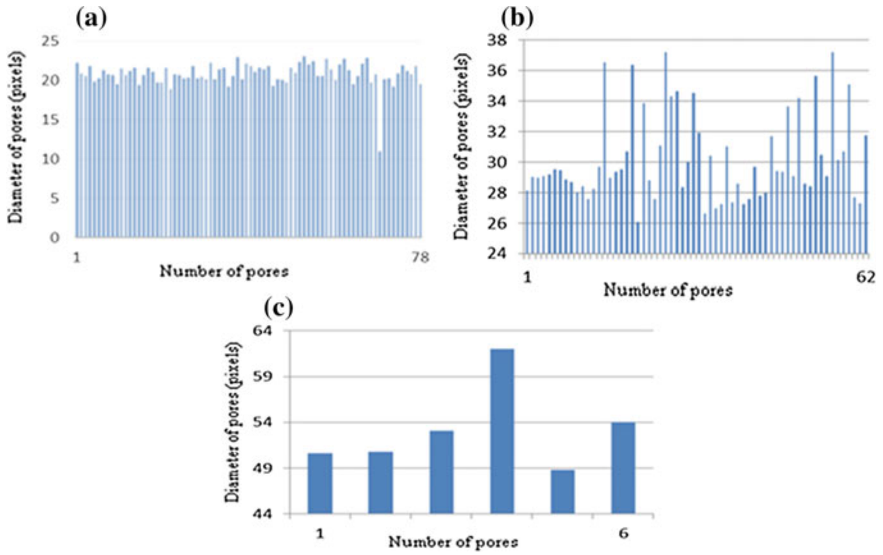


Fig. 5 The distribution of pores in the PS sample image SEM_02. **a** Pores with diameter range of 0–24 pixels. **b** Pores with diameter range of 24–44 pixels. **c** Pores with diameter range of 44–62 pixels

Table 1 Pore size distribution analysis of SEM_02

Pore diameter (pixel)	Distribution percentage (%)
0–24	53.42
24–44	42.47
44–62	4.11

Table 2 Pore size distribution of the SEM_01

Pore diameter (pixel)	AFKM		BW	
	No. of pores	Porosity (%)	No. of pores	Porosity (%)
0–24	41	64.06	47	65.28
24–44	14	21.88	13	18.06
44–62	9	14.06	12	16.67

Table 3 Pore size distribution of the SEM_06

Pore diameter (pixel)	AFKM		BW	
	No. of pores	Porosity (%)	No. of pores	Porosity (%)
0–24	139	74.33	159	72.60
24–44	40	21.39	51	23.29
44–62	8	4.28	9	4.11

5 Conclusion

From the results achieved, the alternative method proposed to obtain the porous density characteristics of the PS sample has been proven to be a unique tool for SEM image processing to analyze and extract information quantitatively. Both the AFKM and BW techniques are useful tools in image processing. Porous material characteristics like the radius of each of the pores, the pore size distribution, and the porosity of the material can be efficiently obtained. The process to obtain the data is also made simpler and faster using this method and is non-destructive to the sample material while it can also provide a more reliable result. By only using an SEM image, all of these data can be efficiently obtained using the image processing technique and image segmentation. In future developments however, several adjustments could be made to the detection and the image segmentation and data acquisition to improve the feasibility of the technique's implementation.

Acknowledgements This work was partially supported by the Universiti Teknologi MARA, Penang Campus. Malaysia and Ministry of Education Malaysia, Grant No: 600-RMI/FRGS 5/3 (81/2014) Entitled "A Novel Random-valued Impulse Noise Removal Based on Adaptive Switching Filter and Local-preserving Scheme".

References

1. Anglin, E.J., Cheng, L., Freeman, W.R., Sailor, M.J.: Porous silicon in drug delivery devices and materials. Adv. Drug Deliv. Rev. **60**, 1266–1277 (2008)
2. Vaccari, L., Canton, D., Zaffaroni, N., Villa, R., Tormen, M., diFabrizio., E.: Porous silicon as drug carrier for controlled delivery of doxorubicin anticancer agent. Microelectron Eng. **83**(4), 1598–1601 (2006)
3. Abd Rahim, A.F., Hashim, M.R., Ali, N.K.: High sensitivity of palladium on porous silicon msm photodetector. Phys B **406**, 1034–1037 (2011)
4. Ahmed, N., Ramizy, M.A., Hassan, Z., Amer, A., Omar, K., Al-Douri, Y., Alattas, O.S.: Nano and micro porous gan characterization using image processing method. Optik-Int. J. Light Electron Opt. **123**(12), 1074–1078 (2012)
5. Peckham, J., Andrews, G.T.: Simple optical method to determine the porosity of porous silicon films. Thin Solid Films **520**(7), 2526–2531 (2012)
6. Halimaoui, A.: Porous silicon formation by anodization. In: Canham, L.T. (ed.) Properties of Porous Silicon. Institution of Engineering and Technology, London (1997)
7. Lehmann, V.: Electrochemistry of silicon: Instrumentation, science, materials and applications instrumentation. Science, Materials, and Applications. Wiley-VCH, Weinheim. (2002)
8. Sulaiman, S.N., Mat-Isa, N.A.: Adaptive fuzzy k-means clustering algorithm for image segmentation. IEEE Trans. Consum. Electron. **56**(4), 2661–2668 (2010)
9. Mat-Isa, N.A., Salamah, S.A., Ngah, U.K.: Adaptive fuzzy moving k-means clustering algorithm for image segmentation. IEEE Trans. Consum. Electron. **55**(4), 2145–2153 (2009)

A Review of Client-Side Toolbars as a User-Oriented Anti-Phishing Solution

Melad Mohamed Al-Daeef, Nurlida Basir and Madihah Mohd Saudi

Abstract Phishing is a cybercrime in which, Internet users are delivered, commonly through emails, to simulated websites where they could be lured to disclose their personal information for attackers' benefit. Phishing attacks were increased about 60 % in the second half of 2013 over what was seen in first half of the same year. Although the number of proposed anti-phishing solutions, phishers still able to bypasses anti-phishing systems, in many cases, through users' inattention behaviour. Therefore, phishing becomes a layered problem that require addressing issues at both of technical and non-technical (human) layers. Numerous of client-side toolbars were proposed as a technical solution to combat phishing attacks at user's layer. Anti-phishing toolbars however, still unable to completely protect users from phishing attacks. This paper has reviewed theses toolbars to provide a clear understanding about their performance and limitation points. Such a review is required to draw clear directions of future solutions to alliviate these limitations.

Keywords Phishing · Anti-phishing methods · Warning principls · Client-side toolbars

1 Introduction

Internet users nowdays can carry out variety of online activities. They can send/receive emails, sell/buy goods, carry out banking transactions, and others. These services unfortunately are susceptible to phishing attacks that commonly start

M.M. Al-Daeef (✉) · N. Basir · M.M. Saudi
Fakulti Sains dan Teknologi, Universiti Sains Islam Malaysia (USIM),
71800 Nilai, Negeri Sembilan, Malaysia
e-mail: meladmohalda@gmail.com

N. Basir
e-mail: nurlida@susim.edu.my

M.M. Saudi
e-mail: madihah@susim.edu.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_36

by sending simulated emails to targeted victims in order to direct them to phishing websites. At simulated phisher's website, the phisher may success in taking the control over the victims' personal information [1, 2]. Phishing is a social engineering attack that usually bypasses security systems through inattention behaviour of Internet users. Such nature of attack makes phishing a dual-layer problem that require effective solutions to address issues not only at technical level, but to address issues at non-technical (human) level as well.

The first protection level against phishing is to detect the attack, after that, further actions of protection can be taken [3]. Proposed anti-phishing solutions are generally classified in two categories, *technical,* and *non-technical*. Technical solutions include, network level protection, authentication, server side classifiers, prevent against duplication, and client-side toolbars. Non-technical solutions on the other side include, user awareness, legislation and law enforcement approaches [4, 5]. Although there are a lot of numbers of implemented technical solutions, no solution yet has proved as an optimal. It is widely argued that, to be optimized, technical solutions need to be complemented by suitable non-technical ones, especially those solutions that meant to deal with human vulnerabilities [4, 6–8].

Since Internet users are one of phishing problem layers, it is therefore an important need to find a way in that; both of technical and non-technical user-oriented solutions complement each other. Due to their user-relation particularity, client-side toolbars as a technical solution are more likely adaptable to be complemented by a user training method as a non-technical user-oriented solution. This paper reviews the client-side toolbars and highlights the points of where and why they have failed to perfectly protect Internet users from being victimized by phishing attacks.

The rest of this paper is structured as follows; Sect. 2 discusses the types of phishing attacks to show that, most of phishing attacks are a user-oriented and benefit from users' inattention and unawareness factors. Section 3 discusses the types of phishing websites that client-side toolbars were designed to resist. Anti-phishing methods that commonly used in building client-side toolbars are discussed in Sect. 4. In section 5, the most important factors behind the success of phishing attacks are highlighted. Section 6 shows how client-side toolbars are categorized based on implemented anti-phishing method(s). Alarming and warning principles that used by client-side toolbars are discussed in Sect. 7. Examples of anti-phishing client-side toolbars are given in Sect. 8, the limitations of these toolbars were highlighted based on their functionality and utilized warning principles. The conclusion of the paper is presented in Sect. 9.

2 Types of Phishing Attacks

Many approaches have been used by phishers to victimize Internet users. Their approaches are generally include: *deceptive phishing, spear phishing, exploit-based phishing,* and *search engine phishing*.

Deceptive phishing is related to the social engineering scheme in which, phishers utilize forged emails that claim sent from legitimate institutions or banks [9]. Such emails usually contain fake URLs (links) to take unaware victims to phishers' websites when they click these fake URLs. In *spear phishing* attacks, email-based phishing is also used to send forged emails to a specific groups of Internet users, such as employees at a particular organization [1]. *Exploit-based* phishing involves more technically sophisticated trick that benefit from security vulnerabilities in Internet browsers and defense systems to install malicious code at users' machines when they click on fake URLs in emails [9]. In *Search Engine Phishing* approach, phishers do not have to contact the victims, victims instead will search for phishing websites that designed to be highly ranked by search engines [10]. Phishing attacks can be also performed through other channels of communication such as, VOIP, SMS, IM, Wi-phishing and even multiplayer games [11]. Any inattention by users, may make them deceived by phishers, especially if phishers urge them to update their login information to avoid their accounts being suspended or terminated [2].

3 Types of Phishing Websites

Phishing websites are always claiming to be owned by legitimate providers of online services. The main aim of such websites is to deceive both of Internet users and search engines; they include spam, concocted, and spoof sites. *Spam sites* are designed to deceive search engines by increasing their rank scores. *Concocted sites* are designed to appear as legitimate commercial sites with the objective of failure-to-ship fraud. Such sites disappear after collecting customers' money without providing promoted goods or services. *Spoof sites* are a looks-like to real commercial sites. Spoof sites are designed to lure victims in disclosing their sensitive information such as passwords and credit card numbers for phishers' benefit. Commonly known spoofed sites include eBay, PayPal, and various banking sites [12].

4 Anti-phishing Methods

To perform their attacks, phishers have to bypass anti-phishing systems that installed either at client or server sides, or at both sides in some cases. Since this study focuses on client-side toolbars, this section discusses the advantages and disadvantages of anti-phishing methods that commonly used to build these toolbars. Client-side toolbars are usually built using either one or a combination of anti-phishing methods that include, blacklists, whitelists, and heuristics. As a consequence, client-side toolbars have inherited the advantages and disadvantages of employed anti-phishing methods. The functionality of anti-phishing methods is mainly based on typical phishing patterns of phishing URLs and pages' content.

Phishing patterns found in URLs include the presence of specific symbols in URLs, IP address in URLs, many dots in URLs, and using of long URLs. Whereas the common phishing patterns found in phishing pages are based on using some tricks to hide the real content of phishing pages from users, this can done by using JavaScript technique and HTML text entry forms [5].

- **Blacklists** are an updated DNS databases of previously known phishing URLs, IP addresses, and keywords. They are usually maintained by online communities such as APWG [1]. The content of blacklists is usually generated by using three common mechanisms that either used in an individual or a combination manner with varying degrees of success and accuracy [8], these mechanisms are:

  – Suspicious URLs are automatically classified based on some phishing patterns.
  – Systems' administrators can manually classify suspicious URLs.
  – Some of anti-phishing systems enable their clients (users) to report and rate suspicious URLs to be recorded in the blacklist.

  Although the high detection accuracy of blacklists with low false positive rate (the number of legitimate instances that wrongly classified as phishing) in many scenarios [13, 14], this method however cannot identify zero-hour phishing instances. That is due to the required time for the list to be updated, and hence recording the fresh suspicious URLs. This limitation gives the phisher a great chance of success before its phishing site being detected [8, 12]. Blacklists are also limited to their requirement of human intervention and verification to be updated [14].
- **Whitelists** are designed to hold known safe URLs that have previously been visited by the user to allow the access only to trusted sites. Whitelists are efficient in detecting zero-hour phishing attacks. If the list content has no entry error, a high true positive results are produced (the number of correctly detected phishing instances). Whitelists however are less commonly implemented than blacklists due to the possibility of containing error data which may entered by users. When users are repeatedly prompted to add URLs of visited sites to the list, over time they will give up and may change to the automatic pattern of update, or may disable this function [15]. Whitelist content is usually short and precise since it based on user's wish of visit (i.e. sites going to visit). Collecting such predefined data however is a difficult, if not impossible process. In addition, it is impossible for whitelists to hold "all" trusted URLs across the entire cyber world [16, 17]. Whitelists can be used however to complement blacklist and heuristic methods as a first level of verification. Thus, saved URLs have not to be unnecessarily re-verified and may be mistakenly misclassified [18].
- **Heuristics, or Rule-Based** method is commonly used to classify a given page by examining one or more of its characteristics or features. This method may include visual similarity, using of search engines, and analysing of certain anomalies found in URLs and pages' source code [18]. Heuristics can be used

either as a solo detection method, or can be combined with blacklist and whitelist methods. As opposed to blacklist-based tools which need to wait for list updating time, heuristic-based tools can immediately catch phishing URLs. Phishers however, still able to bypass this method of detection [14]. Text-based heuristic solutions for example can be easily tricked by making image-based instead of text-based phishing pages or URLs. These solutions can also be tricked by matching text color with background color of the page [15]. Heuristic-based tools suffer from the high false positive rates (the rate of legitimate instances that incorrectly classified as phishing). In addition, heuristic-based tools require to be manually adjusted to identify future phishing tricks [14, 19].

5 Factors Behind the Success of Phishing Attacks

Although there are a lot of numbers of proposed anti-phishing solutions, phishers however, still able to deceive a big proportional number of Internet users and bypass implemented defense techniques. To draw clear direction for future anti-phishing solutions, designers have to clearly identify the factors behind the success of phishing attacks. Most such factors are highlighted in this section.

- Several anti-phishing methods such as blacklists, whitelists, and heuristics are widely used to build most of anti-phishing systems. Each of these methods however, has its limitations that definitely affect the performance and thus, results' quality of the systems by which these methods were employed. Because of that, no solution yet has proved as an optimal, particularly against fresh (zero-hour) attacks [9]. Hybrid solutions that combine two or more of anti-phishing methods can improve the results. Hybrid systems however, suffer the increased complexity and high consumed time and resources [9].
- Some solutions are based on analysing email's header or content to extract features that used in classifying emails as either legitimate or phishing. Researchers in [20] have argued that, classification features in many cases are arbitrarily selected by authors without evaluating their potential efficiency. As a consequence, systems based on unevaluated features will eventually produce inaccurate results. Authors in [21] have proposed an attempt for such evaluation process.
- Anti-phishing systems fail in most scenarios since phishers still able to bypass these systems through human unawareness and inattention behavior [3, 8]. Several techniques were designed for example to warn users when they visit phishing sites. Users in common cases however, do not pay attention to such warnings [5, 22].
- In email-based phishing scenarios, phishing emails are usually composed in a way of playing with victims' psychological factors to instill a sense of urgency and fear in their minds, and hence, push them to click URLs in phishing emails

to update login information and avoid their accounts being suspended or terminated [2].

- Naïve users in many cases do not know the nature of phishing websites or even their existence, and they do not care about websites' legitimacy [23]. Many studies have stated that, users' awareness about phishing is an important defense line which need to be enhanced [4, 6–8]. Thus, variety of users' training methods have been proposed for that purpose.

6 Types of Anti-phishing Client-Side Toolbars

Client-side toolbars are widely used as the last line of defense against phishing attacks, and they generally fall into two categories, *lookup systems* and *classifier systems*.

- **Lookup Systems** in most cases are *a client-server* architecture in which, the server maintains a blacklist of known fake sites, whereas the client-side toolbar checks the authenticity of visited site against the content of the blacklist, and hence warns the user if the checked site poses any threat. To enhance the performance quality, some of lookup systems employ whitelists method coupled with blacklists [12].
- **Classifier Systems** are generally known as heuristics or rule-based techniques that used to analyze emails' or pages' content, and domain registration information based on some of phishing characteristics [22, 24]. They are used to overcome human-caused mistakes and ignorance of phishing situations [3].
- **Hybrid systems** combine both of lookup and classifier mechanisms to alleviate the limitations of each mechanism. Such systems are generally used to analyze the content and domain registration information as a complement mechanism to black and white lists. In such cases, classifier's heuristics are only applied to inspect URLs that not in the black and/or white lists []. eBay toolbar [25] is an example of hybrid systems.

7 Warning Principles in Client-Side Toolbars

Besides the browser indicators that used to warn users about phishing pages, client-side toolbars are also implement their own warning techniques. They usually take some actions and present warning messages at users' interface when they visit suspicious sites. Such warning notifications are used to influence users' behavior and reactions against phishing attacks [12, 26]. Warning messages that displayed by security toolbars are generally classified as passive or active warnings. *Passive warnings* warn users about the potential risk without interrupting browsing task.

Active warnings on the other side are designed to interrupt users' browsing task to warn them about phishing risk [11].

Passive warnings include, colored icons or tabs that used to indicate the degree of the risk. SpoofGuard toolbar [27] for example, uses a traffic light that turns between green, yellow and red colors as the user navigates webpages, and display a warning message if the user tries to provide sensitive information to a spoof site. Researchers in many studies have argued that, most of users do not pay enough attention to warning messages and colored icons that displayed by anti-phishing toolbars. Thus, such warnings can be easily missed when they are frequently displayed to users. Users' behavior of ignorance will raise especially when they frequently receive false alarms [5, 22].

Warning pop-ups are an example of commonly used active warnings that implemented by several security toolbars such as SpoofGuard [27] and eBay Toolbar [25]. Warning pop-ups usually appear at browser centre and block browsing progress until the user acknowledges them. Pop-ups are likely more effective than passive warnings, however, they become less effective when they repeatedly appear. Users over time may tend to either disable or indiscriminately confirm these pop-ups [22].

To be more effective, warning messages should show users a useful security information at the right time [22]. To help users in making wise decisions and correct reactions towards phishing attacks, warning messages should be enough understood by users rather than just asking them for simple responses of "Ok" or "Cancel" [28].

8 Examples of Client-Side Toolbars

This section discusses some examples of client-side toolbars based on what anti-phishing method they implement. A summary of these tools is provided in Table 1 that shows the limitations of these toolbars from warning principles and operational points of view.

- Several client-side toolbars were built using the blacklist method. GoldPhish [15] for example is a content-based tool that uses the optical character recognition OCR technique to read the webpage text. This text then is submitted to Google search engine to retrieve its ranking result compared to top ranked domains. Mozilla Firefox's FirePhish [29] is another example of blacklist-based tools. It checks visited URLs against a blacklist that downloaded and maintained by Firefox browser.
- Using whitelists is another anti-phishing method which is rarely used as a solo method by anti-phishing toolbars. This method is usually used to complement the blacklist and heuristics methods. In such cases, the tool falls into the category of hybrid systems that use a combination of anti-phishing methods. Examples of such toolbars are given later as hybrid systems.

Table 1 Summary of anti-phishing client-side toolbars

Toolbar name	Warning principle limitations	Functionality limitations
GoldPhish	• Relies on a passive warning principle to warn users about phishing sites	• A web page may do not include enough text, logos, or images to verify its domain name • GoldPhish is limited to read only English text • Attacks on Google's PageRank algorithm may advance a phishing site in Google's search results, thus, phishing site is provided to GoldPhish as a valid option
Mozilla Firefox's FirePhish	• Uses a passive warning principle and display a warning pop-up to warn users about phishing sites • Users are promoted to either leave the site or ignore the pop-up. Unaware users may simply choose to ignore the pop-up • Adds a lock icon in the address bar to differentiate between HTTP and HTTPS connections, most users will not notice that	• Stores only URLs of spoof sites, thus, ineffective against concocted sites • Users may turn off or set a long interval to update the content of downloaded blacklist
SpoofGuard	• Only uses a traffic light and display a warning message to warn users about detected phishing sites	• Users are required to adjust some of its threshold values. It is difficult task for naïve users
Microsoft IE phishing filter	• Uses passive warning messages to warn users about detected phishing sites. Users however, are given a choice to continue with the site or leave it, unaware users may simply choose to continue	• Stores only URLs of spoof sites, thus, ineffective against concocted sites • Works only on Microsoft Internet explorer
eBay Toolbar	• Relies on colored tab to warn users about phishing sites	• Only applicable for eBay and PayPal websites, phishing attacks not limited to only eBay and PayPal
PhishNet	• Not mentioned how users are warned about phishing sites. Thus it's assumed that, no new warning principle was employed	• Generates many variations of each blacklisted URL • Many of generated URLs may be either innocent or not exist • Generated URLs require increased bandwidth demands to be checked • Some threshold values need to be adjusted in advance

- Heuristics based toolbars are widely used to detect phishing attacks. SpoofGuard toolbar [27] for example, employs a series of heuristics by performing several tests on the domain and URL, the page content, and the images in visited page to identify it as either phishing or legitimate.

- Hybrid systems or toolbars employ combinations of anti-phishing methods. Microsoft IE phishing filter [30] for example, utilizes a client-side whitelist and a server-side blacklist, and also applies some heuristics if checked URL not found on the black or white lists. Another example of hybrid systems is the eBay Toolbar [25] which uses a combination of a content similarity heuristics and blacklist method to check sites that known to be operated by eBay or PayPal. PhishNet [31] is another example of hybrid systems. It provides a prediction method to reduce the limitations of blacklist method. PhishNet combines the heuristics and blacklist methods in a different way to generate many variations of each URL found in the blacklist to detect the variations of URLs that may be produced by phishers.

9 Conclusion

Client-side toolbars are developed using either one or more of anti-phishing methods that generally include, blacklists, whitelists, and heuristics. Although they help Internet users in many scenarios, client-side toolbars still unable to provide a complete protection against phishing attacks, particularly against fresh (zero-hour) ones. Many of these toolbars have their own functional limitations. In addition, they have inherited the limitations of implemented anti-phishing methods. Blacklist-based toolbars for example still unable to detect fresh attacks due to required time and effort for the list to be updated. Some of toolbars that employ whitelist method suffer some of functional difficulties. They require the users to manually add new trusted URLs to their own whitelists to avoid being blocked from visiting new (not yet added) websites. As a consequence of such annoying process, users will eventually tend either to disable or skip this function. Many of heuristic-based toolbars require the users to adjust some of threshold values of the toolbar, SpoofGuard is an example of such toolbars. This tuning process is not a simple task, especially for naïve users. Other types of toolbars are dedicated to protect only a specific group of users. eBay toolbar for example, is designed to check only websites that operated by eBay or PayPal. Hybrid systems that employ more than one of anti-phishing methods are suffering increased complexity and highly consumed time and resources.

In addition, most of anti-phishing toolbars failed in steering users away from phishing websites due to employed warning approach. Many of toolbars fire alarms only after users have accessed phishing pages. Users in common cases will simply ignore such alarms, thus, they become closer to the danger.

This paper has reviewed a number of client-side toolbars based on their theme of utilized anti-phishing methods and warning principles. This review is important to put a light spot on the limitations that need more research efforts. Next phase of research in this area should focus on finding a way to enhance users' awareness about phishing phenomenon by implementing methods that tighten security concepts with something that users are continually practice. Users' training is a promising approach which can be used to complement the functionality of client-side toolbars since the training and client-side toolbars are both a user-oriented approaches.

References

1. Anti-phishing Working Group. http://www.antiphishing.org/
2. Yu, W.D., Nargundkar, S., Tiruthani, N.A.: Phishing vulnerability analysis of web based systems. In: IEEE Symposium on Computers and Communications. ISCC 2008, pp. 326–331 (2008)
3. Khonji, M., Iraqi, Y., Jones, A.: Phishing detection: A Literature Survey. Commun. Surv. Tutorials IEEE **15**(4), 2091–2121 (2013)
4. Mohebzada, J., El Zarka, A., BHojani, A.H., Darwish, A.: Phishing in a university community: Two large scale phishing experiments. In: IEEE International Conference on Innovations in Information Technology (IIT), IEEE 2012, pp. 249–254 (2012)
5. Ramanathan, V., Wechsler, H.: Phishgillnet—phishing detection methodology Using probabilistic latent semantic analysis, adaboost, and co-training. EURASIP J. Inf. Secur. **2012**(1), 1–22 (2012)
6. Aloul, F.A.: The need for effective information security awareness. J. Adv. Inf. Technol. **3**(3), 176–183 (2012)
7. Kirlappos, I., Sasse, M.A.: Security education against phishing: A modest proposal for a major rethink. IEEE Secur. Priv. Mag. **10**(2), 24–32 (2012)
8. Wilson, C., Argles, D.: The fight against phishing: Technology, the end user and legislation. In: IEEE International Conference on Information Society (i-Society), 2011, pp. 501–504 (2011)
9. Almomani, A., Gupta, B., Atawneh, S., Meulenberg, A., Almomani, E.: A survey of phishing email filtering techniques. Commun. Surv. Tutorials IEEE **15**(4), 2070–2090 (2013)
10. Devmane, M., Rana, N.: Security issues of online social networks. In: Advances in Computing, Communication, and Control, pp. 740–746. Springer, New York (2013)
11. Hong, J.: The state of phishing attacks. Commun. ACM **55**(1), 74–81 (2012)
12. Abbasi, A., Chen, H.: A comparison of tools for detecting fake websites. IEEE Comput. **42** (10), 78–86 (2009)
13. Zeydan, H.Z., Selamat, M.S.: Current state of anti-phishing approaches and revealing competencies. J. Theor. Appl. Inf. Technol. **70**(3) (2014)
14. Zhang, Y., Hong, J.I., Cranor, L.F.: Cantina: A content-based approach to detecting phishing web sites. In: Proceedings of the 16th International Conference on World Wide Web, 2007. ACM, pp. 639–648 (2007)
15. Dunlop, M., Groat, S., Shelly, D.: Goldphish: Using images for content-based phishing analysis. In: IEEE Fifth International Conference on Internet Monitoring and Protection (ICIMP), pp. 123–128 (2010)
16. Cao, Y., Han, W., Le, Y.: Anti-phishing based on automated individual white-list. In: Proceedings of the 4th ACM Workshop on Digital Identity Management, 2008. ACM, pp. 51–60 (2008)
17. Reddy, V.P., Radha, V., Jindal, M.: Client side protection from phishing attack. Int. J. Adv. Eng. Sci. Technol. (IJAEST) **3**(1), 39–45 (2011)
18. Chaudhary, S.: Recognition of phishing attacks utilizing anomalies in phishing websites (2012)

19. Suriya, R., Saravanan, K., Thangavelu, A.: An integrated approach to detect phishing mail attacks: A case study. In: Proceedings of the 2nd International Conference on Security of Information and Networks, 2009. ACM, pp. 193–199 (2009)
20. Toolan, F., Carthy, J.: Feature selection for spam and phishing detection. In: eCrime Researchers Summit (eCrime), IEEE 2010, pp. 1–12 (2010)
21. Al-Daeef, M.M., Basir, N., Saudi, M.M.: A method to measure the efficiency of phishing emails detection features. In: International Conference on Information Science and Applications (ICISA), 2014. IEEE, pp. 1–5 (2014)
22. Wu, M., Miller, R.C., Garfinkel, S.L.: Do security toolbars actually prevent phishing attacks? In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2006. ACM, pp. 601–610 (2006)
23. Dhamija, R., Tygar, J.D., Hearst, M.: Why phishing works. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2006. ACM, pp. 581–590 (2006)
24. Zhang, Y., Egelman, S., Cranor, L., Hong, J.: Phinding phish: Evaluating anti-phishing tools. In 2006. ISOC (2006)
25. Ebay Toolbar and Account Guard. http://pages.ebay.com/Help/Confidence/account-guard.html
26. Madhavan, P.: Handbook of Warnings. In: Michael, S.W. (ed.) 2006, 841 Pages, $260.00, Lawrence Erlbaum Associates, Mahwah, NJ. ISBN 0–8058–4724–3. Ergonomics in design: The quarterly of human factors applications **15**(3):32–33 (2007)
27. Chou, N., Ledesma, R., Teraguchi, Y., Mitchell, J.C.: Client-side defense against web-based identity theft. In: NDSS, 2004 (2004)
28. Downs, J.S., Holbrook, M., Cranor, L.F.: Behavioral response to phishing risk. In: Proceedings of the Anti-phishing Working Groups 2nd Annual eCrime Researchers Summit, 2007. ACM, pp. 37–44 (2007)
29. FirePhish. https://addons.mozilla.org/en-Us/firefox/addon/firephish-anti-phishing-extens/
30. Ie Phishing Filter. http://support.microsoft.com/Kb/930168
31. Prakash, P., Kumar, M., Kompella, R.R., Gupta, M.: Phishnet: Predictive blacklisting to detect phishing attacks. In: Proceedings IEEE INFOCOM, 2010. IEEE, pp. 1–5 (2010)

Quantitative Analysis of Hand Movement in Badminton

Alvin Jacob, Wan Nurshazwani Wan Zakaria and Mohd Razali Bin Md Tomari

Abstract This paper focuses on the development of a mobile measurement device to evaluate the bending and flexion of a Badminton player's hand. Badminton can be classified as a fast racket based sports. A human eyes alone is not capable of capturing all these movements. The usage of high processing optometric system is expensive and large in size. Thus, this paper proposes development of wearable measurement device to analyse badminton player's movements. This project involves the designing of a wearable sensor system that measure the players grasping action. A number of flex sensors is attached to a glove to acquire readings from the movement of the players hand and fingers. Experiments were conducted to determine the feasibility of each sensor towards the design of the new measurement device. The experiment results exhibit the capabilities of the bend sensors to provide accurate information regarding hand and finger movement activity.

Keywords Flex sensor · Monitoring device · Badminton · Flexion · Extension

1 Introduction

Badminton is a fast and dynamic sport that can be classified as one of the fastest racket sports. It can be characterized by the different type of movement that consist of hard smashes, short drop and long clears where all this movements force the

A. Jacob (✉) · W.N.W. Zakaria · M.R.B.M. Tomari
Advanced Mechatronics Research Group (ADMIRE), Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia
e-mail: alvinjacob90@gmail.com

W.N.W. Zakaria
e-mail: shazwani@uthm.edu.my

M.R.B.M. Tomari
e-mail: mdrazali@uthm.edu.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_37

player to act and react in an extremely rapid manner. About 20 % of the attacks performed during a game are smashes or jump smashes [1]. After 100 years of vigorous development, badminton has become a very popular sports played throughout the world [2] and also can be considered the most played game in Malaysia.

To win a badminton game, the proper use of tactics is an important game changer. Players need to be aware of the amount of energy they are exerting during a play to be able to perform well during a game. Scientific studies on tactics, strategy, or playing patterns of international level badminton are however, very limited [1]. The purpose of this study is to profile individual badminton player's way in grasping the badminton racket and to establish a comprehensive database of the player's badminton tactics.

The measurement and evaluation process is necessary for players to be motivated to do better and also very helpful for the coaches to analysis the performance of their players [3]. The data collected can also be used by the players and coaches to predict and prevent injuries caused by improper playing technique and fatigue. Since badminton is considered as a fast sports, it is certain that human eyes alone is not capable to interpret the fast movement of badminton players.

Generally high dynamics movements are normally analysed with using high speed optometric systems such as high speed video. However, due to some technical limitations (e.g. high amount of light) that exists when using this method. These measurements are often performed in a laboratory setting, which does not comply well with the real competition or real training conditions [4]. In addition, the method of using optometric systems can be relatively expensive and requires high speed equipment to operate.

Miniature sensors allow data collection with a high sample rate and a wide measuring range when connected to a microcontroller, this is because the light weight and small size of the sensors will not restrict the performance of the badminton player [5]. Thus, the best method that available to successfully capture the badminton player's data while still offering flexibility and keeping the cost low will be by using wearable sensors. Wearable sensors are being used in many fields to measure human movement and dynamics, where it can also be used to monitor the body's physiological response and also the kinematic aspects of performance. To monitor this in a natural way, there is a need for integrated sensors to be straightforward to use, comfortable to wear and wearable [6].

2 Kinematic Model of the Hand

Kinematics is used to describe motion without consideration given to its mass or the forces acting on it, where a humans motion is projected in three dimensional position and all points of the body move along parallel path and have the same velocity and acceleration at any given instance [7]. Athletic movement is based on flowing dynamic movement rather than static postures, a good athletic posture is actually a series of postures associated together to produce an efficient movement [6]. Sports clothing that are usually close fitting is an ideal method that has been used to capture the kinematics of an athlete. By obtaining this feedback information the athlete reinforces body awareness which may help to improve playing technique.

The human hand skeleton consists of three basic parts, namely phalanges, metacarpus and wrist. In total the human hand has 27 bones together as shown in Fig. 1. By this structure, human hand are capable of performing three basic gripping tasks, namely grasping, holding and squeezing [8]. The skeleton hand is shown in Fig. 2, can be modelled or represented by 19 links and 24 DOFs corresponding to the bones and articulations respectively of a human hand as illustrated in Fig. 2 [9].

Each finger such as the index finger, middle finger, ring finger and little finger except the thumb, can be modelled as a kinematic chain composed of four links and five DOFs. Three out of the four joints have almost parallel axes, involved in flexion and extension movements. The thumb is defined by three links and four DOFs and is much more complex because its movements also strongly involve the metacarpal bone.

By analysing and studying the parameter used by A.M. Mohd Ali [10] in their system, it can be determined which link is to be evaluated for the system being proposed. The smart glove developed by A.M. Mohd Ali consists of 5 flex sensors measuring the whole length of the finger, since the little finger is less utilised when grasping the badminton racket this system focuses on the other three finger and the thumb for the measurement. Further analysis included some video footage of badminton trainer, Nghia Tran teaching his student from the University of

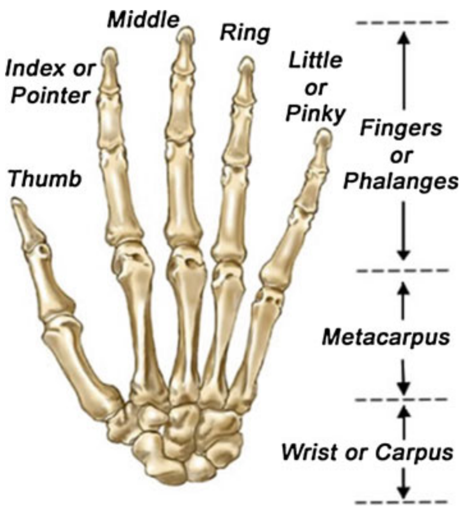


Fig. 1 Human hand skeleton [8]

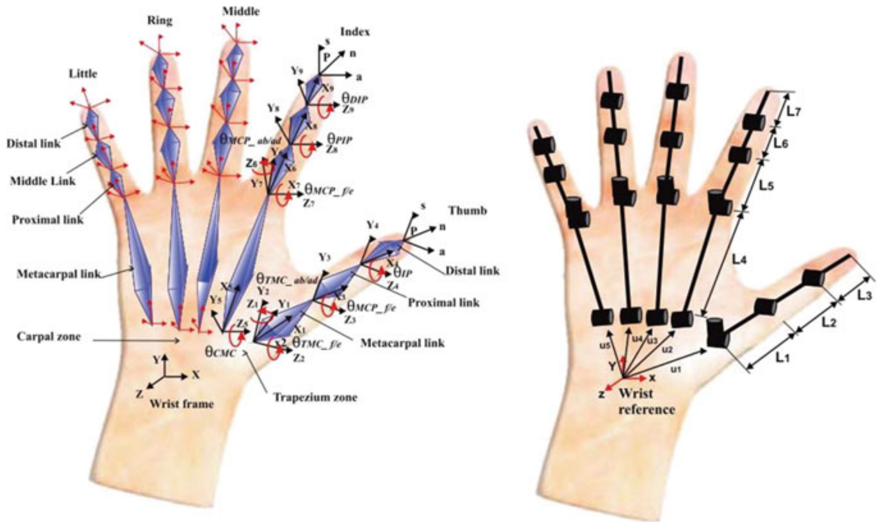


Fig. 2 Kinematic configuration of the human hand [9]

California on how is a proper racket grip [11]. This videos suggested that the wrist is an important joint for badminton players and this joint will be mainly analysed throughout this system since a badminton player uses this joint the most.

3 Hand Movement and Motion Measurement Method

On the basis of previous studies, this paper conducts analysis on the method to read and analysis the grasping action of a badminton player holding and playing with the badminton racket. Previous studies done on this subject are however, very limited as not much focus is given to the hand joint only. Thus methods of measuring the hand and finger movement has been adopted from other fields like Biomedical engineering, which range from rehabilitation to the development of prosthetic hand.

A more general idea can be seen where Akhil Mohan designed a sensorized glove for monitoring hand rehabilitation [12]. In which, a sensorized glove designed using optical linear encoder (OLE), flex sensor and a 3-axis accelerometer to measure wrist flexion, finger flexion and forearm movement as well. Furthermore, a sensorized ball also was constructed using a rubber ball to exercise the hand which is used to measure the grasp force of stroke patients, by embedding pressure sensors into the ball. Flex sensor was used to measure the bending of the finger joint (finger flexion and extension); the sensors property of varying resistance when bent or flexed makes it suitable to be used in this glove. Op-amp circuit was used to minimize the error occurred due to the source impedance of flex sensor. This developed system is able to detect movement of little finger, middle finger, index finger and thumb.

A.M. Mohd Ali proposes an artificial hand gripper controller which is controlled by a smart glove utilizing the flex sensor [10]. The aim of this research is to assist handicap individual in providing them with an enhanced version of prosthetics hand that can be used in rehabilitation process, which is economical and affordable. They focused mainly in the control of multi finger grippers with priority on the finger tips and joint. Controlling a multi finger gripper can be viewed in terms of controlling an object's pose and the forces between the object and its environment. For this they developed a glove using two types of sensors which are flex and flexi-force sensors. The flex sensor is attached to the back of the smart glove in order to detect the finger flexion while the flexi force sensor is to measure the force exerted on the finger tip. From the experiments done, it is said that when the flex sensor is bent inward resistance value increase significantly as the angle of flex sensor is bend further. However, when it is bent outward, the resistance value decreased gradually. These preliminary finding suggest that flex sensor is clearly suitable to be used in the system that will be developed to detect finger bending angle of a badminton player by utilizing inward bend of the flex sensor.

3.1 Hardware Configuration

The main purpose of this project is to design and develop a functional and user friendly device to measure the bending and flexion of the hand and finger of a badminton player.

Figure 3 shows the overall hardware setup diagram and the operation of this system is shown in Fig. 4. First, the Flex Sensor by Spectra Symbol [13] is used to measure the bending and flexion of the players hand and fingers. From the experiment done in [14] and [15], it is seen that when the flex sensor is bend inward, resistance value increased significantly as the angle of flex sensor is bend further. However, when it is bent outward, the resistance value decreased gradually.

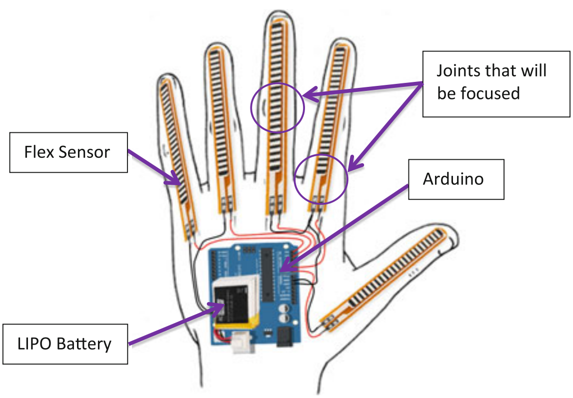


Fig. 3 Illustration of the overall system

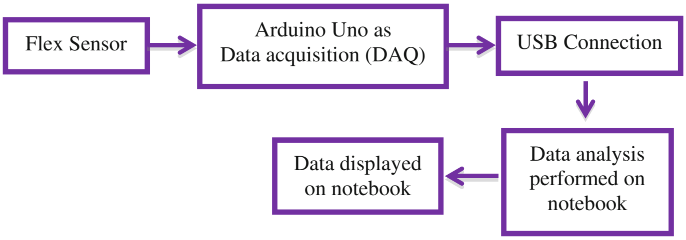


Fig. 4 Flow diagram of the proposed system

These results from previous works suggest that flex sensor is suitable to detect the finger bending angle. The output of this flex sensor is analog, to read the sensor; its variable resistance needs to be converted to variable voltage as this process creates electronic noise, a simple RC filter is used the smooth out the output.

Secondly is the microcontroller, Arduino Uno which performs the recording and measuring activity of the flex sensor. Since this system needs to be user friendly and easy to use, keeping in mind of this reason the Arduino Uno, which is an inexpensive and less hassle microcontroller, is chosen rather than the usual PIC. To use PIC microcontroller, one have to decide types of board, circuitry, language, compiler for the language, hardware programmer and etc. Arduino provides a complete, flexible, easy-to-use hardware and software platform that is widely used by artists, designers and even hobbyists [14]. It contains everything needed to support and simplify the usage of this microcontroller; simply connect it to a computer with a USB cable or power it with an AC to-DC adapter or battery to get started [16].

Lastly is the notebook which displays the recorded data to be analysed. The processed data from the microcontroller will be transmitted to a notebook via a USB cable to display the bending angle of the player's fingers.

3.2 Data Acquisition

Arduino software is free and open source which includes full development environment that can be easily downloaded from the internet. Arduino is programmed in C/C++ language where the IDE (Integrated Development Environment) is used to write sketches containing program code, to be uploaded into Arduino [15]. The process of getting the players data is a continuous process and loops itself to process the data. The data acquisition flow chart is shown in Fig. 5.

First step will be to set up the serial transmission speed and to set input/output for the microcontroller. Second will be to read the analog value from the pin where the Flex Sensor is connected. A delay of 20 ms is implemented to allow the microcontroller to stabilize.

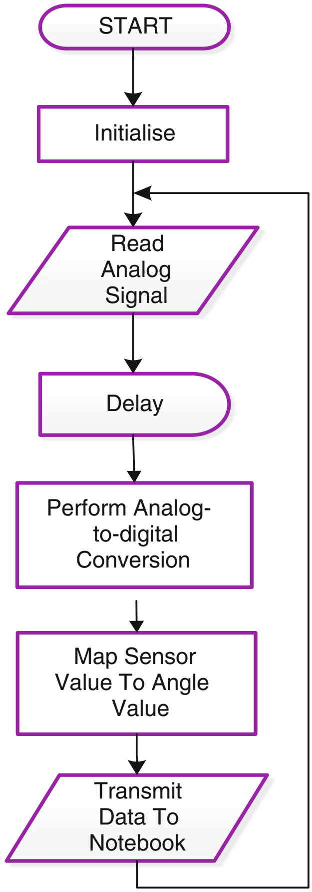


Fig. 5 Data acquisition flow chart

Third process will be where the microcontroller performs Analog-to-Digital conversion of the data received from the sensor devices. After the data is converted into digital form next will be to map the obtained analog values into the angular value. This is done using the Arduino map library. Lastly, is to transmit the data obtained from the player to a notebook where data processing will take place to interpret all this data and display it to the player and coach.

4 Flex Sensor Characteristics Test Procedure

To test the flex sensor characteristic, three different test methods were carried out. In this first experiment, the sensor characteristics are verified manually by using a multi-meter to monitor the resistance value changes as the flex sensor is bent forward. After examining the results from previous studies [10], it can be said that

Table 1 Resistance value for flex sensor

Finger	Resistance value for angle (KΩ)					
	0°	45°	90°	110°	135°	160°
Middle	10.60	13.20	18.60	–	24.40	32.78
Index	9.16	13.60	19.20	–	27.10	33.45
Thumb	9.50	12.04	16.40	19.80	–	–

the more the sensor is bent (from 0° to 160°), the higher the resistance value increases. Next, is to determine the analog voltage value for the flex sensor resistance at certain angle: 0°, 45°, 90°, 135° and 160°. These angles are chosen because they represent the finger bending states when a player is grasping the racket.

The second test is reliability test in which the flex sensor is connected to RC Filter, using a 10 KΩ resistor and a 0.1uF capacitor. Besides smoothing out any unwanted noise, the RC filter also convert the resistance to voltage value for further analysis.

The third test is the condition in which the Arduino microcontroller is used to collect the flex sensor data and subsequently convert the voltage value obtained from previous test to and angular value. The Arduino microcontroller will perform the Analog to Digital conversion and calculate the output angle for the input voltage. The results for each test will be described later in this paper.

4.1 Flex Sensor Test Results

From the experiments conducted, resistance value for the flex sensor (which range 10–50 KΩ) can be plotted as shown on Table 1, for this test a total of 5 samples were taken and the mean value is calculated. The mean of the reading taken for the thumb, index finger and the middle finger is plotted into a graph to investigate the error in the reading taken for the sensors as shown in Fig. 6. Noted that the thumb can only be measured until 110°.

As the second experiment was conducted, resistance value for the flex sensor is converted to voltage and plotted against angle in Fig. 7. For each test 5 samples were taken and the mean of the samples is plotted into a graph to investigate the error in the reading taken for the sensors.

The experiment shows that when flex sensor is bend inward, resistance value increases as the angle of flex sensor increases. However, when the sensor it is bent outward, the resistance value decreased gradually and steadily.

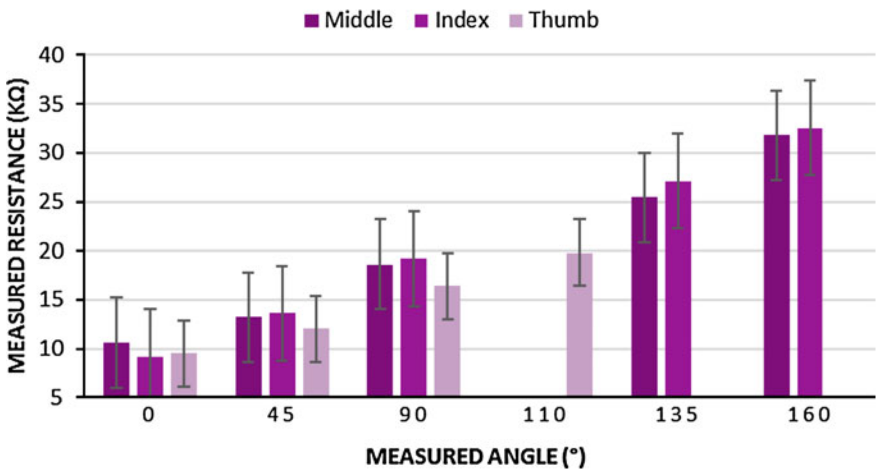


Fig. 6 Measured resistance against measured angle

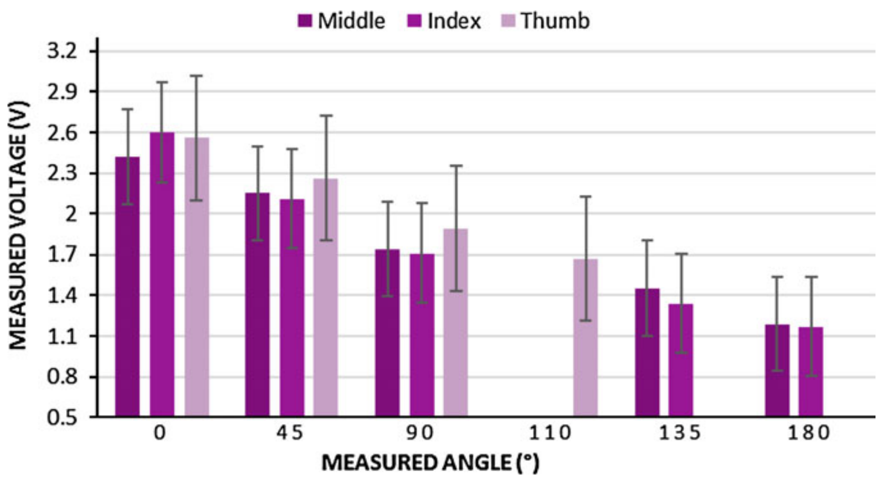


Fig. 7 The graph of output voltage against measured angle

5 Conclusion

The proposed system was tested and we found that the data refining is essential before it can be integrated to the glove worn by the player. The flex sensors were able to detect finger flexion/bending, but the output deteriorates from the maximum value measured when it is kept in the flexed position (e.g. hand fully closed) and has to be ruled out. A miniature RC circuit or the usage of op-amps for better accuracy and reliability is on the design phase and has to be tested. The evaluation in first real hardware experiment shows a reliable performance and the capability of flex sensor to follow the bending of hand finger is quite accurate.

Acknowledgment This work was supported by Ministry of Education and Universiti Tun Hussein Onn Malaysia under Research Acculturation Collaborative Effort Grant Vot 1438.

References

1. Tong, Y.M., Hong, Y.: The playing pattern of world's top single badminton players. In: 18 International Symposium on Biomechanics in Sports, pp. 1–6 (2000)
2. Yang, N.: Research of badminton forehand smash technology based on biomechanical analysis. J Chem. Pharm. Res. **5**(11), 172–177 (2013)
3. Jackson, A.W., Disch, J.G.: Measurement and Evaluation in Human Performance. Human Kinetics, Champaign (2011)
4. Jaitner, T., Gawin, W.: A mobile measure device for the analysis of highly dynamic movement techniques. Procedia Eng **2**(2), 3005–3010 (2010)
5. Jaitner, T., Gawin, W.: Analysis of Badminton Smash With a Mobile Measure Device Based on Accelerometry, pp. 282–284. XXV ISBS Symposium, Ouro Preto, Brazil (2007)
6. Coyle, S., Morris, D., Lau, K.: Textile-based wearable sensors for assisting sports performance. In: Sixth International Workshop on Wearable and Implantable Body Sensor Networks, 2009 BSN 2009, pp. 307–311 (2009)
7. Zatsiorsky, V.M.: Kinematics of Human Motion, 1st ed. Champaign (1998)
8. Virgala, I., Kelemen, M., Mrkva, Š.: Kinematic analysis of humanoid robot hand. Am. J. Mech. Eng. **1**(7), 443–446 (2013)
9. Cobos, S., Ferre, M., Ortego, J.: Efficient human hand kinematics for manipulation tasks. In: Intelligent Robots and Systems, 2008 IROS 2008 IEEE/RSJ International Conference on, pp. 2246–2251 (2008)
10. Jamil, M.M.A, Mohd Ali, A.M., Ambar, R.: Artificial hand gripper controller via smart glove for rehabilitation process. In: IEEE International Conference on Biomedical Engineering (ICoBE), no. February, pp. 300–304 (2012)
11. Tran, N.: Badminton : Proper racket grip in badminton. University of California, United States, YouTube (2010)
12. Mohan, A., Devasahayam, S.R., Tharion, G., George, J.: A sensorized glove and ball for monitoring hand rehabilitation therapy in stroke patients. In: India Educators' Conference (TIIEC), 2013 Texas Instruments, pp. 321–327 (2013)
13. Spectrasymbols.: Flex sensor: How it works (2014)
14. Mohd Ali, M.A.J.A.M., Wahi, A.J.M., Ambar, R.: Development of artificial hand gripper by using microcontroller. Int. J. Integr. Eng. **3**(3), 47–54 (2011)
15. Ambar, M.M.A.J., Ahmad, R.M.S.: Design and development of a multi-sensor monitoring device for arm rehabilitation. Int. J. Integr. Eng. **3**(3), 55–62 (2011)
16. Arduino—ArduinoBoardUno. Available: http://arduino.cc/en/Main/ArduinoBoardUno. Accessed 21 Jan 2015

A Proposed Strategy for Secure and Trusted Environment in e-Government

Tri Kuntoro Priyambodo and Yudi Prayudi

Abstract The Internet as a primary means of the implementation of e-Government is an insecure channel and can provide a loophole, which allows the emergence of various types of threat and vulnerability. This becomes an obstacle in the efforts to increase the participation of the community and gives impacts in decreasing the trust in the system. Therefore, a strategy that involves technological as well as conceptual aspects needs to be carried to realize a secure and trusted environment on e-government. This paper gives an overview of the strategy based on security and trust in one comprehensive solution that can be applied to achieve the goal through the integration of five components, namely: security and standard, security policy, trusted computing, defense-in-depth strategy and human factor. Through this paper then all parties involved in e-government may reconsider a strategy that has been set up to give attention to the issue of secure and trust in the implementation of e-government system.

1 Introduction

E-government is the form of implementation of public services based on the utilization of information and communication technologies. When the implementation of e-government is run properly, there will be some benefits obtained. However, there are also a number of obstacles and challenges in the implementation of e-government. In this case, according to [1], there are 9 factors which generally become the constraints and challenges in the implementation of e-government; one

T.K. Priyambodo (✉)
Department of Computer Science and Electronics, Gadjah Mada University, Yogyakarta, Indonesia
e-mail: mastri@ugm.ac.id

Y. Prayudi
Department of Informatics, Universitas Islam Indonesia, Yogyakarta, Indonesia
e-mail: prayudi@uii.ac.id

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_38

of which is a problem of security and privacy. The key factors in e-government is a security system. This is a consequence of the use of the Internet as the main medium in e-government in which the Internet itself is very susceptible to threats and vulnerability. The Internet is an insecure channel. The Internet as a public channel is the most effective medium that can be used to reach out all walks of life. Therefore, technically a system must have a good security standard, so that threat and vulnerability issues on e-government can be prevented.

It is in line with the opinion of Stephen Smith and Rodger Jamieson in [2] that the key factors in e-government is a security system. This is a consequence of the use of the Internet as the main medium in e-government in which the Internet itself is very susceptible to threats and vulnerability. Regarding this issue, [2, 3] specifically have discussed a wide range of potential vulnerability that is often found in the implementation of e-government.

This is in line with the new trend of the Internet architecture, known as information centric networking (ICN), which is an Internet architecture that is content-focused networking paradigm rather than host-to-host communication. According [4], ICN architecture secures the contents itself instead of securing the communication links. Although the future ICN concept would be more appropriate in the interest of securing information within the scope of e-government, but in the discussion of this paper is to use the security paradigm host to host.

In addition to security issues, the other important thing to note in the implementation of e-government is building the trust to the system, so that it will enhance public participation in utilizing all of the services provided in e-government, including the transaction of confidential data. Hassan [5] argue that in a society, trust is a basic requirement of the technology adoption process. Users tend to avoid using a particular technology in a perceived lack of concern when the demand for security in the technology is not met. The lack of participation of the citizens in employing e-government system is because of lack of trust towards the e-government system. Nevertheless, the data from Security Document [6], e-government services in the future are used for the benefit of healthcare claims, to vote or sign a digital transaction, pay taxes or services, where the digital identity will become ever more significant, for that reasons, the trust in e-government environment becomes a determinant of the success of the system. Also, [7] reveal that trust makes citizens comfortable when sharing personal information, making online government transactions, and acting on e-government advices.

According to [8], trust is not just merely received from a document that contains claims of guarantee given by e-government service providers, but must also be proven and verified by a third party. In this regard, Ideler [9] mentioned that the main problem for increasing the participation of users of information service is the way to make the degree of trust reach an acceptable level. The existence of a number of threats and vulnerabilities on e-government system certainly will lower the degree of trust from the society to the e-government system. In principle, citizens expected high-quality e-government services and full access to information with a possible security and trusted system that can be offered. Hence, a strategy to build an e-government environment which is secure and trusted is demanded, so that the function and goals of e-government can be achieved through maximum participation from citizens to take advantage of all available e-government services.

Issues about security and trust environment for e-government have not been much discussed by previous researchers. Initial researches are focused more on the security issue. According to [7], most of the existing publications on trust in e-government focus on technical perspectives such as PKI. The author himself has previously been doing research about information security strategy on e-government based mobile device [10]. The result is a solution to information security strategy to keep a balance between security and convenience, namely: selection of data and services, appropriate policy, adoption of technology and human education aspect. The review is limited only to a number of issues about security in mobile eGovernment. For this reason, to extend the previous study, a broader study is conducted regarding the issue of secure and trusted environment on e-government.

2 Vulnerability, Secure and Trusted Issue

According to [11], services provided by e-government to citizens, enterprise, a public officer, government administration and agencies via the Internet and mobile connections are vulnerable to a variety of threats. Meanwhile, [12] argues that vulnerability refers to flaws or weaknesses in system security procedures, design, implementation, and internal controls that could be exploited by threat-sources. Once exploited, it could result in a security breach, consequently causing harm to e-government information assets and services. The assets must be protected to ensure secure e-government include client computers, the messages traveling on the communication channel, and the Web and e-government servers—including any hardware attached to the servers [4].

In any system, including the system of e-government, it is known there are four main regions of threat in any given system: programs, peripherals, communications, input and output. According to Ali (2007) in [13], there are many factors that trigger the occurrence of vulnerability. Among those factors are Technical and Technology, Human, Social, Political factors of the Countries, Economic, and Networking.

According to [3], e-commerce is one example of the application of a good security system. However, the use of The Public Key Infrastructure (PKI) that is applies to e-commerce is not fully applicable within the scope of e-government without a thorough analysis of what the new trust model should be. The trust calculation for commerce is based on monetary issues, while government solutions involve important infrastructure, society, and privacy issues.

Trust is part of humanity and social interaction. There are a lot of principles and definitions of trust. In this case [10] has made a list of definitions of trust from a number of sources. While according to [7], trust is defined as an individual's belief or expectation that another party (e-government) will perform a particular action important to trustor in the absence of trustor's control over trustee's performance.

Furthermore, Santos [14], adds that the solutions to improve trust is through two aspects, namely:

- Enforcing the security properties required by the users, that is to provide protection against data users, as well as security of the computing platforms used.
- Giving users guarantees that the desired security properties are being enforced. Considering users are not directly involved in the control process of security and do not know how power computing platforms are used, the users need to be given guarantees that the infrastructure being run is completely safe. In this case, the guarantees that can be given are through trusted computing hardware and trusted certifier that is offline.

Meanwhile according to [15], a system is categorized as a trust if it meets three criteria, namely:

- Protected capabilities, the presence of a set of orders having exclusive permission to access a specific location where sensitive data are stored or a location where a particular activity can be run.
- Integrity measurement, the existence of metrics from the platform characteristics that contain things affecting the integrity of the platform.
- Integrity reporting serves as informing the specific storage location of integrity measurements as well as providing a legal authentication from the stored value based on trusted platform identities.

Trust in the government agency has a strong impact on the adoption of a technology. Colesca [10] reveals that a high level of trust in the government's ability, motivation and commitment to the e-government programs coupled with a high level of trust on enabling technologies leads to a synergy between the government and citizens. In addition, [13] mention that in an e-government system, trusting believes an e-government website will act responsibly when a citizen visits or transacts with it. The existence of threat and vulnerability especially malware would be a factor that can eliminate trust from the system.

3 A Proposed Strategy

To make information available to those who need it and who can be trusted with it, a robust defense requires a flexible strategy that allows adaptation to the changing environment, well-defined policies and procedures, the use of robust tools, and constant vigilance. Thus, it is helpful to begin a security improvement program by determining the current state of security at the site.

A number of researchers have proposed some solutions to address the problem of security and trust in e-government. Among the researchers are [16] who give a strategy solution to strengthen security through security policy, security practices, security procedure and security technology. Other solutions are delivered by [7] that trust on e-government is built by employing nine theoretical constructs that delineate the concept of citizens' trust in e-government. While [10] has conducted research on identifying the relation between trust and e-government services, as well as the main factors affecting the attitude of trust in e-government. In this case, the research findings indicated that citizen's higher perception of technological and organizational trustworthiness, the quality and usefulness of e-government services, the Internet experience and propensity to trust, directly enhanced the trust in e-government. Age and privacy concerns have a negative influence on trust. The other study about trust in e-government was done by [13] who examined the extent to which the effect of information quality, system quality and service quality contribute to building trust towards e-government system.

Unfortunately, those studies did not address the issue of security and trust in one comprehensive solution. The studies also did not include how to improve the strategy and realize a secure and trusted environment in e-government. A strategy is important because it will give an idea of how the unity in point of view and components can be arranged from the beginning or gradually so that the requirements of a secure and trusted environment can be met. The strategy will also provide an overview of the linkages between components that affect the attainment of the goal expected.

To realize a secure and trusted environment, several approaches from a variety of viewpoints need to be adopted. In the previous research, [11] have discussed specifically the security aspect with a mobile apps-based solution. Using the base of the proposed model on the research, this study proposes the inclusion of five components as a strategy to realize a secure and trusted environment on eGovernment. The five components are security standard, security model and trust management, defense strategy, trusted computing and human factor. An explanation of the five components is given in the following description.

3.1 Security and Trust Standard

Standard is the highest level of policy that shows transparency to the public that all the processes carried out in an institution have been in accordance with the provisions. Security standard features a high-level concept. According to [17], ideally every institution has a policy as the guideline to communicate their goal that contains a set of basic principles to be a reference for technical and operational levels. Policy will provide an overview of culture and value built into the institution. Although policy is solely a guideline, given the development of a more advanced technology as well as feedback from the operational experiences and practices on a daily basis, the policy must also be responsive to follow such developments.

Security and trust standard that can be applied in e-government is using the approach of Information Security Governance, that is governance of organizations/institutions that provides a guiding strategy, ensures that the goal of the company is attained, manages risk, utilizes resources of the organization responsibly, and oversees the success or failure of security programs.

3.2 Security Policy, Model and Trust System

A secure environment is strongly influenced by the application of security policy, security model as well as trust management system.

- **Security Policy**. According to [18], security policy is a statement that clearly specifies what should and what should not be in the field of security. In the lower level, security policy will contain a set of policies regarding authorization and secure states. In general, security policy is a set of statements and requirements of system behavior that will ensure the realization of a secure system. Meanwhile, [19] stated that in the coverage of law enforcement, security policy must also include policies about confidentiality of classified data. In this case, all classified data/information should be protected and only users with a certain level who have the right to access such data and information. In addition, there must be rules and obligations that bind users who utilize the classified data.
- **Security Model** is an abstraction that provides a conceptual language that will be used by the administrator to implement the security policy. Security model will define the hierarchy of access or modification of rights that can be owned by users from the institution.
- **Trust Management System** is a framework to determine whether the security policy expressed through logic and abstraction as well as implemented through programming or system setting has completely complied with the policy that should be followed. Trust management system is applied to policy language and compliance checker.

According to [20], for a simple environment, the use of several security models are adequate. For future development, [20] predicts that a security model will be increasingly complex, and the approaches that can be done as solutions are Application-Centric Access Control Models and Technology-Centric Access Control Models. Even according to [6], additionally the issue security and convenience will be more balanced, one of which is through applying the concept of a context-aware authentication.

3.3 Defense-In-Depth Strategy

Based on the idea of [4], a common security system is currently designed and developed based on Defense-In-Depth (DID) model. The system refers to the unification of management and technology used. This model differs from the layered defense that has only one layer of defense to cope with all threats. The implementation of security system is divided into three classifications i.e. prevention technologies in order to protect the system from any intruder and threat at the level of system or storage, detection technology to detect and track the condition of information system when running abnormal or other interruptions in the network or system and an integrated technology that is to integrate important functions for information security of core assets, such as prediction, detection and tracking interruptions. Applying defense-in-depth strategy technically becomes the standard solution in a secure and trusted environment.

3.4 Trusted Computing

Currently, an infrastructure security solution uses software-based approach. In fact, software based approach to handling security still brings a number of gaps for certain parties [21]. One of the problems encountered is the inability of the software to do preventive actions when apparently the attacker performs assault directly on the hardware. If this approach is done, it will cause changes in the integrity or even modification on the security application itself. If the endpoint/client is connected in an open system, such as the Internet, it is very difficult to determine the security level of endpoint/client, when only relying on the software based security mechanism. One drawback of software is, when there is an increase in security level; it will decrease the comfort and ease in using the computer. The vendors of the software have been aware of it.

Trusted computing is selected as the hardware component to support the strategy to build secure and trusted environment. Trusted computing can also provide solutions regarding the authenticity and integrity through the ability to perform verification toward an infrastructure platform. One of the modules from the trusted computing is trusted platform module (TPM), which becomes one of the keys to the application of trusted computing. According to [22], Germany's government is the one that has a high commitment to implement trusted computing in a variety of strategic infrastructures owned by the government.

3.5 Human Factor

In any environment, there is a human between internal and external system. Nikolakopoulos [23] calls it as linking the human factor. Human factor contributes to the emergence of vulnerability that causes the decline in security of an environment. Human error, bad behavior in interacting with the system, a low level of skill, knowledge and education open the possibility for human factor vulnerability. That is why Mitnick and Simon (2002) in [24] mention that "*humans are the weakest connection in information security*". Unfortunately, all attempts conducted

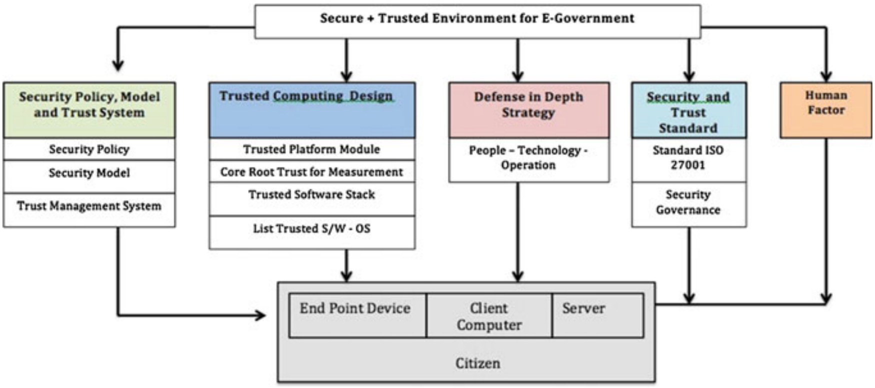


Fig. 1 This image is an illustration of how the relationship of five components as a proposed strategy for secure and trusted environment in e-government

by any institution to increase security are concentrated on the hardware and software rather than on peopleware. Therefore, there must be a mechanism in the institution to focus on the handling of peopleware as unity in the attempts to increase security and trust from the environment. In this case, [25] mention the efforts to improve security must be followed by increasing feedback from the human factor. The feedback is obtained through various methods, such as through modeling to determine the characteristics of human factor in a security system.

According to [26], a human factor will form a group and eventually become the organization's culture. Therefore, human factor in security issue is a complex and dynamic issue because it could be related to various aspects of human. However, all scholars agree that in any security system, a human factor remains as a crucial part. Even according to [25] in the field of security, all technology based solutions can be easily designed and implemented, but not for a solution to the human factor.

The five components are proposed as one unified strategy to realize a secure and trusted environment on e-government. The relationship between those five components is illustrated in Fig. 1.

4 Discussion

One of the illustrations of security solution for e-government is a secure e-government architecture proposed by [27]. Another solution is given by [12] in the form of a framework for securing e-government services that integrate IT security services into e-government maturity models. In this regard, the proposed framework addresses both the quantity of offered e-government services and the quality of security services by aligning strategic objectives between e-government services and security services.

Both alternative solutions are more likely as a technological framework that becomes the basic architecture of e-government system. Whereas, Maria Wimmer and Bianca von Bredow in [2] reveal that security aspects on e-government do not solely concern with technical aspects. In this case, e-government must be established from a non-technical viewpoint as well. That is why, this paper proposes a more comprehensive concept as a solution to provide secure and trust through the incorporation of technical and nontechnical aspects. Solution and discussion regarding trust in an e-government focus more on the user's perspective toward trust in the e-government system as done by [2, 7].

In this paper, we proposed the strategy for secure and trust environment for e-government that contains five components. An explanation of the reasons why the five components were selected have been given at previous sub section. From the explanation, it can be understood why the fifth component is the most crucial thing to be able to realize the concept of security and trust in e-government.

Among those five components, the two components, which are technology-based, are trusted computing and defense-in-depth strategy while the other three components are conceptually based, namely standard, security policy, and human factor. The five components of the strategy proposed in this paper are only recommendation. Ideally, those strategy components can be implemented to realize the concept of a secure and trusted environment in e-government. However, if all the components are not met, there must be at least one of the technology-based components and conceptually based components, which can be implemented. Components of a defense-in-depth strategy and human factor are recommended as a strategy to satisfy at least the minimum level of secure and trusted environment of e-government.

To be able to implement a strategy that is proposed in this paper, the steps that can be performed are:

- Conduct a review of the e-government strategy, whether the five components that proposed in this paper has been part of the strategy. If it hasn't, then it is recommended to refine the strategy to include components that are proposed in this paper.
- A strategy must be a unity between policy makers and implementers. The issue of secure and trust on e-government should be a shared concern among policy makers and implementers. For that, there needs to be a forum to deliver an update on the issue and the technology to improve the concern for security and trust on e-government.

5 Conclusion and Future Research

The Internet as the main medium for the implementation of e-government turns out to have a number of threats and vulnerability. It is becoming an obstacle for the implementation of e-government especially in terms of public participation. One of the solutions is to set up a secure and trusted environment on e-government system in order to increase the adoption of a technology toward the synergy between the government and citizens through e-government services.

The solution proposed in this paper is the development of the concept of secure e-government proposed earlier by [11] with the extension of trusted environment issue. The proposed strategy contains five components, namely: standard, security policy, trusted computing, defense-in-depth strategy as well as a human factor. The five components of the strategy proposed in this paper are for a recommendation. Principally, those strategy components must be implemented to realize the concept of a secure and trusted environment in e-government. However, if all of those components cannot be met, their have to be at least one of the hardware/software based components and conceptually based components that can be realized. Components of defense strategy and human factor are recommended as a strategy to achieve the minimum level of secure and trusted environment of e-government.

The description in this paper is still in the high conceptual level. To determine whether the proposed strategy has met the expectations for a secure and trust environment of e-government, there must be further researches up to the implementation level. Each strategy component proposed in this paper requires further study on the low-level aspects of implementation.

References

1. Alshehri, M., Drew, S.: E-Government fundamentals. In: International Conference ICT, Society and Human Beings (IADIS), 2010, no. 2001, pp. 35–42
2. Hadi, F., Bin Muhaya, F.T.: Essentials for the E-Govemment security. In: International Conference on Information Society (i-Society), 2011, pp. 237–240
3. Moen, V., Klingsheim, N., Inge, K., Simonsen, F., Hole, K.J.: Vulnerabilities in E-Government web portals. Int. J. Electron. Secur. Digit. Forensics **1**(1), 89–100 (2007)
4. Setiawan, A.B.: Implementasi Tata Kelola Keamanan Informasi Nasional Dalam Kerangka e-Government, Jakarta, 2011
5. Hassan, S., Aziz, Z., Nisar, K.: On the cache performance of the information centric network. In: Proceedings of 2013 International Conference on Computing, Electrical and Electronics Engineering. Research Makes a Differ. ICCEEE 2013, pp. 477–481 (2013)
6. Security Document World.: The Role of Trusted Digital Identity in Enabling the eGovernment 2020 Vision, 2014
7. Alsaghier, H., Ford, M.: Conceptualising citizen's trust in e-Government: application of Q methodology. Electron. J. e-Government **7**(4), 295–310 (2009)
8. Paladi, N.: Trusted Computing and Secure Virtualization in Cloud Computing. Lulea University of Technology, Lulea (2012)
9. Ideler, H.A.W.: Cryptography as a Service in a Cloud Computing Environment. Eindhoven University of Technology, Eindhoven (2012)
10. Colesca, S.E.: Understanding trust in e-Government. Inz. Ekon. Econ. **3**, 7–15 (2009)
11. Priyambodo, T.K., Prayudi, Y.: Information security strategy on mobile device based eGovernment. ARPN J. Eng. Appl. Sci. **10**(2), 652–660 (2015)
12. Karokola, G.R.: A framework for securing e-Government services the case of Tanzania. Stockholm University, Sweden (2012)

13. Teo, T.S.H., Srivastava, S.C., Jiang, L.: Trust and electronic government success: An empirical study. J. Manag. Inf. Syst. **25**(3), 99–131 (2009)
14. Santos, N.M.C.: Improving Trust in Cloud, Enterprise, and Mobile Computing Platforms. Universitat de Saarlandes, Saarbrücken (2013)
15. Burmester, M., Mulholland, J.: The advent of trusted computing : implications for digital forensics. In: SAC, 2006, pp. 23–27
16. Singh, S., Karaulia, D.S.: E-Governance : information security issues. In: International Conference on Computer Science and Information Technology (ICCSIT), 2011, pp. 120–124
17. Wada, K., King, P.: IT policy: an essential element of IT infrastructure. Educause Review, no. June, pp. 14–15, Jul-2001
18. Gikas, C.: Information Systems Security : A General Comparison of FISMA, HIPAA, ISO 27000, 2010
19. Taylor, C., Endicott-Popovsky, B., Frincke, D.A.: Specifying digital forensics: a forensics policy approach. Digit. Investig. **4**, 101–104 (2007)
20. Sandhu, R.: Security Models : Past, Present and Future. Institute for Cyber Security, UTSA USA, San Antonio, TX, USA, pp. 1–28 (2010)
21. Amin, M., Khan, S., Ali, T., Gul, S.: Trends and directions in trusted computing : Models, architectures and technologies. In: International Multiconference Of Engineers and Computer Scientist, 2008, vol. I, pp. 19–21
22. Mohaideen, Z., Mubarak, M.F., Ahmad, Z.: Advisory Paper for Trusted Computing Technology, 2013
23. Nikolakopoulos, T.: Evaluating the Human Factor in Information Security. University of Oslo, Oslo (2009)
24. Soltanmohammadi, S., Asadi, S., Ithnin, N.: Main human factors affecting information system security. Int. J. Contemp. Res. Bus. **5**(7), 329–354 (2013)
25. Gonzalez, J.J., Sawicka, A.: A framework for human factors in information security. In: WSEAS International Conference on Information Security, Hardware/Software Codesign, E-Commerce and Computer Networks, 2002, pp. 1871–1877
26. Parsons, K., Mccormac, A., Butavicius, M., Ferguson, L.: Human Factors and Information Security : Individual, Culture and Security Environment, Edinburgh, 2010
27. Dlamini, I.Z., Ngobeni, S.J., Mutanga, M.B.: South African EGov: Secure E-Service. In: eChalange e-2010 Conference, 2010

Microcontroller-Implemented Artificial Neural Network for Electrooculography-Based Wearable Drowsiness Detection System

Keith Marlon R. Tabal, Felicito S. Caluyo and Joseph Bryan G. Ibarra

Abstract Various methods have been explored to develop an effective drowsiness detection system to give drivers a warning of impending drowsiness. The present work has successfully developed an electrooculagraphy-based wearable drowsiness detection system in the form of a visor cap by implementing an artificial neural network (ANN) into an Arduino LilyPadUSB microcontroller. As a result, a stand-alone and wearable system that does not require a computer was achieved. The performance of the system for drowsiness detection has an overall accuracy of 90.00 %, precision of 88.00 %, sensitivity of 91.67 % and a training mean squared error (MSE) of 2.70×10^{-3}.

Keywords Artificial neural network · Electrooculography · Drowsiness detection · Microcontroller implementation

1 Introduction

For many years, drowsiness has been associated with vehicular accidents and fatal crashes. Drowsiness, which is linked with fatigue tend to reduce reaction time, vigilance and attentiveness of a person resulting in poor performance on attention-based activities [1, 2]. If the person happened to be a driver, the occurrence

K.M.R. Tabal (✉) · F.S. Caluyo · J.B.G. Ibarra
Mapúa Institute of Technology, Intramuros, Manila, Philippines
e-mail: km.tabal@gmail.com

F.S. Caluyo
e-mail: fscaluyo@mapua.edu.ph

J.B.G. Ibarra
e-mail: jbgibarra@mapua.edu.ph

K.M.R. Tabal
Camarines Sur Polytechnic Colleges, Nabua, Camarines Sur, Philippines

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_39

of drowsiness while driving may lead to a road accident resulting in damage to properties or worse, may cause the loss of lives.

Various methods using physiological signal analysis such as electroencephalography (EEG) and electrooculography (EOG) along with different classification algorithms have been used to develop effective drowsiness detection systems (DDS). For example, a real-time EEG-based DDS was constructed using a novel six-channel electrode system, a digital signal processing chip and a separate controller chip for the drowsiness detection [3]. Another study employed one-channel sensor incorporated in a headband harness on a DDS [4]. The signals acquired by the hardware were sent via Bluetooth to a computer where MATLAB was used to detect drowsiness. In a similar context, an EEG-based DDS was also developed by employing a fast ICA-based detection algorithm [5]. On the other hand, an EOG-based DDS was used as an alternative to video-based systems in detecting eye activities caused by drowsiness [6]. In a separate study using EOG, eye movements of drivers were observed using a moving-base driving simulator to assess driver drowsiness. A device called ActiCap™ was used to measure EOG signals while an adaptive detection approach was used simultaneously to detect not only eye blinks, but also other driving-relevant eye movements [7]. Another EOG-based drowsiness detection method was proposed using convolutional neural network (CNN) [8]. They used a linear dynamic system to smoothen the EOG signals recorded by the NeuroScan™ system before being fed to the CNN for drowsiness classification. Hybrid systems have also been developed to detect drowsiness. These include a portable wireless device capable of detecting drowsiness in real-time using a combination of EEG and EOG. The signals from the forehead, acquired by silver fabrics installed in a headband, are transmitted using XBee® into a stand-alone microcontroller for analysis [9]. In addition, a similar system for vehicle drivers was also studied using an algorithm based on time duration measurements between blink behaviors. However, the processing of the signal is performed using a signal processing tool kit in MATLAB [10].

The above-mentioned studies are only some of the many methods attempted by researchers to develop effective drowsiness detection systems. However, the development of wearable and stand-alone systems that does not require a computer or other external devices for detecting signs of drowsiness were not explored and thus remained a major challenge in the field of accident prevention systems. To address this problem, an EOG-based drowsiness detection system in the form of a visor cap was explored in the present study by implementing an artificial neural network (ANN) in a microcontroller to achieve a wearable and stand-alone drowsiness detection system. The ANN was trained to classify signs of drowsiness through blink duration patterns acquired from a test subject. Moreover, this study aimed to benefit workers who have long working hours such as long distance drivers and machine operators by warning them of an impending drowsiness through an alarm system.

2 Methodology

2.1 Hardware Design

The design of the hardware was based on the block diagram as shown in Fig. 1. Four (4) silver/silver-chloride (Ag/Ag-Cl) electrodes were used as bio-potential sensors and an instrumentation amplifier with built-in filters was constructed using a single quad operational amplifier [11]. A TLC1079 micropower precision operational amplifier was used because of its minimum supply voltage of 1.4 V, suitable for battery operated applications. The first stage of the amplifier composed of two (2) op-amps is configured as a high-pass filter with cut-off frequency at 0.8 Hz and a voltage gain of 51. The second stage is a combined difference amplifier and low-pass filter with a cut-off frequency at 36.17 Hz and a voltage gain of 83.33. Compared to the suggested and most commonly used range of 5000–7000 [12], the overall gain of the amplifier is only about 4250.

The actual printed circuit board (PCB) implementation of the signal conditioning circuit (Instrumentation Amplifier and Filters) is shown in Fig. 2 along with the rest of the hardware components. It has a dimension of 3.90 cm by 2.60 cm. Combinations of through-hole and surface mount devices (SMD) were used to achieve a compact design that would fit in the interior of the visor cap.

For the main controller, a LilyPadUSB Arduino based on the ATmega32u4 was used because of its 32 KB flash memory, enough to hold the artificial neural network designed for the present application. Moreover, the LilyPadUSB is commonly used for wearable technologies because it can be sown into a fabric. The board contains a MCP73831 Lithium Polymer (LiPo) battery charging chip, making it ideal for portable battery operated applications.

The alarm system is composed of a piezoelectric disk beeper and a micromotor vibrator which is commonly found in cellular phones. The piezoelectric disk beeper and vibrator are connected to pins 3 and 9 of the microcontroller, respectively.

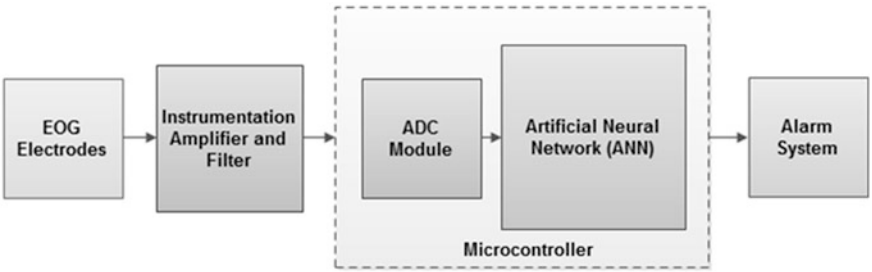


Fig. 1 Block diagram of the drowsiness detection system

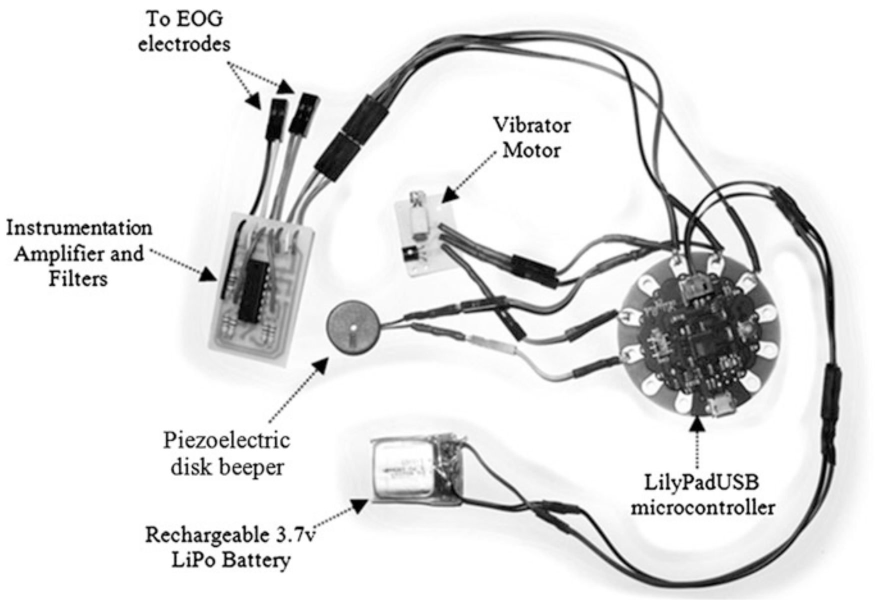


Fig. 2 Actual hardware implementation

2.2 *Electrode Placement Setup Selection*

In order to determine the best electrodes setup, each of the following configurations in Fig. 3 were tested to acquire EOG signals for comparison. The first setup is shown in Fig. 3a where the first electrode which serves as a ground is placed at the center of the forehead; a second electrode is placed above the left eye; and a third electrode is placed below the left eye. The second electrode setup as shown in Fig. 3b has an additional electrode which is placed at the upper right corner of the forehead. This electrode is connected to the reference used by the instrumentation amplifier. The last electrode setup as shown in Fig. 3c is similar to the second electrode setup except for the interchanged ground and reference electrodes.

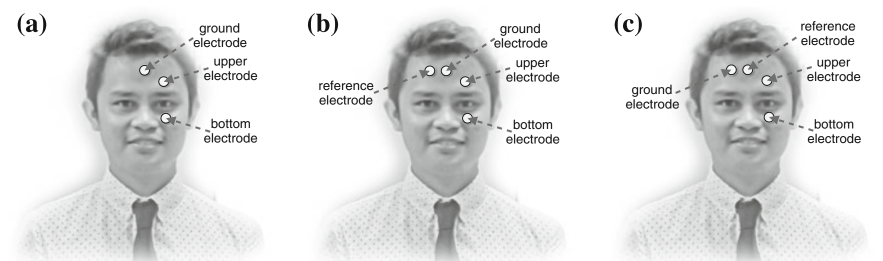


Fig. 3 Electrodes setup

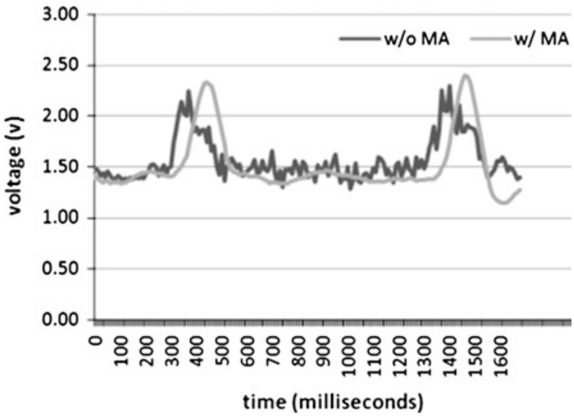


Fig. 4 Noise reduction using moving average (MA)

When the first electrode setup was tested, traces of noise appeared in the signal. The solution used in eliminating the noise was the implementation of moving averages of the signal.

Two separate signals taken without the moving average and with an implemented moving average is shown in Fig. 4. Significant reduction of noise is evident in the signal acquired when an MA was implemented. The MA was therefore implemented in signal acquisition using all three (3) electrode placement setup.

Figure 5 shows a pair of blink signals acquired from the three (3) electrode placement setup. For the first electrode setup, a peak-to-peak voltage (V_{PP}) of 1.14 V with a baseline at approximately 1.56 V is acquired. The second electrode setup acquired a V_{PP} of 1.8 V with a baseline at approximately 1.27 V. The third and last electrode setup acquired a V_{PP} of 1.07 V with a baseline at approximately 1.4 V.

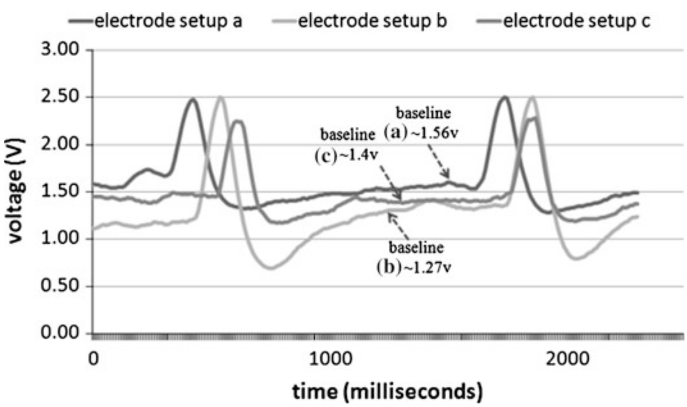


Fig. 5 Blink signals from the three (3) electrode placement setup

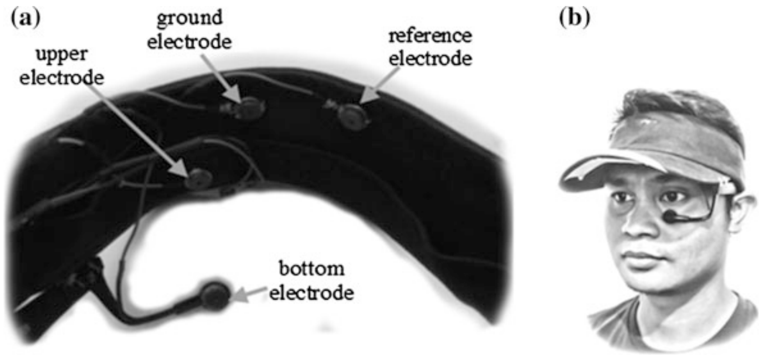


Fig. 6 Actual Electrode Setup Implementation (**a**) interior view (**b**) the visor cap

The factors from which the best electrode setup was based were the V_{PP} and baseline voltage produced by each configuration. Considering the blink signals produced by each electrode setup, the 2nd configuration provided the most pronounced characteristic because of its high V_{PP} and low baseline which were a must for the current application. These characteristics, reduces the possibilities of false blink detections due to the subjects head movements and other facial muscle artifacts. The actual implementation of the electrode setup in the visor cap is shown in Fig. 6.

2.3 EOG Data Acquisition

The EOG data acquisition was performed using a visor cap-to-computer setup. In this process, a USB cable was used to connect the cap into the computer. To identify a blink signal, Eq. 1 was employed to detect if the amplitude (x) of the EOG signal exceeds the 1.60 V threshold voltage.

$$f(x) = \begin{cases} 1, x > 1.60\,\mathrm{V} \\ 0, x \leq 1.60\,\mathrm{V} \end{cases} \tag{1}$$

To collect EOG signals, the user performs a series of normal blinks and mimics the eye blink behavior of a drowsy person by slowly closing and opening the eyes [13, 14]. Figure 7 shows two overlapped blinks during a 700 ms window. Notice that the normal blink has a very short duration pattern compared with a drowsy blink signal.

For every blink signal detected, five (5) raw EOG samples were collected with 50 ms interval and were digitized by the analog-to-digital converter (ADC) module of the microcontroller. The raw data acquired from the ADC ranges between 0 for 0 V to 1023 for 3.3 V representing a 10-bit ADC resolution. One hundred thirty (130) batches of EOG data were collected for the training datasets.

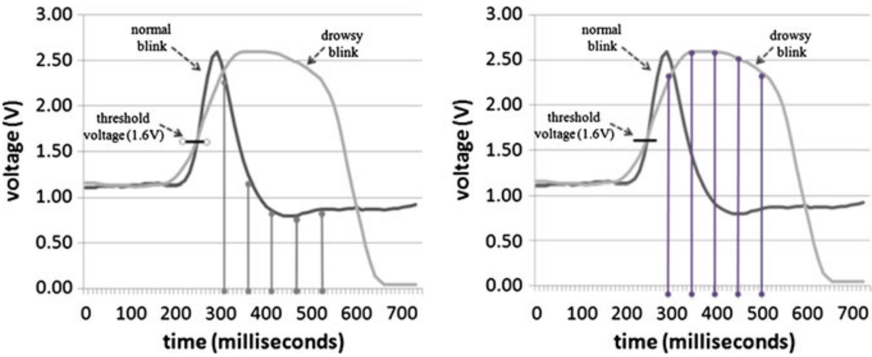


Fig. 7 ANN Architecture

2.4 ANN Network Design

A feed-forward neural network with one hidden layer was used in the system. It is composed of five (5) inputs, five (5) hidden neurons and one (1) output neuron. The five (5) inputs correspond to the 5 EOG samples. Each neuron implements sigmoid function for activation. Biases were also included in each neuron resulting to thirty six (36) weights. The design of the ANN from which the microcontroller implementation and training program was based on is shown in Fig. 8.

After the ANN was created, the network was trained using back-propagation algorithm (BPA) [15]. The neural network was trained using a supervised learning where the inputs along with its corresponding output are embedded in an array form in the program. Batch mode was also implemented in the training where the adjustments of weights were made on an epoch-by-epoch basis. Each epoch consists of the entire set of training examples.

2.5 Neural Network Training Using BPA

The network training was done by creating an ***m*** file in MATLAB. The program was a straight-forward implementation of the standard equations for back-propagation algorithm. In the m file, separate arrays for the input and output training datasets were initialized along with the random values for weights. The learning rate (α) and iterations were set at 0.7 and 250, respectively.

For each hidden neuron, the sum of the product inputs and their corresponding weights were computed using Eq. 2 were x'_i is the weighted sum, x_i is the input, w_{ik} is the weight of the input of the neuron (k) and $n = 5$ is the number of inputs in the network.

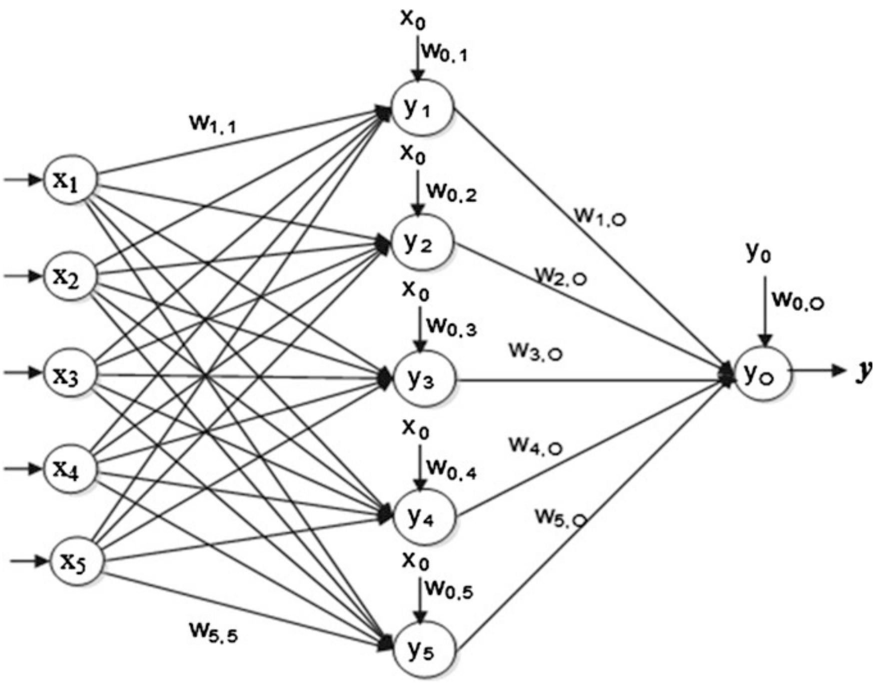


Fig. 8 ANN Architecture

$$x_i' = \sum_{i=0}^{n} x_i w_{ik} \tag{2}$$

The weighted sum has a total of six (6) inputs because of the additional biases in each neuron.

After x'_i was computed, the result was passed on to the sigmoid activation function using Eq. 3. This function provides a smooth transition between inputs because of its curvilinear behavior [16]. It also constricts the values between 0 to 1.

$$y_i = \frac{1}{1 + e^{-x_i'}} \tag{3}$$

Equations (2) and (3) were repeated for the output neuron except that the inputs used were the results of the activation function (y_i) of the hidden neurons. The next step was to compute for the error gradient on the output neuron using the Perceptron Delta Rule which was mathematically expressed in Eq. 4, where δ_O referred to the error gradient, y is the activation function of the output and y' as the desired output.

$$\delta_O = y(1 - y)(y' - y) \tag{4}$$

To determine the error gradient (δ_k) of each hidden neuron, the computed δ_O was propagated back to the hidden layer using Eq. 5, where w_{iO} were the weights of the activation functions (y_i) of each hidden neurons.

$$\delta_k = y_i(1 - y_i)\delta_O w_{iO} \tag{5}$$

Upon determining the error gradient of each neuron in the network, they were used to update the weights. This was done using Eq. 6, where w_{ik} referred to the original weight of input i on neuron k, α is the learning rate, and δ_k was the error gradient of hidden neuron k.

$$w_{ik} = w_{ik} + \alpha w_{ik}\delta_k \tag{6}$$

2.6 *Microcontroller Implementation*

The program for the drowsiness detection system was written using the Arduino Integrated Development Environment (IDE). The weights and biases produced during the training of the neural network is shown on Table 1, where the cell at which the columns (W_i) intersects with the rows (k) represents the weights (w_{ik}) shown in Fig. 7. These weights and biases were permanently assigned to their corresponding variables in the program where Eqs. 2 and 3 were employed.

The program includes the EOG data acquisition through the ADC module of the microcontroller. A moving average is performed on the acquired signals. The result of the process is polled to regularly check if it exceeds the blink threshold. When a blink signal is detected, the microcontroller starts to save the next five (5) EOG samples in an interval of 50 ms per sample. These samples would now be fed to the artificial neural network for blink signal classification. When the blink signal is classified by the neural network as a drowsy signal, the piezoelectric disk beeper and the vibrator would be activated for a duration of one (1) second.

Table 1 Weights and biases

Neuron (k)	W_0	W_1	W_2	W_3	W_4	W_5
0	1.3336	1.4461	−2.8720	−5.5188	14.0919	−2.7838
1	1.6792	−1.1447	0.0315	0.0329	−0.6078	−0.1028
2	0.1832	1.4266	1.8759	0.9566	1.0933	0.5970
3	−6.4900	−2.2150	−2.2792	−2.2295	−1.9414	−1.0666
4	27.0604	8.7401	9.6387	10.1908	8.5782	9.7013
5	−2.1513	−2.2658	−1.7050	−1.5543	−1.7032	−1.2856

The program has a total of 10.302 KB binary sketch size. At this stage, the microcontroller could now function on its own without the help of a computer. The microcontroller could process every blink signal that is fed into it.

2.7 Drowsiness Detection System Testing

After the ANN was uploaded into the microcontroller, the system was tested for detecting normal and drowsy blinks. A test subject wore the visor cap and was asked to perform a series of blinks for preliminary testing. After some adjustments on the cap, the subject performed a series of normal blinks as instructed by an observer. Among the fifty (50) normal blinks performed by the subject, the system was able to positively identify forty-six (46) with only four (4) misclassifications. After the test for normal blinks, the subject was now asked to mimic a drowsy blink by slowly closing and opening his eyes [13]. In doing so, the blink duration becomes longer which can occur due to drowsiness [14]. The subject was instructed by the observer to perform a series of drowsy blinks. After fifty (50) drowsy blinks, the system was able to positively identify forty-four (44) and only six (6) misclassifications.

3 Results and Discussion

The result of the test for the performance of the artificial neural network for blink classification was evaluated using the Confusion Matrix [17] as shown on Table 2. The true positive (TP) refers to the actual number of correctly classified drowsy blinks which were predicted to be drowsy blinks, the false positive (FP) refers to the actual number of incorrectly classified normal blinks which were predicted to be drowsy blinks, the true negative (TN) refers to the actual number of correctly classified normal blinks which were predicted to be normal blinks, and the false negative (FN) refers to the actual number of incorrectly classified drowsy blinks which were predicted to be normal blinks.

The test for the performance of the ANN for blink classification revealed that the system has an overall accuracy of 90.00 %, precision of 88.00 %, sensitivity of 91.67 % and a training mean squared error (MSE) of 2.70×10^{-3}.

Table 2 Classification performance of the ANN

Confusion matrix		Predicted blink signal	
		Drowsy	*Normal*
Actual blink signal	*Drowsy*	TP = 44	FN = 4
	Normal	FP = 6	TN = 46

4 Conclusion

In this study, an electroocculography-based wearable and stand-alone drowsiness detection system utilizing a microcontroller-implemented artificial neural network (ANN) was achieved. The feed-forward network used in the system was trained to classify drowsy blink signals from normal blink signals using back-propagation algorithm through supervised learning. The ANN was successfully implemented on a LilyPadUSB Arduino microcontroller. Moreover, the system does not require additional processing devices such as computers in detecting drowsiness.

Acknowledgment The authors would like to thank the Engineering Research and Development for Technology (ERDT) scholarship program of the Department of Science and Technology-Science Education Institute (DOST-SEI) for funding this study.

References

1. Driver Fatigue and Road Accidents.: A Literature Review and Position Paper. The Royal Society for the Prevention of Accidents (RoSPA), February (2001)
2. Drowsy Driving and Automobile Crashes.: NCSDR/NHTSA Expert Panel on Driver Fatigue and Sleepiness, Report HS 808 707, (1998)
3. Tsai, P.Y., W. Hu, T.B.J. Kuo, L.Y. Shyu.: A portable device for real time drowsiness detection using novel active dry electrode system. In: 31st Annual International Conference of the IEEE EMBS, USA, pp. 3775–3778 (2009)
4. Park, L., Xu, L., Sridhar, V., Chi, M., Cauwenberghs, G..: Wireless dry EEG for drowsiness detection. 33rd Annual International Conference of the IEEE EMBS Boston, Massachusetts USA, pp. 3298–3301 (2011)
5. Chae, M.S.: Portable wireless drowsiness-detection system. 2012 IEEE International Conference, pp. 762–765 (2012)
6. Chieh, T.C., Mustafa, M.M., Hussain, A., Hendi, S.F., Majlis, B.Y.: Development of vehicle driver drowsiness detection system using electrooculogram (EOG). In: 1st IEEE International Conference on Computers, Communications, & Signal Processing with Special Track on Biomedical Engineering, pp. 165–168 (2005)
7. Zhu, X., Zheng, W.L., Lu, B.L., Chen, X., Chen, S., Wang, C.: EOG-based drowsiness detection using convolutional neural networks. In: IEEE International Joint Conference on Neural Network (IJCNN) Beijing, China, pp. 128–134 (2014)
8. Ebrahim, P., Stolzmann, W., Yang, B.: Eye movement detection for assessing driver drowsiness by EOG. In: 2013 IEEE International Conference on Systems, Man and Cybernetics. Manchester, United Kingdom, pp. 4142–4148 (2013)
9. Arnin, J., Anopas, D., Horapong, M., Triponyuwasi, P., Yamsa-ard, T., Iampetch, S., Wongsawat, Y.: Wireless-based portable EEG-EOG monitoring for reat time drowsiness detection. 35th Annual International Conference of the IEEE EMBS Osaka, Japan, pp. 4977–4980 (2013)
10. Desai, Y.S.: Driver's alertness detection for based on eye blink duration via EOG & EEG. Int. J. Adv. Comput. Res. **2**(7), 93–99 (2012)
11. Bárcia, J.C.: Human electrooculography interface. Master's Thesis: Unibersidade Tecnica de Lisboa (2010)

12. Usakli, A.B., Gurkan, S., Aloise, F., Vecchiato, G., Babiloni, F.: On the Use of Electrooculogram for Efficient Human Computer Interfaces. Hindawi Publishing Corporation, Cairo (2010)
13. Kircher, A.: General Information Vitaport II. VTI, Swedish National Road and Transport Research Institute, Linköping (2001)
14. Muzet, A., Pbayle, T., Langrognet, J., Otmani, S.: AWAKE pilot study no.2: Testing steering grip sensor measures. Technical Report IST-2000–28062, CEPA (2003)
15. Rumelhart, D., Hinton, G., Williams, R.: Learning Internal Representations by Error Propagation, Vol. 1. MIT Press Cambridge, MA, USA, pp. 318–362 (1986)
16. Larose, D.T.: Discovering knowledge in data: an introduction to data mining. Wiley-Interscience, Online ISBN: 9780471687542 (2005)
17. Kohavi, R. Provost, F.: On Applied Research in Machine Learning. In Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process, vol. 30. Columbia University, New York (1998)

A Classification of an Audio Signal Using the Wold-Cramer Decomposition

Abdullah I. Al-Shoshan

Abstract Audio signal classification has been approached by many researchers, and the purpose of the classification process is needed to build two different libraries: speech library and music library, from a stream of sounds. In this paper, an approach for audio signal classification is proposed using the Wold-Cramer decomposition. Some simulation using the proposed approach is presented.

1 Introduction

Humans can discriminate speech from music easily in their mind without any influence of the mixed music. Due to the new techniques of analysis and synthesis of speech signals, the musical signal processing has gained particular weight, and therefore, the classical sound analysis techniques are used in processing music signals. Music art has a long and distinguished history. It goes back to the time of Greek and is developed through centuries in both the musical instruments and melodies. The problem of audio signal classification serves as the fundamental step towards the rapid growth in audio data volume [1–3]. There are many kinds of music such as: Classical, Rock, Pop, Disco, Jazz, Country, Latin, Electronic, Arabic, etc. [4–8]. Audio signals change continuously and non-deterministically with time [12]. Consequently they are usually characterized as time averages, and their relative amplitude and frequency contents can be easily specified. As an example, speech and music typically have strong low-frequency energy and progressively weaker high-frequency content [9–12]. The maximum frequency, $f_{\max}$, of an audio signal varies according to audio signal kind; $f_{\max}$ equals 22 kHz in CD quality recording, 11 kHz in FM stereo broadcasting, 6 kHz in stereo or multi-loudspeaker recording, 5 kHz in mono-loudspeaker recording, and 4 kHz in the traditional telephone transmitting quality.

A.I. Al-Shoshan (✉)
Computer Engineering Department, College of Computer, Qassim University, P.O. Box 6688, Qassim 51452, Saudi Arabia
e-mail: drshoshan@gmail.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_40

Audio signals can be classified into the following classes:

1. Speech signal compounded of single talker in specific time period.
2. Completely music signal without any speech component.
3. Mixture of single talker speech and background music.
4. Songs; mixture of music with a singer voice.
5. Singing without music.
6. Abnormal music; uses acclaim cadence, single word cadence, human whistle sound, opposite reverberation or any non-music sound that been inserted as a basic tone of the music melody. These cadences cannot be generated by any of the ordinary musical instrument except modern Organ and mainly processed by a major help of computers.
7. Speech signal compounded of two or more speakers talking simultaneously in a specific time period. A good algorithm for separating the sounds of two talkers taking simultaneously is using the cepstrum analysis.
8. Non-speech and non-music signals: like car, motor, fan sounds, etc.
9. Complex sound mixture like multi-speakers or multi-singers with multi-music sources.

Although all fields of signal processing have attracted a considerable number of researchers and have enjoyed success in recent years, and as new software techniques ride the surging waves of ever faster computers, the problem of classification of music and speech signals is still an important area of research. The overlap in speech and music signal is, in general, very strong so that there is no ordinary filter that can separate them from each other. Speech covers the spectrum from near zero to 3.5 kHz with an average dominant frequency = 1.8747 kHz. However, from the classical theorem of music, the lowest fundamental frequency (A_1) is about 27.5 Hz and the highest tone C8 is around the frequency of 4186 Hz. Therefore, musical instrument manufacture tries to bound music frequency to human's sound limits to achieve strong consonant and also strong frequency overlap. Moreover, music propagates over all the audible spectrum and cover more than audible band (20 kHz), with an average dominant frequency = 1.9271 kHz. Also, a music instrument in general has many fundamental frequencies while speech of specific person has only unique fundamental frequency, and we can assign more accurately a dominant frequency for a specific person speech while it is not the case in specific music instrument. For a music instrument every tone kind (i.e., A) has special fundamental and for a piece of music played by many instruments we will get too many fundamentals.

2 Audio Signal Classification

Audio signal classification approaches can be classified into three categories:

1. Time-domain.
2. Frequency domain.
3. Time-Frequency domain types.

In [7], a two-level music and speech classifier has been developed used long-term features such as differential parameters, variance, time-averages of spectral parameters, and zero crossing rate (ZCR). Saunders [13] has also proposed a two-level algorithm for classification based on the average ZCR and the short-time energy (STE) features, and applied a simple threshold procedure. They have designed their model based on human cochlea functional performance. Using long-term features, like cepstrum pitch or spectral centroid, consumes large delay without worth increase in overall classification precision. It was observed that the most powerful classification features in time domain are the ZCR and the STE [14], however, they fail when noise is involved. In general, the music and speech classification process found in literature can be classified into the following algorithms [7]:

I. Time-domain approaches: ZCR, STE, ZCR and STE positive derivative, Variance of the Roll-Off feature, Pulse Metric, Number of silent segments, Hidden Markov Model (HMM), Neural networks, Number of silent segments
II. Frequency-domain approaches:
   (1) Spectrum: Spectral Centroid, Mean and Variance of the Spectral Flux, Mean and Variance of the Spectral Centroid, Variance of the Spectral Flux, Roll-Off of the Spectrum, Bandwidth of signal, Amplitude, and Delta Amplitude.
   (2) Cepstrum: Cepstral Residual, Variance of the Cepstral Residual, Cepstral feature, Pitch, and Delta Pitch.
III. Time-Frequency domain approaches: Spectrogram, Wavelets, and Evolutionary Spectrum.

3 Evolutionary Spectrum

For a stationary signal, its spectral representation may be viewed as an infinite sum of sinusoids with random amplitudes and phases

$$e(n) = \int_{-\pi}^{\pi} e^{jwn} dZ(w) \tag{1}$$

where Z(ω) is a process with orthogonal increments i.e.

$$E\{dZ^{*}(w)dZ(\Omega)\} = \frac{S(w)dw}{2\pi}\delta(w - \Omega) \tag{2}$$

and $S(w)$ is the spectral density function of $e(n)$. The family of constant amplitude sinusoids is however not appropriate for characterizing non-stationary processes, like audio signal. In the Wold-Cramer decomposition, a discrete-time non-stationary process $\{x(n)\}$is considered the output of a casual linear, and time-variant (LTV) system with a zero-mean, unit-variant white noise input $e(n)$, i.e.,

$$x(n) = \sum_{m=-\infty}^{n} h(n,m)e(n-m), \tag{3}$$

where $h(n,m)$ is the impulse response of the LTV system. Substituting $e(n)$ from (1) into (3) ($S(\omega)$ = 1 for white noise) we get

$$x(n) = \int_{-\pi}^{\pi} H(n,\omega)e^{jwn}dZ(\omega) \tag{4}$$

where the generalized transfer function of the LTV system is defined as

$$H(n,w) = \sum_{m=-\infty}^{n} h(n,m)e^{jwn} \tag{5}$$

A non-stationary process can thus be expressed as an infinite sum of sinusoids with random, time-varying amplitudes and phases. Since the instantaneous variance of $x(n)$ is given by

$$E\left\{|x(n)|^2\right\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |H(n,\omega)|^2 d\omega \tag{6}$$

then, using the Wold-Cramer decomposition, the ES is defined as

$$S(n,\omega) = \frac{1}{2\pi}|H(n,\omega)|^2 \tag{7}$$

Although the time-frequency distributions, like the spectrogram, the wavelets and the ES, are good in discriminating music from speech signals, the main disadvantage of these tools is the cost of their computations; therefore, they may be used in off-line analysis.

4 Simulations

In this section, we will use the Wold-Cramer representation of a non-stationary signal. When the speech signal is examined over a sufficiently short period of time "between 5 and l00 ms", its characteristics are fairly stationary; however, over long periods of time (on the order of 1/5 s or more) the signal characteristics change to reflect the different speech sounds being spoken. Figure 1 shows the ES of a speech

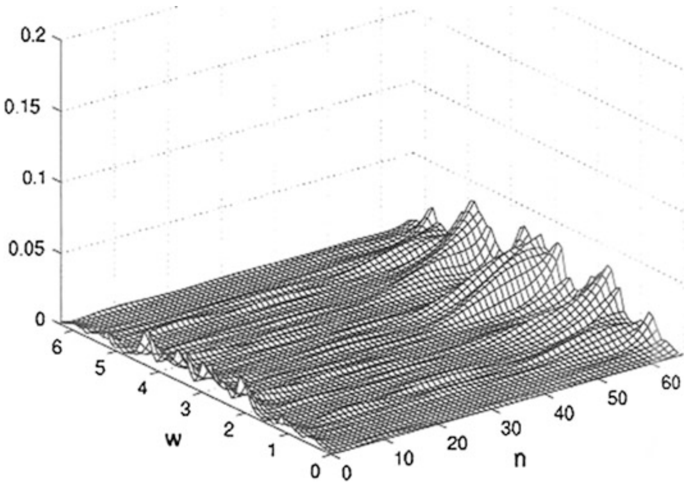


Fig. 1 ES of speech signal

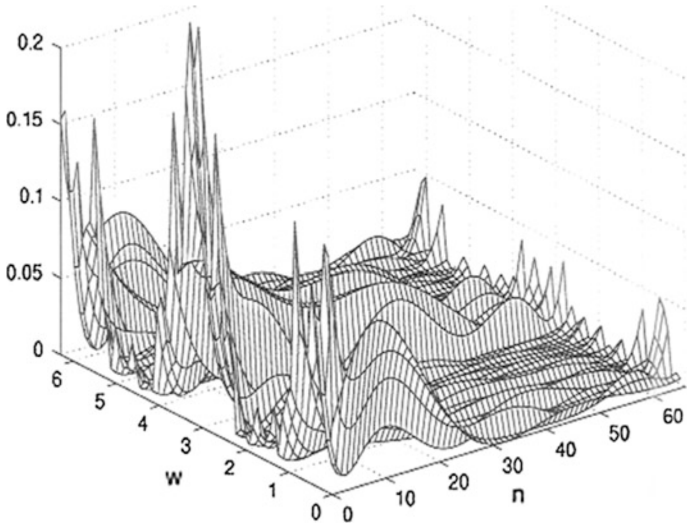


Fig. 2 ES of music signal

signal, however, Fig. 2 shows the ES of music signal, and however, Fig. 3 shows the block diagram for speech/music/mixed classification. The suppression of the amplitude for speech might due to Gaussianity, and from the figures, we observe that the energy of music signal is distributed clearly over the time-frequency domain, and is not the case of speech signal.

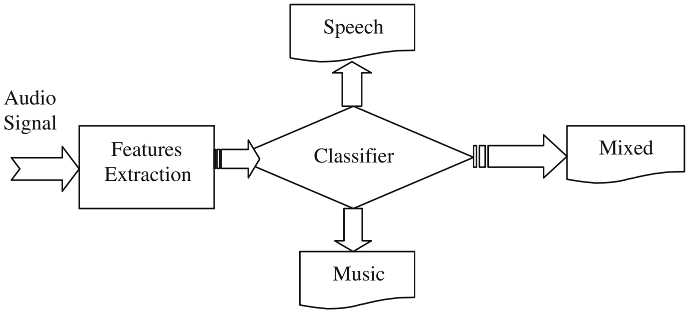


Fig. 3 Block Diagram for speech/music/mixed classification

5 Conclusions

In this paper, an approach for audio signal classification was proposed using the evolutionary spectrum. The purpose of the classification process was needed to build two different libraries: speech library and music library, from a stream of sounds. Some simulation for classification using the proposed approach was presented and has shown clearly some differences between the two signals in the time-frequency domain when using the evolutionary spectrum approach. For future research, the evolutionary bispectrum of speech/music may also give new results, especially when one of the two signals is corrupted with a symmetrical noise, like Gaussian.

References

1. Fu, Z., Lu, G., Ting, K.M., Zhang, D.: A survey of audio-based music classification and annotation. IEEE Trans. Multimedia **13**(2), 303–319 (2011)
2. Herrera-Boyer P., Amatriain, X., Batlle, E., Serra, X.: Towards instrument segmentation for music content description: a critical review of instrument classification techniques. In: Proceedings of ISMIR, 2000
3. Srivivas, M., Roy, D., Mohan, C.K.: Learning sparse dictionaries for music and speech classification, ICDSP2014, pp. 673–675 (2014)
4. Biagetti, G., Crippa, P., Turchetti, C., Morici, A.: Efficient synthesis of piano tones with damped bessel functions. In: IEEE Proceedings of 15th International Conference on Digital Signal Processing (DSP 2007), Cardiff, Galles, UK, pp. 539–542, 1–4 July 2007
5. Lu, M., Dou, W.: Butterfly-like D-tree fusion strategy for real-time speech and music classification, ICMEW2014, pp. 1–4 (2014)
6. Al-Shoshan, A.I.: Speech and music classification and separation: a review. J. King Saud Univ. Eng. Sci. **19**(1), 95–133 (2006)
7. El-Maleh, K., Klein, M., Petrucci, G., Kabal, P.: Speech/music discriminator for multimedia application. In: Proceedings of IEEE ICASSP pp. 2445–2448 (2000)

8. Ramalingam, T., Dhanalakshmi, P.: Speech/music classification using wavelet based feature extraction techniques. J. Comput. Sci. **10**, 34–44 (2014)
9. Ghoraani, B., Krishnan, S.: Time-frequency matrix feature extraction and classification of environmental audio signals. IEEE Trans. Audio Speech Lang. Process **19**(7), 2197–2209 (2011)
10. Jain, A.K., Duin, R.P.W., Jianchang, Mao: Statistical pattern recognition: a review. IEEE Trans. Pattern Anal. Mach. Intell. **22**(1), 4–37 (2000)
11. Gianfelici, F., et al.: A non-probabilistic recognizer of stochastic signals based on KLT. Sig. Process. **89**(4), 422–437 (2009)
12. Biagetti, G., Crippa, P., Turchetti, C.: Modeling of speech signals based on bessel-like orthogonal transform. In: Proceedings of Interspeech 2006—ICSLP Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, U.S.A., 2478–2481, 17–21 Sept 2006
13. Saunders, J.: Real-time discrimination of broadcast speech/music. IEEE ICASSP’96, pp. 993–996 (1996)
14. Al-Shoshan, A.I., Al-Atiyah, A., Al-Mashouq, K.: A three-level speech, music, and mixture classifier. J. King Saud Univ. Eng. Sci. **16(**2), 319–332 (2004)

Model of Drugs Penetration Through Biological Membrane

Martin Augustynek, Adela Zemanova, Jan Kubicek and Marek Penhaker

Abstract This work deals with pharmacokinetics modeling and compilation of functional models. There are three mathematical models representing actions beginning intravascular drug administration until the elimination from the body. Graphics dependences are outputs of these models, provide a quick idea about behavior of drugs in the body and time period, when is medicine completely eliminated. The main model illustrates the penetration of drug across biological membrane using the same principles, but the model tells of what happens to take place on bio membranes and its surrounding area.

Keywords Pharmacokinetic model · Compartment · Biological membrane · Intravascular application

1 Introduction

In many other studies provides information about the transmission of various substances through various types of membranes [1–3]. These penetrations are using various mathematical modeling methods and thanks to these models we can further

M. Augustynek (✉) · A. Zemanova · J. Kubicek · M. Penhaker
Department of Cybernetics and Biomedical Engineering,
Faculty of Electrical Engineering and Computer Science,
VSB—Technical University of Ostrava, Ostrava, Czech Republic
e-mail: martin.augustynek@vsb.cz

A. Zemanova
e-mail: adela.zemanova@vsb.cz

J. Kubicek
e-mail: jan.kubicek@vsb.cz

M. Penhaker
e-mail: marek.penhaker@vsb.cz

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_41

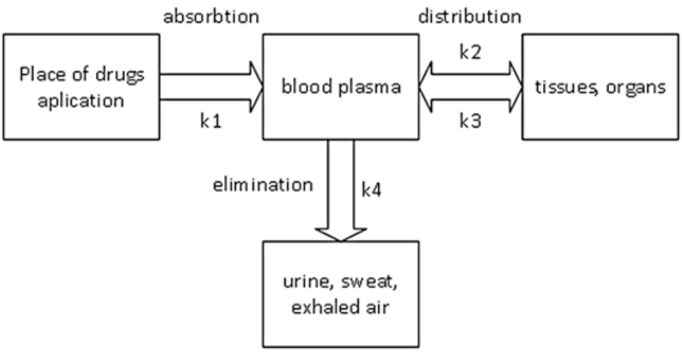


Fig. 1 Pharmacokinetic diagram fate of the drug in the organism [4]

study the behavior of various substances in the body. No study, however, does not address the fundamental principles of pharmacokinetic modeling.

Pharmacokinetics deals with the fate of drugs in the body after administration and studying the processes affecting time during concentration of the drug via an appropriate description of the appropriate pharmacokinetic parameters. Pharmacological response is then developing appropriate mathematical model for interpreting these data [4–6].

Substance (drugs or any other substance), which is in a living organism in a certain way being injected, for some time remains, either unchanged or as metabolites, and then is gradually eliminated from the body out. This is done in various ways, e.g. kidney, lung, skin, intestine, to a lesser extent the mammary glands of breastfeeding women, sweat glands etc. Fate of drugs can be expressed in terms of the kinetic scheme in the next picture (see Fig. 1) [7].

In pharmacokinetics is often used compartmental approach that considers the organism as a set of well interconnected blocks separated by each other biological membranes. Over the compartment is considered the body section of a single drug concentration and uniform conditions for pharmacokinetic processes [4, 8, 9].

The influence of the body to the active substance lies in its influence on individual sections of the movement in the body, i.e. influence on the pharmacokinetics. This organism decides on how fast and in what quantity the drugs reach the target tissue, and whether it will create and maintain the concentration needed to effects on receptors [10].

Interaction of the active substance with the organism starts its administration in a specific dose, specific dosage form and determine the route of administration [11]. In terms of pharmacokinetic drug administration substances can be divided into two basic types:

- Intravascular applications
- Extravascular applications

2 Materials and Methods

Formulation of compartmental pharmacokinetic model is based on the fact that the movement of medicinal substances in the body considerably hampered by the existence of various barriers, which must the drug molecule penetrate. Among the important barrier are undoubtedly biological membranes [12].

Because the distribution of drugs in the body happens very complicated, the pharmacokinetics considerably simplified description of this process through the creation of an abstract model (compartment). Mentioned compartments forme in the organism individual systems, assuming a uniform dispersion of drug in these sections.

The description of most drugs used most frequently one compartment, two compartment, possibly more compartment pharmacokinetic model [4, 13, 14].

2.1 *Model Penetration of Medicaments Trough Biological Membrane*

To the drugs got to your destination in the body, it is essential for overcoming biological barriers, which include biological membranes. The most common mechanism of this permeation is simple diffusion.

By mathematical simulation is necessary to imagine two compartment breathing space membrane, where one compartment presents set out before and second compartment on the contrary set out behind by this one membrane. This phenomenon it is possible to model mathematically by using Eq. 1. Fick's law [14, 15].

$$\frac{dX}{dt} = -D \cdot S \frac{dC}{dt} \tag{1}$$

where:

- dX is quantity of the substance which diffuses during the time interval dt,
- S is size of the area through which the substance diffuses (membrane)
- dC/dx is gradient of the concentration, or change in the concentration dC, which is attributable to the distance dx,
- D is the diffusion coefficient (constant of proportionality); negative sign expresses the loss substances from the point where the substance diffuses.

It should be considered that the drug in the individual compartments homogeneously dispersed and its instantaneous concentration are c_1 and c_2. The first Fick's law therefore follows:

$$\frac{dX_1}{dt} = -PS(c_1 - c_2) \tag{2}$$

where:

- X_1, X_2 is instantaneous amount of drug in compartments 1 and 2,
- S is a surface of the membrane,
- c_1–c_2 is the concentration ratio of the difference of the distance of the thickness of the membrane (substituted for the concentration gradient dC/dx),
- P is permeability coefficient.

Then

$$\frac{dX_1}{dt} = -\frac{PS}{V_1}X_1 + \frac{PS}{V_2}X_2 \quad and \quad \frac{dX_2}{dt} = -\frac{PS}{V_1}X_1 + \frac{PS}{V_2}X_2 \tag{3}$$

By pharmacokinetic analysis are mostly not known size volumes, in which the substance is distributed, therefore they are enter into one constant, along with P and S called penetration rate constant $k_{i,j}$ (from compartment i-th to j-th).

$$\frac{PS}{V_1} = k_{12} \quad and \quad \frac{PS}{V_2} = k_{21}$$

Linear equation then becomes:

$$\frac{dX_1}{dt} = -k_{12}X_1 + k_{21}X_2 \quad and \quad \frac{dX_2}{dt} = k_{12}X_1 + k_{21}X_2 \tag{4}$$

2.2 One Compartments Model Endovascular Dosage of Medicinal Substance

This is a model with a single compartment, where elimination plays a crucial role. When modelling one compartment system can emerge from the differential equation:

$$\frac{dX}{dt} = -k_e X \tag{5}$$

2.3 Two Compartments Model Endovascular Dosage of Medicinal Substance

When creating a two compartment model must take into account one central and one peripheral compartment. In this particular case it is a type of elimination from the central compartment. You can come out of two differential equations:

$$\frac{dX_1}{dt} = -(k_e + k_{21})X_1 + k_{21}X_2 \tag{6}$$

$$\frac{dX_2}{dt} = -k_{21}X_1 - k_{21}X_2 \tag{7}$$

where:

- X_1, X_2 instantaneous amount of drug in compartments 1 and 2,
- k_{12}, k_{21} are penetration rate constant, k_e is the elimination rate constant.

3 Results

The above models have been implemented in Matlab Simuling and for possibility of better control and interpretation of results was created GUI applications [14, 16–18].

3.1 One Compartments Model Endovascular Dosage of Medicinal Substance

For imagination of what is happening with the drug after its administration to the organism, it is possible to assemble again compartmental pharmacokinetic models (see Fig. 2), which are a significant output characteristic depending on the concentration of the drug over time. To create this model is started from one Eq. (7).

From the graph (see Fig. 3) one can see how the drug concentration declines over time after its administration [19]. Taking initial drug concentration at time zero was numerically equal to 4.015 mg/l.

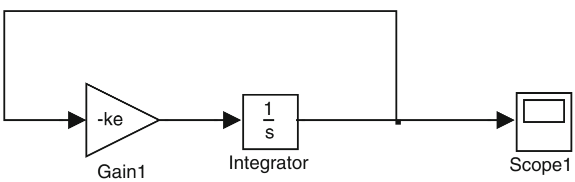


Fig. 2 One compartment model diagram representing the intravascular administration of the drug

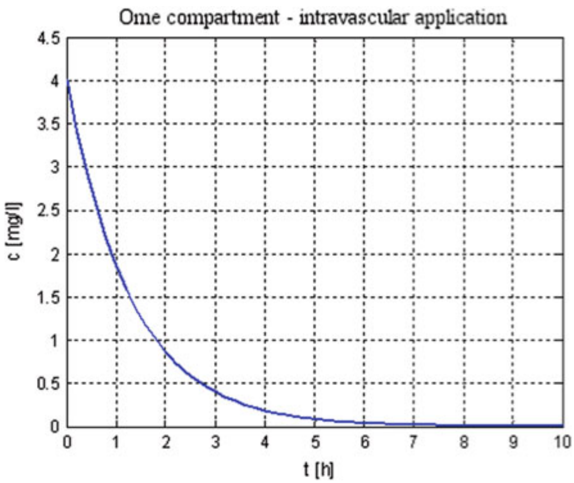


Fig. 3 Resultant graph of one compartment model by intravascular administration of the drug

3.2 Two Compartments Model Endovascular Dosage of Medicinal Substance

The block diagram of the model (see Fig. 4) based on the differential equations given above (Eqs. 6 and 7).

In the resulting graph (see Fig. 5) we can see the blue curve, which represents the decreasing concentration of the central compartment and green curves that characterize the course of concentration in the peripheral compartment.

The peripheral compartment at time zero, there are no amount of the drug, but over time, even in this portion receives. When the green curve reaches its maximum, is at this point completed the process of distribution and elimination occurs alone.

3.3 Model Penetration of Medicaments Trough Biological Membrane

Schematic representation of the transfer of substances membrane corresponds to two compartment model, where it is not included the elimination phase (see Fig. 6). This is purely a penetration of the drug through the membrane compartment from X_1 to X_2 and vice versa. To create the model is started from the differential Eq. (6).

The resulting graph (see Fig. 7) represents the passage of drug through membrane, where the active substance passes from the central compartment to the peripheral

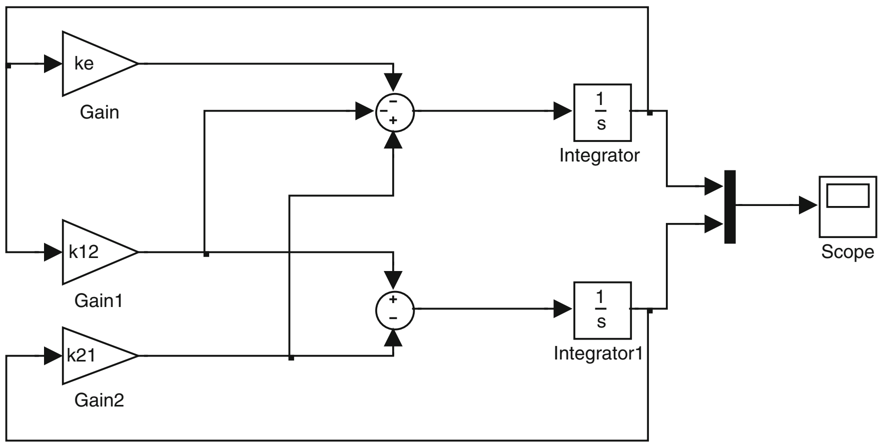
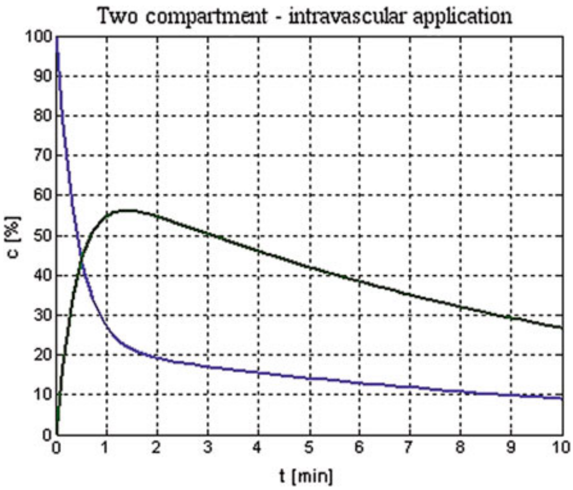


Fig. 4 The two compartment model diagram representing the intravascular administration of the drug

Fig. 5 The resulting graph of a two compartment model intravascular administration of the drug



compartment. The blue curve represents a substance brought into the central compartment and the green curve represents the substance gets into the peripheral compartment from the central. In zero time you can see a difference in concentrations on the membrane. From the side where the substance administered is the maximum, while on the other side of zero. The zero concentration increases, while the maximum value decreases. At the time, however, the concentrations values stabilize.

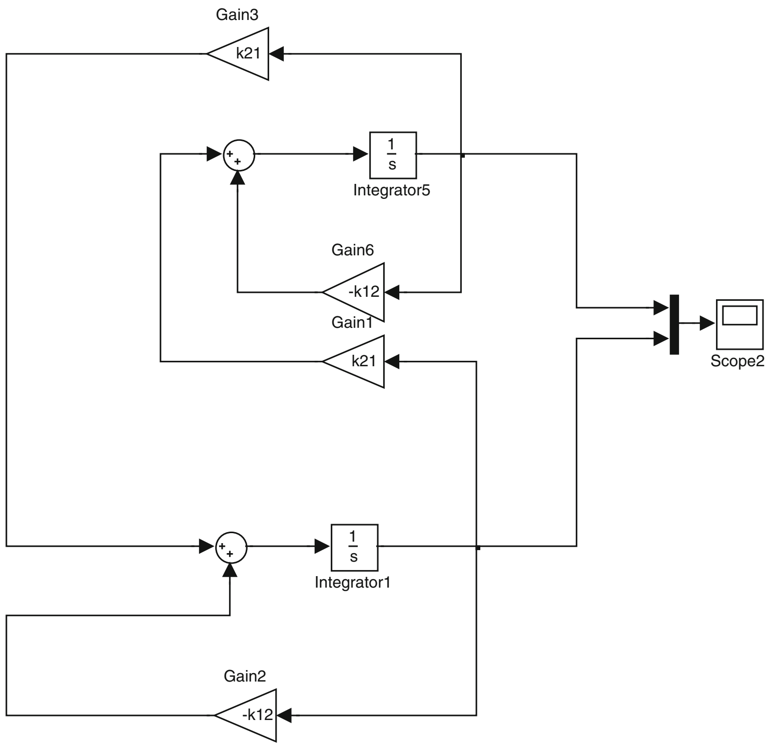
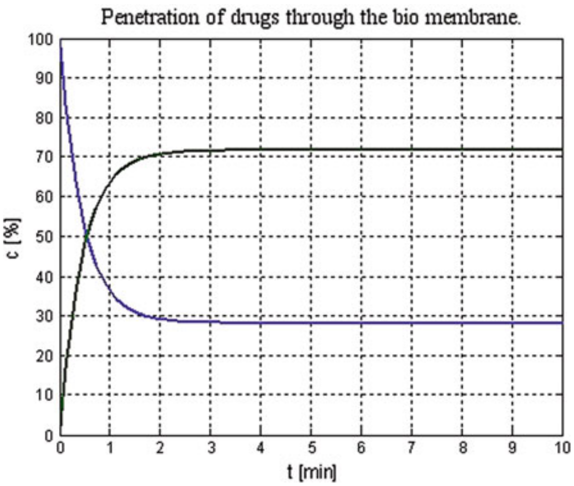


Fig. 6 Diagram of penetration model of drugs through the bio membrane in Simulink

Fig. 7 The resulting graph model of penetration of drugs through the bio membrane



4 Models Verification

Rendering of individual dependence concentration on the drug on time was used program Simulink, where were compiled particular models of differential equations. Since all models operate on the same basis, using modelling equations can be through a different view of addiction compare the accuracy of the results.

This comparison was chosen Microsoft Excel spreadsheet. On the basis of the same input parameters in Matlab, Simulink and Microsoft Excel spreadsheet can be drawn depending compare [20].

For verification, we used an example where a patient received intravenously 200 mg dose of drug. The concentration of the drug varies over time (see Table 1 and Fig. 8).

For using the equation of regression was established elimination rate constant (Fig. 9).

Elimination rate constant was set to a value of 0.7718 h^{-1}. Was also inserted into o compartment model in a control environment GUI, and subsequently drawn (Fig. 10).

Table 1 The concentration of 200 mg administered drug over time

t [h]	0	1	2	3	4	5
c [mg dm^{-3}]	4.015	1.850	0.858	0.397	0.184	0.039

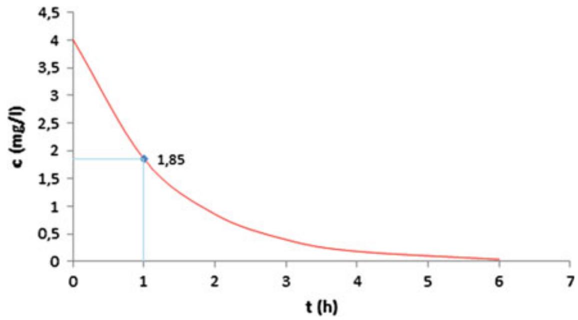


Fig. 8 Part of verification for the concentration at 1 h of drug administration by spreadsheet

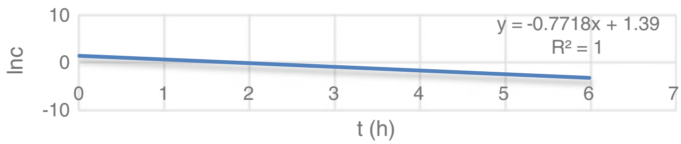


Fig. 9 Logarithmic dependence with equation of regression

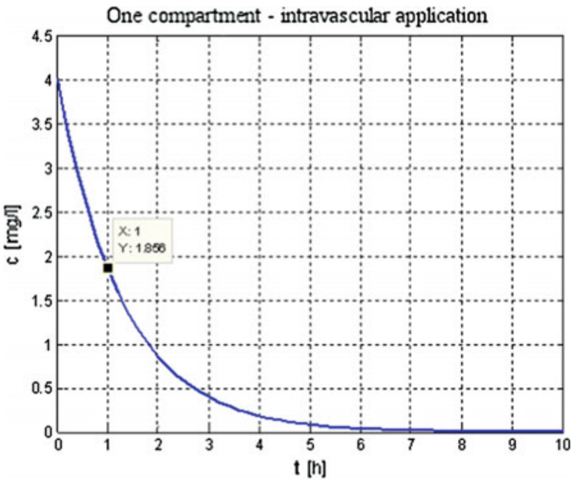


Fig. 10 Part of verification for the concentration at 1 h of drug administration by Matlab

As can be seen above to verify, in the time 1 h from administration of the drug in the individual case the concentration value is in the same range. In Microsoft Excel spreadsheet, it is 1.85 mg/l and through Matlab rounded to one decimal place above 1.856 mg/l. It can therefore be concluded that the functionality of the models processed in Matlab, Simulink is correct.

5 Conclusions

Unlike other studies that have been mentioned in the introduction, the objective of this study verify the basic principles of pharmacokinetics. For this purpose was created three models solving the situation where is a drug into the body injected and we follow its concentration in the body from the actual filing until after removal from the body out. Fourth, the main model of this work deals with the fate of the drug on the biological membrane and in the surrounding area. These models were combined into a single application user interface (see Fig. 11), which makes it easy to use models and easy interpretation of the results in the form of graphical outputs. Functionality of models were validated on the type of the drug of known concentration [18].

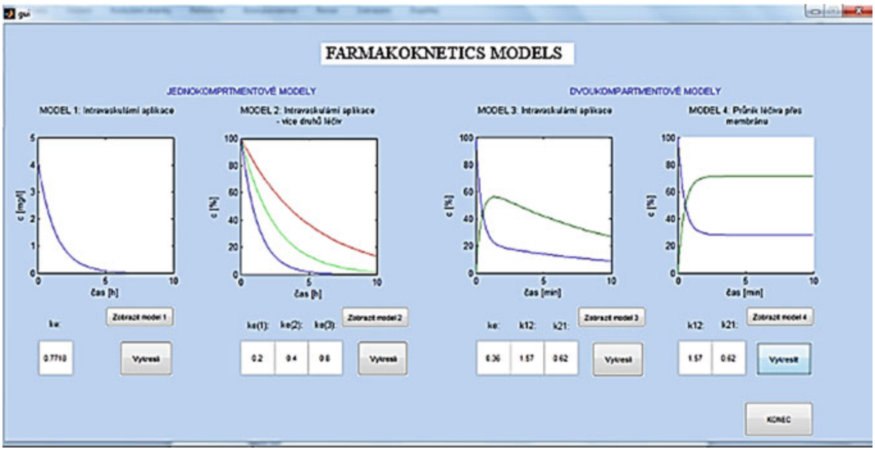


Fig. 11 Preview of the control environment in the GUI status of enabled models

Acknowledgements The work and the contributions were supported by the project SP2015/179 'Biomedicínské inženýrské systémy XI' and This work is partially supported by the Science and Research Fund 2014 of the Moravia-Silesian Region, Czech Republic and this paper has been elaborated in the framework of the project "Support research and development in the Moravian-Silesian Region 2014 DT 1—Research Teams" (RRC/07/2014). Financed from the budget of the Moravian-Silesian Region.

References

1. Ladokhin, A.S.: Measuring membrane penetration with depth-dependent fluorescence quenching: Distribution analysis is coming of age. Biochimica Et Biophysica Acta-Biomembranes **1838**, 2289–2295 (2014)
2. Bolzinger, M.A., Briancon, S., Pelletier, J., Chevalier, Y.: Penetration of drugs through skin, a complex rate-controlling membrane. Curr. Opin. Colloid Interface Sci. **17**, 156–165 (2012)
3. Christ, A., Szurkowski, J., Hanh, B.D., Wartewig, S., Kopycinska, M., Neubert, R.H.H., Cobet, U.: Drug penetration into a membrane investigated by photoacoustic and FTIR-ATR spectroscopy. Anal. Sci. **17**, S371–S373 (2001)
4. Cídlová, H., Štěpán, J.: Řešené příklady z fyzikální chemie. Masarykova univerzita, Brno (2008)
5. Zathurecký, L.: Biofarmácia a farmakokinetika. Vydavateľstvo Osveta, Martin (1989)
6. Cerny, M., Penhaker, M.: The HomeCare and circadian rhythm. 2008 International Special Topic Conference on Information Technology and Applications in Biomedicine, vols. 1 and 2, pp. 110–113 (2008)
7. Cerny, M., Pokorny, M.: Circadian rhythm evaluation using fuzzy logic. Stud. Comput. Intell. **457**, 289–298 (2013)
8. Eck, V., Razím, M.: Biokybernetika. ČVUT, Elektrotechnická fakulta, Praha (1996)
9. Penhaker, M., Tiefenbach, P., Kobza, F.: Lékařská kybernetika. VŠB - Technická univerzita Ostrava, Ostrava (2007)
10. Cerny, M., Penhaker, M.: Wireless body sensor network in health maintenance systems. Elektronika Ir Elektrotechnika 113–116 (2011)

11. Penhaker, M., Darebnikova, M., Jurek, F., Augustynek, M.: Evaluation of electrocardiographic leads and establishing significance intra-individuality. In: Innovations in Bio-inspired Computing and Applications, pp. 295–303. Springer International Publishing (2014)
12. Augustynek, M., Penhaker, M., Korpas, D., Society, I.C.: Controlling peacemakers by accelerometers. 2010 Second International Conference on Computer Engineering and Applications: Iccea 2010, Proceedings, vol. 2, pp. 161–163 (2010)
13. Komers, K., Čegan, A., Komersová, A.: Základy farmakokinetiky. Univerzita Pardubice, Pardubice (2011)
14. Wijk van Brievingh, R.P.V., Möller, D.: Biomedical modeling and simulation on a PC: a workbench for physiology and biomedical engineering. Springer, New York (1993)
15. Potůček, J.: Metodologie modelování biologických systémů. České vysoké učení technické, V Praze (2009)
16. Augustynek, M., Penhaker, M.: Finger Plethysmography classification by orthogonal transformations. In: Computer Engineering and Applications (ICCEA), 2010 Second International Conference on, pp. 173–177. IEEE
17. Augustynek, M., Penhaker, M., Semkovic, J., Penhakerova, P., Cerny, M.: Measurement and diagnosis assessment of plethysmographycal record. In: 5th Kuala Lumpur International Conference on Biomedical Engineering (BIOMED 2011), pp. 320–323
18. Zemanova, A.: Model of drugs penetration through biological membrane (2013)
19. Penhaker, M., Cerny, M., Rosulek, M.: Sensitivity analysis and application of transducers. 2008 5th International Summer School and Symposium on Medical Devices and Biosensors, pp. 85–88 (2008)
20. Augustynek, M., Penhaker, M., Cerny, M., Cernohorsky, J., Bryjova, I.: Education of biomedical engineering in the Technical University of Ostrava. The 15th International Conference on Biomedical Engineering, pp. 531–534. Springer International Publishing (2014)

Power Fading Effects in Millimeter-Wave Radio Over Fiber (RoF) Link

Norliza Mohamed, Suriani Mohd. Sam and Nor Hafizah Ngajikin

Abstract The use of fiber-based wireless system has offered a reliable and cost effective communication system. Such system greatly provides superior possible bandwidths for both fiber and free-space applications. For that reason, radio over fiber (RoF) technology is the best preference to deal with the improved capacity and mobility mainly at higher frequency. However, RoF link might suffer from the power fading effect especially when millimeter-wave (mm-wave) signal transmission is involved. Therefore, this work aims to investigate the mitigating of power fading for the proposed configuration of high optical carrier frequency signal generation at three different conditions based on stimulated Brillouin scattering (SBS) technique.

1 Introduction

Enormous growth of wireless communication system in the last decade has resulted in the significant increase in the demand for high user capacity and high data rate services. In particular a wider radio frequency spectrum is very much needed over a radio link. It is essential for radio link to employ higher frequency carriers because spectrum congestion occurs at low frequencies. Numerous research works have been conducted in mm-wave signal generation with optical mm-wave production

N. Mohamed (✉) · S. Mohd.Sam
Universiti Teknologi Malaysia, Razak School of Engineering and Advanced Technology, Jalan Sultan Yahya Petra, 54100 Kuala Lumpur, Malaysia
e-mail: norlizam.kl@utm.my

S. Mohd.Sam
e-mail: suriani.kl@utm.my

N.H. Ngajikin
Universiti Teknologi Malaysia, Lightwave Communication Research Group, Faculty of Electrical Engineering, 81310 Skudai, Johor Darul Takzim, Malaysia
e-mail: nhafizah@utm.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_42

being an essential technique in RoF system [1, 2]. Signal transmission in mm-wave radio communication systems by using optical fiber has been extensively investigated [3–5] since they provide high bandwidths and pico-cell sizes.

By using RoF, the capacity of optical networks can also be integrated with the flexibility and mobility of wireless access networks. Considering these conditions, the combination of wireless and optical networks could provide a solution for the increasing capacity and mobility as well as reducing the costs in the access network. In this study, the concept of RoF has been implemented due to several advantages offered such as the reduction of the complexity at the antenna site and the allocation of the radio carriers to the different antenna sites (frequency reuse).

On the other hand, RoF link might suffer from the dispersion effect when higher frequency transmission like mm-wave signal is involved. Even though the dispersion effect can be compensated with the use of dispersion-shifted fiber (DSF) where zero dispersion wavelengths occur, such fiber is quite expensive and could increase the cost of creating a new fiber link or replacing the existing link. In addition, due to the zero dispersion wavelength, the attenuation coefficient of the fiber is slightly increased which, might degrade the performance of the signal.

Considering these issues, several techniques were proposed by number of works in avoiding or minimizing the dispersion effects. Many research works have been published and reported in RoF system by applying various techniques as well as configurations for the signal generation and transmission. To ensure that the mm-wave transmission system is tangible, appropriate techniques and configurations are required to adopt the mm-wave in any RoF system. There are many techniques can be applied for the mm-wave signal generation either at the central station (CS) or at the base station (BS) subject to the requirements of the system [6, 7].

Wei et al. in [8, 9] has demonstrated a harmonic RF carrier generation and broadband data up-conversion technique with single mode and single sideband (SSB) modulation for RoF systems using stimulated Brillouin scattering (SBS) with power penalty was less than 1 dB. In SSB modulation technique, it offered a feasible and more robust against fiber dispersion when an unmodulated RF carrier is contemplated. In addition, the integration between single mode modulation (SMM) and SSB has been proven as an effective way to alleviate the dispersion induced power penalty (DIPP) for RoF [10, 11].

In this work, the front-end optical receiver utilizing remote optical local oscillator (ROLO) for RoF system is proposed and it is illustrated in Fig. 1. The figure depicts the transmitter and receiver configurations of the proposed RoF system connected with standard single mode fiber (SSMF) as the transmission medium. Two different signal generation schemes are employed at both transmitter and receiver sides respectively. High optical carrier frequency is generated at one part of the transmitter based on SBS technique while up-conversion technique is done utilizing heterojunction bipolar transistor (HBT), which consequently, higher frequency signals are generated locally at BS. The mixing technique and frequency conversions draw high attention in the application to analog fiber communication.

In addition, by generating the RF carrier signal as a local oscillation signal which replaces the conventional local oscillator (LO), it could reduce the cost of high

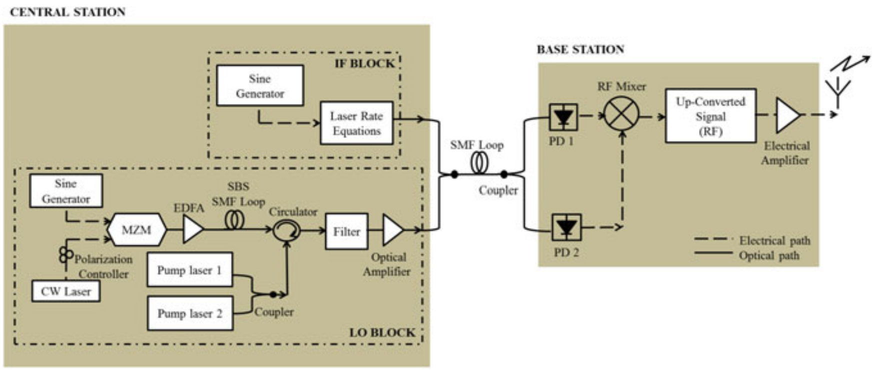


Fig. 1 Block diagram of the RoF-ROLO System

frequency LO and it could be done remotely at CS. This combination of techniques provides a lower cost system and reduces the dispersion effect in terms of the system performance and transmission distance limitation [6, 7, 12, 13]. Therefore, the mitigating power effects of three different conditions in RoF signal generation configuration based on the RoF-ROLO system will be presented and discussed in this paper.

2 Signal Generation Based on SBS Technique

This section explains the development of signal generation based on SBS technique and the simulation work has been conducted utilizing OptiSystem software. The effect of different optical fiber loop length, the effect of different optical carrier input power and the effect of different Responsivity value of the p-i-n photodiode (PD) are considered. The required SBS parameters are appropriately set including the SBS simulation circuit model prior to the simulation of the optical signal generator. In this model, the important parameters of the SBS simulation model are listed in Table 1. The results of the simulation and the performance of the model are studied for three different configurations. The most significant parameters that are considered in this work are the optical input power of the continuous wave (CW) laser, the length of the SBS fiber loop and the responsivity, R of the p-i-n PD, while the other parameters were remained constant.

A CW laser and an electrical signal with frequency of 10 GHz was modulated by the Mach Zehnder modulator (MZM). The CW polarization input was fixed due to the polarization controller (PC) attached before the MZM. The 1550 nm CW laser wave with 1 MHz narrowband linewidth was nonlinearly modulated with the frequency of the electrical generator by the MZM with a sufficient amount of bias voltage. Optical sidebands frequency separated by 10 GHz can be seen at the optical output spectrum of the MZM. These signals were then conveyed to a 25 km

Table 1 Parameters setting of SBS simulation model

Parameters	Values
Bit rate	10 Gbps
Time window	1.28e−0.08 s
Sample rate	640 GHz
Sequence length	128 Bits
Sample per bit	64
Optical Power of CW laser	−30 to 10 dBm
Wavelength frequency of CW laser	1552.52 nm
Linewidth of CW laser	1 MHz
Frequency of Sine Generator, f_{LO}	10 GHz
Frequency of Pump Laser 1 (PL1)	193.09 THz
Frequency of Pump Laser 2 (PL2)	193.11 THz
Responsivity, R of the p-i-n PD	0.1–1.0 A/W
EDFA length	5 m
Optical Fiber Loop Length	1–50 km
Dispersion of Optical Fiber	17 ps/nm/km
Attenuation of Optical Fiber	0.2 dB/km
Brillouin Gain, g_B	4.6e−11 m/W

of SSMF after passed through a 5 m of erbium doped fiber amplifier (EDFA) to control the power of the signal wave.

For the Brillouin to take effect, two pump lasers which are the pump laser 1 (PL1) and pump laser 2 (PL2) were used to generate a shared output signal and coupled them into the fiber of another side. The wavelength of each pump laser was set to be about 11 GHz higher than the modulated signal to avoid any distortion of the signal. Consequently, the optically generated signal could be obtained at the output of the circulator due to the counter-propagating pump waves of the pump lasers. Two chosen frequencies of the sidebands were amplified by the SBS whereas all other frequency components were attenuated due to the natural attenuation in the fiber [14]. These chosen signals were then transmitted to another SMF before they were detected by p-i-n PD at the receiver side.

The circulator allows any signals entering the ports in its rotation. In this situation, the modulated signal from MZM was initially amplified by an EDFA before it entered the first port of the circulator, while the pump signals entered the other input port. The SBS signals were generated when the pump waves of the pump lasers were counter-propagating in the fiber. Hence, the output spectrum of the circulator consists of both counter-propagating signals and the amplified modulated signal from the MZM. Since in this work two pump lasers were used, thus a set of sidebands of optical frequency which are the lower sidebands (LSBs) and upper sidebands (USBs) were obtained at the output port of the circulator as depicted in Fig. 2.

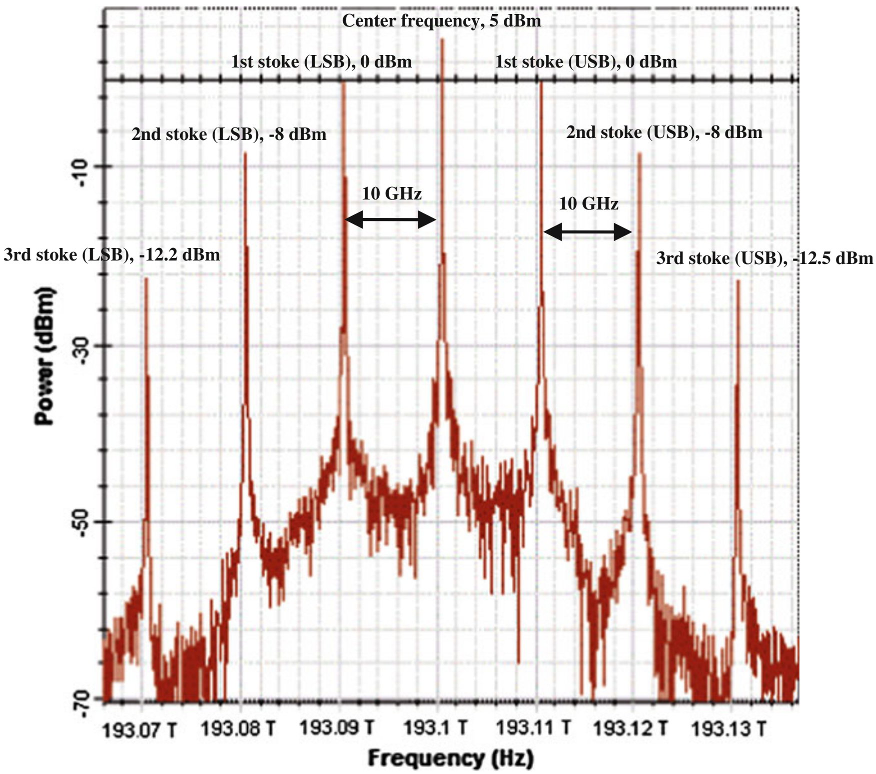


Fig. 2 Optical spectrum of the circulator

3 Performance Analysis and Discussion

Analysis on the performance of the SBS model will be mainly based on the changing effects of some parameters which are by varying the SBS fiber loop length, different optical power of the CW optical laser carrier and the effects of different responsivity value of the p-i-n PD.

3.1 Effect of Different Optical Fiber Loop Length

In observing the effect of different optical fiber loop lengths, fix value of other parameters were initially determined. The effects were based on specified optical power of the CW laser at wavelength of 1552.52 nm. The SBS fiber loop length was varied from 1 to 50 km and the optical amplifier gain was set to be at 20 dB. While, at the receiver, the responsivity, R of the p-i-n PD was fixed at 0.8 A/W. Figure 3 shows that as the frequency getting higher the detected signal getting more

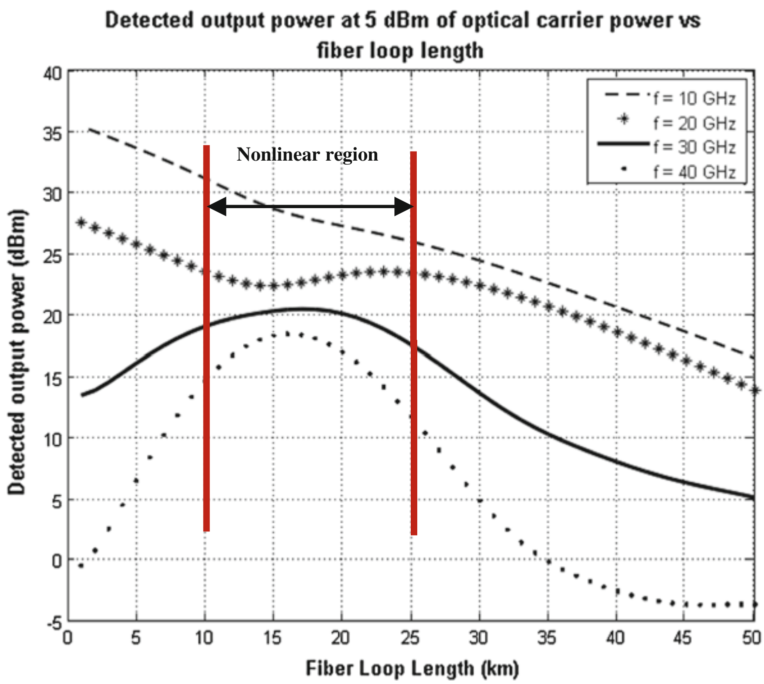


Fig. 3 Detected output power at 5 dBm of optical input power of CW laser for different generated frequencies as a function of fiber loop length

nonlinear due to the nonlinearity effects of the SMF. Obviously, at 20 GHz the nonlinear property had existed at the fiber length of 15 km and at 30 GHz, the nonlinearity was appeared at 10–25 km. While, at 40 GHz, the detected signal was at the highest point when the fiber loop was about 16 km. Based on this finding, it was found that the optimum fiber loop length that can be used in generating the optical signal based on the SBS technique of the model are in the range of 10–25 km with the peak value at about 17 km. In addition, as the distance increased thus the detected power was consequently faded.

3.2 *Effect of Different Optical Carrier Power*

Another important characteristic in the SBS configuration model is by studying the effect of different optical carrier power of the CW laser. Therefore, some parameters had to be constant. In this condition, the wavelength of the CW laser was remained at 1552.52 nm with the optical carrier power level in the range of −30 dBm up to 10 dBm. Meanwhile, the SBS fiber loop length was fixed at 25 km and the optical amplifier gain was set to be at 20 dB. The responsivity of the PD was varied at 0.6,

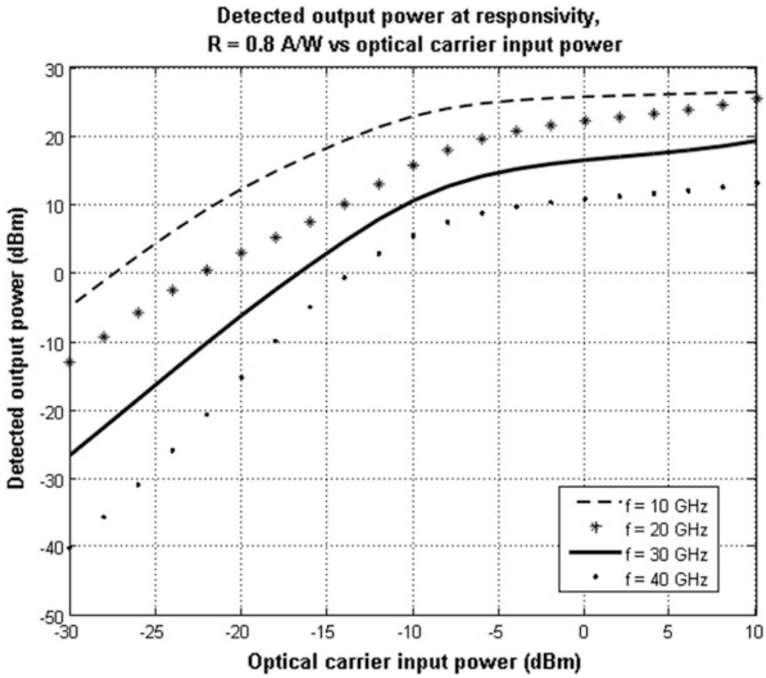


Fig. 4 Detected output power at 0.8 A/W of responsivity for different generated frequencies as a function of optical carrier power

0.8 and 1.0 A/W. Figure 4 depicts the output power that was achieved at 0.8 A/W for different generated frequencies as a function of the optical carrier power.

As can be seen, when higher frequency is generated, the output power has faded at a certain proportion. In that case, at 0 dBm of input power of 10 GHz generated signal, the detected power was about 25.57 dBm whilst at 20 GHz, the detected power was reduced to 22 dBm that is about 14 % decrement. The detected power continues to decrease to about 25.7 and 34.7 % when the 30 and 40 GHz signals were generated respectively. Additionally, as has been mentioned earlier, there is a region where the detected signal started to be almost constant which falls in the range of −12 to −4 dBm. This region has been classified as a transition area. Hence, for this SBS setting and configuration, the maximum optical carrier input power that may be utilized is more or less −4 dBm. If higher input power is consumed, it is a waste of power since the effect to the output power level is very small or the effect to the output power is nearly horizontal. In most devices, the input power is limited at a certain value in which at higher input power might cause impairment to the device.

In addition, the higher the responsivity of the PD, the higher the output power level was detected. Nevertheless, the responsivity has its optimum region and will be discussed in the next subsection. At frequency of 10 GHz, for instance, the detected power is increased about 18.7 % from 0.6 to 1 A/W at 10 dBm of input power. On the other hand, at frequency of 40 GHz, the increment percentage with the same parameter condition is 42.1 % which is more than double from the increment percentage of 10 GHz signal. Therefore, it was found that when the frequency increases, the detected power would rise tremendously from 0.6 to 1 A/W. For this reason, this study was continued by the changing effect of the responsivity of the PD in the next subsection.

3.3 Effect of Different Responsivity Value of the p-i-n PD

Responsivity value of the photodetector also plays an important role during the photodetection of the generated signal. Basically, responsivity, R and quantum efficiency, η are the main characteristics of a PD which contribute to the sensitivity of an optical receiver that is the minimum input power that is required for the receiver to an optimum performance. Responsivity of a typical p-i-n PD is in the range of 0.6–0.9 A/W and relationship between η and R is given as [15]:

$$\eta = R \times \frac{hc}{\lambda q} = \frac{1.24R}{\lambda}\ \text{[A/W]} \tag{1}$$

Therefore, this subsection discusses the effect of different responsivity values for the p-i-n PD that is used in the SBS configuration model. The SBS fiber loop length is predetermined at 25 km as well as the wavelength of the CW laser still remained at of 1552.52 nm. The optical carrier power level is set at −10, −5, 0 and 5 dBm while the optical amplifier gain is retained at 20 dB. As has been discussed in previous subsection, when the input power levels are increased, the detected power levels have remained constant even though the responsivity values are changed. Furthermore, as the responsivity values of the PD are increasing, the detected output power increases at a certain proportion. At input power of −10 dBm, from 0.1 to 0.2 A/W, the output power is increased tremendously about 128.7 %. However, from 0.2 to 0.3 A/W, the output power is increased just about 32.9 %, while from 0.3 to 0.4 A/W, the increment percentage is 17.6 % and so on as can be found in Fig. 5. It was found that as the responsivity increases, the increment rate is decreasing. Hence, from Fig. 5, it was proven that, the responsivity of a typical p-i-n PD was in the appropriate region that was between 0.6 and 0.9 A/W where the trend shows very small increment in the detected power level at that region. It is expected that the increment will remain steady as the responsivity increases.

In Fig. 6, it exhibits the detected output power at −5 dBm of input power for different generated frequencies. The trend of the graph at each frequency shows that as the generated frequency increases, the detected power declines. However, the decrement percentage has raised approximately 40 % of the previous frequency. For instance, at 0.8 A/W of the responsivity, the detected power has dropped about 19.4 % from 10 to 20 GHz. While, from 20 to 30 GHz, the detected power has

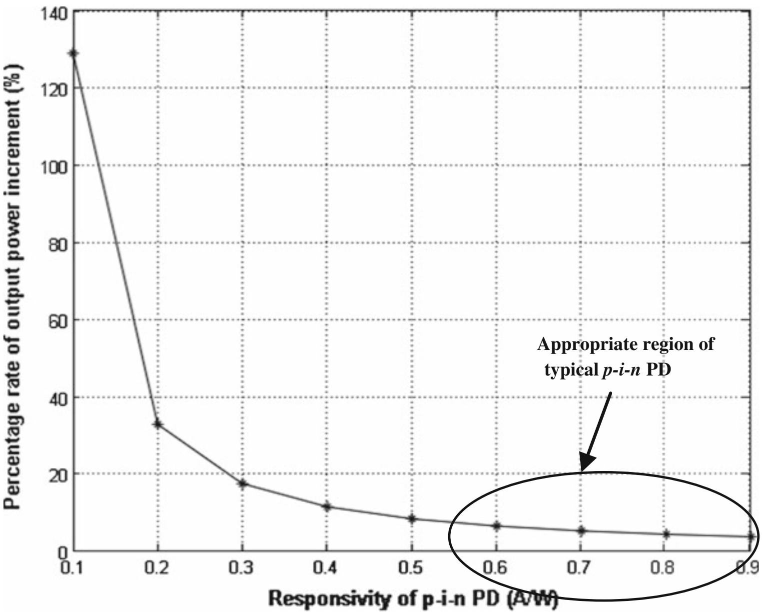


Fig. 5 Percentage rate of output power increment at 10 GHz generated signal as a function of responsivity of p-i-n PD

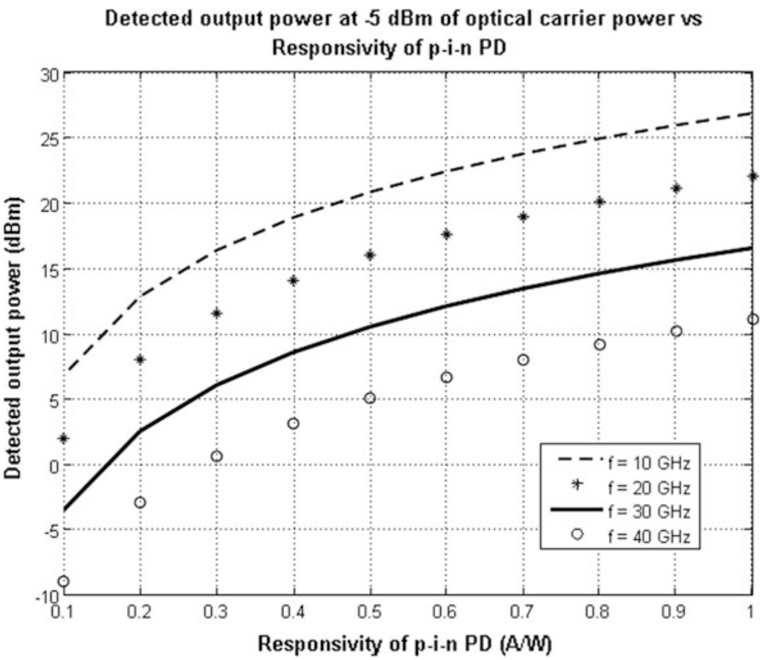


Fig. 6 Detected output power at −5 dBm of optical input power of CW laser for different generated frequencies as a function of responsivity of p-i-n PD

lessen about 27.3 % which the percentage is increased about 40 % of the previous frequency. Similarly, from 30 to 40 GHz, the detected power has reduced around 37.3 %. With this condition, it was disclosed that when the generated frequency increases, the detected power decreases about 40 % more than the previous frequency.

4 Conclusion

This paper has presented the RoF-ROLO system with the integration of SBS signal generator and HBT mixer model. This system has exhibited to a ROLO system that meet the requirements needed in any RoF environment. The performance of the RoF-ROLO system was analyzed and evaluated according to the effect of different input power levels in order to investigate the power fading effect in the RoF link. The simulation of the system was carried out by considering the setting of key parameters which were the optical carrier power of the CW laser, the length of the SBS fiber loop and the responsivity of the p-i-n PD, while the other parameters were remained constant. The analysis on the simulation performance of the SBS model was presented which emphasis on the discussion of the effects of those parameters. It was found that the output power has mitigating effects when those three different conditions were considered.

Acknowledgments The authors acknowledge that this work was supported by Universiti Teknologi Malaysia (UTM) for the financial support through Research University Grant Scheme, Vot. No. 09J82 to the main author.

References

1. Gulistan, A., Ghafoor, S.: Self-phase modulation-based multiple carrier generation for radio over fiber duplex baseband communication. Photon Netw. Commun. **29**(2), 133–137 (2015)
2. Khan, Y., et al.: Generation and transmission of dispersion tolerant 10-Gbps RZ-OOK signal for radio over fiber link. Front. Optoelectron. **5**(3), 306–310 (2012)
3. Chattopadhyay, T.: A millimeter-wave radio-over-fiber system for overcoming fiber dispersion-induced signal cancellation effect. Optoelectron. Lett. **8**(4), 293–296 (2012)
4. Lecoche, F., et al.: Transmission quality measurement of two types of 60 GHz millimeter-wave generation and distribution systems. J. Lightwave Technol. **27**(23), 5469–5474 (2009)
5. Kuri, T., Kitayama, K., Takahashi, Y.: 60 GHz band full-duplex radio-on-fiber system using two-port electroabsorption transceiver. IEEE Photonics Technol. Lett. **12**(4), 419–421 (2000)
6. Han, Xiuyou, et al.: A new method for fiber chromatic dispersion measurement with microwave interference effect. Optical Fiber Technol. **19**(4), 319–324 (2013)
7. Kun, Xu, et al.: Heterodyne mixing and polarization diversity techniques in radio over fiber system with high sensitivity and dispersion tolerance. Sci. China Inf. Sci. **54**(2), 236–243 (2011)

8. Wei, L., et al.: Tunable carrier generation and broadband data upconversion for RoF systems based on stimulated Brillouin scattering. IEEE Trans. Microw. Theor. Tech. **59**(9), 2350–2356 (2011)
9. Li, W., Zhu, N.H., Wang, L.X.: Harmonic RF carrier generation and broadband data upconversion using stimulated Brillouin scattering. Optics Commun. **284**(13), 3437–3439 (2011)
10. Park, C.S., Lee, C.G., Park, C.S.: Experimental demonstration of 1.25-Gb/s radio-over-fiber downlink using SBS-based photonic upconversion. IEEE Photonics Technol. Lett. **19**(22), 1828–1830 (2007)
11. Cheng, H., et al.: Single-sideband modulation based on an injection-locked DFB laser in radio-over-fiber systems. IEEE Photonics Technol. Lett. **22**(7), 462–464 (2010)
12. Martin, E., Barry, L.: Side mode suppression and dispersion compensation analysis of a 60 GHz radio-over-fibre system based on a gain switched laser. Optics Commun. **313**(15), 36–41 (2014)
13. Lu, Z., et al.: Effect of chromatic dispersion and initial chirp on the DOP feedback signal in PMD compensation. IEEE Photonics Technol. Lett. **17**(2), 342–344 (2005)
14. Schneider, T., Junker, M., Hannover, D.: Generation of millimetre-wave signals by stimulated Brillouin scattering for radio over fibre systems. Electron. Lett. **40**, 1500–1501 (2004)
15. Agrawal, G.P.: Fiber-Optic Communication Systems. Wiley, Hoboken (2002)

A New SMS Spam Detection Method Using Both Content-Based and Non Content-Based Features

Nurul Fadhilah Sulaiman and Mohd Zalisham Jali

Abstract SMS spamming is an activity of sending 'unwanted messages' through text messaging or other communication services; normally using mobile phones. Nowadays there are many methods for SMS spam detection, ranging from the list-based, statistical algorithm, IP-based and using machine learning. However, an optimum method for SMS spam detection is difficult to find due to issues of SMS length, battery and memory performances. Hoping to minimize the aforementioned problems, this paper introduces another detection variance that is based on common characters used when sending SMS (i.e. numbers and symbols), SMS length and keywords. To verify our work, the proposed features were stipulated into five different algorithms and then, tested with three different datasets for their ability to detect spam. From the conduct of experiments, it can be suggested that these three features are reasonable to be used for detecting SMS spam as it produced positive results. In the future, it is anticipated that the proposed algorithm will perform better when combined with machine learning techniques.

1 Introduction

Spam messages are unsolicited messages sent by spammers to known and unknown users for various purposes. Among others are for fraud, advertisement and phishing. Spam messages can occur in many forms; emails, SMSs and social network platforms. This paper focuses on the activity of spamming using SMS or text messages as the increasing usage of SMS to communicate nowadays provides huge opportunities for spammers to do their job. SMS different with email in several aspects

N.F. Sulaiman · M.Z. Jali (✉)
Faculty of Science and Technology, Universiti Sains Islam Malaysia (USIM),
Bandar Baru Nilai, 71800 Nilai, Negeri Sembilan, Malaysia
e-mail: zalisham@usim.edu.my

N.F. Sulaiman
e-mail: fadhilahsulaiman90@gmail.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_43

i.e. it does not contain a mailing list of recipients and the maximum words for messages is up to 160 characters [1].

To the best of researchers' knowledge, the detection of spam messages can be done through content-based features [2], non-content features [3] and machine learning classifiers [4]. Many researchers study techniques and methods for filtering spam messages in email [5–11] but less publication reported on the phone.

The common way to detect SMS spam is using 'keywords of spam'. This approach however resulted in using more space and memory; as well as processing would take time due to the larger set of 'keywords of spam'. In this paper, we focus on detecting spam messages of SMS by combining both content-based and non-content based features. As such, three features are proposed; namely based on the length of messages that is more than 100 characters, messages contains special characters such as symbols and numbers as well as keywords of spam messages itself and our justification of using these are highlighted as following:

1. *Length of message* (i.e. greater than 100 characters).
   From our finding analysing SMS spam datasets, we found SMS spams tend to be longer in size. Besides, SMS spam normally use standard and formal language in order to attract users so that they can understand and put their interest to those particular messages. Therefore, in our work, we assume messages that are more than 100 in length could potentially be classified as spam.
2. *Special Characters* (i.e. numbers and symbols).
   From our study, we can reveal that the usage of special characters do exist and common in SMS spams. For instance, spammers prefer using numbers or digits such like phone number, code number to claim, service code and sum of money (i.e. 300 pounds). In addition to this, special characters or symbols such as ****, $, XXX are commonly found to be used in SMS spams.
3. *Keywords of spam and ham.*
   From our analysis, words that being used in spam messages are common and similar across platforms. Thereby, our approach also considers common keywords for detecting spam. Examples of keywords are like—'Free', 'Call' and 'Claim'. Besides, ham keywords are also investigated and use in our detection algorithms.

By detecting SMS spam using their lengths, used characters and keywords, we postulate it will make the detection rate higher; and at the same time minimizing device resources. To achieve what we have claimed for, three aforementioned detection features were stipulated into five different algorithms; with validation of the used datasets were done using four well-known machine-learning classifiers.

The rest of this paper is organised as follows. Section 2 provides an overview about existing methods and techniques for SMS spam detection. Section 3 then describes our methodology with Sect. 4 presents results and discussions from our conduct of testing. The paper concludes with an outlook for future works.

2 Existing Research Related to Spam Detection

Much research has been put to investigate spam messages especially on the email platform. For example, Pour et al. [11] discussed three techniques for email spam detection; namely *list-based*, *statistical algorithm* and *IP-based*. The *list-based* technique is classified into three categories, namely Blacklist, Whitelist and Greylist. Blacklist blocks the IP address based on complaints from recipient [12], with Greylist rejects mail from unknown sources on the theory that real mailers will retry the mail and spammers won't [13]. The *statistical algorithm* can be categorised into content-based method and rule-based method. Content-based method is commonly used and it filters the content of mail body and headers. It uses machine learning which need to be trained [9]. For example, Chakraborty and Mondal [10] applied different decision tree classifiers to filter spam mail while Amayri & Bouguila [14] used Support vector Machine (SVM) for spam filtering. Rule based method works through certain rules and these rules will decide to pass or block the email [9]. Reverse lookup is an example of method in the *IP-based* technique. It is a method of resolving an IP address into a domain name [15].

Development of technology in mobile phone with sophistication of new invention such as smartphones and tablets lead spammers to change their target to mobile users. This is due to the multifunction of mobile phone to access corporate network and data. The authorities and researchers have to take action to reduce and control SMS spams that are rising year by year. Generally, SMS spams can be detected by examining and reviewing message contents (i.e. *content features*) or the way messages are sent (i.e. *non-content features*). Sohn et al. [2] proposed a method of using stylistic information to the content-based mobile spam filtering. They focus on the way the SMS is written (i.e. stylistic aspect) and four features of stylistic were used—length of messages, function word frequencies, part-of-speech n grams and special character. Tan et al. [16] identified features of SMS spams based on word grams, character grams, alphanumeric and non-alphanumeric characters. They also tested a number of statistical features such as message length, proportion of upper-case letters and proportion of punctuation. Other study is by Mujtaba and Yasin [1], where they used four features—size of the message, the existence of frequently occurring monograms in the messages, existence of frequently occurring diagram in the message and messages class. These features are implemented and trained in machine learning algorithm for better accuracy.

Xu et al. [3] focus on the non-content features such as statistical features, temporal features and network features. Their study showed that the temporal features and network features were more effective as compared to the statistic features. Mosquera et al. [17] analyzed the effectiveness of machine learning filters based on linguistic and behavioral patterns. Shahi and Yadav [4] conducted experiments to compare the performance of machine learning classifiers to detect SMS spam in Nepali language and Bilal and Farooq [18] compared four types of evolutionary learning classifiers to filter SMS spams in order to obtain the best classifiers. Other study that conducted in Nuruzzaman et al. [19] used the same approach.

3 Algorithm and Methodology

From the review of existing works in the Sect. 2, it can be suggested that SMS spam research is still needed and provide wide opportunity for improvement. Having said, this paper proposes another variation for detecting SMS spams using the combination of content and non-content features. For testing purposes, the proposed features are expanded into five (5) different algorithms. The idea of expanding features into five (5) variations of algorithms is to investigate their performance, with the ultimate aim to find the best combination for optimum detection.

The first three algorithms use only two aforementioned features; with the remaining algorithms combine all features. Specifically, Algorithm 1 uses only keywords, Algorithm 2 uses the combination of message length and keywords, with Algorithm 3 uses the features of special characters and keywords respectively. Algorithm 4 and 5 use message length, special characters and keywords. Figure 1 presents pseudo code of the proposed method and Table 1 shows the different level of features used in each algorithm.

Two main phases involved in our experiments conduct. In the phase 1, validation of datasets was conducted. To do this, four (4) data mining classifiers in WEKA (i.e. Naïve Bayes (NB), Support vector machine (SVM), k-Nearest Neighbor (k-NN) and Decision Tree (DT)) were used. The used datasets named UCI Machine Learning (UCI) [20], British English SMS Corpora (BEC) [21] and

```
Input; M: SMS messages
Output; H: Ham SMS or S: Spam SMS

Begin (while TRUE)
     Read all SMS messages, M
     Detect S (i.e. using five algorithms as in Table 1)

     Copy to H or S folder
End
```

Fig. 1 Generalized pseudo-code for five algorithms

Table 1 Different features in each Algorithm

Detection of ham/spam				
Algorithm	Phase 1	Phase 2	Phase 3	Phase 4
1	Spam keywords	Ham keywords	(NA)	(NA)
2	Length + Spam keywords	Ham keywords	(NA)	(NA)
3	Number/digit + Spam keywords	Ham keywords	(NA)	(NA)
4	Length	Number/digit + Spam keywords	Ham keywords	(NA)
5	Length	Number/digit + Spam keywords	Ham keywords	Spam keywords

Table 2 Characteristics of used datasets

	UCI	BEC	DIT
Ham	4825	450	–
Spam	747	425	1353
Total	5572	875	1353

Dublin Institute of Technology (DIT) [22] respectively. Table 2 details out different numbers of spam and ham messages in each dataset.

In the phase 2, testing on the proposed algorithms was conducted. Pre-processing or cleaning process for the dataset is not done because characters (i.e. numbers and symbols) in these messages may help with the detection process. Results for experiments conduct for both phase 1 and 2 were analyzed, compared and reported in the next section.

The measurement of detection performance is based on Correctly Classified Messages, True Positive and True Negative.

- Correctly Classified (CC): SMS is correctly classified as spam or ham message. The rate for accuracy of a classifier or algorithm is determined by the formula as below.

$$Accuracy = (TP + TN)/(TP + FN + TN + FP)$$

- True Positive (TP): SMS messages are correctly classified as ham messages
- True Negative (TN): SMS messages are correctly classified as spam messages.
- False Positive (FP) & False Negative (FN).

4 Results and Discussions

Results are reported and discussed in three parts. Part 1 discusses the results for dataset validation; with part 2 and 3 reported the performance detection of the proposed algorithms.

4.1 Part 1: Dataset Validation

Table 3 presents results of experiment conduct for validating datasets. From the Table 3, it can be found that the k-NN classifier produced the best detection rate where it managed to detect the highest number of spam and ham messages and is it also found that all classifiers managed to detect spam messages (TN) of the DIT dataset. As majority of chosen classifiers perform detection of ham and spam messages in a 'good' manner, it can be suggested that the chosen datasets are appropriate to be used for testing the proposed algorithms.

Table 3 Results in WEKA using four classifiers

Results in WEKA							
Dataset	Classifier	Correctly classified (%)	Incorrectly classified (%)	TP	FN	TN	FP
UCI	NB	5306 = 95.23	266 = 4.77	4646	87	**660**	179
	SVM	5537 = 99.37	35 = 0.63	4825	35	**712**	0
	k-NN	5564 = 99.86	8 = 0.14	4825	8	**739**	0
	DT	5416 = 97.20	156 = 2.80	4807	138	**609**	18
BEC	NB	804 = 91.89	71 = 8.11	413	34	**391**	37
	SVM	868 = 99.2	7 = 0.8	450	7	**418**	0
	k-NN	872 = 99.66	3 = 0.34	450	3	**422**	0
	DT	750 = 85.71	125 = 14.29	434	109	**316**	16
DIT	NB	1353 = 100	0 = 0	0	0	**1353**	0
	SVM	1353 = 100	0 = 0	0	0	**1353**	0
	k-NN	1353 = 100	0 = 0	0	0	**1353**	0
	DT	1353 = 100	0 = 0	0	0	**1353**	0

Bold represents the actual number of spam messages detected

4.2 Part 2: Algorithm 1–3 Results

Results for Algorithm 1–3 towards three different datasets are shown in Table 4. From the Table 4, it can be stated that Algorithm 1 managed to detect more spam messages (TN) as compared to other algorithms. For correctly classified messages

Table 4 Results for algorithm 1, 2 and 3

Simulation results								
Dataset	Algo	Detection feature(s)	Correctly classified (%)	Incorrectly classified (%)	TP	FN	TN	FP
UCI	1	Keywords	4228 = 75.88	1344 = 24.12	3501	1324	**727**	20
	2	Length and keywords	4853 = 87.10	719 = 12.90	4178	647	**675**	72
	3	Characters and keywords	5155 = 92.52	417 = 7.48	4454	371	**701**	46
BEC	1	Keywords	766 = 87.54	109 = 12.46	357	93	**409**	16
	2	Length and keywords	788 = 90.06	87 = 9.94	409	41	**379**	46
	3	Characters and keywords	812 = 92.8	63 = 7.2	425	25	**387**	38
DIT	1	Keywords	1294 = 95.64	59 = 4.36	0	0	**1294**	59
	2	Length and keywords	1135 = 83.89	218 = 16.11	0	0	**1135**	218
	3	Characters and keywords	1221 = 90.24	132 = 9.76	0	0	**1221**	132

Bold represents the actual number of spam messages detected into ham and spam, Algorithm 3 is more accurate for datasets of UCI and BEC while Algorithm 1 is more accurate for DIT dataset.

4.3 Part 3: Algorithm 4 and 5 Results

Algorithm 4 and 5 combine all three (3) proposed features and results are shown as in the Table 5. Table 5 indicates that Algorithm for 4 and 5 perform well in detecting spam messages (TN) for all three datasets, with Algorithm 5 performed better as compared to Algorithm 4. In term of accuracy classifying messages into ham and spam, Algorithm 5 produced more accurate detection for datasets of UCI and DIT.

4.4 Discussions

From the conduct of experiments, it was found that the detection feature that is based on keywords (i.e. spam and ham) is still producing good detection results. For all five (5) algorithms, Algorithm 1 (i.e. using our own keywords of spam and ham) performs well and better in detecting spam messages. Here, we argue that in order to have an optimum detection results, that method needs to have a 'sound' list of keywords. With respect to the clients' mobile environment, it is preferable to have a

Table 5 Results for algorithm 4 and 5

Simulation Results								
Dataset	Algo	Detection feature(s)	Correctly classified (%)	Incorrectly classified (%)	TP	FN	TN	FP
UCI	4	Length, character and keywords	5322 = 95.51	250 = 4.49	4684	141	**638**	109
	5	Length, character and keywords (2x)	5357 = 96.14	215 = 3.86	4670	155	**687**	60
BEC	4	Length, character and keywords	788 = 90.06	87 = 9.94	440	10	**348**	77
	5	Length, character and keywords (2x)	725 = 82.86	58 = 6.63	440	10	**377**	48
DIT	4	Length, character and keywords	1063 = 78.57	290 = 21.43	0	0	**1063**	290
	5	Length, character and keywords (2x)	1177 = 87.00	176 = 13.00	0	0	**1177**	176

Bold represents the actual number of spam messages detected minimum list of keywords due to their limitation. On the other hand, if it meant to be implemented in servers' environment, it should be working perfectly. In addition, detection using keywords is sometimes 'unscrupulous' due to various language styles.

The number of messages in each dataset can also affect the detection rate as different datasets may have different number of spams and hams and also contains different message structures. In term of accuracy, results suggest that Algorithm 5 produced high accuracy for UCI dataset as compared to others but Algorithm 3 performed better for the BEC dataset albeit using only two of the proposed features. The similar condition is occurred for the DIT dataset. In this dataset, Algorithm 1 produced higher accuracy as compared to others although this algorithm used only one feature; which is keywords. We anticipate different detection results are due to the 'immature' of the algorithms and there is a need for improvement. When it combined with the machine learning classifier, we expect it to improve and perform better.

We reported that our model produced a 'fair' detection results and we expect this is caused by two conditions. First is due to the way we do our detection. Specifically, the algorithms work on a 'phase-by-phase' basis and they were non-iterative (i.e. static detection). Using static detection might limit the detected results, as messages that are less than 100 of length were not detected at the first stage (Algorithm 2, 4 and 5). Second is due to the nature of the messages itself, as not all messages contain special characters (Algorithm 3). However, from our second validation, we can confirm that our algorithms managed to detect all spam messages that contain 'spam' special characters, and greater than 100 in length.

5 Conclusions

In this paper, three features that based upon SMS length, special characters and keywords are discussed to detect SMS spams. These features are a combination of content and non-content of the messages. These features were tested using three well-known datasets and from our conduct of experiments, it can be suggested that our preliminary algorithms managed to detect spam messages that contains spam keywords, spam special characters and length. In summary we can suggest that, although different datasets could result in different detection rate, but we expect it will produce better detection rate if combined together into single algorithm.

For the future, we will cover our performance evaluations, which includes ROC curve, F-measure and effect towards battery and processing performances. In addition, we also plan to implement dynamic detection rather than static detection as practiced in the present paper. And finally, the proposed algorithms will be applied into machine learning in order to inject the element of 'learning' for optimum detection.

Acknowledgments Authors wish to thank the Ministry of Education (MOE), Malaysia for funding this research. The grant named RAGS with the grant code USIM/RAGS/FST/STH/36/50913.

References

1. Mujtaba, G., Yasin, M.: SMS spam detection using simple message content features. J. Basic Appl. Sci. Res. **4**, 275–279 (2014)
2. Sohn, D-N., Lee, J-T., Rim, H-C.: The contribution of stylistic information to content-based mobile spam filtering. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pp. 321–324 (2009)
3. Xu, Q., Xiang, E.W., Yang, Q.: SMS spam detection using non-content features. IEEE Intell. Syst. (2012)
4. Shahi, T.B., Yadav, A.: Mobile SMS spam filtering for Nepali text using naïve bayesian and support vector machine. Int. J. Intell. Sci. **4**, 24–28 (2014)
5. Zhou, B., Yao Y., Luo J.: A three-way decision approach to email spam filtering. AI'10 Proceedings of the 23rd Canadian Conference on Advances in Artificial Intelligence, Vol. 6085, 28–39 (2010)
6. Nazirova, S.: Survey on spam filtering techniques. Commun. Netw. J. **3**, 153–160 (2011)
7. Mohammad, N.T.: A fuzzy clustering approach to filter spam e-mail. In: Proceedings of the World Congress on Engineering, vol. 1, 945 (2011)
8. Dasgupta, A., Gurevich, M., Punera, K.: Enhanced email spam filtering through combining similarity graphs. In: WSDM'11 Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, pp. 785–794 (2011)
9. Nosrati, L., Pour, A.N.: Dynamic concept drift detection for spam email filtering. In: Proceedings of ACEEE 2nd International Conference on Advances Information and Communication Technologies (ICT 2011), vol. 2 (2011) 124–126
10. Chakraborty, S., Mondal, B.: Spam mail filtering using different decision tree classifiers through data mining approach—a comparative performance analysis. Int. J. Comput. Appl. **47**, 26–31 (2012)
11. Pour, A.N., Kholghi, R., Roudsari, B.: Minimizing the time of spam mail detection by relocating filtering system to the sender mail server. Int. J. Netw. Secur. Appl. **4**, 53–62 (2012)
12. Ramachandran, A., Dagon, D., Feamster, N.: Can DNS-based blacklists keep up with bots?. CEAS 2006-Third Conference on Email and Anti-Spam (2006)
13. Levine, J.R.: Experience with greylisting. In: Proceedings of second conference on Email and Anti-Spam (CEAS 2005), pp. 1–2 (2005)
14. Amayri, O., Bouguila, N.: A study of spam filtering using support vector machines. J. Artif. Intell. Rev. **34**, 73–108 (2010)
15. Rouse, M.: Reverse DNS (rDNS) definition. (2007). http://searchnetworking.techtarget.com/definition/reverse-DNS. Accessed 01 March 2015
16. Tan, H., Goharian, N., Sherr, M.: $100,000 Prize Jackpot. Call Now! Identifying the Pertinent Features of SMS Spam. SIGIR'12 (2012)
17. Mosquare, A., Aouad, L., Grzonkowski, S., Morss, D.: On detecting messaging abuse in short text messages using linguistic and behavioural patterns (2014)
18. Bilal, J.M., Farooq, M.: Using evolutionary learning classifiers to do mobile spam (SMS) filtering. In: Proceeding GECCO'11 Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation, pp. 1795–1802 (2011)
19. Nuruzzaman, M.T., Lee, C., Choi, D.: Independent and personal SMS spam filtering. In Proceedings of IEEE Conference on Computer and Information Technology, pp. 429–435 (2011)

20. Almeida, T.A., GÃ³mez Hidalgo, J.M., Yamakami, A.: Contributions to the study of SMS spam filtering: new collection and results. In: Proceedings of the 2011 ACM Symposium on Document Engineering (DOCENG'11) (2011)
21. British English SMS Corpora. (2011). Downloaded from http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/. Accessed 07 Aug 2014
22. DIT SMS Spam Dataset. Dublin Institute of Technology (DIT). (2012). Downloaded from http://www.dit.ie/computing/research/resources/smsdata/. Accessed 05 Aug 2014

Ultra Wideband Fat Tissue Fabrication Using Different Cross Linking Agent for Microwave Imaging

Aslina Abu Bakar, Ahmad Rashidy Razali, Normasni Ad Fauzi and Mohd Aminudin Murad

Abstract The paper discusses the fabrication of fat tissue using glyoxal and glutaraldehyde as a cross linking agent to be used in breast phantom for microwave imaging application. The fabricated fat tissue's dielectric constant is measured in the ultra-wideband frequency range of 3.1–10.6 GHz. In brief this paper addresses the two issues of breast phantom fabrication; first to simulate the wide range of permittivity of the breast fat tissue and secondly using different cross linking agent such as glyoxal and glutaraldehyde to replace formaldehyde. The previously reported fat tissue fabrication using formaldehyde as a cross linking agent was not safe for handling and hazardous for health. The fabricated fat tissue using the proposed material has shown that it has a close proximity of dielectric constant when compared with the real fat tissue fitted by using a Debye model.

1 Introduction

Currently, the primary techniques used for breast cancer detection is mammography and magnetic resonance imaging (MRI). However active microwave imaging based technique is an emerging technique that can be considered to be a complimentary technique to mammography and MRI. Thus, there is a need to develop an artificial breast phantom to simulate the interactions of electromagnetic waves with biological tissues in the microwave imaging system before the clinical test on human is done. Many researchers has found the substantial contrast in the dielectric properties of normal and malignant breast tissues across the frequencies that is the basis in developing microwave imaging [1–5].

Usually the breast phantom fabrication should take into consideration the shelf time of the phantom in terms of long term stability. Phantom can be deteriorates in

A.A. Bakar (✉) · A.R. Razali · N.A. Fauzi · M.A. Murad
Fakulti Kejuruteraan Elektrik, Universiti Teknologi MARA, Pulau Pinang, 13500 Permatang Pauh, Penang, Malaysia
e-mail: aslina060@ppinang.uitm.edu.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_44

physical properties and water content that may affect the permittivity of the phantom with time.

The breast contains a high concentration of fat (adipose), dense fibrous, loose connective and glandular tissue. Fat composition is the dominant tissue in the breast of older women but younger women has a high percentage of fibrous tissue [6]. It is found that the tissue is composed of fat cells that are almost spherical and range in size from 50 to 200 mm in diameter. The average compositions are consisting of 80–90 % fat, 10–20 % water as well as a small amount of protein [7].

The paper by Lazebnik et al. [8] who has done a large scale study of real breast tissue measurement has given us a good insight on the heterogeneity issue, dielectric properties of the breast tissue which are primarily determined by the adipose(fat) content of each tissue sample which were divided into three categories and the dielectric properties of some types of normal breast tissues are much higher than the previously reported which has led to a low contrast issue between the cancerous and healthy tissue.

In this paper, an alternative chemical to replace formaldehyde has been proposed and the dielectric constant in the UWB range is measured and compared with the real fat tissue.

2 Fat Tissue Development

In the development of the breast fat it is required to mimicks the material of low water content. The fat tissue is made by mixing 12.0 g propylene glycol and 193 mL milli-Q water. Then 30.8 g gelatine is added slowly and continues stirring. This mixture is then heated gradually to achieve the temperature of 90 °C 200 mL of canole oil is heated to 50 °C and poured to the partially cooled gelatin mixture. About 2.2 mL of commercial dishwashing liquid is included as a surfactant to form oil emulsion. The mixture is then cooled to 40 °C. Finally two different cross linking agent in aqueous solution (about 1.512 g) to replace the formalin (an aqueous solution of the chemical compound formaldehyde) is supplemented with stirring. The emulsion is poured into moulds and allowed to cool at room temperature. Previously, formaldehyde as a cross linking agent has been used [9] which is unfavorable due to the safety issue in handling the sample and the vapour can be hazardous when kept for long term.

3 Complex Permittivity Measurement

The complex permittivity of the fabricated phantom were taken using open ended coaxial line probe HP 85070A in the UWB frequencies. All the measurement of complex permittivity was made at room temperature (approximately 24 °C). VNA calibration is done before the dielctric probe measurement is taken placed, followed

Table 1 Single-pole Debye parameters [12]

	ε_∞	$\Delta\varepsilon$	τ(ps)	σ
Fat-high	3.9870	3.54448	13.0000	0.0803
Fat-median	3.1161	1.5916	13.0000	0.0496
Fat-low	2.8480	1.1041	13.0000	0.2514
Glandular-high	14.2770	40.5152	13.0000	0.6381
Glandular-median	13.8053	35.5457	13.0000	0.7384
Glandular-low	12.8485	24.6430	13.0000	0.2514

by the calibration of the probe using 3 standards: open (where the tip of the probe is leaved in air), short (connected to a short circuited stub) and using deionized water where the probe tip is immersed in a deionised water.

The dielectric properties of fat tissue sample with the size of 3 cm × 3 cm were measured twice and averaged to ensure data accuracy. The tissue sample is kept in a close container to keep its moisture.

In order to verify accuracy of the developed fat tissue sample, a comparison is made with the electrical parameters of real breast fat tissues. This is modeled by using Debye Model and the value of the real tissue is referred as shown in Table 1.

The complex permittivity of materials is represented by the Debye model [10, 11] as given by

$$\varepsilon_r(\omega) = \varepsilon_\infty + \frac{\varepsilon_s - \varepsilon_\infty}{1 + j\omega\tau} - j\frac{\sigma_s}{\omega\varepsilon_0} \tag{1}$$

where ε_s is the value of permittivity that would be observed at low frequencies, ε_∞ the value for 'infinite' frequency, τ is the relaxation time and σ is the conductivity.

In the measurement, error sources were derived from the network analyzer error sources and the dielectric error sources. The network error sources are noise (about 0.0006) and the fixed load/probe directivity contribution (0.05–0.15) for the standard calibration (air, short and load), depending on frequency.

4 Result and Discussion

There were two samples that have been measured; sample A is made of the ingredients namely Glyoxal and sample B is made of Glutaraldehyde. The measured results of the two fat tissue samples are presented in the Figs. 1 and 2 over the ultra-wideband frequencies. Figure 1 shows the dielectric constant of the sample with glyoxal and glutaraldehyde while the measured results for conductivity of the samples are shown in Fig. 2. Figure 1 shows the dielectric constant of the two samples as a comparison which shows that the sample with glyoxal having a lower dielectric constant or relative permittivity compared to the glutaraldehyde.

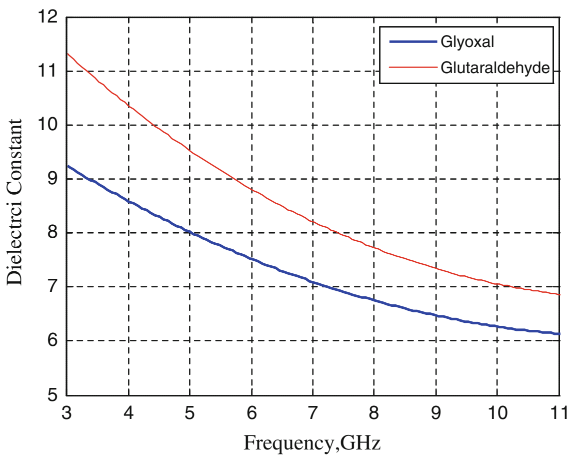


Fig. 1 Dielectric constant versus frequencies for the fat tissues fabricated using two different cross linking agent

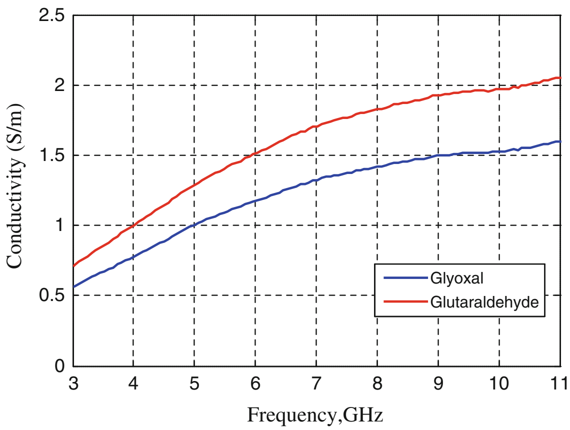


Fig. 2 Conductivity versus frequencies of the fat tissue fabricated using using two different cross linking agent

In [11], the dispersive characteristic in real biological tissue has been reported by using experimental data for normal and malignant breast tissue. Using this data the real normal tissue is represented by the Debye Model and the single pole Debye parameter used in the modeling of the real tissue in this paper is shown in Table 1 [12]. As shown in Fig. 3, the fat tissue samples for different cross linking agent are measured and compared with the real tissue. The proposed sample tissue shows a higher dielectric constant and thus this fat tissue sample can imitate for a fat tissue that has high water content to be used in breast phantom fabrication for microwave imaging application. At the centre frequency, 7.5 GHz it is measured that the dielectric constant of the sample using glyoxal and glutaraldehyde are 7.2 and 8.0 respectively while the real tissue having a dielectric constant value of 6.9. The cross linking agent material; formaldehyde shows a low dielectric constant properties and

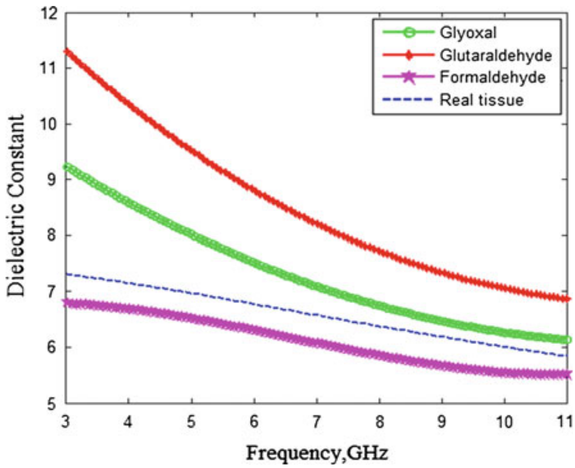


Fig. 3 Dielectric constant versus frequencies for the fat tissues fabricated using two different cross linking agent and real tissue

this is lower than the real tissue. Glyoxal and glutaraldehyde can be a good potential cross linking agent replacing formaldehyde and the measured dielectric constant has shown to give a small difference of 4 and 14.3 % compared to the real tissue taken at centre frequency. By having a wide range of dielectric constant of the fat tissue, thus the breast tissue fabrication using the two cross linking agent can mimics real fat tissue having a high water content(high relative permittivity) and low water content(low relative permittivity) tissue. Figure 4 shows the fabricated sample with glutaraldehyde before the dielectric constant measurement is taken.

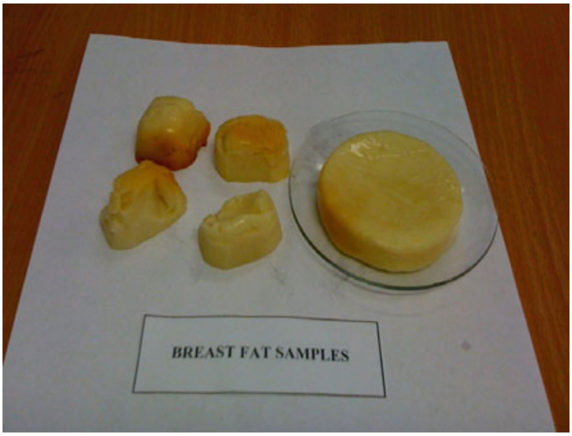


Fig. 4 Fabricated breast fat sample using glutaraldehyde as cross linking agent

5 Conclusion

The paper has reported the measured results of the fat tissue samples using two different cross linking agents. The result also has compared to the previously fabricated sample using formaldehyde which is unfavorable chemical due to health issue in handling phantom for microwave imaging application. The samples constituting the fabricated fat tissue samples have indicated a good emulation of the actual breast fat. For the breast phantom fabrication, the fabricated fat tissue samples is non-toxic and can be embedded in a breast phantom together with other tissue sample such as glandular and tumor tissue to mimics real breast. But undeniably that heterogeneity of the real breast tissue still poses a challenge for accurately representing it in physical breast phantom.

Acknowledgement The authors acknowledge the financial support from Universiti Teknologi MARA, Penang Campus. Malaysia and Ministry of Education Malaysia under Grant No: 600-RMI/RAGS/2013/UITM/TK02/8 Entitled "Time Domain Microwave Imaging in Target Localization".

References

1. Fear, E.C., Hagness, S.C., Meaney, P.M., Okoniewski, M., Stuchly, M.A.: Enhancing breast tumor detection with near-field imaging. IEEE Microwave Mag. **3**, 48–56 (2002)
2. Surowiec, A.J., Stuchly, S.S., Barr, J.R., Swarup, A.: Dielectric properties of breast carcinoma and the surrounding tissues. IEEE Trans. Biomed. Eng. **35**, 257–263 (1988)
3. Campbell, A., Land, D.: Dielectric properties of female human breast tissue measured in vitro at 3.2 GHz. Phys. Med. Biol. **37**, 193 (1992)
4. Sill, J.M., Fear, E.C.: Tissue sensing adaptive radar for breast cancer detection—experimental investigation of simple tumor models. IEEE Trans. Microwave Theory Tech. **53**, 3312–3319 (2005)
5. Klemm,M., Craddock, I.J., Leendertz, J.A., Preece, A., Gibbins, D.R., Shere, M., Benjamin, R.: Clinical trials of a UWB imaging radar for breast cancer. In: Proceedings of the Fourth European Conference on Antennas and propagation (EuCAP), pp. 1–4 (2010)
6. Byrne, C., Schairer, C., Wolfe, J., Parekh, N.: Mammographic features and breast cancer risk: effects with time, age, and menopause status. J. Nat. Cancer Inst. **87**(21), 1622–1629
7. Kyber, J., Hangsen, H., Piquett, F.: Dielectric properties of biological tissue at low temperatures demonstrated on fatty tissue. Phys. Med. Biol. **37**, 1675–1688 (1992)
8. Lazebnik, M., Popovic, D., McCartney, L., Watkins, C.B., Lindstrom, M.J., Harter, J., Sewall, S., Ogilvie, T., Magliocco, A., Breslin, T.M.: A large-scale study of the ultrawideband microwave dielectric properties of normal, benign and malignant breast tissues obtained from cancer surgeries. Phys. Med. Biol. **52**, 6093–6116 (2007)
9. Bakar, A.A., Abbosh, A., Sharpe, P., Bialkowski, M.: Artificial breast phantom for microwave imaging modality. In: IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES), pp. 385–388 (2010)
10. Gabriel, S., Lau, R., Gabriel, C.: The dielectric properties of biological tissues: III. Parametric models for the dielectric spectrum of tissues. Phys. Med. Biol. **41**, 2271–2293 (1996)
11. Lazebnik, M., Okoniewski, M., Booske, J.H., Hagness, S.C.: Highly accurate Debye models for normal and malignant breast tissue dielectric properties at microwave frequencies. IEEE Microwave Wireless Compon. Lett. **17**(12), 822–824 (2007)

12. Zastrow, E., Davis, S.K., Lazebnik, M., Kelcz, F., Van Veen, B.D., Hagness, S.C.: Development of anatomically realistic numerical breast phantoms with accurate dielectric properties for modeling microwave interactions with the human breast. IEEE Trans. Biomed. Eng. **55**(12), 2792–2800 (2008)

Transformation Polynomial Secret Sharing

Tanyaporn Sridokmai, Somchai Prakancharoen and Nalinpat Porrawatpreyakorn

Abstract This research objective is to introduce how the secret sharing algorithm was used to increase the security of your data before distributing to the database. The first arbitrated polynomial function was transformed to be another one in order to enhance more security. Plaintext was reconstructed by reversing two stages interpolation. This algorithm improves more security than ordinary share secret sharing algorithm.

Keywords Secret sharing · Transformations of functions · Interpolation · Algebraic functions

1 Introduction

This research aims to secure the data in distributed databases, because the distributed database is extremely important in many organizations, both public and private sectors such as hospitals, banks or university in order to prevent frauds or destruction of hackers, The purpose of distributed database security is deal with protecting data from deliberately attacks. Current implement for providing secure data storage is to ordinary encrypt the data using an encryption key. If the encryption key is risky thus an unauthorized individual possibly access the data.

T. Sridokmai · N. Porrawatpreyakorn
Faculty of Information Technology, King Mongkut's University of Technology
North Bangkok, Bangkok, Thailand
e-mail: Tanyaporn.Sridokmai@Springer.com; LNCS@Springer.com

N. Porrawatpreyakorn
e-mail: Nalinpat.Porrawatpreyakorn@Springer.com; LNCS@Springer.com

S. Prakancharoen (✉)
Faculty of Applied Science, King Mongkut's University of Technology
North Bangkok, Bangkok, Thailand
e-mail: Somchai.Prakancharoen@Springer.com; LNCS@Springer.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_45

This problem can be solved by split plaintext to multiple parts by secret sharing algorithm. In order to reconstruct the original data, typically more than one of the splitted parts must be retrieved.

Secret sharing refers to techniques for distributing a secret amongst a group of participants, each of whom is allocated with a share of the secret.

Secret Sharing Scheme (SSS) is the method of isolating the secret to shares and distributing them to a group of participants. The collection (set) of groups that are able to reconstruct the secret is termed access structure and its elements are called authorized groups. Secret sharing, which was individually introduced by Blakley [1] and Shamir [2], deals with procedures of distributing secret value among a set of contributors, in such a way that only some qualified secret shares can recover the secret value (ordinary plaintext). The paper objective is to enhance security of SSS by transform arbitrary polynomial equation to a distinct one.

2 Secret Sharing Based Threshold Scheme

In secret sharing [3], the essential thing is the number of participants required to reconstruct the secret. If the number of authorized participants k out of n, we call this scheme (k, n) -threshold access structure and defined as:

$$F = \{A \in 2^p \| A| \geq K\}, \quad 2 \leq K \leq n \tag{1}$$

where F is a method which the dealer can use to distribute shares to each participant any access structure $F \subseteq 2^p$, p is the set of trustees and 2^p is the power set of p satisfies the following condition, A an arbitrary set of participants such that $|A| \geq K$.: if $A \in F$ and $A \subset A' \subset p$ then $A' \in F$, K is the least number of retrieved participants and n is the number of total participants in the system.

2.1 Shamir's Threshold Secret Sharing Scheme [2]

Shamir's secret sharing scheme is a threshold scheme based on polynomial interpolation. It allows a dealer D to distribute a secret value s to n users, such that at least $k < n$ users share are required to reconstruct the secret. The protocol is information theoretically secure, i.e., any fewer than k users cannot gain any information about the secret by themselves. To share the secret s among users $P_1, P_2, \ldots, P_n$ such that k users are required to reconstruct the secret. Dealer D creates a random polynomial $f(x)$ of degree $k - 1$ and constant term s.

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{k-1} x^{k-1} \tag{2}$$

This polynomial is constructed over a finite field, such that the coefficient a_0 is the secret s and all other coefficients are random elements in the field; the field is known to all authorized participants.

Dealer D publicly chooses n random distinct evaluation points: $X_j \neq 0$, and secretly distributes to each user P_j the share $share_j(s) = (X_j, f(X_j)), j = 1\ldots n$. (Remark: The evaluation point X_j could be any publicly known value, therefore for our convenience, we assume $X_j = j$, hence the shares are denoted as) [4]

$$f(1), \ldots, f(j), \ldots, f(n) \tag{3}$$

To reconstruct the secret from each k shares out of n shares, without loss of original plaintext, we have to retrieve at least k shares: $f(1), f(2)\ldots, f(k)$.

By Lagrange interpolation [5], the k participants function value (Range) can be used to recomputed the coefficients a_0 of polynomial equation then we can recall the original plaintext S. The correctness and security conditions are also maintained.

Limitations of the scheme are computationally hard and become impractical with large number of shares K, stack overflow despite Shamir's shares are only pairs (x, y), their sizes increased with the power K.

3 Theory of Graph and Transformations

3.1 The Graph of Element Function

Element functions of graph are algebraic function: The algebraic functions are formed by applying algebraic operations to the identity function

$$y = f(x). \tag{4}$$

3.2 Types of Algebraic Functions

The general polynomial function as shown on Eq. (5)

$$f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_{k-1}x^{k-1} \tag{5}$$

Linear Function: A linear function is a function defined by an equation of the form:

$$y = a_1x + a_0 \tag{6}$$

Table 1 Transformation rules for functions [6, 7]

Equation	To obtain graph
$y = f(x) + k(k > 0)$	Shift graph $y = f(x)$ up k units
$y = f(x) - k(k > 0)$	Shift graph $y = f(x)$ down k units
$y = f(x - h)(h > 0)$	Shift graph $y = f(x)$ right h units
$y = f(x + h)(h > 0)$	Shift graph $y = f(x)$ left h units
$y = -f(x)(c > 0)$	Reflect graph $y = f(x)$ over x -axis
$y = f(x)(c > 0)$	Reflect graph $y = f(x)$ over y -axis
$y = af(x)(a > 1)$	Stretch graph $y = f(x)$ vertically by factor of a
$y = af(x)(0 < a < 1)$	Shrink graph $y = f(x)$ vertically by factor of a
(Multiply y—coordinates of $y = f(x)$ by a)	
$y = f(cx)(c > 1)$	Shrink graph $y = f(x)$ horizontally by factor of c
$y = f(cx)(0 < c < 1)$	Stretch graph $y = f(x)$ horizontally by factor of c
(Divide x—coordinates $y = f(x)$ of by c)	

Cubic Function: A function defined by a polynomial of degree 3. The general form of cubic function is

$$y = a_3x^3 + a_2x^2 + a_1 + a_0 \tag{7}$$

3.3 *The Transformations of Graph Algebraic Function*

In this section we will discuss how the graph of a function may be transformed either by shifting, stretching or compressing, or reflection. The form of the quadratic function in the equation is called vertex form, so named because the form easily reveals the vertex or "turning point" of the parabola [6, 7]. Each of the constants in the vertex that form of the quadratic function plays a role. As you will soon see, the constant a controls the scaling (stretching or compressing of the parabola), [8] the constant h controls a horizontal shift and placement of the axis of symmetry, and the constant k controls the vertical shift. [9] (Table 1)

4 Design of Transformation Polynomial Secret Sharing Algorithm

The whole sequence of practeal solution was shown in Fig. 1.

The Construction of Secret Part.

Step 1: Input- secrete key number of participant
Step 2: Select random values $a_1, a_2..., a_n$
Step 3: Generate polynomial string to share the secret into parts.

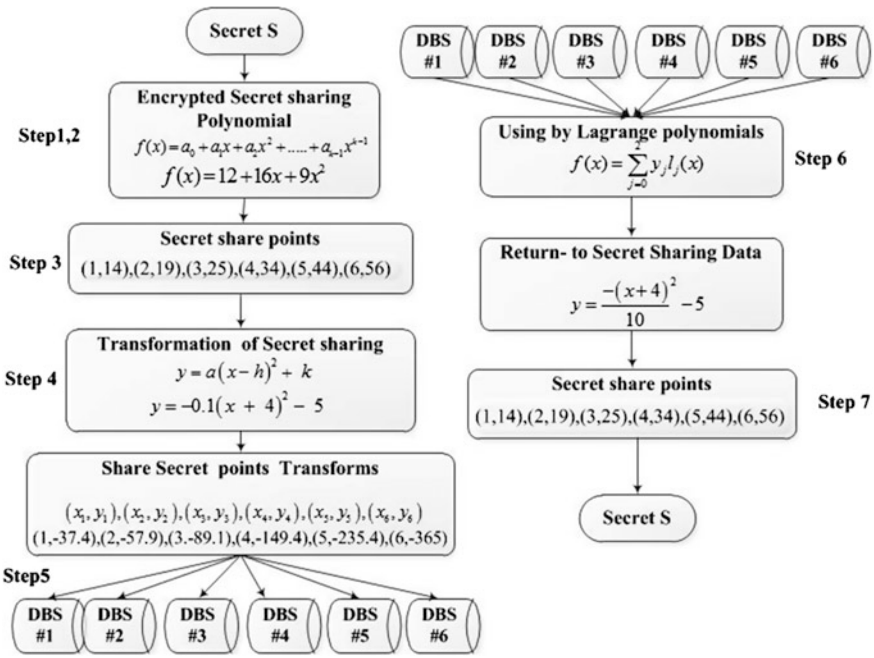


Fig. 1 Transformation secret sharing to database severs (DBS_i)

Transformations of Polynomial Function

Step 4: Transformations of polynomial graph function
Step 5: Sent Transformation Polynomial Secret Sharing to Database

The Reconstruction

Step 6: From Database Server to Secret sharing Data
Step 7: Use by Lagrange Polynomials Return to Secret Sharing

4.1 The Construction of Secret Part (Step1–4)

First, the Data owner (D) has to define the original plaintext (S) which would be assigned as a_0 of the arbitrary polynomial equation. Second, this polynomial function $(f(x))$ will be calculated of $f(x)$ on y_i far each value of x_i these share S_i, on $(x_i, f(x_i))$, were then represent as secret share part of original secret S.

Step 1: D defines $a_0 = s$
Step 2: D select $t - 1$ random, independent coefficients $a_1, \ldots, a_{t-1}, 0 \le a_j \le p - 1$ Defining the random polynomial Z_p,

$$f(x) = \sum_{j=0}^{t-1} a_j \cdot x^j. \tag{8}$$

Step 3: D publicly chooses n random distinct evaluation points: $X_j \neq 0$, and secretly distributes to each users P_j the share

$$share_j(s) = \left(X_j, f\left(X_j\right)\right), \quad j = 1 \ldots n. \tag{9}$$

Example 2 Polynomial Secret Sharing

Let S = 12, n = 6, a_1 = 16, a_2 = 9, k = 3.

Secret share points

$$\begin{aligned} &f(x) = 12 + 16x + 9x^2 \\ &(1, 14), (2, 19), (3, 25), (4, 34), (5, 44), (6, 56) \end{aligned} \tag{10}$$

For each participant a different single point of share $j(S)$; $j = 1, 6$ (both x and $f(x)$).

Let us consider (2, 19), (4, 34), (5, 44).

$(x_0, y_0) = (2, 19), (x_1, y_1) = (4, 34), (x_2, y_2) = (5, 44)$. The graph was illustrated as Fig. 2

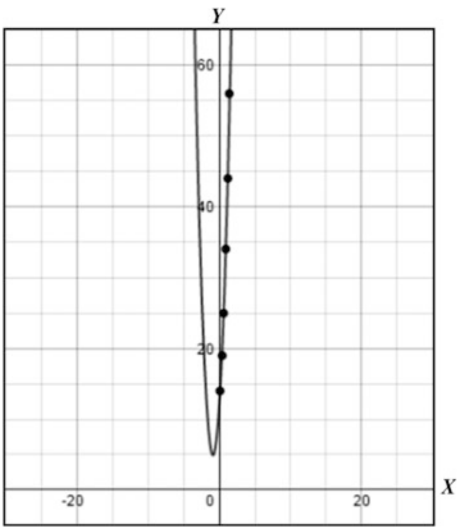


Fig. 2 Graph of polynomial secret sharing (base on Eq. 10)

4.2 Transformations of Defined Polynomial Function (Step4)

The Data owner (D) has to define another polynomial function with has the same order. Former polynomial result $(f(x_i), x_i)$ were then used as in put of new polynomial function then used as input of new value of secret share $(f'(x_i), x_i')$

Vertical scaling and reflection along with horizontal and vertical translations for the graph can all happen in one equation using a, h and k then

$$y = f(x) \tag{11}$$

$$y = x^2 \tag{12}$$

And our equation that contains the scaling's, reflections and translations looks like this

$$y = a(x - h)^2 + k \tag{13}$$

Example 3 Polynomial Secret Sharing and Transformation

Therefore, if $a = -0.1$ and $h = -4$ and $k = -5$, we say that the allusion parabola is reflected across the x-axis and vertically scaled by a factor of 0.1 and horizontally translated −4 units and vertically translated −5 units. Our equation for this would appear as (14): [9]

$$y = -0.1(x + 4)^2 - 5$$

Or, with a bit of simplification:

$$\begin{aligned} y &= (-0.1)(x + 4)^2 - 5 \\ y &= \frac{-(x + 4)^2}{10} - 5 \end{aligned} \tag{14}$$

Hence, the secret share of polynomial transformations are (1, −37.4), (2, −57.9), (3. −89.1), (4, −149.4), (5, −235.4) and (6, −365).

Transformations of polynomial graph function as shown in Fig. 3.

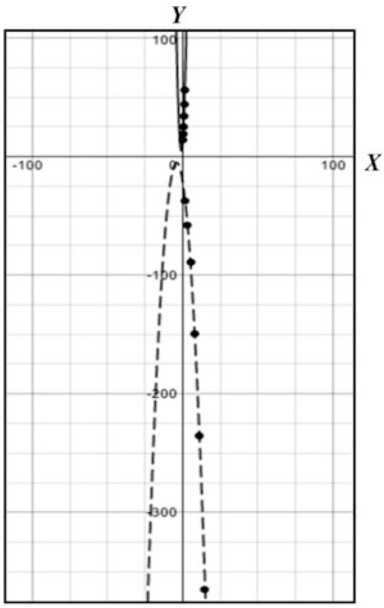


Fig. 3 Graph of polynomial secret sharing and transformation graph (based on Eqs. 10 and 14)

4.3 Sent Transformation Polynomial Secret Sharing to Database Servers (Step 5)

After 4.2, these secret shares should be sent to be kept in distrusted database engines (DB_i). Data base administrator could not understand it since it was only secret share under double polynomial secret sharing algorithm performing.

4.4 The Secret Reconstruction by Lagrange Polynomials and Return to Secret Sharing (Step 6)

First (Step 6), in situation that *D* would like to retrieve the original secret *S*, *D* has to identify ID of the target tuple. This ID will be sent to all *DBS* then *DBS* has to perform file retrieval then sent found tuple back to *D*

Second, sent back tuples were interpolated by Lagrange interpolation which should give a transform polynomial function.

Third (Step 7), transform polynomial function were used to compute of former polynomial with Lagrange interpolation then the value a_0 form this equation was represent as original plaintext under more confidence than ordinary *SSS*.

To reconstruct the secret from each subset of *k* shares out of *n* shares, without loss of generality we will mark this subset: $f(1), \ldots, f(k)$ [4].

Use Lagrange interpolation to find the unique polynomial $f(x)$ such that degree of $f(X) < t$ and $f(j) = share_j(s)$ for $j = 1, 2, \ldots, t$.

Follow to position of the original equation.
The three transformation of secrets point were

$$(x_0, y_0) = (2, -37.4), (x_1, y_1) = (4, -149.4), (x_2, y_2) = (5, -235.4)$$

We will compute Lagrange basis polynomials:

$$\begin{aligned}
l_0 &= \frac{x - x_1}{x_0 - x_1} * \frac{x - x_2}{x_0 - x_2} = \frac{x - 4}{2 - 4} * \frac{x - 5}{2 - 5} = \frac{1}{6x^2} - \frac{11}{2x} + \frac{31}{3} \\
l_1 &= \frac{x - x_0}{x_1 - x_0} * \frac{x - x_2}{x_1 - x_2} = \frac{x - 2}{4 - 2} * \frac{x - 5}{4 - 5} = \frac{-1}{2x^2} - \frac{31}{2x - 5} \\
l_2 &= \frac{x - x_0}{x_2 - x_0} * \frac{x - x_2}{x_2 - x_1} = \frac{x - 2}{5 - 2} * \frac{x - 4}{5 - 4} = \frac{1}{3x^2} - 2x + \frac{22}{3} \\
&= -37.4\left(\frac{1}{6x^2} - \frac{11}{2x} + \frac{31}{3}\right) + (-149.4)\left(\frac{-1}{2x^2} - \frac{31}{2x} - 5\right) \\
&\quad + (-149.4)\left(\frac{1}{3x^2} - 2x + \frac{22}{3}\right)
\end{aligned} \tag{15}$$

And return to transformation polynomial secret sharing

$$y = \frac{-(x + 4)^2}{10} - 5 \tag{16}$$

The Secret sharing 3 points after return to Polynomial Secret Sharing

$$\begin{aligned}
&(2, 19),\ (4, 34),\ (5, 44) \\
&(x_0, y_0) = (2, 19), (x_1, y_1) = (4, 34), (x_2, y_2) = (5, 44)
\end{aligned}$$

We will compute Lagrange basis polynomials:

$$\begin{aligned}
l_0 &= \frac{x - x_1}{x_0 - x_1} * \frac{x - x_2}{x_0 - x_2} = \frac{x - 4}{2 - 4} * \frac{x - 5}{2 - 5} = \frac{1}{6x^2} - \frac{11}{2x} + \frac{31}{3} \\
l_1 &= \frac{x - x_0}{x_1 - x_0} * \frac{x - x_2}{x_1 - x_2} = \frac{x - 2}{4 - 2} * \frac{x - 5}{4 - 5} = \frac{-1}{2x^2} - \frac{31}{2x - 5} \\
l_2 &= \frac{x - x_0}{x_2 - x_0} * \frac{x - x_2}{x_2 - x_1} = \frac{x - 2}{5 - 2} * \frac{x - 4}{5 - 4} = \frac{1}{3x^2} - 2x + \frac{22}{3} \\
f(x) &= \sum_{j=0}^{2} y_j l_j(x) \\
&= 19\left(\frac{1}{6x^2} - \frac{11}{2x} + \frac{31}{3}\right) + 34\left(\frac{-1}{2x^2} - \frac{31}{2x} - 5\right) \\
&\quad + 44\left(\frac{1}{3x^2} - 2x + \frac{22}{3}\right)
\end{aligned} \tag{17}$$

We could recover original polynomial equation with '12' as an original

$$f(x) = 12 + 16x + 9x^2 \tag{18}$$

5 Performance and Efficient Analysis [10]

In this part analyzes this scheme's performance and efficiency by comparing it with the existing schemes. In order to compare with the existing project, we classify the existing scheme based on the distribution mechanism of secret share. There is a secure channel in the Shamir scheme [2] Table 2 shows a simple comparison between them.

Following from Table 2:

1. Secrecy: Shamir scheme construct secret shares from an original data based on arbitrary defined polynomial equation. These secrets shares should be used to reconstruct original data. For our scheme, the polynomial equation was increase more secrecy by perform a transformation it to the new one. This technique could increase secrecy level more than ordinary Shamir scheme.
2. Extensible: When polynomial is kept fixed, pieces can be dynamically added or deleted without affecting the other pieces. Two polynomial equations need not to produce the same amount of secret shares
3. Reliability: at least a number of users, greater than the threshold, are required for reconstruction the original data.
4. Flexible: we can supply each participant different number of pieces according to administration defined.
5. Minimal: The size of each piece does not exceed the size of the original data.

In time aspect, as shown in Fig. 4, our scheme consumes the highest time for splitting and recovering data file. Compared with Shamir's scheme, because it has attach, Transformations of Polynomial Function in order to higher security.

Table 2 Comparision of other schemes

Scheme	Secrecy	Extensible	Reliability	Flexible	Minimal	Performance
Shamir scheme [2]	Yes	Yes	Yes	Yes	No	Low
Our scheme	More	More	Yes	Yes	No	A little bit more

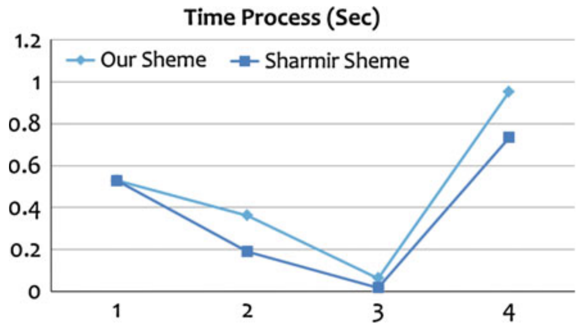


Fig. 4 Split and recover time for two schemes. *1* The construction of secret part. *2* Sent secret sharing to database. *3* From database server to secret sharing data. *4* The reconstruction part

6 Summary and Future

This paper presents a secret sharing scheme enhancement by transformation of first defined polynomial function to the second polynomial. The disadvantage of this method was encryption activity that the plaintext must be encrypted and decrypts all the time there were referenced. It should be good if we can performance file manipulation without decryption on it. Homomorphic algorithm should be considered as the next research. Cryptographer should use Homomorphic algorithm to perform encryption on plaintext before and after secret sharing without any encrypt the original plaintext.

References

1. Blakley, G.R.: Safeguarding cryptographic keys, p. 313. International Workshop on Managing Requirements Knowledge. IEEE Computer Society (1899)
2. Shamir, A.: How to share a secret. Commun. ACM **22**(11), 612–613 (1979)
3. Al-Ssulami, A.M.: A novel threshold secret sharing scheme using FFT algorithm. Int. J. Inf. Secur. Sci. **2**, 19–27 (2013)
4. Dahlia M.: Secret sharing, Lecture in an advance course in computer and network security. The Hebrew University of Jerusalem (2004)
5. Weisstein, E.W., Branden, A.: Lagrange interpolating polynomial—from Wolfram MathWorld, http://mathworld.wolfram.com/LagrangeInterpolatingPolynomial.html
6. Zobel, E.A.: Function transformations | zona land education. http://zonalandeducation.com/mmts/functionInstitute/functionTransformations/functionTransformations.html
7. Jones, J.: Shifting, reflecting, and stretching graphs. https://people.richland.edu/james/lecture/m116/functions/translations.html
8. Wortman, K.: Graph transformations. In: Math 1050. http://www.math.utah.edu/~wortman/1050/

9. Johnson, J.: Parent functions and transformations—she loves math. http://www.shelovesmath.com/algebra/advanced-algebra/parent-graphs-and-transformations/
10. Senanayake, K.: What I'm learning: Adi Shamir's secret sharing scheme. http://mytechnicaldocs.blogspot.jp/2011/09/how-to-share-secret.html

Ant Colony-Dijkstra's Algorithm for Evacuation Preparedness in High Rise Buildings

Nor Amalina Mohd Sabri, Abd Samad Hasan Basari, Burairah Hussin, Khyrina Airin Fariza Abu Samah and Yuseni Wahab

Abstract The critical tasks during evacuation process is how to find the right ways in order to escape from the danger place to a safe place. In process of finding the right ways, most of the evacuees are panicked. Subsequently, make the process more difficult. With that occurrence, the main objectives of this research study are to identify the suitable shortest path algorithm for evacuation in high rise building, then design and develop an evacuation route via shortest path algorithm in order to obtain an exit route to evacuate by using Optimization and Artificial Intelligence Technique. The objectives that involved are to help the evacuees to find the best routes during evacuation process. Six phases of methods are raised to accomplish the objectives by utilizing the Dijkstra and Ant Colony Optimization Algorithm. The first step is started from the original building layout. Then transform the layout into 2D layout plan. After that, import the matrix data to generate graph theory. Next step is utilizing the both approaches to achieve the shortest path. The preliminary result has shown positive result which can deliver the shortest path to help evacuees.

N.A.M. Sabri · A.S.H. Basari (✉) · B. Hussin · K.A.F.A. Samah
Faculty of Information and Communication Technology,
Centre for Advanced Computing Technology (C-ACT),
Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya,
76100 Durian Tunggal, Melaka, Malaysia
e-mail: abdsamad@utem.edu.my

N.A.M. Sabri
e-mail: noramalinamohdsabri@gmail.com

B. Hussin
e-mail: burairah@utem.edu.my

K.A.F.A. Samah
e-mail: rinarinza@gmail.com

Y. Wahab
Faculty of Business Innovation and Accounting, Kolej Universiti Islam Melaka (KUIM),
Kuala Sungai Baru, 78200 Masjid Tanah, Melaka, Malaysia
e-mail: yuseni@kuim.edu.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_46

Keywords Ant colony optimization · Dijkstra's algorithm · Evacuation · Shortest path

1 Introduction

Evacuation is a process of moving people from danger place to safe place during an emergency. Unexpected event occurs such as natural disaster may cause an emergency situation which brings damage and losses. Though, evacuation planning needs to be prepare to maximize the evacuation effectiveness [1] and also minimize the property losses [2].

The main problems in evacuation process are the difficulties to find the right route in order to escape towards safe place [3] and the behaviour of evacuees during the evacuation especially in indoor place such as high rise building [4]. Moreover, the building itself become more complex with the high population density [5] effect the time of evacuation [6].

An evacuation planning is needed to carry out, in order to improve the efficiency of evacuation process. During emergency situation, evacuees mostly panicked hard to make a decision to find the best path to escape especially the evacuation take place in the closed building. Moreover, the building tend to be complex in structure and design subsequently effect the process of finding the best path and trigger various behaviour of evacuees during the critical situation [7]. Planning the evacuation is important requirement to evacuate people safely when emergency occurs [8] and the planning can ensure the evacuation process done quickly and carefully [9].

A good evacuation management is able to catch the attention of evacuees in guiding them toward exit route quickly and help them make a good decision. Hence, this research design and develop the evacuation preparedness model via shortest path algorithm and Artificial Intelligence Technique on exit route selection to evacuate.

This research is expecting to produce independent evacuation preparedness algorithm for exit route and also guide evacuees escape safely from the building. The evacuee especially those who are unfamiliar with the building can find the best path by following the guide provided and make the evacuation process more efficient by reducing the evacuation distance and time with the improvement of evacuees' behaviour.

The contribution of this research study are guide evacuees to find the best path easy and smooth in a safest way and in consequences reduce the fatalities and injuries of evacuees during the evacuation.

2 Research Methodology

This research methodology utilise Dijkstra's Algorithm and Ant Colony Optimization in finding the best path in the evacuation process in a high rise building.

Dijkstra's Algorithm efficient in generating the shortest path for selection of route and claim to be the best techniques in solving the simple shortest path problem by providing the shortest path from any evacuation node [10, 11]. This algorithm also able to provide safe evacuation plan [12, 13, 14] and solve single source shortest path problem in a graph search algorithm [15, 16] to find the shortest distance between a node and all other nodes to suit the target of Dijkstra's Algorithm [17, 18].

The other technique is Ant Colony Optimization (ACO) Algorithm which is the Artificial Intelligence technique. This algorithm is proved to be the alternative for Dijkstra's Algorithm in solving the problem of finding the shortest path. The ACO algorithm is represents the way of real ants optimize their route while searching for a food. Ants are able to find the shortest pathways from their nest to food place by exchanging information among them via pheromones [19]. Based on the presence and concentration of pheromones, the ants guide their direction and tend to move to high concentration of pheromones to find the optimal path while looking the food [20]. In adapting the ants with the evacuation process, the food is representing the exit node while the ant nest represents the current location of the evacuees.

This implementation of Dijkstra's Algorithm and Ant Colony Optimization involves 6 steps as shown in figure below (Fig. 1. Implementation of Dijkstra's Algorithm and Ant Colony Optimization Steps).

In Fig. 1 the first step is derived from building floor layout. Then convert into 2D floor layout in order to design the evacuation preparedness model by setting up each

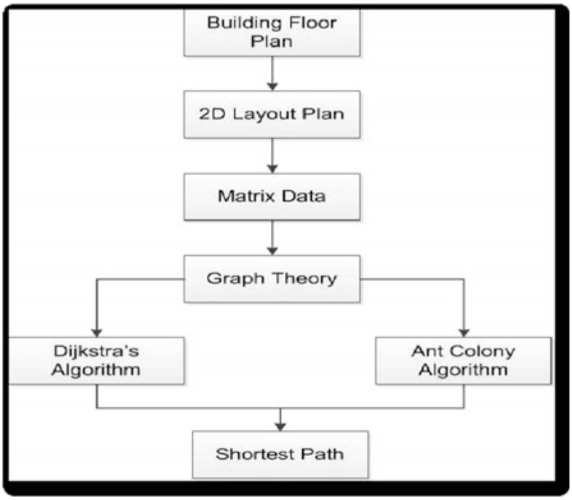


Fig. 1 Implementation of Dijkstra's algorithm and ant colony optimization steps

door in the floor as a node. The distance from each node are also obtain from the 2D layout plan and are collected into a matrix table as a matrix data format to generate the graph theory. Finally, utilize the two selected algorithm.

3 Result and Discussion

Simulation program using MATLAB has been built to find the shortest path using Dijkstra's Algorithm and Ant Colony Optimization. Dijkstra's Algorithm calculate the shortest path using the given distance based on the real building layout, the total distance is from the source node to destination node [21]. Meanwhile, the Ant Colony Optimization obtains the shortest path from the coordinate of each node in the building layout. From the coordinate of node, the ACO calculate the total distance which is the shortest. The result are compared and proved to be the same, it is means that ACO algorithm can be the alternative technique of Dijkstra's Algorithm in finding the shortest path for evacuation.

According to the 2D layout plan Fig. 2, there are a few nodes for door and staircase. Through the simulation model, the shortest path is obtained between nodes 38 and 47.

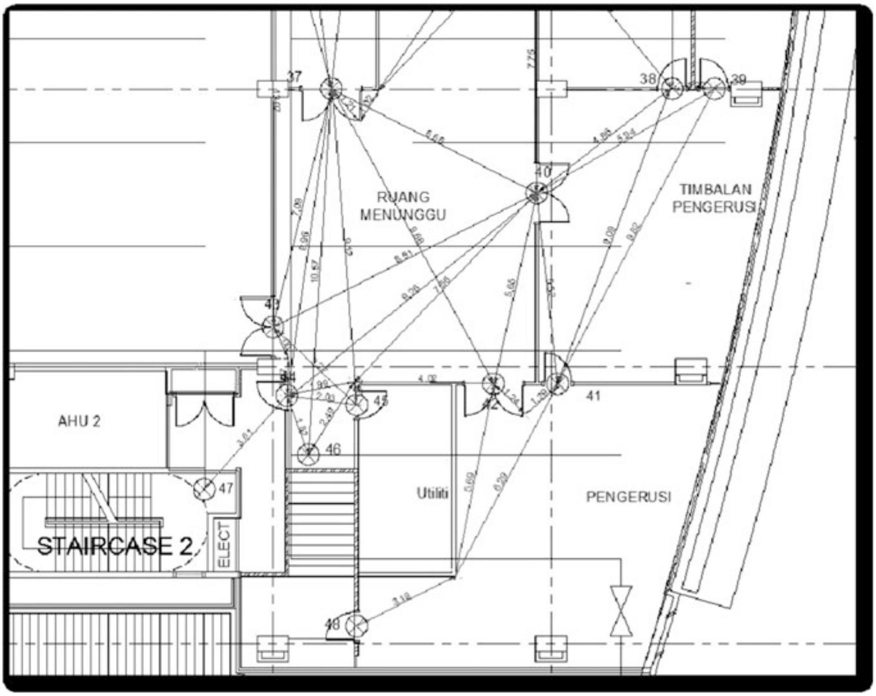


Fig. 2 2D layout plan with nodes and path between nodes along with distance

The first calculation is based on Dijkstra's Algorithm (DA), the starting step is to collect the number of nodes in the 2D layout plan above into the matrix table. The matrix table as shown in Table 1 consist of the node and the weight or the distance between the connected nodes.

Based on this matrix table, the calculation is done by the Dijkstra's Algorithm (DA) to obtain the shortest path between the nodes 38 to 47. However, before the calculation is performed the matrix table are read as sparse format and the node are renamed start from one in ascending order. The sparse format is in Table 2.

There are 10 nodes and 27 edges created based on the data imported. The calculated shortest path is between the nodes 1 to node 10 in the simulation model as the nodes are renamed (as shown in Table 3).

The result obtain by DA is shown in the graph (as shown in Fig. 3)

The path taken is through node 1 > 3 > 7 > 10 with total distance is 17.73 meter. This result is compared with the result of the other technique which is Artificial Intelligence technique, Ant Colony Optimization (ACO). The result of shortest path using ACO can be referred in Fig. 4.

The result shows that the total distance is 17.26 in meter. The difference is due to the way of calculation between DA and ACO. ACO use coordinate of the node to calculate the shortest path based on the 2D layout plan (Fig. 2). Furthermore, the node are also different as the ACO rename the node into new node (Table 4).

The total of shortest distance and the direction of path are same with the result of Dijkstra's algorithm. The differences of the number of nodes can be summarized in the Table 4. From the result that has been obtain, we can prove that Ant Colony Optimization can be the alternative substitute of Dijkstra's Algorithm in finding the shortest path.

Table 1 Matrix table

Node	38	39	40	41	42	43	44	45	46	47
38	0	0	4.86	9.09	0	0	0	0	0	0
39	1.22	0	5.94	9.62	0	0	0	0	0	0
40	4.86	0	0	5.52	5.65	8.51	9.26	0	0	0
41	9.09	0	5.52	0	2.53	0	0	0	0	0
42	0	0	5.65	2.53	0	0	6.01	0	0	0
43	0	0	8.51	0	0	0	2.07	0	0	0
44	0	0	9.26	0	6.01	2.07	0	0	0	3.61
45	0	0	8.55	0	5.01	3.27	2.03	0	0	0
46	0	0	0	0	0	0	1.83	0	0	0
47	0	0	0	0	0	0	0	0	0	0

Table 2 sparse format

Node	Distance
(2,1)	1.2200
(3,1)	4.8600
(4,1)	9.0900
(1,3)	4.8600
(2,3)	5.9400
(4,3)	5.5200
(5,3)	5.6500
(5,4)	2.5300
(3,5)	5.6500
(4,5)	2.5300
(7,5)	6.0100
(8,5)	5.0100
(3,6)	8.5100
(7,6)	2.0700
(6,3)	8.5100
(7,3)	9.2600
(8,3)	8.5500
(1,4)	9.0900
(2,4)	9.6200
(3,4)	5.5200
(8,6)	3.2700
(3,7)	9.2600
(5,7)	6.0100
(6,7)	2.0700
(8,7)	2.0300
(9,7)	1.8300
(7,10)	3.6100

Table 3 Rename node DA

2D layout plan	DA
38	1
39	2
40	3
41	4
42	5
43	6
44	7
45	8
46	9
47	10

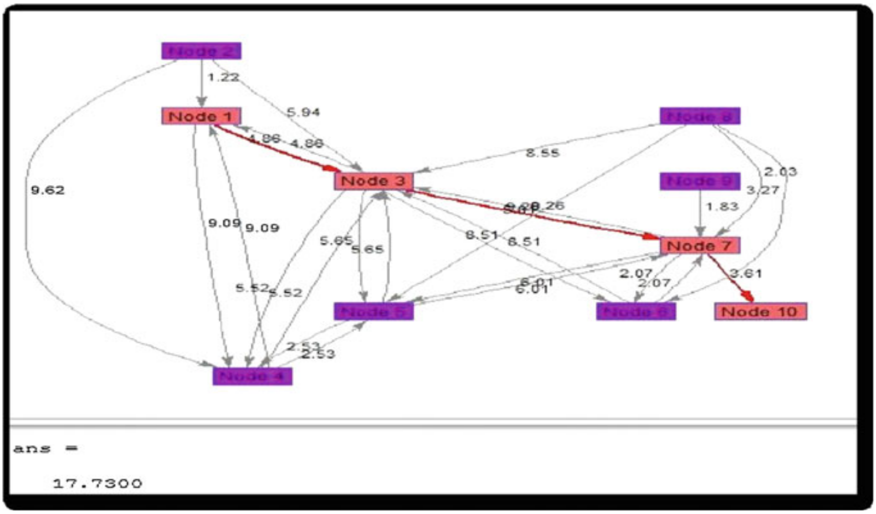


Fig. 3 Result DA in graph theory

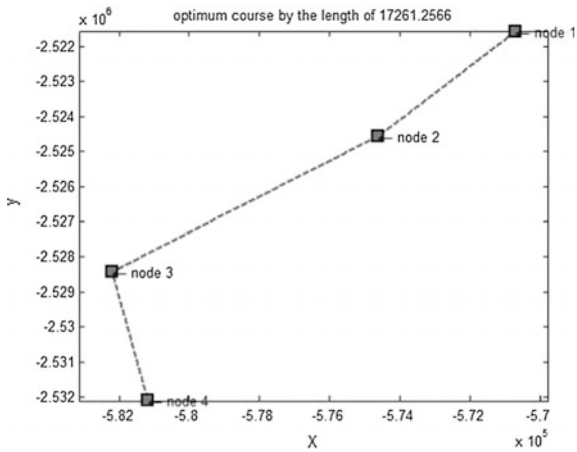


Fig. 4 Result of ant colony optimization

Table 4 Rename node ACO

2D layout plan	DA	ACO
38	1	1
40	3	2
44	7	3
47	10	4

4 Conclusion

In this paper, finding the shortest path using Dijkstra's Algorithm and Ant Colony Optimization Algorithm were discussed by using evacuation preparedness simulation model via MATLAB program. To obtain the result, six steps of method had been applied. The first step is converting original building layout into 2D layout plan, second is designing the 2D layout plan of building followed collecting the matrix data to generate graph theory. Then, calculate the shortest path using the two selected techniques which are Dijkstra's Algorithm and Ant Colony Optimization Algorithm. For future improvements is to obtain the safest path using Ant Colony Optimization.

Acknowledgements The authors would like to thank the Faculty of Information Technology and Communication (FTMK), Universiti Teknikal Malaysia Melaka (UTeM) for supporting this research. This research also part of Master by research in Information and Communication Technology and funded under Fundamental Research Grant Scheme (FRGS) number FRGS/1/2013/ICT07/02/2/F00162.

References

1. Li, Q., Fang, Z., Li, Q., Zong, X.: Multiobjective evacuation route assignment model based on genetic algorithm. In: 18th International Conference on Geoinformatics, pp. 1–5, June 2010
2. Wenxue, C., Zihui, Z.: Path selection for emergency transport vehicles. In: International Conference on Logistics System and Intelligent Management, pp. 1050–1054, Jan 2010
3. Goodwin, M., Granmo, O.: Ant colony optimisation for planning safe escape routes, pp. 53–62 (2013)
4. Khyrina, A.F.A.S., Burairah, H., Abd Samad, H.B.: A systems thinking in natural disaster management: evacuation preparedness, pp. 384–389 (2013)
5. Li, W.: A new risk assessment method of escape route based on Monte Carlo method and event tree to emergency evacuation. J. Inf. Comput. Sci. **11**(1), 219–227 (2014)
6. Jirasingha, W., Patvichaichod, S.: Modeling fire evacuation of a library building based on the numerical simulation. Am. J. Appl. Sci. **8**(5), 452–458 (2011)
7. Sabri, N.A.M., Basari, A.S.H., Husin, B., Khyrina, A.F.A.S.: Simulation method of shortest and safest path algorithm for evacuation in high rise building. Appl. Math. Sci. **8**(104), 5163–5172 (2014)
8. Fang, Z., Zong, X., Li, Q., Li, Q., Xiong, S.: Hierarchical multi-objective evacuation routing in stadium using ant colony optimization approach. J. Transp. Geogr. **19**(3), 443–451 (2011)
9. El-hamied, S.S.A., Saleh, A.A.E.: Survey on using GIS in evacuation planning process. Int. J. Comput. Sci. Inf. Secur. **10**(8), 40–43 (2012)
10. Wu, C.-H., Chen, L.-C.: 3D spatial information for fire-fighting search and rescue route analysis within buildings. Fire Saf. J. **48**, 21–29 (2012)
11. Lämmel, G., Grether, D., Nagel, K.: The representation and implementation of time-dependent inundation in large-scale microscopic evacuation simulations. Transp. Res. Part C Emerg. Technol. **18**(1), 84–98 (2010)
12. Kang, W., Xi, B.: A heuristic implementation of emergency traffic evacuation in urban areas, pp. 40–44 (2013)
13. Wang, T.., Huang, R., Li, L., Xu, W., Nie, J.: The application of the shortest path algorithm in the evacuation system. Appl. Shortest Path Algorithm Evacuation Syst. 250–253 (2011)

14. Zhang, H., Liu, H., Zhang, K., Wang, J.: Modeling of evacuations to no-notice event by public transit system, pp. 480–484 (2010)
15. Jaiswal, N., Chakrawarti, R.K.: Increasing no.of nodes for Dijkstra algorithm without degrading the performance. Int. J. Eng. Comput. Sci. **2**(3), 569–573 (2013)
16. Bu, F., Fang, H.: Shortest path algorithm within dynamic restricted searching area in city emergency rescue. In: 2010 IEEE Int. Conference on Emergency Management and Management Sciences, pp. 371–374, Aug 2010
17. Xu, Y., Wang, Z., Zheng, Q., Han, Z.: The application of Dijkstra's algorithm in the intelligent fire evacuation system. In: 4th International Conference on Intelligent Human-Machine Systems and Cybernetics, pp. 3–6, Aug 2012
18. Kaitao, C., Quanbo, G., Shenan, D.: Research and design for emergency evacuation instructions system based on wireless sensor network. In: 24th Chinese Control Decision Conference, pp. 3849–3854, May 2012
19. Zhou, Y.: Runtime analysis of an ant colony optimization algorithm for TSP instances. IEEE Trans. Evol. Comput. **13**(5), 1083–1092 (2009)
20. Qing-quan, C., Kun, Y.A.O.: Research on the model of pedestrian evacuation in large public building based on ant colony algorithm. International Conference on Automatic Control and Artificial Intelligence (ACAI 2012), pp. 1682–1685 (2012)
21. Sabri, N.A.M., Basari, A.S.H., Husin, B., Khyrina, A.F.A.S.: The utilisation of Dijkstra' s algorithm to assist evacuation route in higher and close building (2015)

All Optical Signal Restoration for 10G DPSK System

Bhagwan Das, M.F.L. Abdullah and Nor Shahida Mohd Shah

Abstract Ultrafast Kerr nonlinearity, offers various signal processing functions to be used. Several nonlinear transmission impairment such as phase noise, amplitude noise and power spectral efficiency loss become proficient to degrade the performance of optical systems, when intense power is launched in optical fiber. In modern communication system, all optical regeneration is one of the solution to mitigate transmission impairments instead of O-E-O conversion. In this paper, all optical regeneration is demonstrated for 10G DPSK system using 3R regeneration and phase sensitive amplification to mitigate amplitude and nonlinear phase noise form 10G noisy DPSK transmission system. Bit error rate of 10^{-12} is achieved at power penalty of 5 dBm. In developed system the novelty exist in compensating two noise simultaneously; ASE noise and phase noise using integrated algorithm developed using 3R and PSA. The system is developed and tested using commercial software package optisystem to check the feasibility of the system design in real time implementation. The developed all optical regeneration system is very demanding for long distance high-speed communication systems. The result conclude the positive feasibility for real time implementation of designed system.

Keywords Optical regeneration · DPSK system · Bit error rate · 3R regeneration · Phase sensitive amplifier

---

B. Das (✉) · M.F.L. Abdullah · N.S.M. Shah
Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia (UTHM), Parit Raja, 86400 Batu Pahat, Johar, Malaysia
e-mail: he130092@siswa.uthm.edu.my

M.F.L. Abdullah
e-mail: faiz@uthm.edu.my

N.S.M. Shah
e-mail: shahida@uthm.edu.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_47

1 Introduction

The current research focus, to increase the system range and stoutness for transmission systems [1]. Ultra-high speed communication is achieved using all-optical signal regeneration, which is one of the promising techniques. High power transmission inside the optical fiber produce several transmission impairments for example linear, nonlinear, and scattering etc. due to interaction of fiber material and refractive index [2]. The intensity of noise produce is based on level of interaction between signal and fiber material [3, 4]. Mainly these impairments eithers are amplitude or frequency/phase dependent [5]. Out of these impairments nonlinear phase noise is much more dominant than amplitude noise and handled distinctly [6]. Signal regeneration perform by electrically or all optically, both have their own importance in optical communication system. In modern optical system all-optical regeneration is widely used, because electronic regeneration requires conversion of information from optical to electrical and vice versa [7, 8], while all optical is free from O-E-O conversion. An Ideal all optical regeneration remove both amplitude and phase noise [9, 10]. However, use of the complex optical-field dominantly produces the nonlinear phase that degrade system performance [11–14].

2 Literature Review

Long-haul communication system require optical amplifier to compensate fiber losses [15]. Erbium-doped and Raman fiber amplifiers are currently being used [16]; these amplifiers are examples of phase-insensitive amplifiers (PIAs), which produce signal gain that is independent of the signal phase [17]. In principle, one could also use phase-sensitive amplifiers (PSAs) in communication systems [18]. The potential advantages of PSAs include, but are not limited to, noise reduction [19], the reduction of noise-induced frequency and phase fluctuations [20], dispersion compensation and the suppression of the modulation instability [1]. All optical regeneration is previously investigate through many techniques such as single pump configuration (Non-degenerated) [2], dual pump configuration (degenerated) using phase sensitive amplifier (PSA) for single or multilevel optical regeneration each has its own advantages and disadvantages [3, 4]. Furthermore, all-optical regenerations are also investigated theoretically, numerically with 2R (Re-amplify, reshaping) regenerator, 3R (Re-amplify, reshaping, retiming) configuration, phase extraction, pump dithering, Semiconductor optical amplifier (SOA), format conversion and with black box model [5–7]. The PSAs configuration is power consumptive and complex to use [8]. All optical regeneration still requires consideration to improve existing techniques in terms of improvement in BER, Eye diagram, power consumption and Q-factor of the system [9]. To improve the performance of existing all optical regenerative systems, we have demonstrated the all-optical regeneration for 10G noisy DPSK to improve the BER using 3R and PSA.

3 Problem Statement

When high power is launched inside fiber, nonlinear impairments became prominent and distort the signal during transmission. This produce spectral efficiency loss. Electronic regeneration is not adequate to be used for this problem because of its speed limitations and losses at high transmission. All optical regeneration is proposed to mitigate the nonlinear phase and ASE noise from noisy transmission systems.

4 Methodology

All optical regeneration is performed for 10G DPSK systems using PSA with 3R regeneration. With this technique the nonlinear phase noise mitigate with PSA and other noises such as ASE are removed using 3R regeneration process. Figure 1 shows, the system model for the all-optical regeneration system. The model contains the optical transmitter of 10G DPSK, noise emulator to introduce the external noise in the system to perform the regeneration. At the end, BER and eye diagram is compared before and after applying our all optical regenerative for 10G noisy DPSK system.

4.1 Optical Transmitter

In optical communication system, to recover lost spectral efficiency of the system, signal in phase and in amplitude is required. Therefore, in such type of systems the in phase transmission are preferred over amplitude only formats. In this regard, DPSK offers advantage in terms of resilience to transmission impairments [9]. In this work, 10G DPSK system is designed using (1) which describes the mathematical expression for the DPSK optical transmitter. Equation (1) states that we can

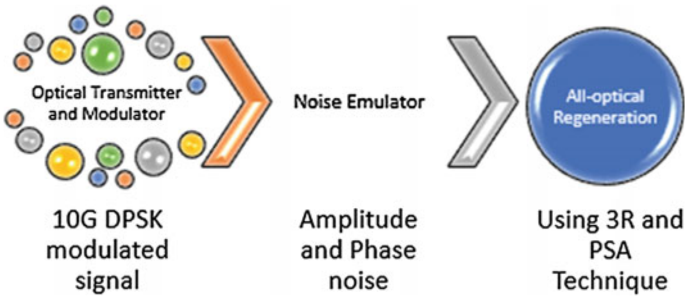


Fig. 1 System Model for 10G DPSK noisy all optical regenerative system

Table 1 Parameters for DPSK transmitter

Pseudo noise random sequence	RZ pulse generator	Continuous wave laser	Modulator
Bit Rate: 10G Operation mode: Log sequence Number of leading zeros: Time window * 5/100) * Bit rate Number of trailing zeros: (Time window * 5/100) * Bit rate	Pulse shape: Rectangular with 5Vp-p Duty cycle: 50 % Encoding scheme: RZ	Frequency 1552.5 nm Power: 10 dbm Linewidth: 10 MHz Initial phase: 0°	Extinction ratio: 20 dB Negative signal chirp: 1 Symmetry factor: −1

transmit different amount of power P(t), depending upon the encoding scheme and sampling time Ts [9]. The DPSK transmitter requires the pseudo random bit sequence, laser source, pulse generator and optical modulator. Table 1 shows, the parameters used in optical transmitter. Figure 2 shows the schematic design of DPSK transmitter in Optisystem.

$$P(t) = \begin{pmatrix} E & \mathit{for\ RZ}\ 100\ \% \\ E\ *\ \cos\left(\frac{\pi}{2}\cos^2\left(\frac{1.5\pi t}{Ts}\right)\right) & \mathit{for\ RZ}\ 66\ \% \\ E\ *\ \cos\left(\frac{\pi}{2}\cos^2\left(\frac{3\pi t}{Ts}\right)\right) & \mathit{for\ RZ}\ 50\ \% \end{pmatrix} \tag{1}$$

The system contain Pseudo noise random sequence with bit rate of 10G, CW laser as the source and RZ pulse generator as an electrical input. All these signals feed into LiNbO3 MZIMs modulator, which produce the 10G DPSK output. The optical transmitter also consist of narrow-line width laser to generate a lightwave whose wavelength conforms to the ITU grid [10, 12, 19].

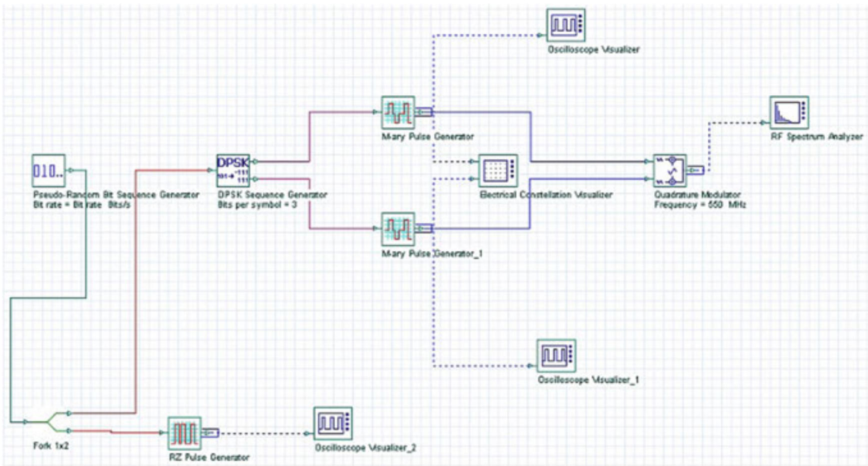


Fig. 2 DPSK optical transmitter in optisystem

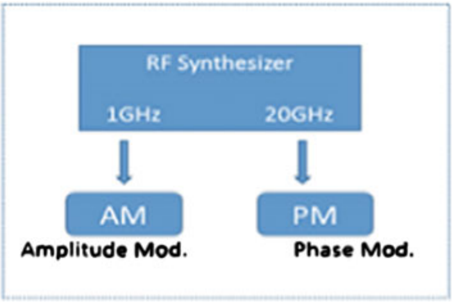


Fig. 3 Model of noise emulator

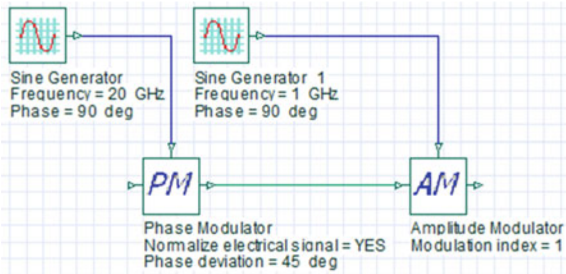


Fig. 4 Noise emulator in optisystem

4.2 Noise Emulator

In 10G DPSK, system noise is externally inserted to perform all optical regeneration for noisy system. Amplitude and phase noise are included using respective modulators. The phase modulator is driven at a frequency of 15 GHz and amplitude modulator driven at frequency of 2 GHz; both modulators are shown in Fig. 3.

The amplitude and phase noises are controlled by varying the corresponding depths to emulate different noise levels allowing control of the phase and amplitude noise levels, respectively. Figure 4 displays, the noise emulator design in optisystem. Both modulator are modulated with 1 GHz and 20 GHz RF signal tone frequency using sine wave generator respectively.

5 Experimental (Simulation) Description All Optical Regeneration Using 3R (3 Regeneration) and Phase Sensitive Amplifier

In this work, all optical regeneration is performed over 10G noisy DPSK system. 3R regeneration is technique, which is used to recover the signal from noises, using reamplification, reshaping and retiming. Equation (2) states that, when power is launched inside optical fiber due to change in phase the signal is varied, by neglecting dispersion [20].

$$U(L,t) = u(0,t)\ \exp[i\gamma PL|u(0,t)|^2] \tag{2}$$

Regenerator performance depends mainly on parameter $\phi = i\gamma PL$. Optimum value of ϕ_{NL} (nonlinear phase noise) is close to $3\pi/2$ because spectrum exhibits two peaks with a sharp dip at the original wavelength [20]. The main interest of 3R optical regeneration is to prevent from noise accumulation during transmission. Phase sensitive amplifier is, relying on the second and third order fiber nonlinearity and Four-wave mixing (FWM) [20]. Phase sensitive amplification using 3rd order nonlinearity of medium occurs when a pump wave interacts with signal and idler waves of lower frequency. If the signal and idler frequencies are identical, the process is said to be degenerate [20]. Degenerate PA is governed by the frequency-matching condition $\omega 2 = 2\omega 1$, where $\omega 2$ and $\omega 1$ are the pump and signal frequencies, respectively, and the amplitude as shown in (3) and (4) [20];

$$d_z A_1 = i2\overline{\gamma} A_2 A_1^* \exp(i\beta_z) \tag{3}$$

$$d_z A_2 = i2\overline{\gamma} A_1^2 \exp(i\beta_z) \tag{4}$$

where $\beta = 2\beta 1 - \beta 2$ is the linear wave number mismatch and γ is the non-linear coupling coefficient, which is proportional to $\chi^{(2)}$. One can choose the amplitude units in such a way that $|Ai|$ is proportional to photon flux. Suppose that wave 2 is a strong pump and wave 1 is a weak signal. Then, in the small-signal (undepleted-pump) approximation, the growth rate will be k is given by $(|\gamma|^2 - \delta^2)^{1/2}$. The transfer functions satisfy the auxiliary Eq. (5) [1].

$$E = a\cos(\omega t + \phi) + b\sin(\omega t + \phi) \tag{5}$$

This is how, the 3R and PSA are collectively used to mitigate ASE and nonlinear phase noise respectively as shown in (5) [2]. In (5) two quadrature amplitudes are a and b; phase of the pump signal is to be amplify using sin*wt* component and deamplify using cos *wt*, the two-kinds of input signals with isotropic (phase insensitive) noise are transformed to the squeezed state [3]. BER for 3R regeneration and PSA for signal accumulation scales are shown in (6) [4]. In (6) SNR is the signal

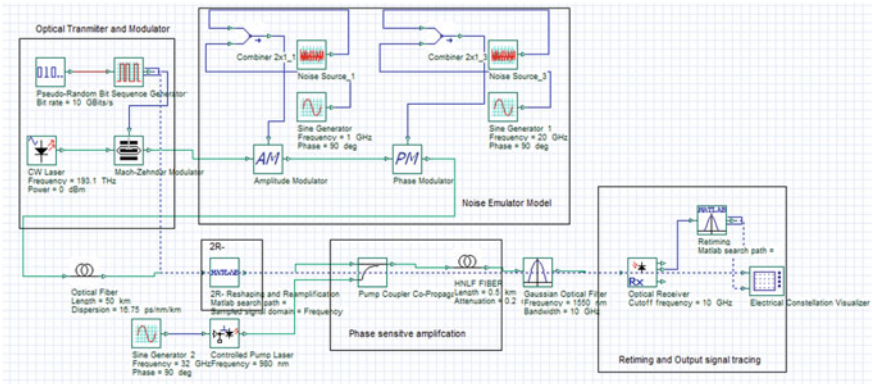


Fig. 5 Developed model for all optical regeneration

to noise ratio, N is the rank of the regenerator and C a suitable constant. One can obviously see the advantage of avoiding noise accumulation with a regenerator [5].

$$BER \approx N \exp(-CSNR) \tag{6}$$

Figure 5 describes, that first 10G DPSK signal is generated, after that amplitude and phase noise is inserted in the system externally [6]. The all optical regeneration is performed over this noisy signal using 3R and PSA as discussed earlier. After that, BER is calculated before and after the all-optical regeneration to validate the functionality of regenerator [7]. In all optical regeneration algorithm using 3R and PSA, 3R regenerator only can mitigate the ASE noise [8], synchronize the pulses [9], recover the signal for amplitude regeneration not for phase as much efficiently [10], overall 3R regeneration does not improve the BER of the optical system [11]. The main advantage of this technique is that with the help of 3R regenerator the ASE noise and other amplitude related noise can mitigate and PSA will control the phase response of the system [12]. Finally, we have achieved the all-optical regeneration using the 3R and PSA for 10G noisy DPSK transmission system with improved BER.

6 Results and Discussion

Figure 6 describes the optical DPSK transmitter response using electrical constellation diagram. The global parameters set for transmitter are sample bit 64, sample rate 60 kHz, sample length 128 with Bit rate of 10 GB/s. Figure 7 represents the electrical constellation diagram for 10G DPSK transmitter after inserting the amplitude and phase noise in the system. The amplitude modulator includes 1Vp-p amplitude and phase shift of 90° and having modulation index of 1. The Phase modulator includes the 1Vp-p amplitude, bias of 0.5 and phase shift of 90°, having

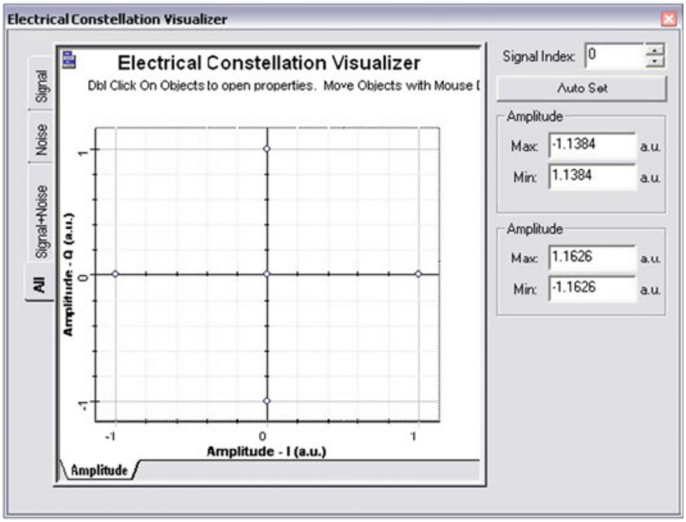


Fig. 6 Response of DPSK transmitter

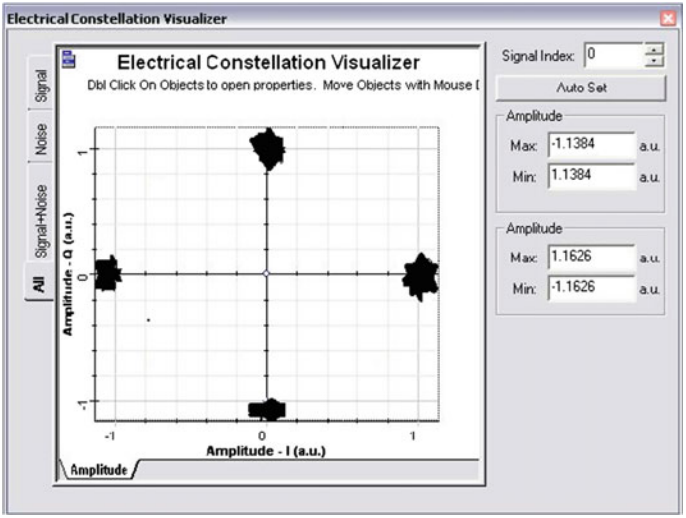


Fig. 7 DPSK response with noise

normalized electrical signal with phase deviation of 90°. Overall, 12 dB noise is included in 10G DPSK system using both modulators. Figures 8 and 9 represent BER and eye diagram of 10G noisy DPSK system before and after all optical regeneration using 3R and PSA. The ideal BER ism10^{-15} for 10G DPSK system and ideally eye diagram should be opened and having no jitters. Figure 8 shows, the

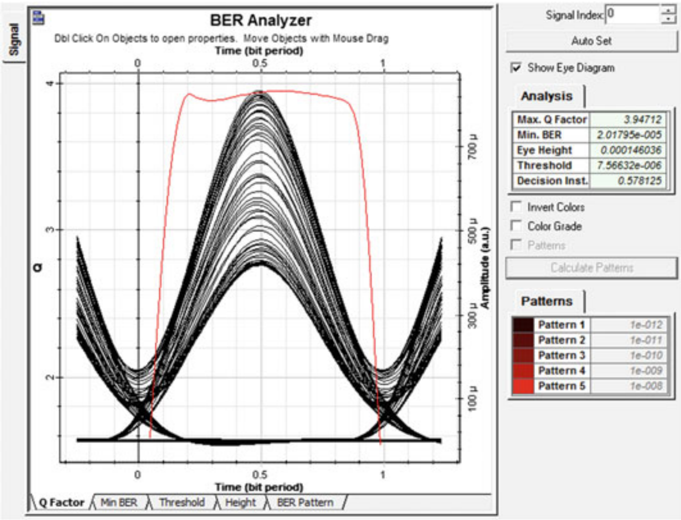


Fig. 8 BER, Eye diagram and Q-factor before all optical regeneration

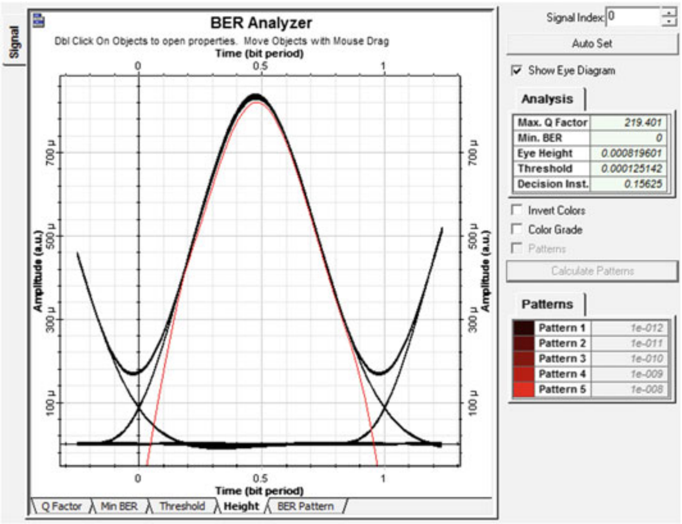


Fig. 9 BER, eye diagram and Q-factor after all optical regeneration

BER and eye diagram before the all optical regeneration, the BER at this level is 10^{-4}, and eye diagram is also not eye opened and having jitters and in this case the Q-factor for the system 48. Figure 8 also contains several FWM components, which clearly shows that the all optical regeneration is performed adequately. Using our developed techniques, we have improved BER up to 10^{-11}, with eye-opened

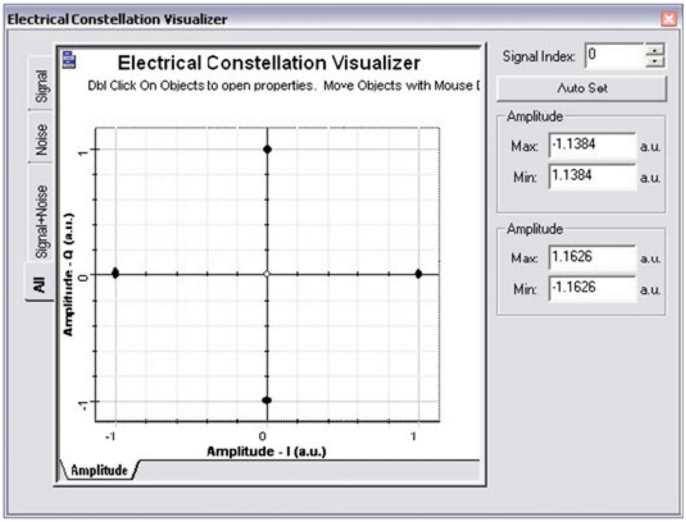


Fig. 10 Response of DPSK transmitter after applying all optical regeneration

diagram, having some jitters, all the other FWM components are filtered, only the 10G transmitted signal is regenerated with Q-factor of 135 as shown in Fig. 9. The results accomplishes that, we have achieved the significant improvement in BER and 64 % improvement in the Q-factor of the system with eye opened diagram, having minimum jitters. Overall, the noise remain in the system is 2.1 dB; this states that that all optical regeneration has removed the noise from the 10G noisy DPSK system significantly using our designed technique. Important thing to discuss is that all these significant improvement in BER, eye diagram and Q-factor are achieved at modulation index of 1. The change in modulation index also produces the changes in output, we have encountered have this problem during our findings; in future, there is a need to design such system that are independent from this modulation order.

Figure 10 defines the response of 10G noisy DPSK system after applying all optical regenerative technique using 3R and PSA. It can be clearly seen in the Fig. 9, that ASE noise and nonlinear phase noise is significantly removed. The external noise was included in the system was 12 dB and our designed techniques has reduced the both noise up to 2 dB at power penalty of 2 dB for overall system.

7 Conclusion

The results of research work determines that all-optical regeneration is practicable using combination of 3R and PSA technique. The proposed technique mitigates ASE noise and nonlinear phase noise from 10G DPSK and provide significant improvement in BER and Q-factor of the system. The result of this study suggest a number of new avenues for research, such as using similar technique 3R and PSA, the regeneration can be performed for PSK, DQPSK in optical communication system. Furthermore, the energy efficient optical communication system design using FPGA design will produce green optical communication system, which is the future demand of high-speed communication.

Acknowledgments We are thankful to Universiti Tun Hussein Onn Malaysia (UTHM), Malaysia that encourages us to contribute in research. This work is supported by Research Acculturation Collaborative Effort (RACE) Grant [vot1437] & Postgraduates Incentive Grant [U168] (GIPS) UTHM.

References

1. Richardson, D.J., et al.: Advances in optical signal processing based on phase sensitive parametric mixing. In: Advanced Photonics Congress, pp. 27212–27219 (2012)
2. Jones, L., et al.: All-optical broadband phase noise emulation. In: Advanced Photonics Congress, pp. JM5A. 19 (2012)
3. Roethlingshoefer, T., et al.: All-optical multilevel amplitude and phase regeneration. In: IEEE Photonics Society Summer Topical Meeting Series, pp. 89–90 (2014)
4. Yu, C., et al.: Wavelength-shift-free 3R regenerator for 40-Gb/s RZ system by optical parametric amplification in fibe. IEEE Photonics Technol. Lett. **18**(21–24), 2569–2571 (2014)
5. Lian, J., et al.: All-optical non-conjugated wavelength multicasting of QPSK signal with capability of phase regeneration. Opt. Express **22**(19), 22996–23006 (2014)
6. 刘汝斌, et al.: 全光 3R 系统再生能力的测试方法研究. 光电子. 激光 **7**, 1–5 (2013)
7. Slavík, R., et al.: All-optical phase and amplitude regenerator for next-generation telecommunications systems. Nat. Photonics **4**(10), 690–695 (2010)
8. Matsumoto, M.: All-optical DQPSK signal regeneration using 2R amplitude regenerators. Opt. Express **18**(1), 10–24 (2010)
9. Shah, N.S.M., Sato, M., Matsumoto, M.: Adaptive delay control for time-interleaved multi-channel amplitude limiter based on saturation of four-wave mixing in a fiber. Opt. Express **19**(22), 21246–21257 (2011)
10. 张婧, et al., 基于光纤自相位调制多波长全光再生的色散管理优化. 物理学报 **59**(10), 7002–7007 (2010)
11. Kakande, J., et al.: All-optical processing of multi-level phase shift keyed signals. In: Optical Fiber Communication Conference, paper OW1I-3 (2012)
12. Willner, A.E., et al.: All-optical signal processing. J. Lightwave Technol. **32**(4), 660–680 (2014)
13. Lei, G.K., Marhic, M.E.: Amplification of 800 Gb/s PDM-DPSK signals in a one-pump fiber optical parametric amplifier. In: European Conference on Optical Communication (ECOC), pp. 1–3 (2014)
14. Ettabib, M.A., et al.: All-optical phase regeneration in a highly nonlinear lead-silicate fiber. In: Advanced Photonics Congress, paper JM5A-11 (2012)
15. Mohd Shah, N.S., Matsumoto, M.: Analysis and experiment of all-optical time-interleaved multi-channel regeneration based on higher-order four-wave mixing in a fiber. Optics Commun. **284**(19), 4687–4694 (2011)
16. Gruner-Nielsen, L., Jakobsen, D., Herstrom, S., Pálsdóttir, B., Dasgupta, S., Richardson, D., et al.: Brillouin suppressed highly nonlinear fibers. In: 38th European Conference and Exhibition on Optical Communications (ECOC), pp. 1–3 (2012)

17. Lundstrom, C.: Applications of fiber-based phase-sensitive parametric amplifiers. In: IEEE Photonics Society Summer Topical Meeting Series, pp. 93–94 (2014)
18. Olonkins, S., Supe, A., Lyashuk, I., Porins, J.: Demonstration of binary PolSK to OOK modulation format conversion using a single-pump fiber optical parametric amplifier. In: 9th International Symposium on Communication Systems, Networks & Digital Signal Processing (CSNDSP), pp. 955–959 (2014)
19. Slavik, et al.: Coherent all-optical phase and amplitude regenerator of binary phase-encoded signals. IEEE J. Sel. Top. Quantum Electron. **18**(2), 859–869 (2012)
20. Slavík, R., Kakande, J., Phelan, R., O'Carroll, J., Kelly, B., Richardson, D.J.: Chirp reduction and on/off contrast enhancement via optical injection locking and coherent carrier manipulation. In: SPIE Optics + Optoelectronics, 87810G (2013)

Real-Time Segmentation and Tracking Module of Target of Interest from Video Sequence in Object Recognition Systems

Slavomir Matuska, Robert Hudec, Miroslav Benco and Patrik Kamencay

Abstract This paper proposes a real-time objects segmentation and tracking module from video sequences, which can be effectively used in real-time object recognition systems. The module is based on background subtraction method in combination with CAMshift (Continuously Adaptive Mean shift) algorithm. In the first step, background subtraction method is applied to determine pixels of moving objects in video stream. Then, foreground pixels are used as starting point for CAMshift algorithm. CAMshift finds optimal size, position and orientation of moving objects. After that, key frame extraction method is applied in order to choose only relevant frame in later objects classification.

1 Introduction

Real-time segmentation and fast objects tracking in video sequences are an essential part of many application, especially in traffic monitoring, surveillance systems, smart cities, human-machine interface and different kind of security solutions. Over the last few years, there were published many approaches focusing on achievement of good and robust real-time objects segmentation and tracking, for example in [1, 2].

The main task of this module is to detect moving objects in the scene and then track theirs movement across the scene. For people is easy to see moving object and

S. Matuska (✉) · R. Hudec · M. Benco · P. Kamencay
Department of Telecommunication and Multimedia, University of Zilina,
Univerzitna 8215/1, 01026 Zilina, Slovakia
e-mail: slavomir.matuska@fel.uniza.sk

R. Hudec
e-mail: robert.hudec@fel.uniza.sk

M. Benco
e-mail: miroslav.benco@fel.uniza.sk

P. Kamencay
e-mail: patrik.kamencay@fel.uniza.sk

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_48

separate it from the rest of the scene. In computer vision, one of the simplest way how to separate moving object from scene background is to use background subtraction method. This method is widely used in computer vision and involves calculating a reference image, subtracting each new frame from the reference image, thresholding and evaluating the result. These steps result in creation of binary segmented image with highlighted moving objects. The simplest form of reference image is a time-averaged background image. This method suffers from many problems and requires a training period to calculate reference image [3, 4]. There had been also a problem with rapid illumination changes which was solved later in [5]. After successful determination and detection of moving objects, CAMshif algorithm is applied to find optimal object size, position and orientation. CAMshift is very efficient in tracking object with similar color and was used in many approaches [6, 7]. Object color should be different from the color of the background. In our approach, this problem is solved during background subtraction phases and in CAMshift phase only foreground pixels are used. CAMshift can be summarized whit these steps [8]:

1. Choose the initial region of interest which contains the object we want to track.
2. A color histogram computation from the region of interest.
3. A probability distribution computation using color histogram.
4. Based on the probability distribution image, find the center mass of the search window using mean-shift method.
5. Center the search window to the point taken from step 4 and repeat step 4 until convergence.
6. Process the next frame witch search window position from the step 5.

In the last part of the module, key frame extraction method is applied on target of interest determinated in CAMshift phase. This method uses early and late filtration to mark only relevant targets of interest in order to achieve best results in objects classification.

This paper is focusing on creation simple and efficient method of segmentation and tracking objects in video sequence using background subtraction method and CAMshift algorithm. The paper is organized as follows: detailed method description is given in the second part. Experimental results are described in the third part and conclusion and future work are shown and discussed in last parts.

2 Proposed Method

As it was mentioned before, the main task of real-time segmentation and tracking module is detection of moving objects and marking relevant target of interest for object classification.

Figure 1 shows block diagram of our proposed method. Video sequence in resolution 1920 × 1080 and frame per second rate at 25, using H.264 codec serves as input data stream. Moreover, video resolution is reduced to final resolution

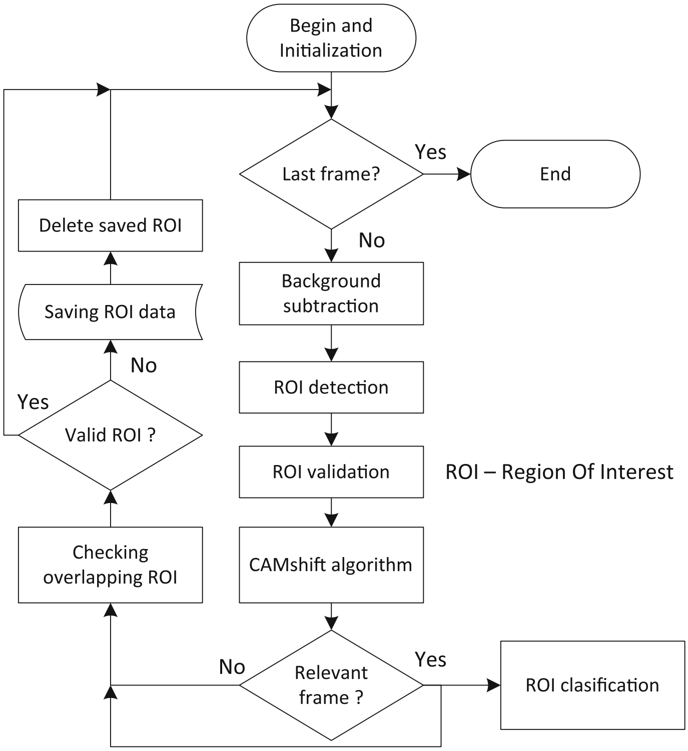


Fig. 1 Real-time segmentation and tracking module block diagram

640 × 380 pixels, i.e. divided factor 3. Data stream is processed frame by frame following block diagram in Fig. 1. In the beginning, background subtraction method needs at least 50 frame to initialization.

After that period, each frame is blurred with Gaussian with 2 × 2 kernel and then background subtraction method is applied on the processed frame. Output of this phases is shown in Fig. 2, where actual frame is shown in Fig. 2a, binary segmented image with highlighted foreground pixels is given in Fig. 2d and background image can be found in Fig. 2b. In the next step, moving object candidates are detected using binary segmented image which is scanned by sliding window with size 20 × 20 pixels and with 10 pixels step. Sum of white moving pixels is count in each sliding window and when this sum is greater than threshold value, actual sliding window is marked as moving object candidate. Valid candidates are show as gray rectangles in Fig. 2c. Candidates are checked for relevance and best of them is marked as a region of interest (ROI) with moving object.

If they are more ROI in one scene, algorithm of checking overlapping ROI is used to evaluate possible ROI overlapping and in this case, there is need to merge two ROIs into one. In the last step of the module, ROI are checked for their presence in the scene and data of non-valid ROI are saved and afterwards ROI

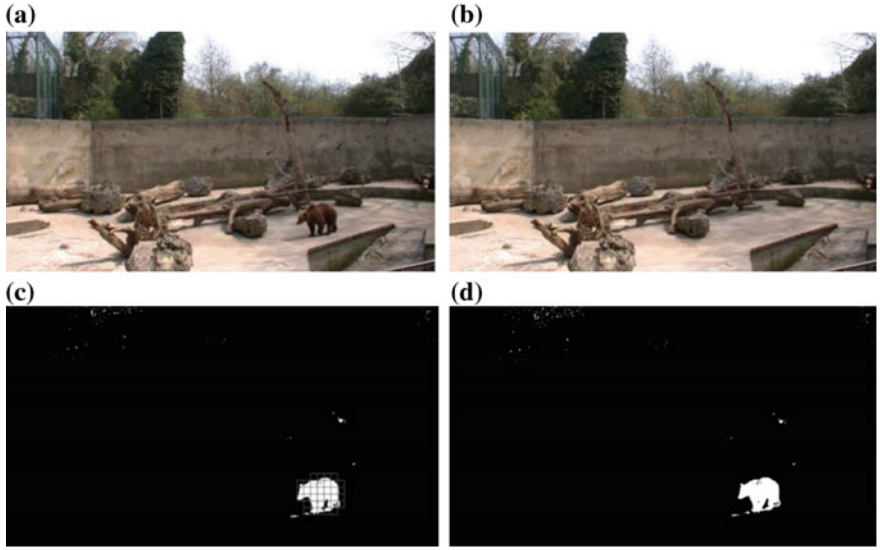


Fig. 2 Background subtraction images

tracking is stopped and erased from valid regions of interest. All these steps are repeated until last frame of the video sequence is processed or live stream is stopped.

2.1 Key Frame Extraction Method

In order to achieve best accuracy in object classification, key frame extraction method is used to filtrate regions of interest. Proposed method consist of two parts:

- Early filtration.
- Late filtration.

Early Filtration. Early filtration is applied on the moving ROI in real-time and its task is to do first filtration of valid regions of interest. ROI needs to fulfill two conditions in order to participate in objects recognition:

- ROI size must be more then threshold value.
- ROI must consist at least of 40 % moving pixels.

Threshold value was determined empirically from testing video sequences and its value was set to 10000 pixels. This condition should filtrate ROI where unknown object is not clearly visible in ROI, for example when object is entering or leaving the scene or is partially overlapped by other objects. The second threshold value was also determined empirically. In testing video sequences average percentage of moving pixels in regions of interest was about 50 % and therefore the threshold was set to 40 %.

Later Filtration. Later filtration is applied on a set of the regions of interest belonging to the tracking object, after this object has left the scene. Later filtration consist of two conditions:

- Number of edges in ROI.
- Number of frame with tracking object in video sequence.

Canny detector is used to detect edges in each ROIs. Edges are represented by white points in output binary image from Canny detector. Number of white points is counted and average value of white points in images is calculated using the formula 1:

$$edge_{mean} = \frac{\sum_{i=0}^{n} edgeCount_i}{n}. \tag{1}$$

where n is number of images which belong to single tracking object and $edgeCount_i$ is number of white points in individual ROI. If the number of edges in ROI is lower than average value, particular ROI is used in object classification. Idea of this condition is to remove these ROIs, where unknown object is overlapping by some other objects from background, such as tree, stone etc. The objects which are not at least 40 frames at video sequence are removed from valid detected objects according to the second condition. This condition remove false detected objects in video sequences.

3 Experimental Results

The module was programmed and developed in C ++ language using OpenCV libraries. For verification of the module robustness and accuracy, the module was tested on static video sequences with resolution 1920 × 1080. In video sequences, moving objects were represented by animals in their natural conditions or in zoo. These videos were created as a part of the international project E!6752—DETECTGAME. There were 5 different kinds of animal, namely wolf, brown bear, fox, deer and wild boar. Information about video length and number of frame in video sequences per animal kind are given in Table 1.

Real-time segmentation and tracking module was tested and recall, precision and F1 measure were evaluated.

Table 1 Video sequences information

	Wild boar	Brown bear	Wolf	Fox	Deer
Video length (min)	08:55	07:23	04:52	01:21	09:24
Frame count	13,375	11,075	7300	2025	14,100

3.1 Precision, Recall and F1 Measure

Precision. Precision is defined as a proportion between the number of valid frames specified by the module and the number of all frames classified as valid frames. Precision is defined by formula 2:

$$Precision = \frac{t_p}{t_p + f_p}. \tag{2}$$

where t_p defines the number of frame correctly classified as valid class and f_p defines the number of frame incorrectly classified as valid class.

Recall. Recall is defined as a proportion between the number of valid frames specified by the module and the number of all valid frames. Precision is defined by formula 3:

$$Recall = \frac{t_p}{t_p + f_n}. \tag{3}$$

where t_p defines the number of frame correctly classified as valid class and f_n defines the number of frames which were not classified as valid, but they should be.

F1 measure. F1 measure combines precision and recall in harmonic mean and F1 measure is defined by formula 4:

$$F1 = 2 * \frac{precision * recall}{precision + recall}. \tag{4}$$

3.2 Results

Recall, precision and F1 measure were evaluated for every video sequence and results are shown in Table 2.

From Table 2 is evident that high recall, precision and F1 measure were achieved. Only in two cases, namely deer_0 and deer_3, precision achieved low values. This was caused by the nature of the movement of deer animal. Deer is often static and therefore background subtraction method after some time includes deer into background pixels. Graphical results for recall and precision for wild boars and deers are shown in Fig. 3 and for wolfs, foxes and brown bears are shown in Fig. 4.

Table 2 Results of recall, precision and F1 measure on video sequence database

Video name	Recall	Precision	F1 measure	Video length (min)
Bear_0	1.00	0.86	0.91	3:30
Bear_1	1.00	0.81	0.89	0:39
Bear_2	1.00	0.56	0.72	0:37
Bear_3	0.92	0.97	0.95	2:37
Wolf_0	1.00	0.88	0.94	1:18
Wolf_1	0.67	0.99	0.80	0:52
Wolf_2	1.00	0.98	0.99	0:12
Wolf_3	0.97	0.84	0.90	0:15
Wolf_4	0.89	1.00	0.94	0:30
Wolf_5	0.88	0.86	0.87	1:45
Deer_0	1.00	0.32	0.49	1:55
Deer_1	0.99	0.99	0.99	1:55
Deer_2	1.00	0.58	0.73	4:57
Deer_3	0.95	0.31	0.46	1:27
Boar_0	0.96	0.94	0.95	0:38
Boar_1	0.83	0.89	0.86	0:42
Boar_2	1.00	0.69	0.82	1:00
Boar_3	1.00	0.71	0.83	1:20
Boar_4	1.00	0.54	0.70	0:36
Boar_5	0.83	0.96	0.89	4:39
Fox_0	1.00	0.88	0.93	0:16
Fox_1	1.00	0.89	0.94	0:17
Fox_2	1.00	0.70	0.83	0:10
Fox_3	1.00	0.92	0.96	0:38

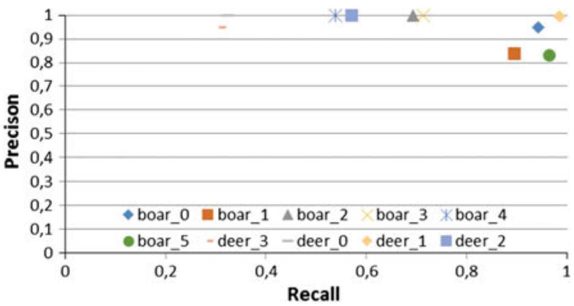


Fig. 3 Recall and precision for video sequences with boars and deers

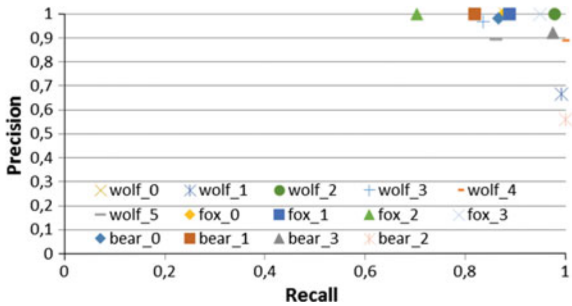


Fig. 4 Recall and precision for video sequences with wolfs, foxes and bears

4 Conclusion and Future Work

In this paper, real-time and robust module for objects segmentation and tracking was introduced. The main task of this module is detection, segmentation and tracking of multiple moving objects in video sequences. The module of object segmentation and tracking involves two main common techniques, namely background subtraction method and CAMshift algorithm. From the realized experiments is evident, that our approach has achieved high value of recall, precision and F1 measure. The module is suitable to be used in many application areas, for example in smart cities, where can be used in traffic monitoring and intelligent traffic control. Moreover, it can be also used in any type of surveillance systems, human-machine interface or security solutions. In the future, this module will be used as a part of the intelligent system for animal recognition from video sequences. Therefore, key frame extraction method was implemented in this module to insure the most relevant results in objects classification. The main task of the intelligent system is to create a migration corridors for wild animals in designated areas.

Acknowledgments The work presented in the paper was supported by the University Science Park of the University of Zilina (ITMS: 26220220184) supported by the Research &Development Operational Program funded by the European Regional Development Fund and EUREKA project no. E! 6752—DETECTGAME: R&D for Integrated Artificial Intelligent System for Detecting the Wildlife Migration.

References

1. Intachak, T., Kaewapichai, W.: Real-time illumination feedback system for adaptive background subtraction working in traffic video monitoring. In: International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), 7–9 Dec 2011
2. Jenifa, R.A.T., Akila, C., Kavitha V.: Rapid background subtraction from video sequences, pp. 1077–1086. International Conference on Computing, Electronics and Electrical Technologies, 21–22 Mar 2012
3. Zivokovic, Z.: Improved adaptive gaussian mixture model for background subtraction, vol. 2, pp. 28–31. In: Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004) (2004)

4. Kaewtrakulpong, P., Bowden, R.: An improved adaptive background mixture model for real-time tracking with shadow detection, vision and virtual reality group. In: 2nd European Workshop on Advanced Video Based Surveillance Systems (AVBS01), Sept 2001
5. Sajid, H., Cheung, S.S.: Background Subtraction under sudden illumination change, pp. 384–391. In: Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 29 Aug 2010–1 Sept 2010
6. Kassir, M.M., Palhang, M.: A region based CAMShift tracking with a moving camera, pp. 451–455. In: Second RSI/ISM International Conference on Robotics and Mechatronics (ICRoM), 15–17 Oct 2014
7. Cong, D., Shi, P., Zhou, D.: An improved camshift algorithm based on RGB histogram equalization, pp. 426–430. In: 7th International Congress on Image and Signal Processing (CISP), 14–16 Oct 2014
8. Bradski, G.R.: Real time face and object tracking as a component of a perceptual user interface, pp. 214–219. In: Fourth IEEE Workshop on Applications of Computer Vision (WACV '98), Princeton, NJ, 19–21 Oct 1998. Print ISBN: 0-8186-8606-5

Analysis of Electromyography (EMG) Signal for Human Arm Muscle: A Review

A.F.T. Ibrahim, V.R. Gannapathy, L.W. Chong and I.S.M. Isa

Abstract Muscles provide motion in response to nerve impulses. There are two main categories of muscle can be distinguished based on their anatomy and the particular functions they perform. Skeletal muscles are the largest group of muscles control posture, generate heat, and provide motion control and are mostly influenced by the brain in conscious acts. Smooth muscles provide rhythmic motion outside the control of the brain. A nerve cell provides a train of impulses delivered to a group of muscles in which the impulses depolarize muscle cells and cause muscles to contract. Electromyography (EMG) is a technique for evaluating and recording electrical activity produced by skeletal muscles when the muscles are stimulated. Surface electromyography and needle electromyography are two general methods of recording the electrical activities of muscle tissue. In this project, surface electromyography will be used to test the analysis for human arm muscle. Electrodes placed on the skin surface can be used to monitor the coordination of entire muscle groups but will not reveal much about the individual muscle cells. Thorough testing will be done as the future work once the fabrication has been completed.

1 Introduction

Electromyography (EMG) is a technique for evaluating and recording electrical activity produced by skeletal muscles when the muscles are stimulated. Electromyography (EMG) is performed using an instrument called electromyograph

A.F.T. Ibrahim (✉) · V.R. Gannapathy · L.W. Chong · I.S.M. Isa
Faculty of Electronic and Computer Engineering, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
e-mail: fayeez@utem.edu.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_49

to produce a record called electromyogram. The word electromyogram stems from three terms. ‘electro’ means pertaining to electrical activity, ‘myo’ has a Greek root mean muscle, and ‘gram’ stands for recording [1]. By combining this term, electromyography (EMG) refers to recording of a muscle’s electrical activities. Electromyography (EMG) can be defined as a signal that records the electrical activities generated by the depolarization of muscles cells during muscle contraction and the nerve impulses that initiate the depolarization of the muscle. There are two general methods of recording the electrical activities of muscle tissue such as surface electromyography (EMG) and needle electromyography (EMG). Surface electromyography (EMG) in which the electrodes are applied on the skin is non-invasive for measuring electrical activity resulting from contraction and relaxation exercises. Meanwhile, needle electromyography (EMG) is the needles with electrodes are inserted into muscle itself is an invasive technique for measuring muscle electrical activity resulting from contraction and relaxation exercises. In this study, surface electromyography (EMG) will be used to test the signal analysis for human arm muscle. A typical electromyography (EMG) signal is the result of the electrical activities of many motor units that are weighted according to the amount of fat and skin between each motor unit from the electrode [2]. Electrodes placed on the skin surface can be used to monitor the coordination of entire muscle groups but will not reveal much about the individual muscle cells. In general, this technique is used to identify which muscle groups are involved in a particular motion or action.

2 Background Study

Muscles provide mobility in response to nerve impulses. There are two main categories of muscle that can be differentiated based on their anatomy and the particular functions they carry out. Skeletal muscles which are the largest group of muscles control posture, generate heat, and provide motion control and are mostly influenced by the brain in conscious acts. Second group of muscles is the smooth muscles which provide rhythmic motion outside the control of the brain.

In principle, a nerve cell provides a train of impulses delivered to a muscle or group of muscles. These impulses depolarize muscle cells and cause muscles to contract. The frequency of the nerve impulses determines the process of muscle depolarization and muscle contraction. Several muscle fibers are innervated by only a single motor-neuron or motoneuron. The structure containing the motoneuron and its connected muscle fibers is known as the motor unit. Each muscle is composed of a number of motor units, and each motor unit is driven by a neuron. Depending on the collective effects of the information coded in the nerves driving a muscle, a particular action is performed by the muscle. Absence of any electrical impulses will make the muscle relax.

Muscle contraction consists of two methods which is isometric and isotonic. The muscle is not allowed to shorten when it is under isometric contraction. Meanwhile, a constant force is applied by the muscle when it is under isotonic contraction. In a wrap, not all muscle cells contract simultaneously because some remain idle and take over when the contracting muscle cells relax. Distribution of the firing sequence of various motor units controls this operation to provide a desirable action by the muscle.

Before starting the project, it is essential to do some research on the importance and contribution toward the project moreover on the generation of the electromyography (EMG). The electromyography (EMG) is generated when a motor neuron action potential from spinal cord arrives at a motor end plate [3]. Its arrival causes a release of Acetylcholine (ACh) at the synaptic cleft which causes a depolarization or known as action potential. This action potential electrically travels downward from the surface in a transverse tubule. This in turn causes a release of Ca++ causing cross-bridge binding and sarcomere of the muscle to contract as shown in Fig. 1.

After doing research on the generation of the electromyography (EMG), it is important to explore the characteristics of the electromyography (EMG) signal. The

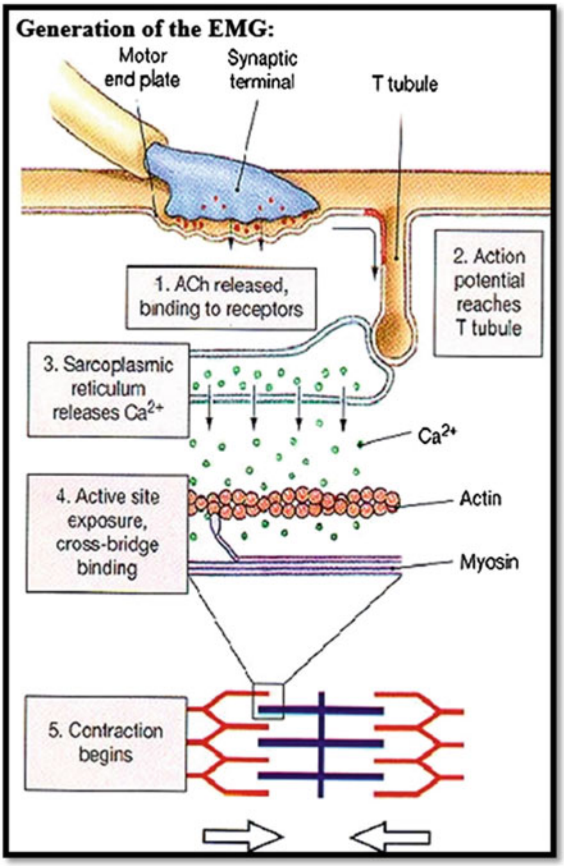


Fig. 1 Generation of electromyography (EMG)

amplitude of the electromyography (EMG) signal is stochastic or known as random in nature [4]. The amplitude is ranged from 0 to 10 mV peak-to-peak. The usable energy of the signal is limited to the 0–500 kHz frequency range with the dominant energy being in the 50–150 Hz range.

Most signals have noise. Electromyography (EMG) signal also have noise since the signal produced is very small. There are some characteristics of the electrical noise produced by the signal such as ambient noise, motion artifact, inherent noise and muscle crosstalk [5]. Ambient noise which corresponds to power line noise lies in wide range of frequencies but its dominant component is 50 Hz or 60 Hz which is the most common source of electrical noise in the electromyography (EMG) signal. Motion artifact is caused by the movement at the interface between the detection surface of the electrode and skin which lies in 1–10 Hz frequency range. Inherent noise comes from electronics instrumentation since all electronics equipment generates electrical noise. This noise which cannot be eliminated has frequency components that range from 0 Hz to several thousand Hz. Muscle crosstalk is a phenomenon in which signal recorded over one muscle was in fact generated by a neighboring muscle and conducted to the recording electrodes. Electromyography (EMG) signal is very important to evaluate the condition of nerves and muscles. However, electromyography (EMG) signal is a complicated signal in which is controlled by nervous system and is dependent on the anatomical and physiological properties of muscle. Therefore, it is hard to detect the signal generated from the muscle since the characteristic of the signal has amplitude range of 0–10 mV prior to amplification and frequency range of 10–500 Hz.

In order to test the EMG signal, most of us cannot afford to have their own EMG machine to test their muscle especially for muscle injury people. EMG machine is very expensive and the machine is only provided by hospital. Most injured people need to refer to hospital in physiotherapy unit to test their contraction of muscle. Every month, the tragedy of accidents occurs and most cases involving injury to the muscle. This will cause a lot of people need to refer to EMG machine to test the contraction of muscle.

Meanwhile, we can see the overview that most muscle injury people do not like to wait for turn just to get their muscle checked [6]. Normally, EMG machine is expensive. Hence, electromyography (EMG) circuit is designed so that it can be used to overcome this problem.

3 Related Work

Wireless sensor networks have become increasingly popular in nowadays applications due to decreasing technology costs and improved product performance, robustness and extensibility [7]. The use of wireless sensors for biomedical research applications has been widely reported in the literature which utilizes wireless sensing capabilities in the form of either body worn or ambient sensing device [8].

Wearable physiological monitoring systems have been developed in a variety of studies. Applying new sensing technology to healthcare maybe part of an alternative way to the financial and demographic crisis facing global healthcare system [9]. Researchers applying new methods to noninvasive patient monitoring and diagnostics are assisted by the features of Sensing Health with Intelligence, Modularity, Mobility and Experimental Reusability (SHIMMER) [10].

SHIMMER EMG expansion module provides a highly economical wireless EMG acquisition system using conventional disposable electrodes as shown in Fig. 2. SHIMMER EMG measures and records electrical activity associated with skeletal muscle contractions and can be used to analyze and measure the biomechanics of human or animal movement [11]. SHIMMER EMG is non-invasive and therefore the activity it measures is a representation of the activity of the whole muscle or group of muscles whose electrical activity is detectable at the electrode site.

It is important to know the overall system of SHIMMER EMG module rather than discuss about it functionality. SHIMMER EMG module is comprised of four system overview such as electrodes, amplifier, bandwidth filter and analogue to digital converter (ADC) as shown in Fig. 3. Electrodes consist of positive, negative and neutral electrode where they are connected from each EMG board. Second part is amplifier which increases amplitude of input signal. Every amplifier has noise when amplify the signal. Therefore, bandwidth filter is added into the system to filter the noise. Bandwidth filter bandpass filter with bandwidth of 5 Hz to 322 Hz. Bandwidth filter is an active filter and therefore it has gain. The final part of the system is analogue to digital converter (ADC) which converts the input analogue signal to a digital representation of this signal by assigning a value between 0 and 4095 to each sample. This is the output from the SHIMMER EMG module.

The biomedical activity inside the muscle of a human body is detected with the help of EMG electrodes. As described in the introduction earlier, there are two main types of EMG electrodes which are surface and inserted electrodes [12]. Inserted electrodes are divided into two types which is needle and fine wire electrodes as shown in Fig. 4[13]. Needle electrodes are widely used in clinical procedures in neuromuscular evaluations. The tip of the needle electrode is bare and used as a

Fig. 2 SHIMMER EMG module

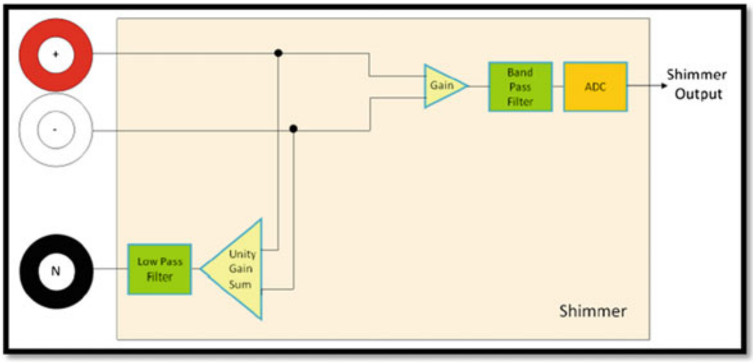


Fig. 3 Basic system overview

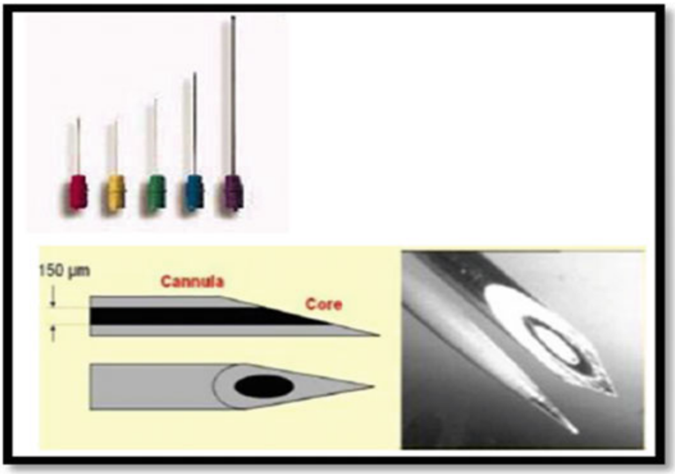


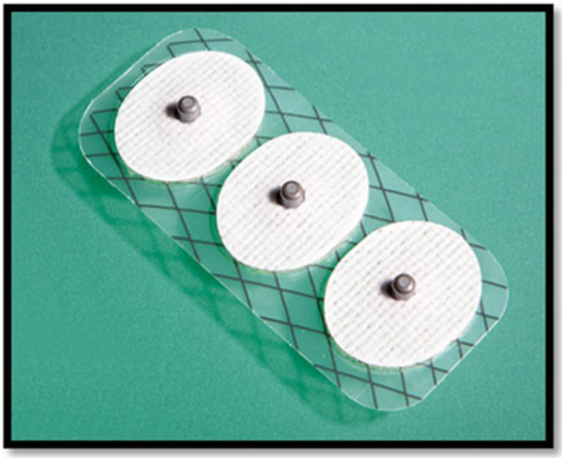
Fig. 4 Inserted electrodes

detection surface in which it contains an insulated wire in the cannula. Second type of inserted electrode is fine wire electrodes in which the wire electrodes are made from any small diameter, highly non-oxidizing, stiff wire with insulation. This type of electrode are extremely fine, they are easily implanted and withdrawn from skeletal muscles, and they are generally less painful than needle electrodes whose cannula remains inserted in the muscle throughout the duration of the test. In the project, inserted electrodes will not be used because the structure itself exposed to the disadvantage of this type of electrodes. Besides, the application of inserted electrodes requires strict medical supervision and certification.

Therefore, surface EMG (sEMG) electrode will be used to test and analyze the small EMG signal throughout the whole project. Surface EMG electrodes provide a non-invasive technique for measurement and detection of EMG signal. The theory behind these electrodes is that they form a chemical equilibrium between the detecting surface and the skin of the body through electrolytic conduction so that current can flow into the electrode. This type of electrodes is simple and very easy to implement. Surface EMG electrode is being increasingly used to detect muscle activity in order to control device extensions to achieve prosthesis for physically disabled and amputated population. There are two types of surface EMG which is gelled EMG electrode and dry EMG electrode. Gelled EMG electrodes contain a gelled electrolytic substance as an interface between skin and electrodes as shown in Fig. 5. The most common composite for the metallic part of gelled electrodes is silver-silver chloride (Ag-AgCl) in which AgCl layer allows current from the muscle to pass more freely across the junction between electrolyte and the electrode. Dry EMG electrodes do not require a gel interface between skin and the detecting surface and this type of electrode may contain more than one detecting surface.

Application of surface EMG electrodes requires proper skin preparation beforehand. In order to have a good quality of EMG signal, the skin should be cleaned with alcohol in order to eliminate any wetness or sweat on the skin. Besides skin preparation, proper EMG electrode placement also affects the quality of EMG signal. In order to acquire the best possible signal, surface EMG electrodes should be placed between the motor unit and the tendinous insertion of the muscle, along the longitudinal midline of the muscle as shown in Fig. 6. The distance between the center of the electrodes should be 1–2 cm. The electrodes should neither place at or near tendon nor at the edge of the muscle.

Fig. 5 Gelled EMG electrode



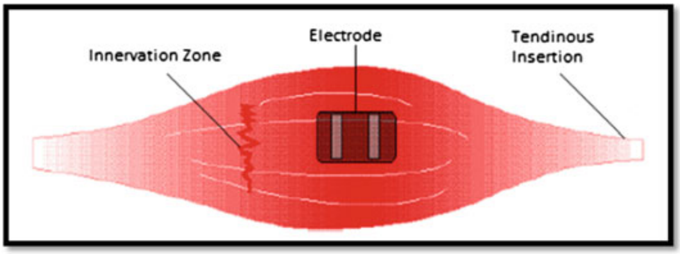


Fig. 6 EMG electrode placement

4 Future Work

Some of the issues were not taken into consideration due to lack of information in Matlab. Therefore, hardware is required to develop to support the evidences of the literature review. The study of hardware must be based on the instrument amplifier and bandpass filter because it is important to build EMG circuit since it sense very low signal. Next stage is the designing and fabrication of the circuit before the testing and analysis.

5 Conclusion

From the observation, the EMG circuit functions excellently in meeting almost entire target aims and objectives requirement although there are some part of LCD display is in unstable condition. From the previous research, surface EMG electrodes are used to attach to human arm body as an agent of detection of human arm muscle contraction. It is proved that the signal will varied when the muscle begin to contract. In the project, the contraction of muscle will show some values on LCD display and no values will show if the muscle relaxed. Muscle contraction normally generates electrical activity with the signal between 1 mV until 10 mV prior to amplification. When muscles contract, the threshold voltage of the muscle contraction will be around 0 V until 3 V after the amplification. Therefore, the reference voltage set on the ADC circuit is 5 V.

Acknowledgement The authors would like to take this opportunity to thank those who contributes directly or indirectly in completion of this article and also for their constructive comments. In addition, the authors also would like to express their gratitude to Faculty of Electronic & Computer Engineering and Universiti Teknikal Malaysia Melaka (UTeM) for the funding, support and encouragement.

References

1. Gupta, S.D., Yusuf, S.A., Ammar, J.K., Hasan, K.: An analysis to generate EMG signal and its perspective: A panoramic approach, Advances in Power Conversion and Energy Technologies (APCET), International Conference, pp. 1–5 (2012)
2. Ravariu C.: Contributions to novel methods in electrophysiology aided by electronic devices and circuits. In: Dr. Gargiulo, G. (ed.) Applied Biomedical Engineering. InTech (2011). doi:10.5772/21443
3. Najarian, K., Splinter, R.: Biomedical signal and image processing, pp. 237–255. Taylor and Francis Group, New York (2006)
4. Poo, T.S., Sundraj, K.: Design and development of a low cost EMG signal acquisition system using surface EMG electrode, Proceedings of the International Conference on Man-Machine Systems (ICoMMS), pp. 24–27 (2009)
5. Salman, A., Iqbal, J., Izhar, U., Khan, U.S., Rashid, N.: Optimized circuit for EMG signal processing, Robotics and Artificial Intelligence (ICRAI), International Conference, pp. 208–213 (2012)
6. Mademli, L.: The physiological background of EMG. Centre of Research & Technology—Hellas Informatics & Telematics Institute. Retrieved from www.iti.gr
7. Khan, M.H., Wajdan, A., Khan, M., Ali, H., Iqbal, J., Shahbaz, U., Rashid, N.: Design of low cost and portable EMG circuitry for use in active prothesis applications, Robotics and Artificial Intelligence (ICRAI), International Conference, pp. 204–207 (2012)
8. Jamal, M.Z.: Signal acquisition using surface EMG and circuit design considerations for robotic prosthesis. Computational Intelligence in Electromyography Analysis—A Perspective on Current Applications and Future Challenges, pp. 427–448 (2012)
9. Boylestad, R.L., Nashelsky, L.: Electronic Devices and Circuit Theory, 10th edn., pp. 711–718 (2009)
10. Burns, A., Doheny, E.P., Greene, B.R., Foran, T., Leahy, D., O'Donovan, K., McGrath, M.J.: SHIMMER: An extensible platform for physiological signal capture. In: 22nd Annual International Conference of the IEEE EMBS, pp. 3759–3762
11. SENIAM Recommendation. Retrieved from www.seniam.org on 16 Oct 2013
12. Fridlund, A.J., Cacioppo, J.T.: Guildelines for human electromyographic research. Soc. Psychophys. Res. **23**(5), 567–589
13. Muscle and Joint Mechanics. Retrieved from smpp.northwestern.edu/bmec66/weightlifting/emgback.html on 9 Oct 2013

ANN Diagnosis for Defect Detection and Classification in Two-Layer Printed Circuit Boards Using Supervised Back-Propagation Algorithm

Rionel Belen Caldo

Abstract In this work, the proponent makes use of Artificial Neural Network (ANN) to visually inspect and classify the defect found in two-layer Printed Circuit Boards (PCBs). The proponent trained and tested the data for pattern recognition using C language. The supervised back-propagation learning algorithm was used for training and testing of PCB patterns. This learning algorithm is suitable for training multi-layered neural network and for generating the deltas of all output and hidden neurons. Considering that training and testing the data would only provide outputs with respect to generated weights, the proponent makes use of another program for defect detection. Excel VBA macro program was used for commonality testing of actual versus expected outputs. Also, it was used in making PCB defect detection possible by marking each defective unit. The proponent modeled a bare PCB circuit with 80 × 44 dimensions. The PCB board was further divided into 32 panel sides, each with 10 × 11 dimensions. There were five defective units modeled in the first layer and there were 14 classified defects used in the second layer. These data were trained and tested successfully, accurately and reliably using ANN.

Keywords Artificial neural network · Pattern recognition · Visual inspection · Defect detection · Defect classification · Printed circuit board · Excel VBA

R.B. Caldo (✉)
College of Engineering and Computer Studies, Lyceum of the Philippines University – Laguna (LPU-L), Laguna, Philippines
e-mail: rionel14_caldo@yahoo.com

R.B. Caldo
Electronics and Communications Engineering, De La Salle University (DLSU), Manila, Philippines

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_50

1 Introduction

Accordingly, PCB manufacturing is the backbone of electronic manufacturing industry [1]. By definition, the Printed Circuit Board (PCB) is a thin board consisted of fiberglass with electrical wires printed onto the board. It could be likened to electrical wires, connecting the central processor with other components on the board. Traditionally, in PCB inspection, the image of the bare PCB is being captured and inspected manually and visually in finding the defects [2]. Manual inspection, which is the traditional method of inspecting PCB, is ineffective due to limited human resources and speed constraints [1]. Considering that inspection of solder joints is a critical process in the electronic manufacturing industry, this manual method seemed to be risky. If the defects were not detected concurrently, the reduction of manufacturing cost, yield improvement, quality products and its reliability are at risks [3]. In giving solution to problems encountered in human's manual inspection method, automated visual inspection is being offered. This method would eliminate subjective judgments and would provide fast, quantitative and accurate dimensional assessments [4]. Significantly, inspection process is a must for modern manufacturing environment. In fact, mass-production manufacturing facilities of various electronic industries would require their people to achieve 100 % quality assurance of all parts, subassemblies and finished goods.

Recently, there have been a lot of works and research endeavors concentrated on PCB defect detection. PCB defects detection is a must to verify PCB characteristics. This defect detection would ensure that the PCB design is in conformity with its desired specifications [5].

In the study entitled, *"Automatic PCB Inspection Algorithms: A Survey"*, the proponents examined different algorithms and techniques used for automated inspection of PCBs. The proponents make use of classification tree in grouping the algorithms according to its classification. The proponents concentrated on image analysis, fault detection strategies and state-of-the-art techniques [6].

The authors in the study, *"Neural Network Diagnosis for Visual Inspection in Printed Circuit Boards"*, presented an automatic optical inspection system in diagnosing PCBs mounted in Surface Mounting Technology (SMT). The proponents make use of neural network and they re-processed PCB tested images in reducing the amount of data to be fed unto the neural net. They concluded that the Automatic Optical Inspection system provides low cost solution with very fast diagnosis and easy set-up, which is beneficial in industrial applications [7]. According to the study conducted by Khalid et al. in [5], in spite of the need to detect defects, it is also a must to classify these defects so that the source of these defects can be identified and prevented. The latter statement has been neglected and not given much attention. Hence, the proponents of the study proposed an algorithm in grouping the defects found on the bare PCB. Using synthetically generated PCB image, their algorithm was able to group 14 commonly known PCB defects into five groups. The algorithm they have proposed include image subtraction, image adding, logical XOR and NOT, and flood fill operator [5].

Mar et al. in [3], proposed two inspection modules: “front-end” and “back-end” inspection modules. The front-end includes illumination normalisation, localisation and segmentation. On the other hand, back-end includes the classification of solder joints using the Log-Gabor filter and classifier fusion [3]. On the other hand, in the study, “*A Visual Inspection System for Surface Mounted Devices on Printed Circuit Board*”, the proponents divided the inspection process into two stages: screening stage and classification stage. In screening stage, one image feature is abstracted from the examined image and is used as a screening index. In the second stage, the proponents make use of neural networks in integrating all available image feature information. Moreover, parting coefficient was used as an index for selecting proper image features. The system was trained using set of revised image data and the images collected from production line were used to test the trained system [8]. In [2], Anu et al. proposed a system where the reference image is obtained from a CAD file. In finding the defects, portion of the reference image is selected and cropped. Using template matching, same portion is obtained from the inspection image. Then, image subtraction is performed and from the subtracted image, the defects were then identified. The defects were extracted using feature extraction method. Using shape analysis, the standard defects are colored and displayed [2]. Ibrahim et al. in [4], improved the PCB inspection process by incorporating a geometrical image registration, minimum thresholding technique and median filtering in solving alignment and uneven illumination problem. There were six types of defects being identified by the system: missing hole, pin hole, under-etch, short-circuit, mouse-bite and open-circuit [4]. According to Choksi et al. in [1], image difference operation is preferred to be used in automated PCB inspection system as well as in many other image processing applications. The inspection of PCB consists mainly of misplaced or missing components in the PCB. If there is any missing electronic component, then it is not so damaging the PCB. But, if any of the component that can be placed only in one way and has been soldered in other way around, then the same will be damaged and there are chances that other components may also get damaged. In eliminating this to happen, PCB inspection is demanded to take care of the missing or misplaced electronic components. In their research work, the proponents proposed an efficient algorithm that detects and locates any defect found on PCBs. It was improved using geometrical image registration, minimum thresholding technique and median filtering. The defect classification operation was used to enhance the image difference operation in terms of computation time using wavelet transform techniques [1]. In this study, the proponent sought to use C program language and Excel VBA macro program in training and testing the data for defect detection and classification using ANN. Literatures, on the other hand, cited the use of image processing, Matlab and CAD in processing their data. Also, most of the studies conducted make use of one layer PCB. In this work, the proponent will make use of two-layer PCB. The bare PCB caters 14 known defects.

2 Methodology

In this work, the proponent makes use of Artificial Neural Network (ANN) in defect detection and classification for two-layer PCBs. Specifically, the supervised back-propagation learning algorithm was used for pattern recognition in this study. This algorithm would require a known desirable output for each input value in obtaining the loss function gradient. The main objective of this algorithm is to train a multi-layered neural net, which could learn appropriate internal representations. Thus, allowing the system to learn varied maps of input to output representations [9]. The general block diagram is shown in Fig. 1 and it was further simplified in Fig. 2.

In this work, the proponent presented a visual inspection system using pattern recognition, for which the physical device to be inspected is the PCB and the classifier is neural network. The purpose of the proposed system is to automatically detect and classify occurrence/s of defect/s, which can be recognized using manual visual inspection. As shown in Fig. 1, the system is consisted of five processes. The first two processes is purely concern in training and testing the data. Upon completing the latter two processes, the proponent will test the data for defective PCB

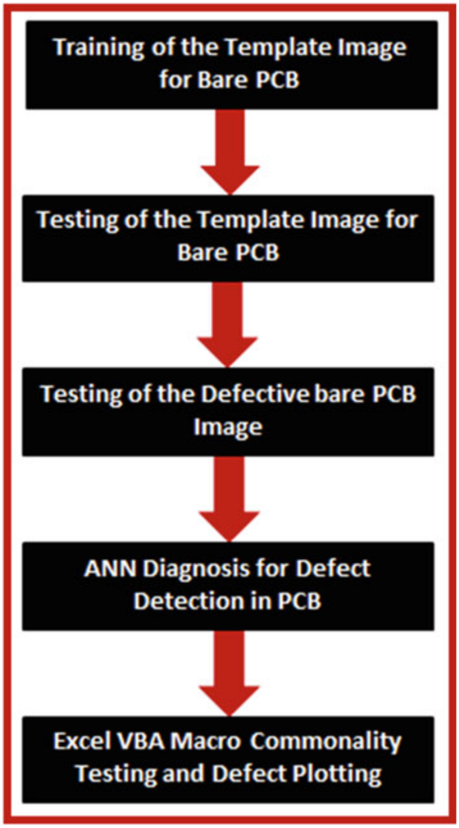


Fig. 1 Proposed ANN PCB diagnostic system

units. The computational tool employed by the proponent in this study is Artificial Neural Network. ANN will be used for detecting defects. Also, the proponent makes use of Excel VBA macro program for commonality testing. This Excel VBA macro program will be used to classify the defects based on the ANN training results. The weights and outputs generated by ANN diagnostic system using C language will be fetched and interpreted directly by the macro program. The macro program will give the correlation between the actual and expected outputs. The statistics can be plotted on a line graph and the defect can be plotted on a spare PCB unit for visual view of the results.

Figure 2 is the simplified block diagram of the system, which is comprised of three major sections namely: Image acquisition/registration, inspection and detection processes. The template and defective bare PCB images will be registered unto the ANN system for training and testing purposes, the learning process takes place

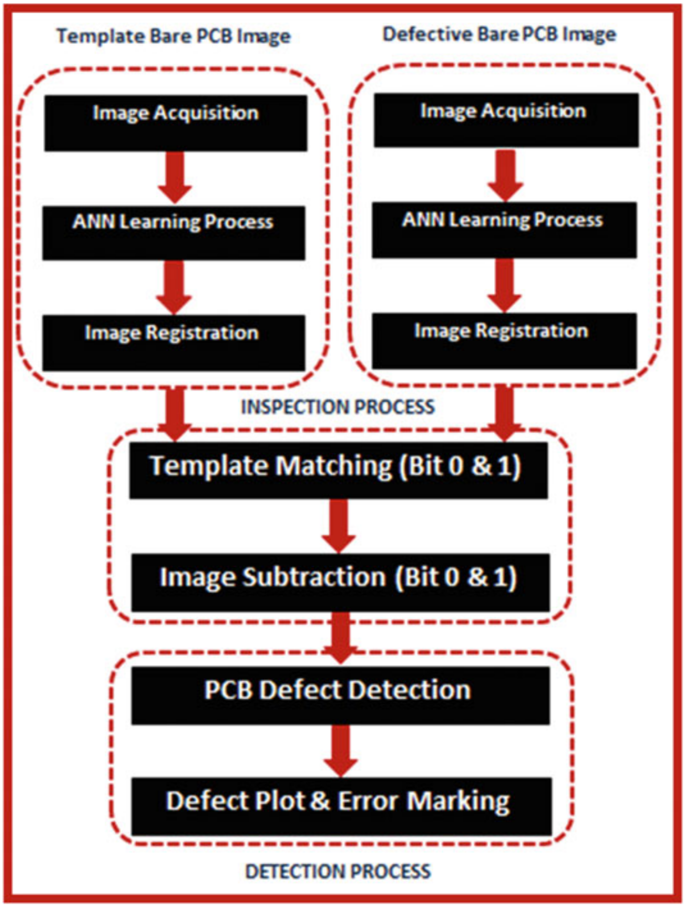


Fig. 2 Simplified block diagram of PCB inspection and detection system

in this stage. The training and testing of data samples is vital and essential for defect inspection and detection. The next stage, which is the inspection process, includes template matching and image subtraction. The NOT operator will be used to match and to subtract the tested image with that of the trained image. Each pattern will be represented with bits ‘0’ and ‘1’. The final stage, which is the detection process, includes defect detection, defect plot and error marking. The defect detection would show statistics for conformity with specs while the defect plot and error marking process would show visually the ANN diagnostic results. The last section incorporates classifying the defects.

3 Design Considerations

Figures 3, 4 and 5 show the template image for PCB bare circuit, PCB pattern for training data, and PCB training data with 32 panels respectively. In this study, the proponent makes use of the template image for PCB bare circuit used by Moganti et al. and Khalid in 2008 [5, 6]. The template image for bare PCB was used for training the data with respect to desired output patterns labeled as panel 1 to panel 32. Panels were labeled horizontally as shown in Fig. 4. The PCB template was further divided into 32 panels for ease of training the data. As for the defective

Fig. 3 Template image for PCB bare circuit

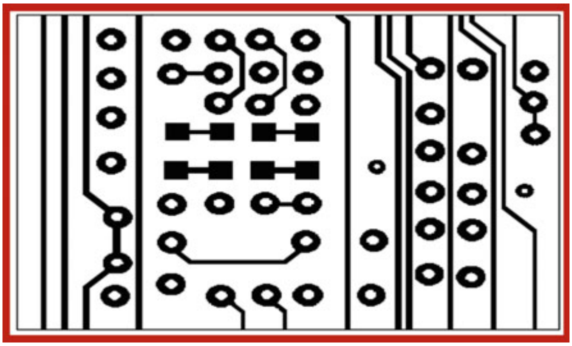
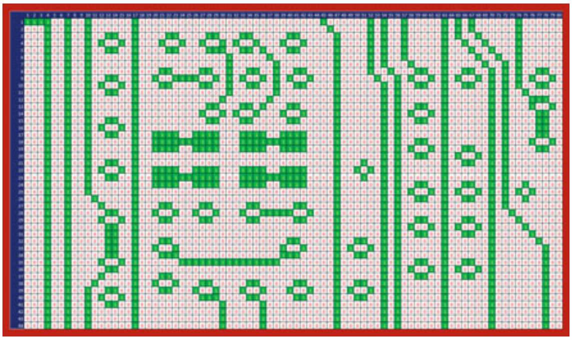


Fig. 4 PCB pattern for training data



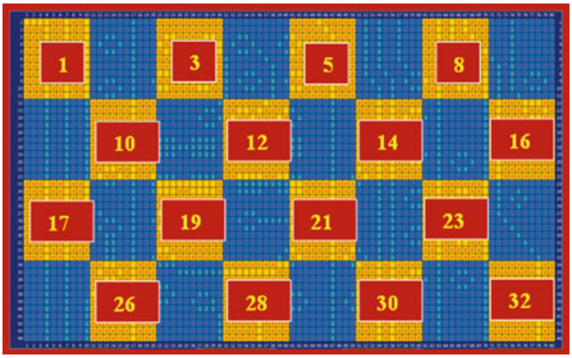


Fig. 5 PCB training data with 32 panels

image of the bare PCB, the proponent constructed five defective images for testing purposes (will be used for layer 1 of the PCB) as elicited in Figs. 6, 7, 8, 9 and 10. The proponent did not consider specifying the classification of defect for layer 1 of the PCB; rather nonconformance with PCB desired specifications were the ultimate basis of testing. The proponent represented image in black with bit of '1' and blank image with bit of '0'.The proponent considered matching the template image for bare PCB with the testing data with defects. Basically, the NOT operator was used in determining the commonality and/or nonconformity of the actual and the expected PCB design.

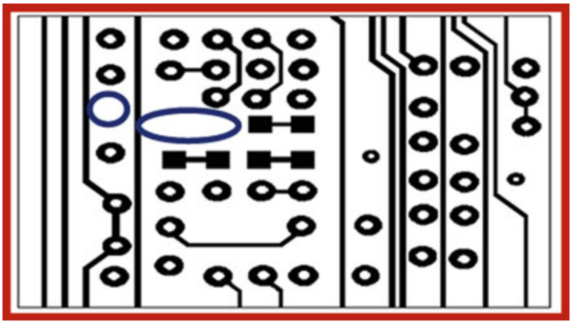


Fig. 6 Defective PCB image 1

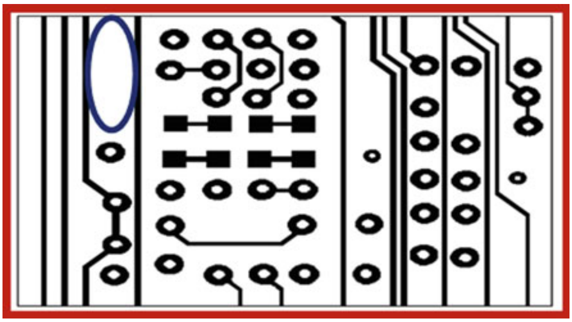


Fig. 7 Defective PCB image 2

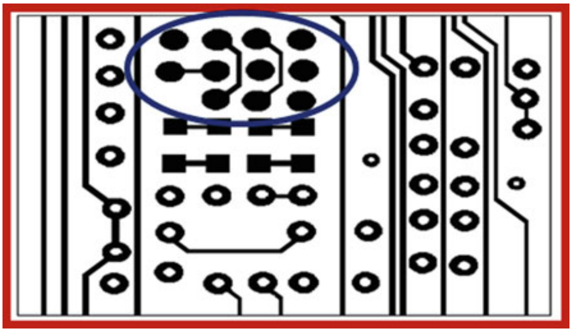


Fig. 8 Defective PCB image3

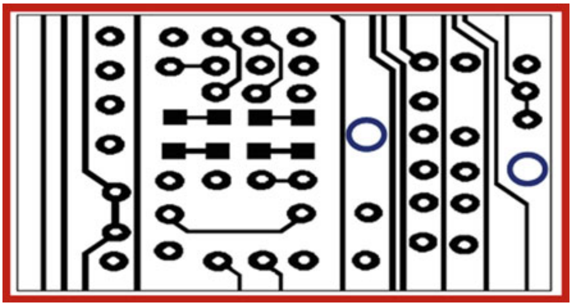


Fig. 9 Defective PCB image 4

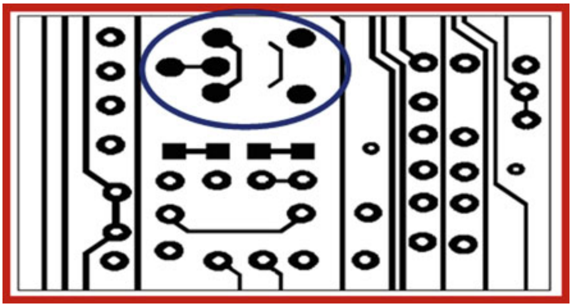


Fig. 10 Defective PCB image 5

For testing purposes, the proponent makes modifications of the bare PCB circuit pattern. As for Fig. 6, the defect is the missing H-pattern and third circle on the upper-left side of the panel. For Fig. 4, there were three missing circles on the top-left side of the panel. On the third elicited defect, there were 11 circles on top, which were soldered unintentionally. It should not be soldered completely, leaving the center open. On the fourth defect, two small circles were missing. Finally, on the last defect exhibited in Fig. 10, there were defects on missing pattern and unexpected soldering. Like the PCB pattern for training data in Fig. 4, these defective images will be represented in terms of bit ‘1’ and ‘0’. Also, these testing defects will be fed to the neural net and the system should be able to recognize these defects.

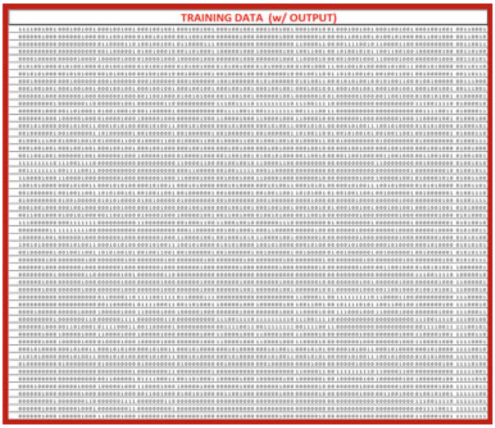
As for the layer 2 of the PCB, the proponent modeled the defective PCB patterns used by Moganti et al. and Ibrahim et al. in 2012 [4, 6] in bits of 0's and 1's. The classifications of defects used were as follows: 1. breakout, 2. pin hole, 3. open circuit, 4. Under-etch, 5. mouse-bite, 6. missing conductor, 7. spur, 8. short, 9. wrong hole, 10. conductor too close, 11. spurious copper, 12. excessive short, 13. missing hole and 14. over-etch. In this study, these defects were used for identification and classification of the defects in the second layer of the PCB.

4 Experiment Results

Table 1 shows the ANN training data used in this study. There were 53 patterns trained for layer 1. The first 32 patterns represent the PCB bare pattern discussed in Figs. 3 and 4. The trailing 21 patterns represent the five defective units (D1 to D5) discussed in Figs. 6, 7, 8, 9 and 10. There were two defective panels in D1 and D2, six defective panels in D3, four defective panels in D4 and seven defective panels in D5. As for testing the data, the proponent makes use of the training data, but without the expected output.

In the second layer, the proponent trained and tested data for 46 patterns. The first 32 patterns represent the PCB bare circuit. The trailing 14 patterns represent the 14 classification of defects as shown in Fig. 11. Moreover, in simulating the ANN diagnosis for defect detection in PCBs, the proponent conducted several trainings and tests for different parameters. Truly, this is the very crucial part in ANN implementation. Based on the obtained results, the proponent was able to get an optimized set of parameters, which would give the best and the ideal ANN simulations as it garnered 100 % statistics when the actual and the expected outputs were compared.

Table 1 ANN training data (input and output)


TRAINING DATA (w/ OUTPUT)

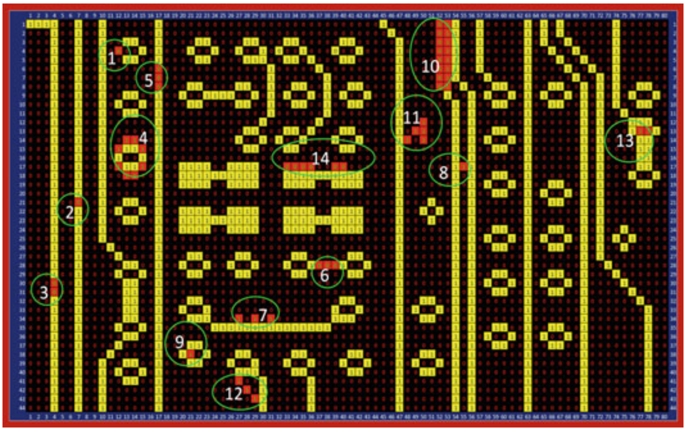


Fig. 11 PCB with 14 classified defects

Listed below are the parameters used by the proponent. Same number of layers and size for each layer parameters were used in testing the bare pattern and in verifying with five defective units (first layer) and 14 classified defects (second layer).

Error tolerance: **0.01**
Learning parameters: **0.1**
*Momentum paramete*r: **0**
Noise factor: **0**
Maximum cycles: **5000**
Number of layers: **3 [has 1 hidden layer]**
Size for each layer: **110 110 8**

As for training data, there were 265,000 patterns formulated. For each input vector, the output vector is represented with comparison to expected output vector as tabulated in Table 2. As for the testing data, on the other hand, there were 53 patterns observed. Similarly, the input vector is represented with respect to its corresponding output vector as tabulated in Table 3.

It could be analyzed in Tables 2 and 3 that the data being trained for layers one and two were accurate. Thus, the training is strong enough to detect occurrence/s of defect/s. The proponent rounded off the actual outputs and these were compared individually with its expected outputs for different samples using Excel VBA macro program. In further assessing the robustness of the system, the percent error per bit was computed (without rounding off) and the mean of 8 bits for each sample were plotted in Figs. 12 and 13. Figures 12 and 13 present the graphical representation of training results. Figure 12 justifies that the training for layer 1 gives reliable results as the highest obtained percent error is only **3.15 %** and the average error is only **0.52 %.** On the other hand, Fig. 13 justifies that the training for layer 2 gives reliable results as the highest obtained percent error is only **4.40 %** and the average percent error is only **0.72 %**.

Table 2 ANN training data for first layer

S.No	ANN training data commonality analysis				
	Code assignment	Panel count (location in the PCB)	Actual output	Expected output	Detection rate 100 %
1	1	1	00110001	00110001	1
2	2	2	00110010	00110010	1
3	3	3	00110011	00110011	1
4	4	4	00110100	00110100	1
5	5	5	00110101	00110101	1
6	6	6	00110110	00110110	1
7	7	7	00110111	00110111	1
8	8	8	00111000	00111000	1
9	9	9	01000010	01000010	1
10	10	10	01000011	01000011	1
11	11	11	01000010	01000010	1
12	12	12	01000011	01000011	1
13	13	13	01000100	01000100	1
14	14	14	01000101	01000101	1
15	15	15	01000110	01000110	1
16	16	16	01000111	01000111	1
17	17	17	01001000	01001000	1
18	18	18	01001001	01001001	1
19	19	19	01001010	01001010	1
20	20	20	01001011	01001011	1
21	21	21	01001100	01001100	1
22	22	22	01001101	01001101	1
23	23	23	01001110	01001110	1
24	24	24	01001111	01001111	1
25	25	25	01010000	01010000	1
26	26	26	01010001	01010001	1
27	27	27	01010010	01010010	1
28	28	28	01010011	01010011	1
29	29	29	01010100	01010100	1
30	30	30	01010101	01010101	1
31	31	31	01010110	01010110	1
32	32	32	01010111	01010111	1
33	*D1*	*10*	*10000001*	*10000001*	*1*
34	*D1*	*11*	*10000010*	*10000010*	*1*
35	*D2*	*2*	*11000001*	*11000001*	*1*
36	*D2*	*10*	*11000010*	*11000010*	*1*
37	*D3*	*3*	*11100001*	*11100001*	*1*
38	*D3*	*4*	*11100010*	*11100010*	*1*

(continued)

Table 2 (continued)

S.No	ANN training data commonality analysis				
	Code assignment	Panel count (location in the PCB)	Actual output	Expected output	Detection rate 100 %
39	*D3*	*5*	*11100011*	*11100011*	*1*
40	*D3*	*11*	*11100100*	*11100100*	*1*
41	*D3*	*12*	*11100101*	*11100101*	*1*
42	*D3*	*13*	*11100110*	*11100110*	*1*
43	*D4*	*13*	*11110001*	*11110001*	*1*
44	*D4*	*14*	*11110010*	*11110010*	*1*
45	*D4*	*22*	*11110011*	*11110011*	*1*
46	*D4*	*24*	*11110100*	*11110100*	*1*
47	*D5*	*3*	*11111001*	*11111001*	*1*
48	*D5*	*4*	*11111010*	*11111010*	*1*
49	*D5*	*5*	*11111011*	*11111011*	*1*
50	*D5*	*10*	*11111100*	*11111100*	*1*
51	*D5*	*11*	*11111101*	*11111101*	*1*
52	*D5*	*12*	*11111110*	*11111110*	*1*
53	*D5*	*13*	*11111111*	*11111111*	*1*

Table 3 ANN training data for second layer

S.No	ANN training data commonality analysis				
	Code assignment	Panel count (location in the PCB)	Actual output	Expected output	Detection rate 100 %
1	1	1	00110001	00110001	1
2	2	2	00110010	00110010	1
3	3	3	00110011	00110011	1
4	4	4	00110100	00110100	1
5	5	5	00110101	00110101	1
6	6	6	00110110	00110110	1
7	7	7	00110111	00110111	1
8	8	8	00111000	00111000	1
9	9	9	01000010	01000010	1
10	10	10	01000011	01000011	1
11	11	11	01000010	01000010	1
12	12	12	01000011	01000011	1
13	13	13	01000100	01000100	1
14	14	14	01000101	01000101	1
15	15	15	01000110	01000110	1
16	16	16	01000111	01000111	1
17	17	17	01001000	01001000	1

(continued)

Table 3 (continued)

S.No	ANN training data commonality analysis				
	Code assignment	Panel count (location in the PCB)	Actual output	Expected output	Detection rate 100 %
18	18	18	01001001	01001001	1
19	19	19	01001010	01001010	1
20	20	20	01001011	01001011	1
21	21	21	01001100	01001100	1
22	22	22	01001101	01001101	1
23	23	23	01001110	01001110	1
24	24	24	01001111	01001111	1
25	25	25	01010000	01010000	1
26	26	26	01010001	01010001	1
27	27	27	01010010	01010010	1
28	28	28	01010011	01010011	1
29	29	29	01010100	01010100	1
30	30	30	01010101	01010101	1
31	31	31	01010110	01010110	1
32	32	32	01010111	01010111	1
33	*D1*	*2*	*10000001*	*10000001*	*1*
34	*D2*	*9*	*10000010*	*10000010*	*1*
35	*D3*	*17*	*11000001*	*11000001*	*1*
36	*D4*	*10*	*11000010*	*11000010*	*1*
37	*D5*	*2*	*11100001*	*11100001*	*1*
38	*D6*	*20*	*11100010*	*11100010*	*1*
39	*D7*	*27*	*11100011*	*11100011*	*1*
40	*D8*	*14*	*11100100*	*11100100*	*1*
41	*D9*	*27*	*11100101*	*11100101*	*1*
42	*D10*	*6*	*11100110*	*11100110*	*1*
43	*D11*	*13*	*11110001*	*11110001*	*1*
44	*D12*	*27*	*11110010*	*11110010*	*1*
45	*D13*	*16*	*11110011*	*11110011*	*1*
46	*D14*	*12*	*11110100*	*11110100*	*1*

5 Discussion and Analysis of Results

In this section, the proponent discusses the testing results of the study. Figures 14, 15, 16, 17 and 18 elicit the ANN Testing for five defective units. As mentioned, for D1 and D2, there were two defects being observed (panels 10 and 11) and (panels 2 and 10) respectively. For D3, there were six defects (panels 3 to 5 and 11 to 13). Four defective panels were traced in D4 (panels 13, 14, 22 and 24). Lastly, there were seven defective panels in D5 (panels 3 to 5 and 10 to 13). As for the defect rate, the

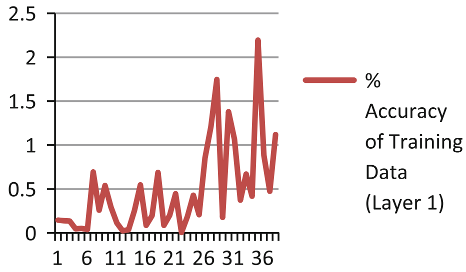


Fig. 12 Percent accuracy of training data with 21 defects (Layer 1)

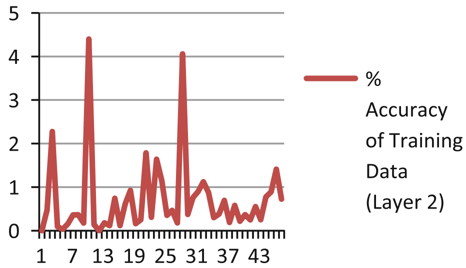


Fig. 13 Percent accuracy of training data with 14 defects (Layer 2)

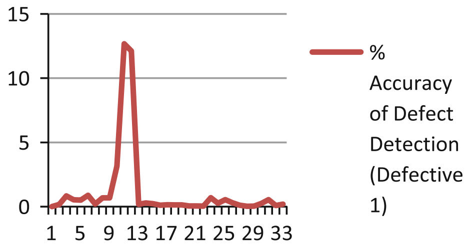


Fig. 14 Percent accuracy of defect detection for testing data (defective 1)

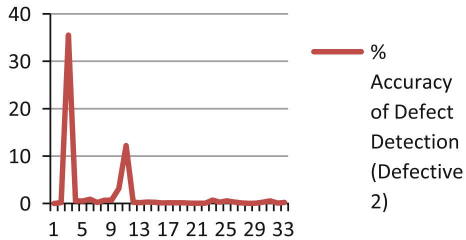


Fig. 15 Percent accuracy of defect detection for testing data (defective 2)

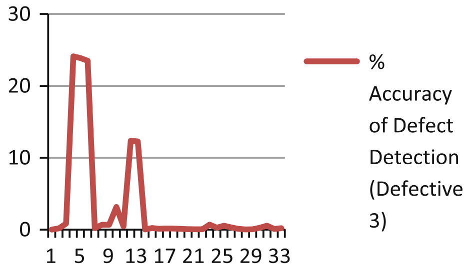


Fig. 16 Percent accuracy of defect detection for testing data (defective 3)

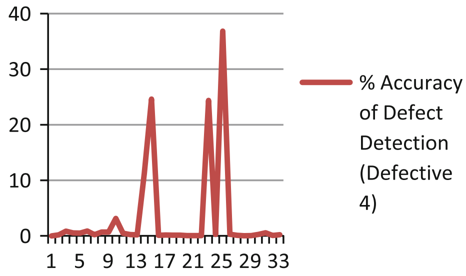


Fig. 17 Percent accuracy of defect detection for testing data (defective 4)

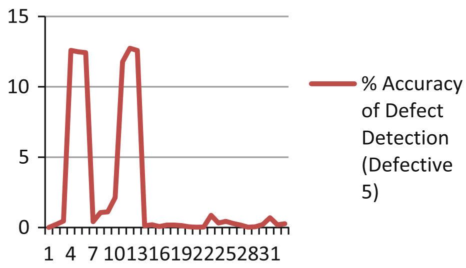


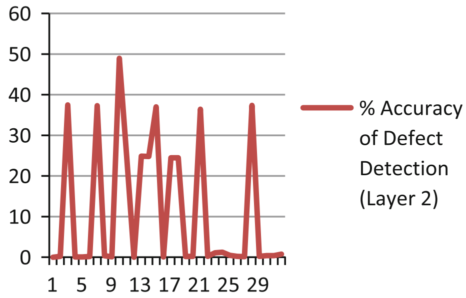
Fig. 18 Percent accuracy of defect detection for testing data (defective 5)

detection for each defective unit is as follows: **93.75 %** for defectives 1 and 2, **81.25 %** for defective 3, **87.5 %** for defective 4 and **78.125 %** for defective 5.

On the other hand, Fig. 19 shows the ANN testing for adopted 14 classifications of defects. As for the testing, for D1, panel 4 is defective. For D2, the defective panel is panel. For D3, panels 8 and 9 were affected. Same defect detection and localization applies to D4 to D14. As shown in the figure, different panels were being affected for different classification of defects. As for the defect rate, the detection for 14 defects is **65.625 %**. This considers two defects affecting panel 2 and three defects affecting panel 27.

Based on the experiments performed, it could be analyzed that the data being tested for layers one and two gives 100 % accurate defect detection. The system was able to detect and classify occurrences of defects. Like in trained data, the proponent assessed the robustness of the system by getting the percent error per bit

Fig. 19 Percent accuracy of defect detection for testing data with 14 defects



(without rounding off) and the mean of 8 bits for each sample. It was realized that the system gives favorable results. Figure 14 justifies that the testing data for defective 1 gives reliable results as it obtained reliable defect detection at **12.67** and **12.12 %** and the average defect detection is **12.40 %**. Figure 15 for defective 2 also gives reliable results at **35.5** and **12.21 %** defect detection and **23.86 %** on the average. Similarly, defective 3–5 gives favorable results, giving average defect detection **48.07**, **48.57** and **37.37 %**. Lastly, for the testing of the second layer, the highest defect detection is **48.96 %**, the lowest detection is **24.46 %** and the average defect detection is **32.54 %**.

6 Conclusions and Future Directives

In this paper, the proponent adopted the supervised back-propagation learning algorithm using C language for training and testing the data of the modeled PCB bare pattern. Also, in providing visualization of results, the proponent makes use of Excel VBA macro program. This high-level language was used in migrating the generated text file to a more refine data representation for ease of analysis. The proponent was able to detect the pre-assigned and trained defective units using the desired parameters being set using trial and error method. There were two layers of PCB studied and tested. The first layer of the PCB makes use of the defective units modeled by the proponent. On the other hand, in the second layer of the PCB, the proponent adopted 14 classified defects. The system was able to automatically detect occurrence/s of defect/s with 100 % accuracy for two layers of PCB using ANN. Also, the robustness and reliability of the system were tested by computing the percent error per bit (without rounding off) and the mean of 8 bits for each data sample. The total average defect detection for reliability tests of layers 1 and 2 are **34.05** and **32.54 %** respectively. This justifies that the system is capable of detecting the occurrences of defects accurately and reliably.

Although there are many works in this field especially the use of PCB, the novelty of the study applies to the use of C program language and Excel VBA macro program in defect detection and classification. As for the new insights on this field, the proponents make use of two-layer PCB circuit. In future researches, the proponent will use actual layers of PCBs and the images will be captured real time using Matlab. The proponent will also try other computational tools such as fuzzy logic and evolutionary algorithms.

Acknowledgments The Lord and Savior, Jesus Christ is gratefully acknowledged in the entire workings of the study. The proponent would want to give thanks as well to De La Salle University for supporting his doctorate study. Lyceum of the Philippines University—Laguna (LPU-L) is acknowledged as well for research funding.

References

1. Choksi, A.H., Vashi, R., Sevak, M., Patel, K.: Printed circuit board defect detection using wavelet transform. Int. J. Curr. Eng. Technol. 2014
2. Anu, A., Anouncia, M., Rajaji, L.: Automatic visual inspection of PCB using CAD information. In: International Conference on Radar, Communication and Computing (ICRCC), Tiruvannamalai, TN., India, 21–22 Dec 2012
3. Mar, N.S.S., Yarlagadda, P.K.D.V., Fookes, C.: Design and development of automatic visual inspection system for PCB manufacturing. Rob. Comput. Integr. Manuf. 2011
4. Ibrahim, I., Abdul Rahman, S., Abu Bakar, S., MohdMokji, M., Ahmed Mukred, J.A., MdYusof, Z., Ibrahim, Z., Khalil, K., Mohamad, M.S.: A printed circuit board inspection system with defect classification capability. Int. J. Innovative Manage. Inf. Prod. **3**(1) 2012
5. Khalid, N.K., Ibrahim, Z., Zainal Abidin, M.S.: An algorithm to group defects on printec circuit board for automated visual inspection. IJSSST. **9**(2) 2008
6. Moganti, M., Ercal, F., Dagli,C.H., Tsunekawa, S.: Automatic PCB inspection algorithms: a survey [Online]. Accessed 6 Feb 2015
7. Fanni, A., Lera, M., Marongiu, E., Montisci, A.: Neural network diagnosis for visual inspection in printed circuit boards [Online]. Accessed 26 Feb 2015
8. Lin, S.-C., Su, C.-H.: A visual inspection system for surface mounted devices on printed circuit board. In: CIS 2006
9. McCaffrey, J.: Neural Network Back-Propagation for Programmers. MSDN Magazine, Oct 2012 Issue

An Intelligent PCB Visual Inspection System for Defect Detection and Localization in Excel VBA Macro

Rionel Belen Caldo

Abstract An intelligent system using Excel VBA macro program was made in this research. The proponent modeled a bare Printed Circuit Board (PCB) pattern used by Moganti et al. and Khalid in 2008. The proponent represented this PCB bare circuit in 80 × 44 dimensions with bits of '0-blank image' and '1-filled/black image'. The PCB pattern was further divided into 32 panels in 10 × 11 dimensions with same bit representation. This bare pattern was compared with defective units for template matching using logical operators. The proponent considered building the system in high-level programming type of language. Considering that the system's algorithm would require the use of Graphical User Interface (GUI) for visualization purposes. Excel VBA macro program, in this accord, was preferred by the proponent as it provides powerful tools both in GUI constructions and in Microsoft Excel applications. As the study aims to provide a more systematic way of inspecting PCB for defect detection, the system is built with functionalities capable of handling and mimicking the tasks of PCB visual inspector. The system would not only detect occurrence/s of defect/s, rather it would provide defect analysis by localizing the exact location of the defect on a specific panel and assessing defect likelihoods using repository and trend chart.

Keywords Artificial intelligence · Pattern recognition · Visual inspection · Defect detection · Printed circuit board · Excel VBA

R.B. Caldo (✉)
College of Engineering and Computer Studies, Lyceum of the Philippines University-Laguna (LPU-L), Laguna, Philippines
e-mail: rionel14_caldo@yahoo.com

R.B. Caldo
Electronics and Communications Engineering, De La Salle University (DLSU), Manila, Philippines

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_51

1 Introduction

Designing and creating an intelligent visual inspection system for automatic Printed Circuit Board (PCB) defect detection and localization is indeed becoming not just a need, but a must for PCB electronic industries. Accordingly, PCB visual inspection in general term contributes the largest manufacturing cost, considering that these industries aimed at producing 100 % quality products. However, production of zero defects or integrity issues in terms of producing non-defective unit is an impossible target. Indeed, no production system is capable of meeting such unrealistic target. As a matter of fact, as of date, PCB designs are becoming more and more complex. In statistical point of view, as we increase the number of components inside PCB, we also increase the defect rate. For these reasons, the proponent considered providing viable solution in the final testing stage, considering that in this stage, defective units are being detected. The PCB will have to be repaired or rework whenever defective units were observed. What will be developed in this research work is a decision-helping tool using an intelligent expert system. Linguistics rules shall be created in analyzing the functional tests of data samples. This study will give focus as well on the localization of the defective component.

The proponent gathered relevant literatures and tabulated reviews as shown in Table 1. In the table, the serial number, research title, proponents' name, year of publication and algorithms used were itemized. As for the synthesis of reviewed literatures, most of the proponents make use of image processing applications in PCB defect detection, localization and classification [1, 2, 3, 4]. Also, computational intelligence such as fuzzy logic [5] and neural network [6, 7] were integrated and made visible on the defect analysis. The program applications used are Matlab and Computer Aided Design (CAD)[2]. Most of the researches focused on the "front end" inspection for detecting and localizing defects [8]. Though, "back end" inspection was considered as well [8]. Classification tree was used as well for image analysis [9]. Most of the researches make use of a bare PCB circuit for template matching. Accordingly, classifying the type of defects is becoming a trend in current researches dealing with PCB visual inspection system [10, 4]. As defined by Gebus et al. in [5], expert system is a program having a huge amount of knowledge about a specific topic. By gathering this knowledge obtained from identified experts, the proponent will be able to create a system as efficient as those identified experts. In this case, the proponent will model the visual inspector as expert for defect detection and localization. Accordingly, knowledge can be divided into two forms: knowledge about the environment and the knowledge about interactions between the previous facts. As for the environment, the proponent considered getting information on different defects. As for the interactions between previous facts, the proponent will generate and establish linguistic rules and algorithms in making decisions for defect detection and localization using Excel VBA macro program. Basically, the proponent will be collecting essential data based from related literatures and he will be using conditional statements in the process in lieu of detecting and localizing defects found in PCB [5]. The highlights and main

Table 1 Tabular representation of literatures for PCB visual inspection system

S. No	Research title	Proponents	Year	Algorithms used
1	Printed circuit board defect detection using wavelet transform	Amit H. Choksi, Ronak Vashi, Mayur Sevak and Kaushal Patel	2014	Wavelet decomposition, thresholding, image difference operation
2	PCB defect detection, classification and localization using mathematical morphology and image processing tools	P.S. Malge, R.S. Nadaf	2014	Morphological image segmentation algorithm, simple images processing theories
3	Automatic visual inspection of PCB using CAD information	A. Anu, Margret Anouncia and L. Rajaji	2012	Reference image, inspection image, template matching, image subtraction and shape analysis
4	Design and development of automatic visual inspection system for PCB manufacturing	N.S.S. Mar, P.K.D.V. Yarlagadda, C. Fookes	2011	Segmentation of solder joint, Classification of solder joint, Log-Gabol filter and classifier fusion
5	A printed circuit board inspection system with defect classification capability	Rahman Syed Abu Bakar, Musa Mohd Mokji, Jameel Abdulla Ahmed Mukred, Zulkifli Md Yusof, Zuwairie Ibrahim, Kamal Khalil, Mohd Saberi Mohamad	2012	Image subtraction, thresholding, image registration
6	An Algorithm to Group Defects on Printed Circuit Board for Automated Visual Inspection	Noor Khafifah Khalid, Zuwairie Ibrahim, Mohamad Shukri Zainal Abidin	2008	Image subtraction, image adding, logical XOR and NOT and flood fill operator
7	Automatic OCB inspection algorithms: a survey	Madhav Moganti, Fikret Ercal, Cihan H. Dagli, Shou Tsunekawa	N.A.	Classification tree
8	Neural network diagnosis for visual inspection in printed circuit boards	A. Fanni, M. Lera, E. Marongiu, A. Montisci	N.A.	Neural network
9	A visual inspection system for surface mounted devices on printed circuit board	Shih-Chieh Lin and Chia-Hsin Su	N.A.	Screening, classification (neural network), parting coefficient
10	Defect localization on a PCB with functional testing	Sebastian Gebus, Sebastien Lorrilard, Esko Juuso	2002	Fuzzy logic or linguistic equations in fault diagnosis

workings of the study are being discussed in different sections. Section 1 elicits the synthesis of related literatures. Section 2 discusses the methodology of the study. Section 3, on the other hand, describes the details of design considerations. Section 4 contains the analysis and experimental results for defect detection and localization while the conclusion part is described in Sect. 5. Finally acknowledgments and references are given.

2 Methodology

An intelligent system to automatically detect and analyse PCB defects is proposed in this study. The proponent considered using Excel VBA as its preferred platform in program applications and simulations of the visual inspection system. In Fig. 1, the proponent elicits the main flow of the system using rectangular blocks. This was further simplified and detailed in waterfall model as represented in Fig. 2. As shown in Fig. 1, there are two major blocks of processes. On the first two blocks is the representation of PCB bare circuit pattern and defective PCB dummies in bits of '0' and '1'. The remaining blocks were considered to be as the most important process blocks. These blocks include the template matching of the two compared PCB layouts and the utilization of Excel VBA macro program for defect detection and localization.

3 Design Considerations

Considering that this research work is in parallel with the work of the proponent entitled, "*ANN Diagnosis for Defect Detection and Classification in Two-Layer Printed Circuit Boards using Supervised Back-Propagation Algorithm*" in 2015, same template image for PCB bare circuit was used. This template image is commonly used by previous researchers in defect analysis. Also, the PCB pattern used for training data was adopted for ease of data representation. Similarly, bit **0's** and **1's** were used in representing blank and black images respectively (Figs. 3 and 4).

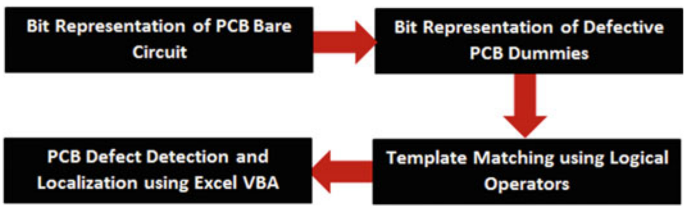


Fig. 1 Main flow of PCB visual inspection system

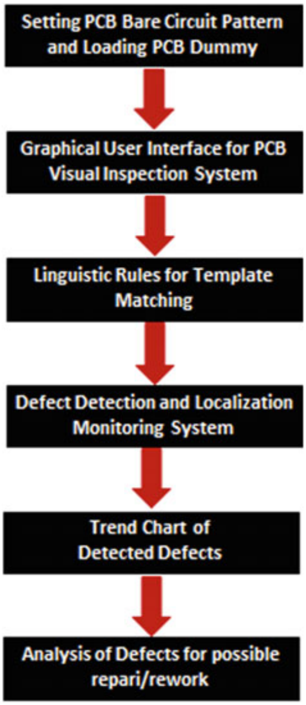


Fig. 2 Process flow of PCB visual inspection system

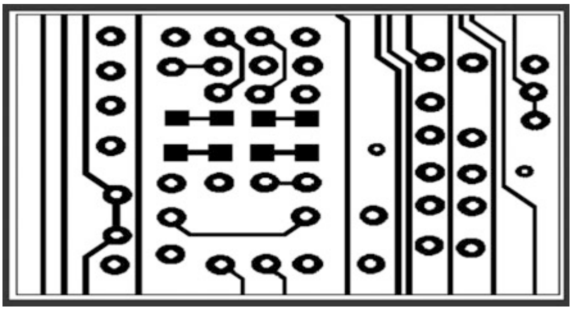


Fig. 3 Template Image for PCB bare circuit

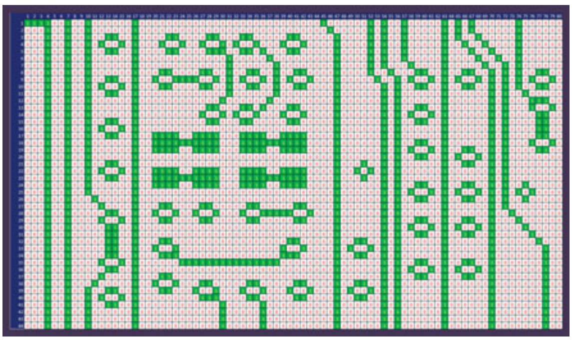


Fig. 4 PCB pattern for training data

Fig. 5 Dummy defective PCB image 1

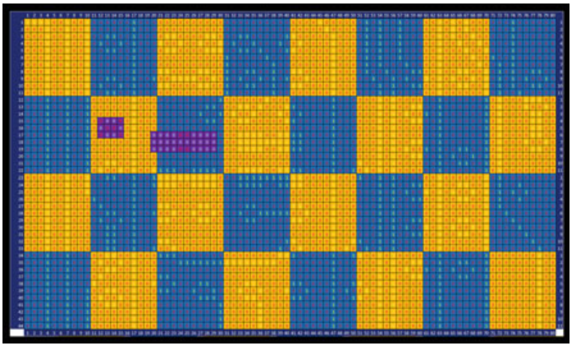
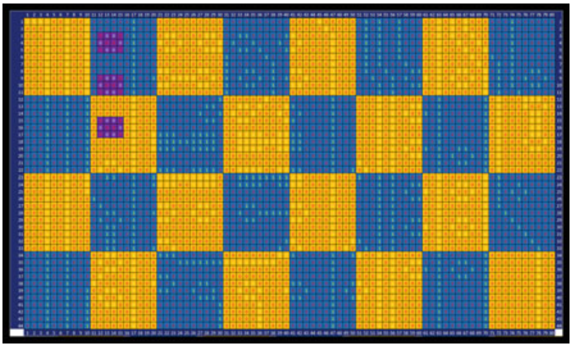
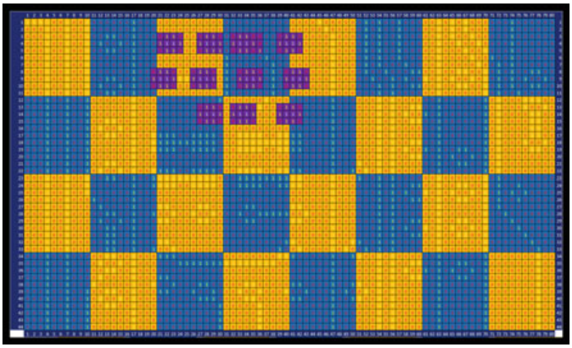


Fig. 6 Dummy defective PCB image 2



The template image for bare PCB used has 80 × 44 dimensions. This was divided in panels 1 to 32. Each panel has equal dimensions of 10 × 11. In this work, the proponent makes use of 10 defective images of bare PCB. The sample or dummy patterns were showcased in Figs. 5, 6, 7, 8, and 9. These dummy defective PCB images are represented in bits of 0’s and 1’s from panel 1 to panel 32. These dummy panels are to be loaded unto the system and it will be compared automatically to the modeled PCB pattern. The defect detection and localization will be

Fig. 7 Dummy defective PCB Image 3



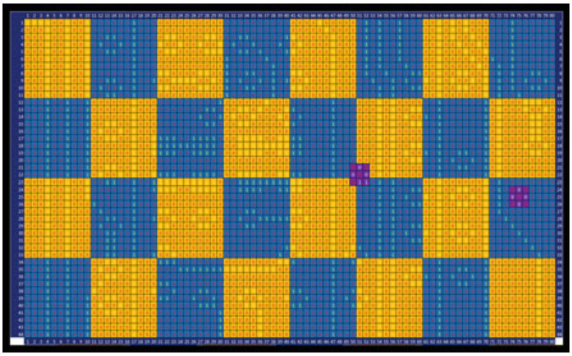


Fig. 8 Dummy defective PCB image 4

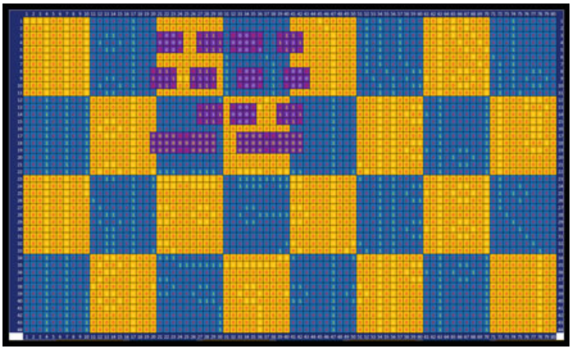


Fig. 9 Dummy defective PCB image 5

stored and will be plotted on a trend chart. Each dummy defective PCB is being controlled using pre-assigned control/lot number. Like in previous work, nonconformance with PCB desired specifications would serve as ultimate basis of testing and analysis. The proponent considered matching the template image for bare PCB with the testing data with defects. Basically, the **logical NOT** and **logical AND** operators will be used in determining the similarity, commonality and/or nonconformity of the actual and the expected PCB designs.

In simulating the system, the proponent provided ten modifications or revisions of the PCB template as tabulated in Table 2. These modifications cater different classifications of defect. This table was updated after dummy defective PCB images were loaded unto the PCB visual inspection system. Like the PCB pattern represented in Fig. 4, these defective images were represented in terms of bit ‘1’ and ‘0’.

Basically, the algorithm to automatically detect and localize defects using Excel VBA macro program is merely a *data comparator*. This data comparator will simply recognize identical values. Here is how the algorithm works. The leftmost XNOR gate compares the template PCB (X_{11}) and the dummy PCB (Y_{11}); if they are the same, Z_{11} gives value of ‘1’. The second XNOR gate compares the template PCB (X_{10}) and the dummy PCB (Y_{10}); if they are the same, Z_{10} gives a value of ‘1’. In turn, the remaining XNOR gates compare the bits that are left, producing a bit of

Table 2 Repository of detected and localized defects

Defect detection and localization monitoring system					
Lot No.	Date	Shift	Operator	No. of defects	Affected panels
P0000001	9-Mar-15	A	Rionel Caldo	8	10 11
P0000002	9-Mar-15	B	Elmer Dadios	8	2 10
P0000003	9-Mar-15	C	Donabel Abuan	9	3 4 5 11 12 13
P0000004	9-Mar-15	D	Imelda Martin	7	13 14 22 24
P0000005	9-Mar-15	N	Rionel Caldo	32	2 3 4 5 10 11 12 13
P0000006	9-Mar-15	A	Elmer Dadios	3	1 21 26
P0000007	9-Mar-15	D	Donabel Abuan	11	1 7 15 21 26
P0000008	9-Mar-15	C	Donabel Abuan	8	13 15 18 26
P0000009	9-Mar-15	D	Imelda Martin	18	13 15 16 18 24 26 32
P0000010	9-Mar-15	N	Rionel Caldo	9	2 10 22 26 29 31

‘1’ output for equal bits and a bit of ‘0’ output for unequal bits. If all data values in **X** and **Y** are identical, all XNOR gates have high outputs and the AND gate has a high *EQUAL*. This means that the dummy PCB is not defective. If words **X** and **Y** differ in one or more bit positions, the AND gate has a low *EQUAL*. This means that the dummy PCB is defective.

The PCB visual inspection system is consisted of five worksheets: the PCB bare circuit, the PCB bare circuit with pattern, the database criteria, the dummy panel and the defect detection monitoring. The PCB bare circuit worksheet elicits the pattern in bits of ‘0s’ and ‘1s’. The PCB circuit with pattern, on the other hand, is simply the extension of the latter worksheet. This extension worksheet provides panel representations of panel 1–32, which will be used in comparison with the dummy panel. Also, this extension worksheet serves as the main template to be used for matching purposes. The database criteria worksheet was used in this study to store the name of the operator and the shifting schedule of data encodes. This was protected and hidden in the program for security reasons. The values stored in this worksheet will be displayed in the main GUI as shown in Fig. 11. The dummy panel and the defect detection monitoring worksheets are the two most important worksheets in this study. Three command buttons are situated in the dummy panel worksheet. This includes ‘Load Dummy PCB’, ‘Dummy PCB Inspection’ and ‘Set by Default the Dummy’, elicited in Fig. 10. The ‘Load Dummy PCB’ is used for loading the dummy PCB panel for template matching. The user will simply place the dummy panel in bits of 1 and 0 in the workspace provided for. On the other hand, the ‘Set by Default the Dummy’, will simply set the workspace in its original state. Lastly, the ‘Dummy PCB Inspection’ seemed to be the busiest among all the buttons, as it does the main workings of the system. Sample source codes for these commands buttons are elicited in the preceding texts.

Fig. 10 Main command buttons for PCB visual inspection system


```
'Sample source code for PCB Visual Inspection System
Sub load_dummy()
actwork = ActiveWorkbook.Name
actsheet = ActiveSheet.Name
mypassword = "2tet"
Workbooks(actwork).Sheets(actsheet).Unprotect mypassword
End Sub
Sub PCB_Inspection()
actwork = ActiveWorkbook.Name
actsheet = ActiveSheet.Name
mypassword = "2tet"
Workbooks(actwork).Sheets(actsheet).Unprotect mypassword
Updating.Show
Workbooks(actwork).Sheets(actsheet).Protect mypassword
End Sub
Sub set_by_default()
actwork = ActiveWorkbook.Name
actsheet = ActiveSheet.Name
mypassword = "2tet"
Workbooks(actwork).Sheets(actsheet).Unprotect mypassword
x = 2
Do While x < 46

y = 2
Do While y < 82
Workbooks(actwork).Sheets(actsheet).Cells(x, y).Select
'This would set the dummy panel in its original state
Call orig_color
y = y + 1
Loop
```

```
x = x + 1
Loop
Workbooks(actwork).Sheets(actsheet).Protect mypassword
MsgBox "You have successfully set by default the dummy panel!", vbInformation,
"PCB Visual Inspection System_CALDO"
End Sub
```

Upon clicking the 'Dummy PCB Inspection', the main GUI automatically appears as shown in Fig. 11. The system would require the user to provide the lot number and date of defect analysis for traceability. Also, the user will have to select the shifting schedule and the name of operator as provided in the database criteria worksheet.

Clicking the command button, 'Detect Defect' is used for PCB defect detection and localization. It links and updates the encoded data in the 'Defect Detection Monitoring' worksheet. This worksheet consolidates all data entries and it graphs the lot number against the no. of defects being detected. The system is equipped with fool-proofing methodologies, so as to prevent wrong data encodes. This includes, but not limited to, wrong lot number, double lot entries and incomplete data entries. Moreover, validations and confirmations using message boxes were observed in this study. These fool-proofs and validations were used for containing data integrity issues. After clicking the 'Detect Defect command button, the data entries were validated for completeness and the lot number was checked for double lot entries. Once completed, the bit contents of the loaded dummy panel were compared with that of PCB bare pattern bit contents. The panel defects 1–32 were initialized to 0. The proponent makes use of the *'strcmp'* operator in comparing the contents of two worksheets. Each panel is provided with individual counter. This counter is incremented whenever similar value is observed. There were 352 patterns for each set. The total number of defects is computed by subtracting the total

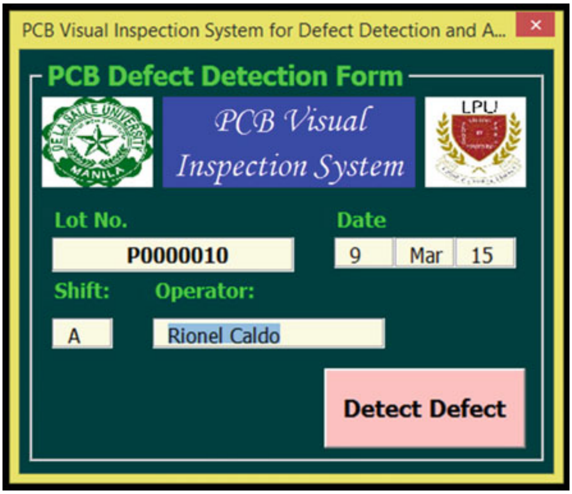


Fig. 11 Main GUI of PCB visual inspection system

number of patterns (352) with the summation of all defects. The system will then display the total number of defects being detected. Remember that each panel is consisted of 11 bits. If the defect counter for each panel is less than 11, it means that the panel is defective. The defective panel is localized by displaying the affected panel number using message boxes. Finally, the data will now be transferred to the main repository and it will be graphed using a trend chart. A confirmation box of successful data encodes and updates will appear once completed.

4 Experiments and Analysis of Results

The main user form is introduced in the previous section, specifically in Fig. 11. As exhibited in the figure, the proponent would require few data inputs from the user for traceability purposes. The template PCB pattern is made available in the system. The user will simply load the dummy PCB for testing purposes. The system will then detect occurrence/s of defects. After detection of defect/s, the system will then localize the affected panel/s. *Refer to Fig.* 12 *for sample validations*. Once, the process of identification and localization of defects were completed, the system

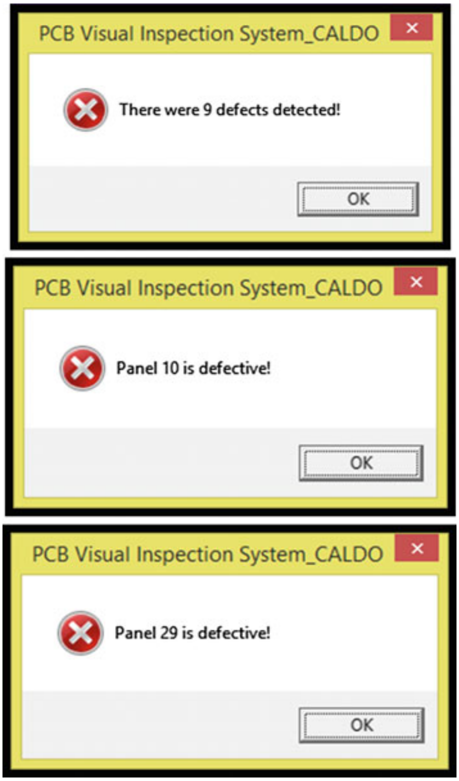


Fig. 12 Sample PCB panel defect detection

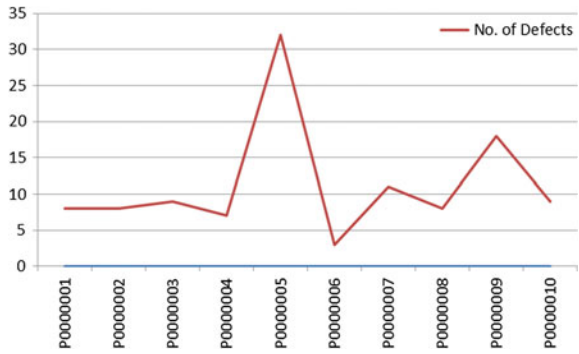


Fig. 13 Trend chart for PCB visual inspection system

automatically generates report pertaining to PCB defect analysis. This report will be stored in a central repository database for monitoring and controlling purposes. The trend chart shown in Fig. 13 is also provided for visualization. This process works for each load of PCB dummy controlled in pre-assigned unique lot numbers.

In simulating the defect detection and analysis in Excel VBA macro, the proponent conducted several trials and tests as enumerated in Table 2. Based on the obtained results, it is observable that the system is able to detect and classify the defects congruently as it gives 100 % detection rate when the actual and the expected outputs were compared.

5 Conclusion

In this paper, the proponent makes use of Excel VBA macro program for intelligent PCB defect detection and localization. In addition, in providing visualization of results, the proponent provided Graphical User Interface (GUI) for traceability, repository for monitoring and trend chart for controlling purposes. In this work, the proponent makes use of 10 defective units (represented in bits of 0 and 1) for simulation purposes and it was able to detect and localize defects accurately by template matching using logical operators. It is realized that high-level programming language like Excel VBA macro program can be used powerfully in modeling an intelligent system, as it provides ease of code in the part of the program design and user-friendliness for end-users.

Acknowledgments Jesus Christ as the Lord and Savior of the proponent is gratefully acknowledge, as the ultimate source of knowledge and wisdom. De La Salle University is acknowledged as well for giving financial assistance for the proponent's doctorate study, and the Research and Statistics Center of Lyceum of the Philippines University-Laguna (LPU-L) for research funding, publication and dissemination.

References

1. Choksi, A.H., Vashi, R., Sevak, M., Patel, K.: Printed circuit board defect detection using wavelet transform. Int. J. Curr. Eng. Technol. 2014
2. Anu, A., Anouncia, M., Rajaji, L.: Automatic visual inspection of PCB using CAD information. In: International Conference on Radar, Communication and Computing (ICRCC), Tiruvannamalai, TN., India, 21–22 Dec 2012
3. Ibrahim, I., Abdul Rahman, S., Abu Bakar, S., MohdMokji, M., Ahmed Mukred, J.A., MdYusof, Z., Ibrahim, Z., Khalil, K., Mohamad, M.S.: A printed circuit board inspection system with defect classification capability. Int. J. Innovative Manage. Inf. Prod. **3**(1) 2012
4. Malge, P.S., Nadaf, R.S.: PCB defect detection, classification and localization using mathematical morphology and image processing tools. Int. J. Comput. Appl. **87**(9) 2014
5. Gebus, S., Lorrilard, S., Juuso, E.: Defect localization on a PCB with functional testing. University of Oulu, Control Engineering Laboratory—Series A, May 2002
6. Fanni, A., Lera, M., Marongiu, E., Montisci, A.: Neural network diagnosis for visual inspection in printed circuit boards [Online]. Accessed 26 Feb 2015
7. Lin, S.-C., Su, C.-H.: A visual inspection system for surface mounted devices on printed circuit board. In: CIS 2006
8. Mar, N.S.S., Yarlagadda, P.K.D.V., Fookes, C.: Design and development of automatic visual inspection system for PCB manufacturing. Rob. Comput. Integr. Manuf. 2011
9. Moganti, M., Ercal, F., Dagli,C.H., Tsunekawa, S.: Automatic PCB inspection algorithms: a survey [Online]. Accessed 6 Feb 2015
10. Khalid, N.K., Ibrahim, Z., Zainal Abidin, M.S.: An algorithm to group defects on printec circuit board for automated visual inspection. IJSSST. **9**(2) 2008

Modeling Academic Achievement of UUM Graduate Using Descriptive and Predictive Data Mining

Fadzilah Siraj

Abstract The selection of data mining approaches is based on the ability of data mining as a powerful tool for academic analysis purposes. In higher educational institution, data mining can be used for the process of uncovering hidden trends and patterns that help the institutions in forecasting the students' achievement. Today, the abilities such as intelligence, skill and CGPA are identified as a main factor for academic achievement. In essence, it is a common practice to use the CGPA as an indicator of students' academic achievement. However, measuring the academic achievement is not an easy task. The purpose of this study is to investigate factors that associated with academic achievement for the undergraduate students of University Utara Malaysia (UUM) based on College of Arts and Sciences (CAS), College of Business (COB), and College of Law, Governance and International Studies (COLGIS) using descriptive and predictive data mining. Prior research indicates that students and faculty shared a common perception of the skills necessary for success in the degree programs. Based on the results extracted from descriptive and predictive data mining, empirical investigation using logistic and neural networks reveal that factors such as family income, race and language skill have significant association with academic achievement.

1 Introduction

Data mining (DM) is defined as computer automated exploratory data analysis of large complex data sets that can be used to discover patterns and relationships in data with an emphasis on large observational databases [1, 2]. DM applies modern statistical and computational technologies to the problem of finding useful patterns hidden within large database [3, 4]. To uncover hidden trends and patterns, it uses a

F. Siraj (✉)
School of Computing, College of Arts and Sciences, Universiti Utara Malaysia,
06010 Sintok, Kedah, Malaysia
e-mail: fad173@uum.edu.m

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_52

combination of an explicit knowledge base, sophisticated analytical skills, and domain knowledge. In effect, these trends and patterns form the basis of predictive models that enable analysts to produce new observations from existing data.

The tasks of DM can be modeled as either Predictive or Descriptive in nature [5]. A Predictive model makes a prediction about values of data using known results found from different data while the Descriptive model identifies patterns or relationships in data. Descriptive DM is used to discover interesting regularities in the data, to uncover patterns and find interesting subgroups in the bulk of data [6]. In education, [7] used Descriptive DM to determine the demographic influence on particular factors.

The selection of DM approaches is based on the ability of DM as a powerful tool for academic analysis purposes [6]. In higher educational institution, DM can be used to uncover the hidden trends and patterns that help the management in forecasting the students' achievement. Academic achievement is the major concern in the universities [7]. The issues of prediction and explanation of academic achievement and a study to identify the key indicators to the academic success and persistence of students are extremely important [8, 9]. The academic achievement is always described in terms of grades or degree completion [10].

Today, the abilities such as intelligence, skill and CGPA are identified as a main factor for academic achievement [11–13]. However, measuring the academic achievement is not an easy task [14]. In Malaysia's education context, the excellent in academic achievement is the upmost target for every student and institutions and also generally, a key indicator for individual success. To date, information regarding national database about IPTA also has been processed. However, only descriptive statistics have been utilized. At UUM, very little attempt has been made to analyze graduates in detail using descriptive statistics and cross tabulations as well as employing Neural Network (NN) to uncover the hidden information within the data. The purpose of this study is to investigate factors that associated with academic achievement for the undergraduate students of UUM based on College of Arts and Sciences (CAS), College of Business (COB), and College of Law, Governance and International Studies (COLGIS). Prior research indicates that students and faculty shared a common perception of the skills necessary for success in the degree programs [15]. Based on results extracted from DM algorithm, empirical investigation using Logistic Regression (LR) and NNs reveal that factors such as family income, race and language skill have significant association with academic achievement.

2 Related Works

Numerous studies have been conducted to determine factors that affect academic achievement. Both students and academicians rated self-motivation [16], aptitude, prior academic performance, effort, motivation [17] and attitude [18] play significant roles in academic achievement. Efforts to increase students' academic performance and achievement have also been linked to family involvement, emotional support, parent's adherence to education, their economic conditions, social background, the way parents communicate with teachers and actively participate in educational program can increase the performance of students [19, 20]. Gender differences exist in academic performance, in fact, female students show higher confidence and better performance in all examined academic subjects [21].

To date, higher education organizations are placed in a very high competitive environment and aiming to get more competitive advantages over the other business competitors. To remain competitive, these organizations need a deep and enough knowledge for better assessment, evaluation, planning and decision making. Since most of the required information can be extracted from the historical and operational data that reside in the organizations' databases, DM can be exploited in extracting the required information that can be used to facilitate competitiveness and also additional insight for education management institutions based on operational data in facilitating decision making. Such a need for understanding large data sets is found in education, engineering, business, medicine and others [22]. The technological development could be used to allow better collection, and more importantly to analyze large amount of data that finally produce knowledge that is meaningful to organizations which needs to be discovered. Failing to do so will inevitably contribute to the loss of useful information, money, time and efficiency. Hence, in this regard, descriptive and predictive approaches in DM could deliver meaningful knowledge that assists the organizations to sustain its competitiveness.

In DM, two most commonly used models in education are logistic regression (LR) and NNs [23, 24]. For instance, LR used to model the examination result indicates that the performance of female students are 2.128 times higher than that of the performance of male students [25]. The results of applying DM to enrolment data of Sebha University in Libya shows that NN obtains the highest results accuracy [26] compared to LR and decision tree [27]. NN is potentially enhancing the effectiveness of a Nigerian Universities admission system and able to predict the performance of more than 70 % of the prospective students [28].

Academic achievement is one of the most important criteria in measuring student success at the university [29]. In fact, it is aligned with other aspects of achievements such as communication, leadership, self-management, attitude and personality [30]. In this, study LR and NN methods were explored in providing more insight in conjunction with academic achievement.

3 Approach

The approach for this study has been adapted from [27] as shown in Fig. 1.

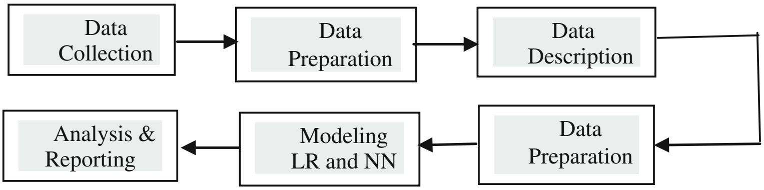


Fig. 1 Steps involved in the study for obtaining empirical results

3.1 Data Collection

The respondents of this study include UUM graduates who have completed their undergraduate study in the year 2006, 2007 and 2008.

3.2 Data Preparation

Data preparation covers all activities for constructing the final data set that are conducted repetitively to analyze the data, choosing suitable attributes to use as predictors as well as transformation and cleaning data.

3.3 Data Description and Data Cleaning

A total of nearly 300 attributes have been explored from the data sets. During this phase, the treatment for the missing values are conducted. However, the way to handle the missing value depends upon the data that has been collected.

3.4 Modeling

Initially, Descriptive DM is carried out to investigate the nature of the dataset and the distribution of each attribute. Frequency tables are generated, and the correlation analysis has also been conducted to determine the relationship between the attributes, including Cross Tabulation Analysis (contingency tables). However, only the Cross Tabulation analysis is reported in this paper. Clustering Analysis is performed based on 4 clusters. For Clustering Analysis, Kohonen network is used (Fig. 2) assuming the clusters are formed at the output layer from patterns that share common features.

To enable the Kohonen layer to group similar patterns, a neighborhood of artificial neurons around the winning artificial neuron (at the output layer) is also altered to be more like the input pattern [30]. As the clusters were generated through Kohonen networks, these clusters are then used as output for LR and NN.

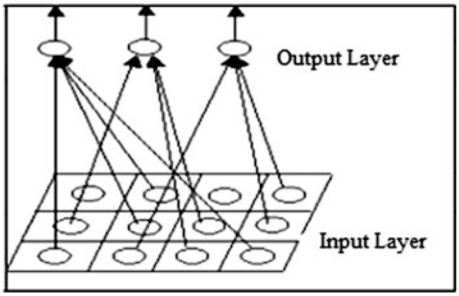


Fig. 2 Kohonen network

In addition, comparison between the predictive approaches has been conducted to get some insight about the strength and weaknesses of each approach since one of the aims of the study was to determine whether these methods were well suited for extracting the required knowledge. As a result, the predictive method will be able to predict in which cluster the future student falls into based on the academic information of the students. This is the main difference between the proposed method and the previous method using LR and NN as the predictive methods.

LR analysis model is also known as one of the most useful tools in quantitative analysis phase of the decision-making process [31]. It allows the prediction of a discrete outcome from a set of variables that may be continuous, discrete, dichotomous, or a combination of any of these by fitting a set of points to a curve [32]. The LR equation can be written as in [31]

$$\theta = \frac{\exp(\beta_0 + \beta_1 + \cdots + \beta_k x_k + \alpha)}{1 + \exp(\beta_0 + \beta_1 + \cdots + \beta_k x_k + \alpha)} \tag{1}$$

where α is the constant from the equation, and β is the coefficient of the predictor variables. Alternatively, the LR equation can be written as

$$\pi = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}} \tag{2}$$

where π is the probability for the outcome of interest or “event”, α is the intercept, ß is the regression coefficient, and $e = 2.71828$ is the base for the system of natural logarithms x can be categorical or continuous, but Y is always categorical.

NN is one of the DM techniques; defined as an information-processing system which is inspired from the function of the human brain whose performance characteristics are somehow in common with biological NN [33]. Multilayer Perceptron (MLP) is one of the most common NN architecture that has been used for diverse applications, particularly in forecasting problems [29]. The network is normally composed of a number of nodes or processing units, and it is organized into a series of two or more layers (Fig. 3).

The training of MLP could be stated as a nonlinear optimization problem aims to find out the best weights that minimize the difference between the input and the

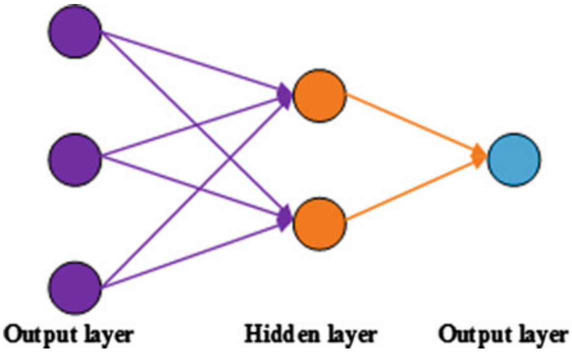


Fig. 3 MLP architecture

output. Its most popular training algorithm is Back propagation (BP), and it has been used in solving many problems in pattern recognition and classification. Some BP parameters such as number of hidden nodes at the hidden layers 'learning rate, momentum rate, activation function and the number of epoch are also investigated to obtain a suitable NN prediction model [34]. Furthermore, these parameters could change the performance on the learning from bad to good accuracy [35]. The accuracy of NN is provided by a confusion matrix. The accuracy of NN is calculated using (3)

$$Percent\ of\ Correct = \left(\frac{Tot.\ of\ correctly\ predicted\ patterns}{Total\ no.\ of\ patterns}\right) * 100\,\% \qquad (3)$$

3.5 Analysis and Reporting

The techniques that produce highest accuracy are considered as the most suitable technique for prediction purposes. Although NN produces higher accuracies than LR in many instances, it still remain a black box to researchers. Hence, the use of LR and Descriptive DM could lead to some insight with regard to correlation and association of the attributes with the academic achievement. In this study, the attributes are selected based on the statistical results and the academic achievement is measured by CGPA. The comparison among graduates is conducted in accordance to the programs undertaken at UUM.

4 Results

The entire population of UUM graduates who had completed their studies between 1 and 6 months upon graduation for the year 2006, 2007 and 2008 were collected from UUM database. A total of 5005, 3605 and 5723 UUM graduates participated in the survey were complete enough to be used in the study. The distribution of the graduates of baccalaureate programs for the year 2006–2008 based on college of Arts and Sciences (CAS), College of Business (COB) and, College of Law, Governance and International Studies (COLGIS). The results indicate that COB has the highest percentage of graduates involved in this study for the year 2006 and 2008. In the year 2007, CAS percentage exceeds COB by 13.9 % [36]. Further analysis on the performance of the graduates based on each college for the three years is also shown in Fig. 4. For CAS, the performance of the graduates with CGPA above 3.67 improves throughout the three years; in fact the year 2008 shows the highest percentage (3.4 %) amongst the three colleges. Although the CGPA of COLGIS graduates with CGPA of 3.00–3.66 decreases to 34.8 % in 2007, the average CGPA obtained by the graduates of this college is the highest among the three colleges. Note that by the year 2008, less than 4 % of the graduates from all colleges obtained CGPA 2.00–2.49. When the academic achievement is cross-tabulated with gender, it is very clear that majority of the female graduates obtained CGPA between 3.00 and 3.66 for three consecutive years (Fig. 4). These findings are consistent for the three colleges in UUM. On the other hand, most male students scored CGPA between 2.50 and 2.99 in 2006 (47.1 %) and (36.6 %).

The Chi-squared test statistic is 470.88 with an associated $p < 0.01$; hence gender is associated with CGPA. In addition, graduates from CAS achieved better results than other colleges (Fig. 5). However, the findings in this study reveal that employment rates are higher for the COB compared to other colleges.

In terms of gender versus employment, there is a significant association between these two variables. Hence, more promotion should be encouraged towards increasing the number of male enrollment to the universities (Fig. 5). The appropriate departments need to recommend programs suitable for improving the academic achievement particularly for the male students.

The results of the experiments using NN and LR are summarized in Table 1.

The same data sets have been used for both predictive methods. Clearly, NNs superseded the performance of LR for all colleges. The highest accuracy obtained by LR is 61 % whilst the NN scores more than 99 % of the classification accuracies.

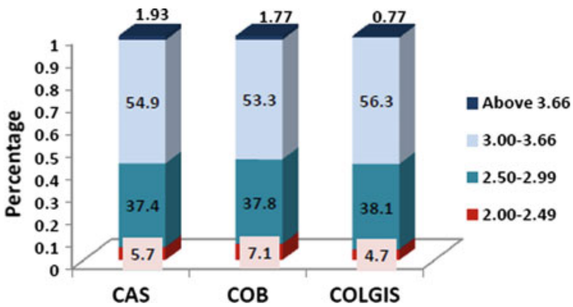


Fig. 4 Distribution of CGPA with respect to college and year

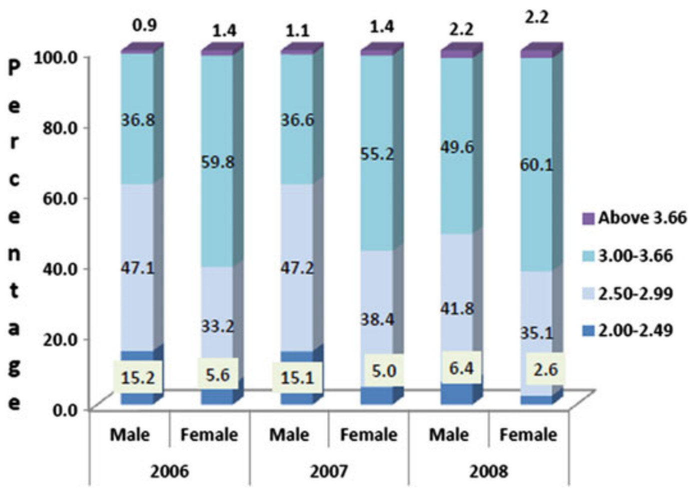


Fig. 5 Distribution of CGPA with respect to gender for the year 2006–2008

Table 1 The performance of NN and LR

College	Neural network					Logistic regression
	Input	Seed no.	Hidden unit	Output	Percentage correct	Percentage correct
CAS	4	5	4	4	99.12	61.0
COB	7	5	4	4	99.29	59.8
COLGIS	4	4	6	4	99.11	60.4

Comparing the likelihood ratio test values, the Wald statistics and the correlation values, the logistic equation for CAS is written as

$$\mathrm{P(event)} = \frac{Exp(5.04 - 0.7 * G + 0.20 * BI - 0.16 * S - 0.18 * FI)}{(1 + Exp(5.04 - 0.17 * G + 0.2 * BI - 0.16 * S - 0.18 * FI))} \tag{4}$$

Similarly, the logistic equation for COB is written as

$$\mathrm{P(event)} = \frac{Exp(1.33 - 0.01 * G - 0.05 * R + 0.34 * BI + 0.02 * S - 0.03 * FI + 0.35 * BS + 0.72 * GK)}{(1 + Exp(1.33 - 0.01 * G - 0.05 * R + 0.34 * BI + 0.02 * S - 0.03 * FI + 0.35 * BS + 0.72 * GK))} \tag{5}$$

Table 2 Summary of significant predictor variables for academic achievement

College	Gender	BI	Sponsor	Family income	Race	BM skill	General knowledge
CAS	✓	✓	✓	✓			
COB	✓	✓	✓	✓	✓	✓	✓
COLGIS	✓	✓			✓	✓	

Finally, the logistic equation for COLGIS is written as

$$\mathrm{P(event)} = \frac{Exp(4.74 + 0.53 * G - 0.09 * R - 0.45 * BI + 0.20 * BS)}{(1 + Exp(4.74 + 0.53 * G - 0.09 * R - 0.45 * BI + 0.20 * BS))} \quad (6)$$

where

G is Gender
R is Race
S is Sponsor
FI is Family Income
BS is BM Skill
GK is General Knowledge

The attributes that are significantly associated with the academic performance for each college indicate that the most common factor for the three colleges that are associated with academic performance are Gender and BI (Table 2). Gender factor may be biased since the intake of female to male is about 3:1 ratio. However, BI factor is in line with the findings from Ministry of Education which indicate graduates from IPTA need to improve their BI skill for both communicative and writing skill.

It also appears that both CAS and COB graduates' academic performance has significance association with the Sponsor and Family Income. For COB and COLGIS graduates', race and BM skill are the two significant factors that are associated with academic performance.

5 Conclusion

This study attempts to model the academic achievement that can be used by the UUM management or educators to analyze the students' performance. The empirical findings reveal and confirm some speculation about the factors that influence academic performance. Further drilling information as a continuation of this paper may reveal more knowledge and that may lead in the formation of policy or framework for academic performance for the students in UUM in general and with respect to colleges in particular. Other Predictive or Clustering may also help to uncover the hidden information that still exist within the UUM graduates data.

The result from this study is expected to be used for identifying the factors influencing students' academic performance. In addition, the prediction model could be used by the management to design special programmes for the 'outstanding' and the 'low' achievers for each degree programme. In this way, students who are expected to do well could be pushed to the limit or reach the excellent level. On the other hand, students who are expected to be low achievers could be assisted to gain better grades upon graduation. This is to ensure the quality of graduates is either sustained or progress in a positive directions. DM can assist in the decision making process or to explain and justify it. Further research on DM is expected to increase due to the efficiency and the level of accuracy obtained through Descriptive and Predictive DM.

Acknowledgment The author gratefully acknowledge the University Utara Malaysia for providing the grant to conduct the research.

References

1. Debreceny, R.S., Gray, G.L.: Data mining journal entries for fraud detection: an exploratory study. Int. J. Account. Inf. Syst. **11**(3), 157–181 (2010)
2. Spangler, S., Ying, C., Kreulen, J., Boyer, S., Griffin, T., Alba, A., Kato, L., Lelescu, A., Yan, S.S.: Exploratory analytics on patent data sets using the SIMPLE platform. World Pat. Inf. **33** (4), 328–339 (2011)
3. Nasseri, M., Tavakol-Davani, H., Zahraie, B.: Performance assessment of different data mining methods in statistical downscaling of daily precipitation. J. Hydrol. **492**, 1–14 (2013)
4. Kim, S.B., Jitpitaklert, W., Park, S.-K., Hwang, S.-J.: Data mining model-based control charts for multivariate and autocorrelated processes. Expert Syst. Appl. **39**(2), 2073–2081 (2012)
5. Luan, J.: Data mining and its applications in higher education. New Dir. Institutional Res. **2002**(113), 17–36 (2002)
6. Delavari, N., Phon-amnuaisuk, S.: Data mining application in higher learning institutions. Inform. Educ. Int. J. **7**(1), 31–54 (2008)
7. Miller, L.K.: Loyola eCommons the Impact of Intrusive Advising on Academic Self Efficacy Beliefs in First-Year Students in Higher Education. Loyola University (2010)
8. Karau, S.J.: Cross-Cultural Differences in the Academic Motivation of University Students in Malaysia and the United States. North Am. J. Psychol. **9**(2), 275–292 (2007)
9. Ervina, A., Md Nor, O.: Undergraduate students' performance: the case of University of Malaya. Qual. Assur. Educ. **13**(4), 329–343 (2005)
10. Ditcher, A., Tetley, J.: Factors influencing university students' academic success: what do students and academics think. HERDSA Annual International Conference, Melbourne (1999)
11. Noor Azina, I., Azmah, O.: Comparing university academic performances of HSC students at the three art-based faculties. Int. Educ. J. **7**(5), 668–675 (2006)
12. Nordaliela, M.R., Zaidah, I., Roziah, M.J.: Predicting students' academic achievement: comparison between logistic regression, artificial neural network, and neuro-fuzzy. In: Proceedings of International Symposium on Information Technology (ITSim 2008), vol. 1, pp. 1–6 (2008)
13. Geiser, S., Santelices, M.V.: Validity of high-school grades in predicting student success beyond the freshman year: high-school record vs. standardized tests as indicators of four-year

college outcomes. http://cshe.berkeley.edu/publications/docs/ROPS.GEISER._SAT_6.12.07.pdf (2007). Retrieved on 6 Jan 2008
14. Ryerson, P.: Policy On Undergraduate Grading, Promotion, And Academic Standing (The 'Gpa Policy') **46**(2), 1–11 (2013)
15. Price, B.A., Randall, C.H.: Technology-based degrees: do faculty and students agree on necessary skills for success in these programs? In: Proceedings of the 5th Conference on Information Technology Education (CITC5 04), pp. 235–240 (2004)
16. Shahmandi, E., Silong, A.D., Ismail, I.A., Samah, B.B.A., Othman, J.: Competencies, roles and effective academic leadership in World Class University. Int. J. Bus. Adm. **2**(1), 44–53 (2011)
17. Benford, R., Gess-newsome, J.: Factors Affecting Student Academic Success in Gateway Courses at Northern Arizona University. Arizona University (2006)
18. Facey-Shaw, L., Golding, P.: Effects of peer tutoring and attitude on academic performance of first year introductory programming students. In: Proceedings of the Frontiers in Education, 35th Annual Conference, pp. S1E–1–S1E–6 (2005)
19. Guiffrida, D.A., Lynch, M.F., Wall, A.F., Abel, D.S.: Do reasons for attending college affect academic outcomes?: A test of a motivational model from a self-determination theory perspective. J. Coll. Student Dev. **54**(2), 121–139 (2013)
20. Shipley, N.L., Jackson, M.J., Segrest, S.L.: The effects of emotional intelligence, age, work experience, and academic performance. Res. High. Educ. J. **1**, 1–18 (2008)
21. Mueller, S.L., Conway, M.: Dato-on: gender-role orientation as a determinant of entrepreneurial self-efficacy. J. Dev. Entrepreneurship **13**(1), 3–20 (2008)
22. Bryant, R.E., Katz, R.H., Lazowska, E.D.: Big-data computing: creating revolutionary breakthroughs in commerce, science, and society motivation: our data-driven world. Comput. Res. Assoc. **8**, 1–7 (2008)
23. Siraj, F., Asman, H., Hasan, M.R.: Modeling information technology competency using neural networks. Int. J. Comput. Appl. **58**(11), 1–6 (2012)
24. Siraj, F., Yusoff, N., Mohd Ali, N.: Exploring hidden relationships within students' data using neural network and logistic regression. In: Proceedings of Knowledge Management International Conference and Exhibition (KMICE), pp. 162–168, Kuala Lumpur, Malaysia (2006)
25. Saha, G.: Applying Logistic Regression Model to the Goutam Saha Govt. Degree College, Kamalpur, Tripura, vol. 4, No. 2, pp. 105–117, 201118 (2011)
26. Siraj, F., Abdoulha, M.A.: Uncovering hidden information within university's student enrollment data using data mining. In: 3rd Asia International Conference on Modelling and Simulation, pp. 413–418 (2009)
27. Siraj, F., Omer, E.A., Hasan, M.R.: Data Mining and Neural Networks : The Impact of Data Representation. Intech, Croatia (2012)
28. Oladokun, O.E., Adebanjo, V.O., Charles-Owaba, A.T.: Predicting students' academic performance using artificial neural network: a case study of an engineering course. Pacific J. Sci. Technol. **6**(1), 72–79 (2008)
29. Iramaneerat, C.: Predicting academic achievement in the medical school with high school grades. J. Med. Assoc. Thai **89**(9), 1497–1505 (2006)
30. Youdrg, F., Redzuan, M., Mansor, M.B., Abu Talib, M., Juhari, R.: Memory as a mediator between depression and academic achievement among Iranian adolescents. Eur. J. Psychol. **4**, 40–51 (2009)
31. O'Connor, M., Marquez, I., Hill, T., Remus, W.: Neural network models for forecast a review. In: IEEE Proceedings of the 25th Hawaii International Conference on System Sciences, vol. 4, pp. 494–498 (2002)
32. Dunham, M.H.: Data Mining Introductory and Advanced Topics. Pearson Education, Inc., Upper Saddle River (2003)
33. Fausett, L.: Fundamentals of Neural Networks Architectures, Algorithms, and Applications. Prentice Hall, Upper Saddle River (1994)

34. Sirikulvadhana, S.: Data mining as a financial auditing tool. M.Sc. thesis in Accounting. The Swedish School of Economics and Business Administration. www.pafis.shh.fi/graduates/supsir01.pdf (2002). Retrieved on 12 Sept 2008
35. Nawi, N.M., Ransing, M.R., Ransing, R.S.: An improved learning algorithm based on the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method for back propagation neural networks. In: Sixth International Conference on Intelligent Systems Design and Applications, vol. 1, pp. 152–157 (2006)
36. Siraj, F., Haris, M.F.: Profiling Of UUM graduates based on academic achievement and colleges. In: Proceedings of the International Soft Science Conference 2011 (ISSC2011), pp. 1–10, Ho Chi Minh, Vietnam (2011)

A Novel Cuckoo Search Based Clustering Algorithm for Wireless Sensor Networks

Md. Akhtaruzzaman Adnan, M.A. Razzaque, Md. Anowarul Abedin, S.M. Salim Reza and Molla Rashied Hussein

Abstract Wireless sensor networks are primarily characterized by inadequate energy supply. Therefore, development of an energy efficient protocol can play an important role in impacting the network lifetime. Typically, communication is the most energy expensive act that nodes perform and limited energy of nodes is the main obstacle. An efficient cluster arrangement might be a solution. Though optimum clustering in wireless sensor networks is an NP-Hard problem, at present, bio-inspired metaheuristic approaches are very popular in solving them. This paper presents a centralized energy-aware clustering algorithm for wireless sensor networks using the novel bio mimic cuckoo search algorithm. The cost function was defined, with the goal of maximizing the network lifetime and minimizing the intra-cluster distance. The performance of the proposed algorithm is evaluated with well-known centralized and decentralized clustering protocols. The results derived from simulations show that proposed solution can enhance network lifetime over its comparatives.

Keywords Cuckoo search · Energy efficiency · Clustering · Wireless sensor networks

Md.A. Adnan (✉) · M.R. Hussein
Department of Computer Science and Engineering, University of Asia Pacific, Dhaka, Bangladesh
e-mail: adnan.iut05@gmail.com

M.A. Razzaque
School of Computer Science and Statistics, Trinity College, Dublin, Ireland

Md.A. Abedin
Pervasive Computing and Communications for Sustainable Development, University of Lorraine, Nancy, France

S.M. Salim Reza
Department of ICT, Bangladesh University of Professionals (BUP), Dhaka, Bangladesh

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_53

1 Introduction

Recent progresses in Micro-Electro-Mechanical Systems (MEMS) technology have brought good news for wireless sensor network (WSN). This technology ensures that multifunctional, low power wireless sensor nodes can be developed at a low cost. At the same time, the technology enables the nodes, which are tiny in size, to communicate within a short distance [1–3]. These sensor nodes are equipped with some amount of sensing, control, data processing, and communicating components [4]. A number of these tiny nodes are deployed inside the sensing area or close to it in order to build WSNs. These WSNs are powerful in a sense that they are capable of supporting a number of diverse applications including environment monitoring [5, 6], medical care [7], precision agriculture [8, 9], military target tracking and surveillance [10, 11], natural disaster relief and monitoring [12, 13] etc.

WSNs are critically resource constrained by limited power supply, memory, processing performance and communication bandwidth [4]. As a result, in almost every scenario, sensor nodes rely on limited energy sources e.g. batteries. Once the nodes are deployed and functioning in the network, replacing the batteries generally is not a feasible task. But at the same time, a WSN must perform for a given period of time or more than that. Hence, most existing works (e.g. clustering, prolonging lifetime) in WSNs area are dealing with energy efficiency. Clustering is an approach that is used to control network energy expenditure efficiently. It minimizes the number of sensor nodes that communicate with the base station (BS) from a long distance and distributes the energy spending uniformly amongst the nodes of the sensor network.

This paper focuses on the development of a centralized, energy efficient clustering algorithm based on CS to extend sensor network lifetime and evaluation of its performance on several WSN simulation environments. The protocol partitions the total network into several clusters using CS algorithm. The algorithm considers energy consumption of the nodes in the optimization process. The evaluation of the CS clustering algorithm against its counterparts was done on the basis of uniform clustering capability, network lifetime, scalability and the amount of data sent to the BS. The developed protocol later was tested against the previous well known and established protocols like LEACH, LEACH-C and PSO—one of the most used bio inspired optimization algorithms.

The rest of the paper is organized as follows: In Sect. 2 we discussed the literature review part. In Sect. 3, the network models and radio energy models utilized in the proposed protocol are described. In Sect. 4 we give a comprehensive explanation of the projected cluster configuration protocol using CS algorithm. Section 5 presents the simulation results of our protocol and in Sect. 6 conclusions are drawn.

2 Literature Review

By adopting cluster based network approach a number of protocols and optimization techniques have been proposed in literature. Among them one of the most famous clustering protocols is Low Energy Adaptive Clustering Hierarchy (LEACH) [14]. It provides significantly better result than conventional multi hop routing schemes, e.g. the minimum transmission energy (MTE) in terms of energy saving and network lifetime [15]. A further improvement in distributed nature of LEACH is known as LEACH-C [16], which forms the clusters using a centralized algorithm executed by the BS. There were some other works based on low energy protocols. In [17] authors presented two approaches of low power consumption by using the information obtainable from a gate level VHDL simulations. In another work [18], an algorithm is presented for the detection of distributed systems by huge scale collaborative sensors in low power environment. Tree or linear hopping network (ToLHnet) [19] is another powerful protocol that supports mixed networks in low complexity.

Apart from these, some researchers started to use bio-inspired optimization algorithms to solve the problem of efficient clustering [20]. A number of bio-inspired algorithms developed for clustering in WSNs namely Ant Colony Optimization (ACO) [21, 22], Genetic Algorithm (GA) [23], Particle Swarm Optimization (PSO) [24, 25] etc. Cuckoo Search (CS) is one of the recent additions in the pool of bio-inspired optimization algorithms. CS was developed in 2009 by Yang and Deb [26]. The motivation of this algorithm came from the breeding behaviour of cuckoo bird, which has a unique characteristic of laying eggs in other bird's nest. CS has shown its ability to render better results than other algorithms of its type against several benchmark tests and statistical analysis showed that the problem solving success of the CS algorithm is better than the PSO and GA [27, 28].

So it was the target of this research work to test its validity in performing better to solve clustering problems in WSNs efficiently. The first work of cuckoo search to improve the network lifetime of the sensor field was proposed by Dhivya in [29]. To the best of knowledge, till date, this is the only work that exactly matches the field of this proposed study, considering only CS algorithm in cluster head (CH) selection. In this work, clusters are formed with the nodes with least energy and data collected by these nodes are transmitted to the designated CH. Therefore the lowest energy nodes are initially exploited. The goal was to reasonably balance the energy usages among the nodes on the basis of their remaining energy and to increase the lifetime of the network. The acquired results are compared with only LEACH protocol to show the effectiveness of the method. The key difference between the proposed work and [29] is the function of CS in selecting the most favourable nodes as CHs on the basis of different cost functions, in order to extend the network lifetime.

3 The System Model

3.1 Network Model

We presume a network model analogous to those used in [14, 16], with the attributes mentioned below: the BS is fixed and located inside the sensor network field. All sensors are stationary, perform sensing tasks periodically and always have data to send to the BS. Each node can operate in both cluster head (CH) mode or sensing mode—depends on its remaining energy level. Data aggregation is used to reduce the amount of sent messages.

3.2 Radio Energy Model

A first order radio energy model is used in [15] and our protocol uses a similar radio model for the sensors. In this model, to achieve a satisfactory Signal-to-Noise-Ratio (SNR) in transmitting an l-bit message over a distance d, the energy expended by the radio is given by:

$$\begin{aligned} \mathrm{E_{TX}}(\mathrm{l},\mathrm{d}) &= \mathrm{l}\cdot\mathrm{E_{elec}} + \mathrm{l}\cdot\varepsilon_{\mathrm{FS}}\mathrm{d}^2, \quad \text{if } \mathrm{d} < \mathrm{d_0} \\ &= \mathrm{l}\cdot\mathrm{E_{elec}} + \mathrm{l}\cdot\varepsilon_{\mathrm{TR}}\mathrm{d}^4, \quad \text{if } \mathrm{d} \geq \mathrm{d_0} \end{aligned} \tag{1}$$

where E_{elec} is the dissipated energy for each bit of data that is used to run the receiver circuit or the transmitter. ε_{FS} and ε_{MP} varies with respect to transmitter amplifier model that is being used. d_0 denotes the threshold value of the transmission distance. Now, to receive l bit data packet, the energy used by the radio is:

$$E_{RX}(l) = l \cdot E_{elec} \tag{2}$$

4 Proposed Algorithm Description

4.1 Cuckoo Search Algorithm

Cuckoo search (CS) is a population based optimization technique and as many other metaheuristic algorithms it starts with random initial population. In order to simplify CS algorithm three idealized rules can be used [23].

- A single cuckoo lays one egg at a time and dumps it randomly in one of the host nest
- Selection of the highest quality egg among the best nests to carry out to the next generation (a sort of eclecticism)
- The number of host nests is fixed. The host bird can discover an egg laid by a cuckoo with the probability $p_a \in [0, 1]$

4.2 Cluster Setup Using Cuckoo Search Algorithm

In general there are two types of communication happening in a wireless sensor network: inter cluster communication and intra cluster communication. The intra cluster communication is communication between nodes which is relatively cheaper than the previous one. By clustering we try to increase this intra cluster communication. One of the ways is to select a definite amount of clusters, select a number of CHs that would represent each cluster and all the non CH nodes will send their information to the respective CHs. After accumulating all the information from non CH nodes, the CH node will forward this message to the BS. Thus the total amount of energy consumption will be decreased immensely. One more issue needs to be addressed here, that is if only one CH is acting always as the head node, after certain amount of time it will lose its energy level. In this case, a new node is required to be the CH. This decision of selecting a new CH will be done by the respective algorithm. Amount of residual energy, distance from non CH nodes etc. are the criteria in defining and nominating a new CH.

Generally, all clustering protocols consist of four major stages and two phases. The four stages are: CH selection, cluster formation, data aggregation, and data communication. The two phases are the setup phase and steady state phase. At the starting of each setup phase, nodes send the information to the BS about their current energy level and locations. The BS calculates the average energy level of all nodes based on the received information. For every round, only those nodes are qualified as CH candidates, who have higher energy level than the average of that cluster. This process ensures that only nodes with an adequate energy level are selected as CHs. Next, the BS runs the CS algorithm to decide the best K number of CHs that can minimize the cost function, as defined by:

$$f_1 = \max_{k=1,2,\ldots,K}\left\{\sum d\left(n_i, CH_{e,k}\right)/\left|C_{e,k}\right|\right\} \tag{3}$$

$$f_2 = \sum_{i=1}^{N} E(n_i)/\sum_{k=1}^{K} E\left(CH_{e,k}\right) \tag{4}$$

$$cost = \beta \times f_1 + (1+\beta) \times f_2 \tag{5}$$

where function f_1 is the highest average Euclidean distance between nodes and their associated CHs and $\left|C_{e,k}\right|$ is the amount of nodes that fit into cluster C_k of egg e. Ratio of total primary energy of every node n_i, i = 1, 2, …, N in the sensor network with all present energy of the CHs candidates of the present round is defined by f_2. A user defined constant β is essential to determine the role of every sub objectives. The fitness function described here has the goal of concurrently reducing the distance of intra cluster between CHs and their nodes, as quantified by f_1. It also optimizes the energy efficiency of sensor network as measured by f_2. As per the cost function defined earlier, a tiny value of f_1 and f_2 implies to compact cluster formations with the most favorable set of sensor nodes that have adequate energy to perform the task of CH.

If a sensor network has N sensors and K predetermined clusters, the clusters for that network can be defined by the following steps:

1. Initialize H eggs to contain K randomly selected CHs among the suitable CH candidates.
2. Evaluate the cost function of each egg *i*.

   (a) For each node n_i, $i = 1, 2, \ldots,$ N

   - Calculate distance $d(n_i, CH_{e,k})$ between node n_i and all CHs $CH_{e,k}$
   - Assign node n_i to cluster head $CH_{e,k}$ where;

$$d(n_i, CH_{e,k}) = \min_{\forall k=1,2,\ldots,K} \{d(n_i, CH_{e,k})\} \tag{6}$$

   (b) Calculate cost function with Eqs. (3)–(5).

3. Find the best nest with high quality eggs.
4. A fraction p_a of worse nests are discarded and new ones are built.
5. Keep the finest solutions (or nests with quality solutions);
   Rank the solutions and find the current best.
6. Map the latest updated position with closest (x, y) coordinates.
7. Repeat steps 2–6 until the highest number of iteration is reached or any other stopping criterion.

After receiving the best set of clusters, CHs and their associated cluster members, the BS transmits the information that contains the CH-ID for each node back to all nodes in the network. The cluster head node acts as the local control centre to coordinate the data transmission. Once the CH completes receiving data from all of its member nodes, it does data aggregation and sends it to BS. We adopt the analogous approach as in [15] to send the data from CH to BS using fixed spreading code and Carrier sense Multiple Access (CSMA). Figures 1 and 2 show the flowchart of CS algorithm and the selection procedure of the best set of cluster heads per round applied in cluster setup phase.

5 Simulations and Analysis

The proposed protocol is implemented and the performance is assessed using MATLAB R2013a. We execute the simulations in environment of 100 nodes in a network area of 500 m × 500 m. The preliminary energy of nodes to is not equal, means it is a heterogeneous network. The BS is located in the middle of the network

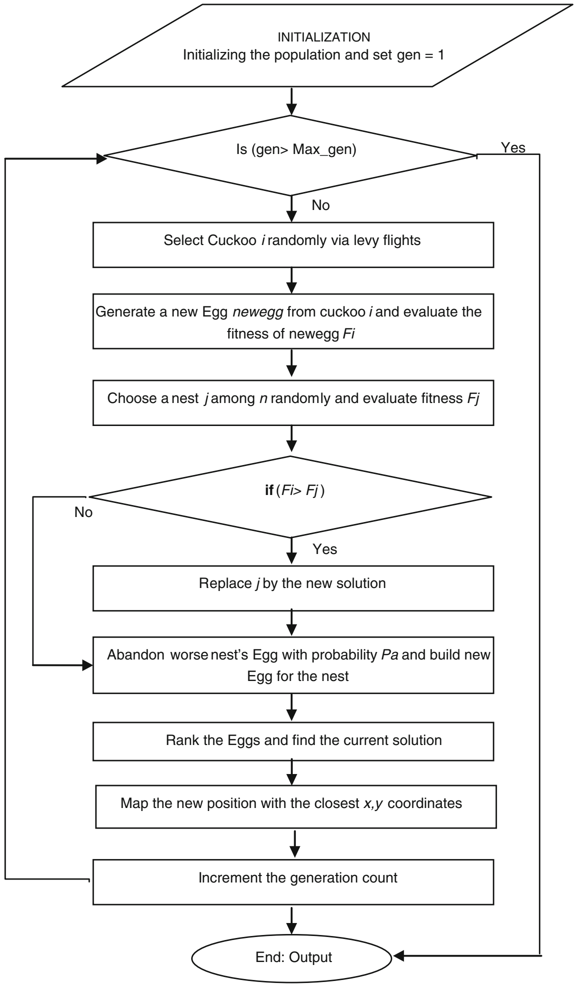


Fig. 1 Flow chart of CS algorithm used in clustering

area with coordinates (250, 250). We decided that 20 % of the sensor nodes will have 5 J of primary energy, whereas rest 80 % of nodes will have 2 J of primary energy. Around 5 % of the total nodes will be selected as CHs (K = 5). The comparison of our protocol was performed with the two benchmark protocols, LEACH, LEACH-C and one bio inspired based approach in clustering using PSO. The simulations were run until every node in the network had died. Moreover, the data message was fixed to 6500 bytes. The length of the control packet was 200 bytes. For the parameters of CS algorithm, we used $H = 30$ eggs, $p_a = 0.25$. We set $\beta = 0.5$ to give identical contribution of every sub-objective. Table 1 shows the network parameters at a glance.

Figures 3, 4, 5 and 6 give a pictorial view about the formation of cluster for a random deployment of 100 nodes during the tenth round of the communication.

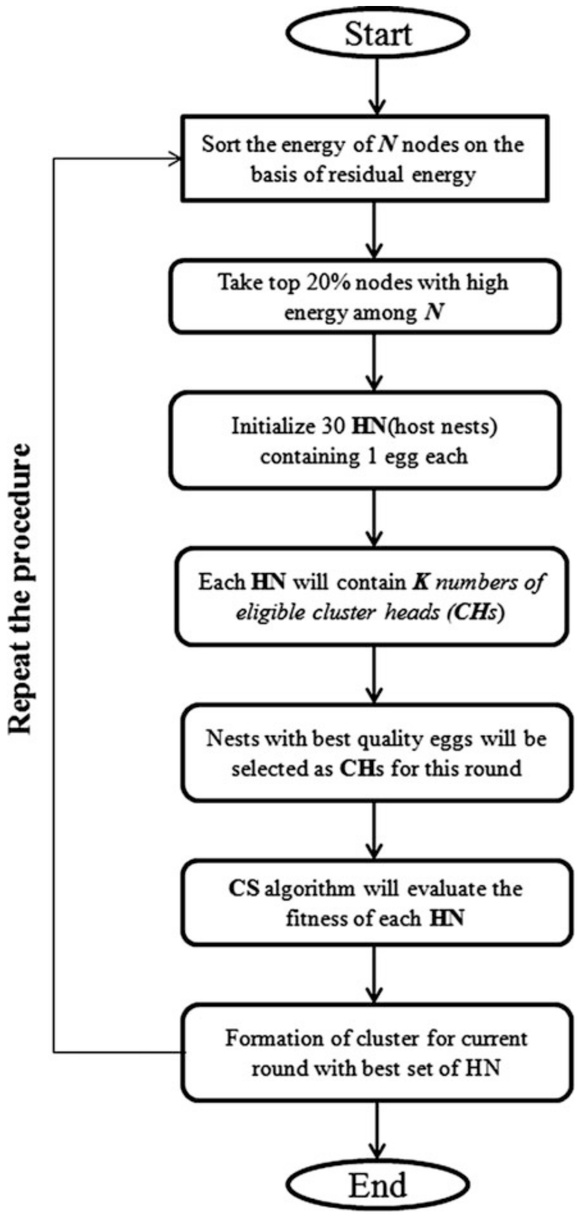


Fig. 2 Selection of best set of cluster heads per round using CS algorithm

From figures, it is clear that our protocol (see Fig. 6) can generate better network partitioning, where all CHs are uniformly positioned across the network and located almost near the centre of each cluster. On the other hand LEACH and LEACH-C, generate an uneven allocation of CHs throughout the network. PSO attempted to show better results than these two. Meanwhile, the projected algorithm which uses

Table 1 Network parameters and their values

Parameters	Values
Number of nodes	100
Area size	500 m × 500 m
Base station position (coordinates)	(250, 250)
Data size	6500 bytes
Control packet size	200 bytes
Number of clusters K = k_{opt}	5 % of total nodes

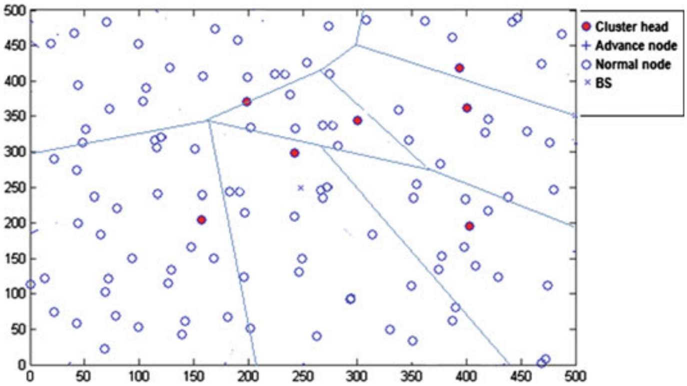


Fig. 3 Network partitioning by LEACH

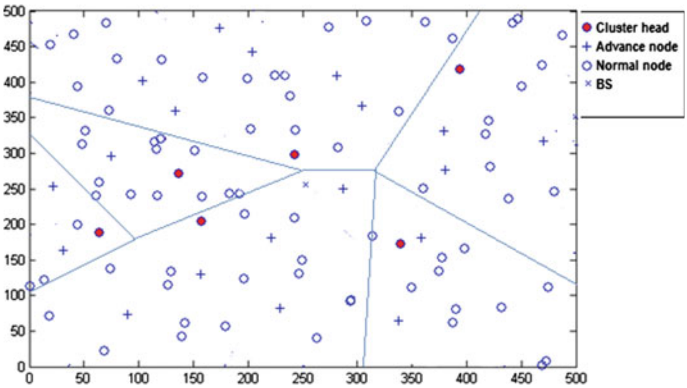


Fig. 4 Network partitioning by LEACH-C

CS, tries to avoid poor cluster formation due to the inherent cooperation of CS that can attain global minimum of intra cluster distance.

Figures 7 and 8 respectively illustrates the lifetime after the death of 50 % of total nodes and the total network lifetime, defined by amount of nodes active over round (100 nodes, 1000 rounds). It shows that the proposed protocol exceeds the

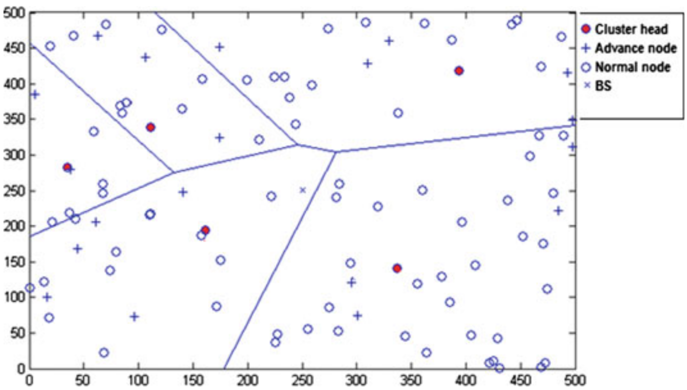


Fig. 5 Network partitioning by PSO

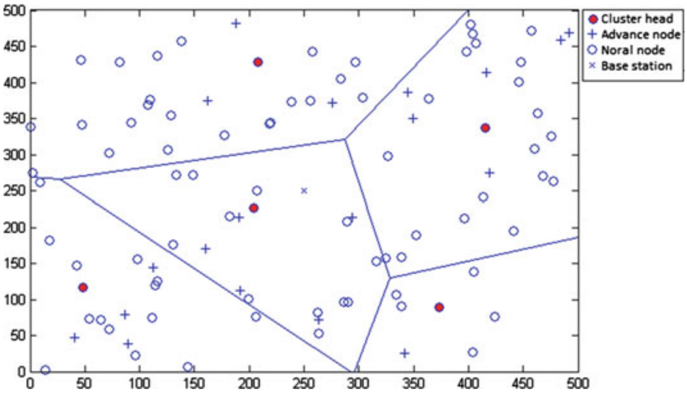


Fig. 6 Network partitioning by CS

total network lifetime of LEACH, LEACH-C, and PSO, by nearly 54, 38, and 9 % respectively. This improvement is based on two reasons. Firstly, with minimum intra cluster distance and optimal CH distribution across the network, the proposed protocol shows improved network partitioning. As a result, the energy dissipated by all nodes for communication is reduced. Secondly, the CS technique adopted in the protocol produces a set of good trade-offs where the values of the cost function are tolerable to the network requirements.

Figure 9 shows the total amount of data messages received at BS by all the protocols and algorithms. The proposed algorithm improves the data delivery by factors of 17 % over PSO, 54 % over LEACH-C and 161 % over LEACH. The idea behind this is, our protocol can take the benefit of picking the high energy node as a CH by taking into consideration the residual energy of the CH candidates and also the minimum distance between the nodes and their CHs by implementing optimized

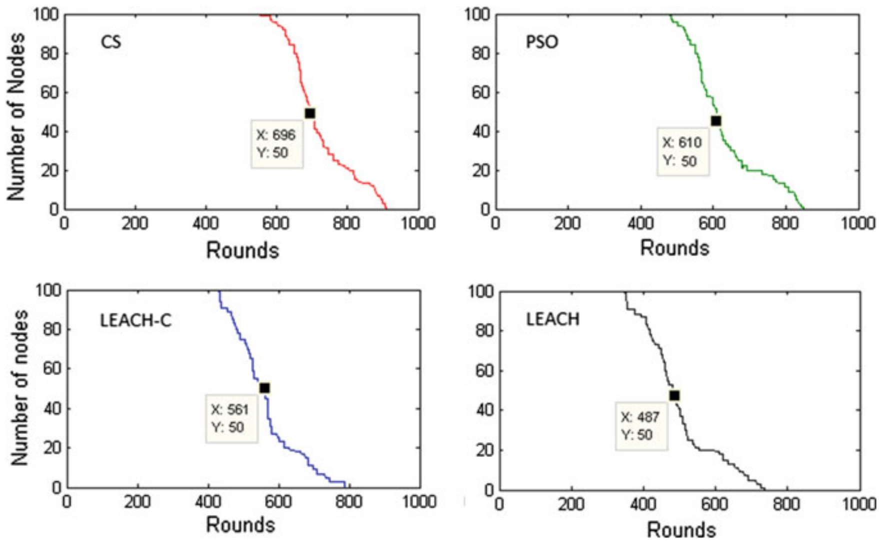


Fig. 7 Network lifetime after 50 % node dies

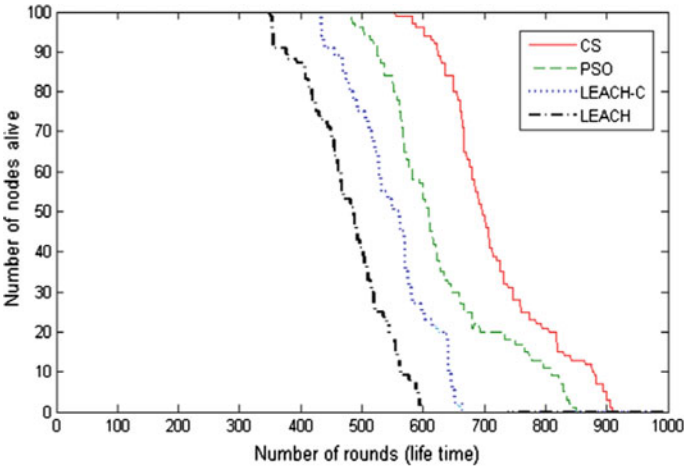


Fig. 8 Network lifetime till all the node dies

cost functions. Hence, more data messages are delivered to the BS. This experiment also facilitates the achievement of one of the objectives of this study.

For the scalability experiment in this work, the evaluation of the CS based clustering algorithm is tested for a number of incidents where the amount of nodes varies from 100 to 500. This simulation showed the limitation and capability of the WSN scalability of our proposed protocol. Previous results in this paper showed

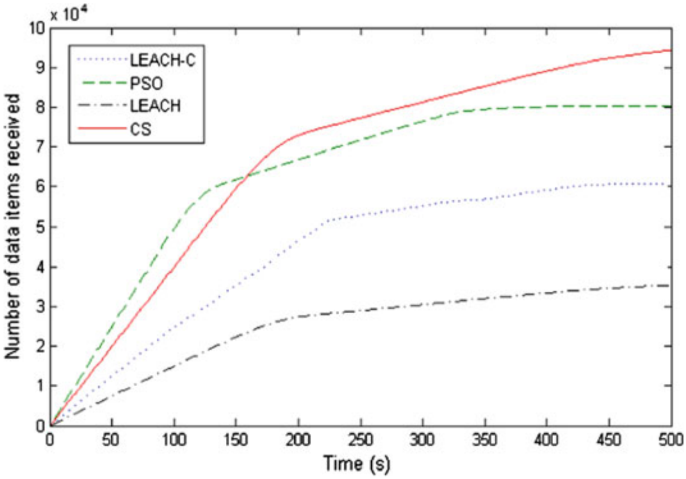


Fig. 9 Total amount of data received at the BS over time

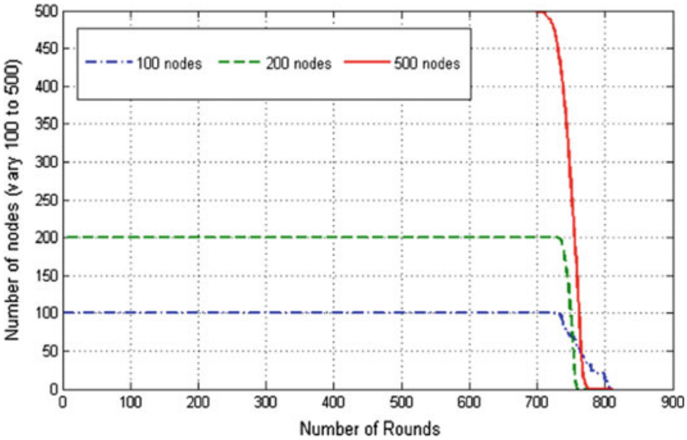


Fig. 10 Scalability of the proposed protocol

that the CS protocol runs well for 100 nodes. From Fig. 10, it is visible that it produces moderate results even when the node numbers are increased from 100 to 200 and 500 respectively. But it is noticed that, the efficiency of the proposed protocol decreases when the network is highly dense.

6 Conclusion and Future Works

In this paper, an energy-aware hierarchical clustering algorithm has been presented for WSNs, which uses CS algorithm. The cost function was defined in a manner, which takes into calculation the highest distance amid the non-CH nodes and its related CH, and the residual power of CH candidates in CH selection algorithm. Simulation results indicate that the proposed CS based protocol gives better network life time and capable of delivering more data to the BS compared to LEACH, LEACH-C and PSO based clustering approaches. In addition, the projected protocol generates better cluster formation by equally allocating the CHs all through the network area. Our future endeavor is to extend our work to the cross layer optimization between query and routing strategies. Furthermore, it can be extended by counting multi-hop communication among CHs to increase energy efficiency. Additionally hybridization of several other learning techniques could be done to determine more energy efficient clusters.

References

1. Kahn, J.M., Katz, R.H., Pister, K.S.J.: Emerging challenges: mobile networking for smart dust. J. Commun. Netw. **2**, 188–196 (2000)
2. Fang, W.C., Lin, T.H.: Low-power radio design for wireless smart sensor networks. In: International Conference on Intelligent Information Hiding and Multimedia Signal Processing 2006 IIHMSP06, pp. 583–586. IEEE (2006)
3. Florence, M.L., Swamydoss, D.: A survey on wireless sensor network architecture, protocols and applications. J. Glob. Res. Comput. Sci. **2**, 149–152 (2011)
4. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless sensor networks: a survey. Comput. Netw. **38**, 393–422 (2002)
5. Vasilescu, I., Kotay, K., Rus, D., Dunbabin, M., Corke, P.: Data collection, storage, and retrieval with an underwater sensor network. In: Proceedings of the 3rd International Conference on Embedded Networked Sensor Systems SenSys 05, vol. 154. San Diego (2005)
6. Werner-Allen, G., Lorincz, K., Ruiz, M., Marcillo, O., Johnson, J., Lees, J., Welsh, M.: Deploying a Wireless Sensor Network on an Active Volcano. IEEE Computer Society (2006)
7. Honeine, P., Mourad, F., Kallas, M., Snoussi, H., Amoud, H., Francis, C.: Wireless sensor networks in biomedical: Body area networks. IEEE (2011)
8. Zhang, Z.: Investigation of wireless sensor networks for precision agriculture. In: American Society of Agricultural and Biological Engineers, pp. 1157–1164 (2004)
9. Ruiz-Garcia, L., Lunadei, L., Barreiro, P., Robla, I.: A review of wireless sensor technologies and applications in agriculture and food industry: state of the art and current trends. Sensors **9**, 4728–4750 (2009)
10. Ozdemir, O., Niu, R.N.R., Varshney, P.K.: Channel aware particle filtering for tracking in sensor networks. In: Fortieth Asilomar Conference on Signals Systems and Computers, pp. 290–294 (2006)
11. Ozdemir, O., Niu, R., Varshney, P.K.: Tracking in wireless sensor networks using particle filtering: physical layer considerations (2009)
12. Wirawan, W., Rachman, S., Pratomo, I., Mita, N.: Design of low cost wireless sensor networks-based environmental monitoring system for developing country (2008)

13. Bahrepour, M., Meratnia, N., Poel, M., Taghikhaki, Z., Havinga, P.J.M.: Distributed event detection in wireless sensor networks for disaster management. In: International Conference on Intelligent Networking and Collaborative Systems, pp. 507–512. IEEE (2010)
14. Heinzelman, W.R., Chandrakasan, A., Balakrishnan, H.: Energy-efficient communication protocol for wireless microsensor networks. In: 33rd Hawaii International Conference on System Sciences, pp. 1–10. IEEE, Hawaii (2000)
15. Rodoplu, V., Meng, T.H.: Minimum energy mobile wireless networks. IEEE (1999)
16. Heinzelman, W.B., Chandrakasan, A.P., Balakrishnan, H.: An application-specific protocol architecture for wireless microsensor networks. IEEE Trans. Wirel. Commun. **1**, 660–670 (2002)
17. Caldari, M., Conti, M., Crippa, P., Nuzzo, G., Orcioni, S., Turchetti, C.: Instruction based power consumption estimation methodology. In: 9th International Conference on Electronics, Circuits and Systems, pp. 721–724. Dubrovnik (2002)
18. Biagetti, G., Crippa, P., Gianfelici, F., Turchetti, C.: Sensor network-based nonlinear system identification. In: 12th International Conference on KES, pp. 580–587. Zagreb (2008)
19. Biagetti, G., Crippa, P., Curzi, A., Orcioni, S., Turchetti, C.: TOLHNET : A low-complexity protocol for mixed wired and wireless low-rate control networks. In: Proceedings of the 6th European Embedded Design in Education and Research, pp. 177–181. Milano (2014)
20. Adnan, M.A., Razzaque, M.A., Ahmed, I., Isnin, I.F.: Bio-Mimic optimization strategies in wireless sensor networks: a survey. Sensors **14**, 299–345 (2014)
21. Salehpour, A., Mirmobin, B., Afzali-kusha, A.: An energy efficient routing protocol for cluster-based wireless sensor networks using ant colony optimization. In: International Conference on Innovations in Information Technology, pp. 455–459. Al Ain (2008)
22. Ziyadi, M., Yasami, K., Abolhassani, B.: Adaptive clustering for energy efficient wireless sensor networks based on ant colony optimization. In: Seventh Annual Communication Networks and Services Research Conference, pp. 330–334. IEEE Computer Society (2009)
23. Mehr, M.A.: Design and implementation a new energy efficient clustering algorithm using genetic algorithm for wireless sensor networks. World Acad. Sci. Eng. Technol. **53**, 430–433 (2011)
24. Shih, T., Member, S.: Particle Swarm Optimization Algorithm for Energy-Efficient Cluster-Based Sensor Networks, pp. 1950–1958 (2006)
25. Latiff, N.M.A., Tsimenidis, C.C., Sharif, B.S.: Energy-aware clustering for wireless sensor networks using particle swarm optimization. In: IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications, pp. 1–5. Athens (2007)
26. Yang, X., Deb, S.: Cuckoo search via Lévy flights. In: Nature and Biologically Inspired Computing. World Congress on NaBIC 2009, pp. 210–214. IEEE, Coimbatore (2009)
27. Civicioglu, P., Besdok, E.: A conceptual comparison of the cuckoo-search, particle swarm optimization, differential evolution and artificial bee colony algorithms. Artif. Intell. Rev. **39**, 315–346 (2011)
28. Adnan, A., A.Razzaque, M.: A comparative study of particle swarm optimization and cuckoo search techniques through problem-specific distance function. In: International Conference of Information and Communication Technology (ICoICT), pp. 88–92. Bandung, Indonesia (2013)
29. Dhivya, M., Sundarambal, M., Vincent, J.: Energy efficient cluster formation in wireless sensor networks using cuckoo search. In: SEMCCO'11 Proceedings of the Second international conference on Swarm, Evolutionary, and Memetic Computing, pp. 140–147. Springer, Berlin (2011)

Smart Location Reminder Based on Position Tracking

Pavel Masek, Jan Dvorak, Ondrej Krejcar and Kamil Kuca

Abstract This work addresses deployment of mobile devices' sensors in improvement of functionality of applications. The application that uses these sensors can offer better services to the user. In this project, the sensor for localization using GPS will be employed. This sensor will extend the possibilities of an application of reminder type with the possibility of notification based on the user's location. It will be possible to parametrize and combine this extension with the already known notifications based on time. The supplemented application is much more capable to carry out all the work and it is capable of more intuitive reminders. At the same time, the application is developed using the platform Apache Cordova and it will be targeted on all the main platforms.

Keywords Geolocation · Mobile application · Apache Cordova · Cross-platform development · Location tracking

1 Introduction

The employment of smart phone sensors is the key element to making a developed application more attractive. If the sensors are used correctly or in a resourceful way they have significantly bigger chance to gain a toehold in the market with mobile applications. In the broader sense, it is possible to regard any source of information

P. Masek · J. Dvorak · O. Krejcar (✉) · K. Kuca
Faculty of Informatics and Management, Center for Basic and Applied Research,
University of Hradec Kralove, Rokitanskeho 62, 500 03 Hradec Kralove, Czech Republic
e-mail: ondrej@krejcar.org

P. Masek
e-mail: Pavel.masek@uhk.cz

J. Dvorak
e-mail: jan.dvorak.2@uhk.cz

K. Kuca
e-mail: kamil.kuca@uhk.cz

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_54

that transmits data to another controlling unit as a sensor. Mobile phones literally abound in a variety of sensors such as accelerometer that measures the acceleration of the user's movement, a gyroscope that defines the orientation of the phone in the space, and many others.

This work uses the sensor for user's localization using GPS. GPS is a military satellite system for localization run by the Ministry of Defence of the United States of America. With its employment it is possible to determine the exact location unit the accuracy up to 10 m. The mobile sensor for localization is plentifully used for the improvement of services of the applications. A good example of these applications are the weather forecast applications. These applications get the user's location and based on that they view information about the weather in given location.

The way the GPS works is described in the articles [1, 2]. The articles introduce several algorithms that the GPS system is based on and that particularize the obtained data. There are uncountable possibilities to use the acquired user's location smartly from the simple display of the user's location on a social network up to a recommendation of the best restaurant in the neighbourhood.

This project will try to improve an application of reminder type using the acquired data from GPS. The application of reminder type can have different uses. For example, in the article [3] there was used an analogue of reminder mobile phone application that notified the patients with diabetes to take their medications. The basic application of reminder type can look as a notebook where the user takes notes of the tasks to do. That kind of an application is not very effective in its functionality. It reminds the user about the tasks only at the moment that the user uses this application. Better applications allow at least setting a time when the event or a task would notify the user. This is a better solution but not sufficient enough.

Using geolocation is an essential extension for providing reminder services. The geolocation allows to set the activation of notifications based on user's location. There is also a possible combination of notification based on user's location and the time in which the user should be notified. An application enhanced this way is more competent of carrying out its purpose. Development process of mobile application as well as wide (cloud) solution development need to be also based on strong economy model where we need to define purpose of solution as well as possible financial income or distribution model [4–6].

The application will be developed using the platform Apache Cordova that allows multiplatform development of mobile applications. The platform uses the technologies HTML, CSS and JavaScript for the application development. That way the application can be developed without writing a native code of specific platform.

2 Problem Definition

There is already quite a competition in the application field for notification administration. It can be assumed from the number of different rankings that compare the functionality of individual applications as found in the articles [7, 8].

For the evaluation of the rival applications, the selection was adapted to those that use notifications based on the user's location. Expected requirements for applications are: accurate notification of a created event according to the location, intuitive user interface, considerate manipulation with system sources, the possibility of editing and creating events using a web interface and the existence of an application on all major mobile platforms.

The first of the competitive applications is an application for Android platform that is called Geobell. Unfortunately, this application is created only for this platform. It notifies of the created events quite precisely. However, the use of this application is rather unintuitive. Among other of its drawbacks, there is also an excessive battery consumption. Also, it does not allow adding and editing events using a web browser.

One very well-known application for the administration of notifications is Google Calendar. It offers the option of adding and editing notifications using the web browser and mobile application and its simple and intuitive user interface. Perhaps its only flaw is the lack of the possibility of creating notifications based on the location.

The application LocationMinder is developed for the platform iOS. This application offers intuitive user interface and an accurate location detection. According to the user rating it can be concluded that this application is very popular. Among its disadvantages is the lack of web interface and the fact that the application exists only for the iOS platform.

The last from the selection of rival applications is the application Location reminder for Windows Phone platform. The application has not a very well arranged interface and overall bad implementation when it comes to using GPS and that results in a lower battery life.

The result of the research and testing of the rival applications is a discovery that none of the applications meet all the conditions that were required. The resulting application will be inspired by the parts that were designed well in the rival application and that way it will bring a combination of all the requirements in one application. Also, the application will be available on all major mobile platforms since we will be using the technology Apache Cordova.

3 New Solution

When developing mobile applications there was one very unpleasant fact for the developers or the developing studio. If they wanted to aim the application on all major mobile platforms, they needed to know three different programming languages for three major mobile platforms. The result was that the same application needed to be written separately for each platform and there was no way to use the written code again for another platform. This brought a very robust and high quality solution, however at the same time it caused the mobile application development to be a very expensive matter. The articles [3, 7–9] deal with the advantages and comparison with native solution. Nowadays, there are two ways that remove this problem.

The first way is provided by a company Xamarin with a commercial product with the same name. Xamarin allows writing multiplatform mobile application in the programming language C# from which a native code of given application for given platform can be generated. The development of applications using the product Xamarin is extensively described for example in literature [7–11]. The literature can guide the developers from the complete beginning up to the implementation of the application like online chat. The big advantage of this technology is the fact that it uses the native code of the given platform and that way it is faster than the following multiplatform solution.

The second possibility is using open source platform Apache Cordova, that stands as a foundation for a very well-known project for the production of multiplatform applications called PhoneGap. Apache Cordova uses HTML, CSS and JavaScript for the production of applications. Using this technology offers several advantages. The first advantage is the same as in the previous technology and that is a single code for all platforms. The second advantage are the used technologies for the application production. As a matter of fact, we could say that these technologies or at least their foundation is known by every developer. That is why it is much easier for a company to hire an employee to work with this technology.

The technology Apache Cordova basically works on the principle of insertion of a full-page component WebView. It is a component that displays the content of websites. A code written in HTML, CSS and JavaScript will be displayed in this component. Later, the functions of individual libraries of the platform Apache Cordova are called by JavaScript based on different events. These functions are mostly added to a global JavaScript object window.

The resulting application will proceed from the pieces of knowledge acquired from the research that is described in the previous chapter. One of the requirements is aiming the new application on all major platforms that are Android, iOS and Windows Phone. This goal will be achieved using this very platform, Apache Cordova. A more detailed procedure on how to develop application on this platform is described in literature [8–12].

Another fundamental point is an exact notification on events related to the user's location. Getting the location will be solved using the library org.apache.cordova.geolocation. The manner and possibility of using the library are introduced in the document [9–11]. A very closely related requirement is a requirement on a careful treatment of the system sources. This requirement is referring to an excessive battery consumption. The excessive battery consumption is usually caused by an overly short span between gaining the locations. By setting a suitable span between repeatedly localization the user, we can prevent this problem. This span would be further possible to edit in the application setting just in case that some user would not find it suitable.

For the intuitivity of user interface its development will follow the article [10] that is related to the development of multiplatform mobile applications in the comparison with the development of user interface written in the native code of the platform.

There is another requirement connected to the intuitivity and generally the simple controllability of the application. In a similar way that the rival solution Google calendar allows editing and synchronization notifications in a web browser, the new application should also have this supplement. The user gets another chance to insert and edit their notifications in an easier way using this supplement. The web application will use Java Spring MVC in the server and MySQL as a database. So that the application would seem as an ordinary application, its frontend part will be written as a single page application (further as SPA) in a same way as Google calendar. For creating SPA we will use an typescript framework to create SPA Bobril TS. The use of Bobril TS is described in online documentation [3, 7–10] including the examples and all the codes.

4 Implementation

Development of multiplatform applications using framework Apache Cordova uses JavaScript as a programming language. The application is then developed as a so called SPA website. Therefore, it is very convenient to already use some JavaScript framework for creating SPA. There is a selection of many popular frameworks such as AngularJS or React.

The firstly named AngularJS uses a standard approach to HTML DOM and its manipulation. As opposed to AnglularJS, React uses the also called virtual DOM. Basically all the HTML elements are generated using JavaScript. A big advantage of this approach is a much higher speed. The very popular social network Facebook is written in React.

Any of the above mentioned frameworks were not used for the application implementation. As a framework we chose Bobril TS. It is a typescript framework for creating SPA applications. Typescript is a superstructure of JavaScript that brings a type control, classes, interface modules, intelligence and code refactoring. Framework Bobril TS is inspired by a library React and so it also uses the virtual DOM. A big advantage is the variety of types that is welcome for creating a more extensive application. Another advantage of this framework is its size. The modified basic version has less than 10 kB.

The main point of the whole application is the implementation of user notifications based on their location. The library org.apache.cordova.geolocation is used for acquiring the user's location. That provides the access to GPS of the mobile device. It allows the implementation of methods not only for getting the coordinates successfully but also unsuccessfully. A detailed and a complete description of the library can be found in the online documentation [11].

Acquiring the user's location occurs every five minutes in the default application settings. That way the most precise notification is emphasized. This time interval can be edited in the application settings.

The user can create two basic types of notifications and those are the geolocation and time notifications. These two options can be later combined. In the case of time notification the start and the end of the notification has to be set. In the case of geolocation notification the setting is a bit more difficult. Inserting the geolocation notification can be done using two different ways. The first one is writing the exact address of the place of notification. The other option is choosing the location using a map. In order to display the map the library Google Maps JavaScript API v3 can be used.

The next parameter is setting notifications when coming to the given place or when leaving the place. The last parameter is choosing the radius for notifications.

By combining the geolocation and time notifications it is possible to achieve a very precise notification. For example, when leaving for a holiday “Remind me on 2.4.2015 from 3:30 to 4:30 in the distance 500 m from home—Have you packed your passports?”

The calculation of the distance between two GPS coordinates is calculated using haversian formula. For this calculation a small JavaScript library haversine.js is used. The use of this library is described in the online documentation [13].

The application always keeps the list of events. The geolocation events are being checked every time the device gets new coordinates to find out whether they occur. So according to the default setting it happens every five minutes. At the same time one of the previous locations of the user are preserved in order to be able to find out whether the user left the place of notification. For each geolocation event there is a calculation for the distance from the user’s location. This distance is later compared with the set radius of the event. In the case of the event of coming to a place the distance for the notification to go off must be smaller than the radius. In the case of leaving the place the actual distance has to be bigger than the radius and at the same time the previous distance has to be smaller than the radius of the given event (Fig. 1).

Entering the event is enabled even through a web interface. Because of that it is necessary to make a synchronization of the event in the mobile database with the database of the web application. An automatic synchronization is in the default settings set in a 15 min time interval. This automatic synchronization occurs only in the moment when the mobile device is not used. Also, the synchronization caused every time the application is started or reopened.

The synchronization as such runs that way that the application sends the time of the last synchronization that is saved in the user parameter to the server. Each user has a parameter of the last modification locally on the mobile device and also on the server.

The server compares the time of the last update sent by the mobile device with the time set in the database. In the moment when these two times differ, the server generates the array of all the active events and it sends them back to the mobile device together with a new time of the last update.

The mobile device updates the time of the last modification in its local storage and then it starts comparing the events. Each event has a unique ID generated by a database of the server part that serves as a definite identifier. Using this ID the events are compared. The events that have the same ID and different figures are updated in the mobile devices accordingly to the events received from the server.

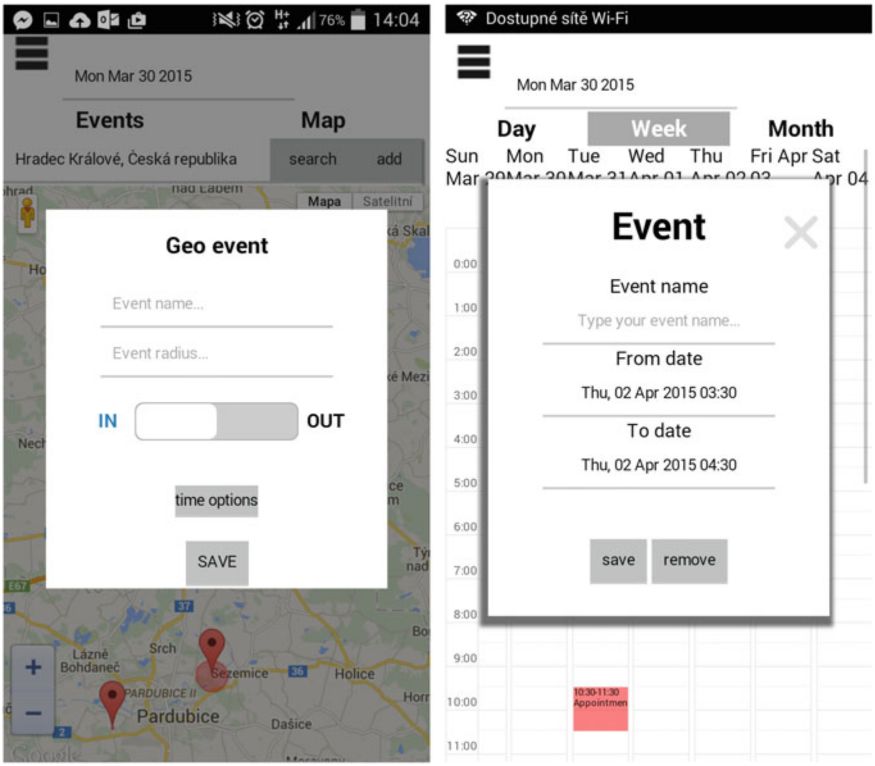


Fig. 1 Final GUI of developed application

The events that are saved only in the mobile device and are not in the event field generated by the server are deleted. And the events that are not saved in the mobile devices and are in the field generated by the server are added as new events.

The persistent events saving in the mobile devices is solved using saving json file. Json file was chosen because it represents the natural construction of the language object JavaScript. For the access to the storage of the mobile device the plugin org.apache.cordova.file is used. Using it can help to read easily and to record files into the mobile device storage. The entire documentation with examples of use is accessible online on [14].

5 Testing of Developed Application

The testing of the application occurred in several phases. The user interface was tested in the beginning of the development; saving data testing was added later together with the testing of getting the location. For the suitable, complex and many times repeating problems we used unit testing and more specifically the testing framework jasmine for testing of the JavaScript.

In the end, a final short UX testing was run together with installing the testing version of the application on three different mobile devices.

The user interface was tested in the first phase of the development. All the elements that cause any interaction were tested. Especially the buttons, opening the left menu, touch events and swipe gestures.

Based on the result of the first phase testing, we abandoned animating opening the left menu using CSS3 transition effect. The animation was not smooth and it did not seem as a functioning application.

The rival applications using native codes of given platforms had no notable problem with the smoothness of the animation.

The next phase consisted of testing the persistent data saving and its loading from the mobile device storage. The result of the testing would be the same as the list of events that it was created before the testing.

The tested cases were: saving during a standard turning off of the application, turning off the application using the task manager or when turning off the phone.

In all the cases of persistent saving testing, the list of events after the testing was the same as the list of events before the testing.

After the successful test of persistent saving, it was time to test event synchronization. Testing sets were created and they were tested for all possible cases that could happen. For every testing set a correct solution was created and that was to be achieved.

Among the tested cases there were these situations: no event is saved in the mobile device (it can be the case of a starting state of mobile application), no event is saved on server (the case when the application started being used), the device has events and the server generates empty space (here there are two possible states, if the time of actualization is different all the events would be deleted in the mobile device, if the time of the actualization is the same there are no events to be actualized, adding or removing), server generates a bigger number of events than the ones saved in the mobile device (here the reason can be an addition of new events through a web interface or addition of new and removing a smaller number of old events than the number of added ones), and the last case is generating a smaller number of events than those that are saved in the mobile device (the case can be a removal of the events or addition of lower number of new events).

For this phase of testing the unit testing was used with the aid of testing framework jasmine. The documentation for the testing framework jasmine including the example of the code is available online on [15].

The penultimate phase of testing was connected with a testing of notification correctness using the user's location. For this testing, Ripple emulator was used. It allows testing the applications like Apache Cordova right in the desktop browser. Ripple emulator can be installed as a package node.js or it is possible to add it as a plugin in the browser google chrome.

Ripple emulator allows simulating many states that the real mobile device can get into including the simulation of phone moving or the access to the network. It also allows simulating user's location that significantly facilitates and accelerates the development and the testing of the application.

During the testing, a testing data set with the correct results was created. After the calculation of notifications the results from the testing set had to match the results of the calculation.

The correctness of alerts was tested also by users on three different mobile devices. The testing devices were mobile phones with the system Android from Samsung—Galaxy S2 (Android 4.2), Galaxy S3 (Android 4.4) and Galaxy S4 (Android 4.4).

For the data collection from the feedback a popup window was edited so that it would alert to the event. Two buttons for a positive and negative evaluation were added to the window.

The test results confirmed an almost immaculate display of the popup window with the notification in the moment when the user reached a certain area. However, at the same time some errors appeared in the notification. In the moment when the user entered a very small radius (namely 100 m and less), the user passed the area before the application would request the location again. For that reason, the time for location requesting will have to be edited to a smallest interval possible but with a respect to the battery life.

The last phase of testing was a user UX testing. The users participated in this user testing. The application was introduced to them and they were explained what the application can operate. Afterwards, the application was installed into their mobile phones. Without any instructions they were given several tasks. Both users got the same tasks. First, they had to create two time events, then four events connected to a certain locations, and in the end two combined events (geolocation together with time event).

The users had no problem with creating a time event. Both accomplished this task correctly and in a very short time. However, in the case of creating a geolocation event a problem arose when using the map to enter the location. In this step, the user interface is not assembled very clearly and so it would be suitable to edit the logic of entering the location using the map. During creating a combined event no new problems arose.

A change of the default application displaying also emerged from the user testing. The users made a remark that even though it is a geolocation application, the first thing that is displayed after the launch is the calendar with time notifications. For this reason, we added a possible setting of the default displaying in the settings section.

Usage of created solution can be emergent also in critical areas e.g. in biomedical or biomedicine where the solution is based on sensors while it produce some kind of relevant information for physicians (patient enter some area—home or flat, etc.) [6, 16, 17]. In such areas where we are dealing with patient data we need to take care also of security issues of such solution [18–26].

6 Conclusions

The application was being developed as an application of Apache Cordova type. Thanks to this approach it was possible to create an application for all three main platforms with a single code. This approach could be a suitable way to lower the costs for a mobile application development and also the time for the development.

The application could supplement the reminder market. Unlike the rival solutions, it contains both the web interface and geolocation, and it is created for all three main mobile platforms, which are iOS, Android and Windows phone.

To a certain extent, the application meets the requirements defined in the beginning of the development. However, in the present state it would not be able to compete with the already created solution. As for the graphic side, it does not make a great first impression. Also, it would be convenient to work on a better alerting in the case of a small radius selection. Last but not least, it would be necessary to edit the user interface to bring it well fitted to the users using another UX testing where we would like to follow existing guidelines [27–28].

Acknowledgments This work and the contribution were supported by project “SP-103-2015—Smart Solutions for Ubiquitous Computing Environments” Faculty of Informatics and Management, University of Hradec Kralove, Czech Republic. Last but not least, we acknowledge the technical language assistance provided by Jirina Cancikova.

References

1. Yu, K., Dutkiewicz, E.: Correction to “geometry and motion-based positioning algorithms for mobile tracking in NLOS environments. IEEE Trans. Mob. Comput. **4**, 704 (2012)
2. Yu, K., et al.: Real-time mobility tracking algorithms for cellular networks based on Kalman filtering. IEEE Trans. Mob. Comput. **4**, 195–208 (2005)
3. Patel, S., et al.: Mobilizing your medications: an automated medication reminder application for mobile phones and hypertension medication adherence in a high-risk urban population. J. Diab. Sci. Technol. **7**(3), 630–639 (2013)
4. Maresova, P., Klimova, B.: Investment evaluation of cloud computing in the European business sector. Appl. Econ. **47**(36), 3907–3920 (2015)
5. Maresova, P., Halek, V.: Deployment of cloud computing in small and medium sized enterprises in the Czech Republic. E & M Ekonomie a Manag. **17**(4), 159–174 (2014)
6. Penhaker, M., Krejcar, O., Kasik, V., Snášel, V.: Cloud computing environments for biomedical data services. In: 13th International Conference on Intelligent Data Engineering and Automated Learning, IDEAL 2012, vol. 7435, pp. 336–343. LNCS, Natal (2012)
7. Kazmucha, Allyson. http://www.imore.com/. iMore. [Online] 14. 7 2014. http://www.imore.com/best-task-and-reminder-apps-mac-goodtask-clear-due-and-more. Accessed 18 Dec 2014
8. http://www.technobezz.com/. TECHNOBezz. [Online] 19. 6 2014. http://www.technobezz.com/6-top-best-android-apps-reminders/. Accessed 18 Dec 2014
9. Heitkötter, H., Hanschke, S., Majchrzak, A., Tim, A.: Evaluating cross-platform development approaches for mobile applications. In: Web Information Systems and Technologies, vol. 120. Springer, Berlin (2013)
10. Charland, A., Leroux, B.: Mobile application development. Commun. ACM **54**, 5 (2011)

11. Kriz, P., Maly, F.: Computer-aided OSPF network design tool. In: 34th International Conference on Information Technology Interfaces (ITI), pp. 105–110. Cavtat, CROATIA, 25–28 June 2012
12. Behan, M., Krejcar, O.: Adaptive graphical user interface solution for modern user devices. In: Lecture Notes in Computer Science 4th Asian Conference on Intelligent Information and Database Systems, ACIIDS 2012, vol. 6592, pp. 411–420. Kaohsiung, Taiwan, LNCS, 19–21 Mar 2012
13. Mobile application development, creating exciting apps for mobile devices using PhoneGap. Shotts, Kerri, Charland, Andre a Leroux, Brian. Birmingham, Packt Pub (2013)
14. Behan, M., Krejcar, O.: Modern smart device-based concept of sensoric networks. EURASIP Journal on Wireless Communications and Networking, vol. 2013, issue 1, no. 155. doi:10.1186/1687-1499-2013-155. Received 9 Oct 2012; Accepted 15 May 2013; Published 6 June 2013. ISSN 1687-1499 (2013)
15. Krejcar, O.: Threading possibilities of smart devices platforms for future user adaptive systems. In: Lecture Notes in Computer Science, 4th Asian Conference on Intelligent Information and Database Systems, ACIIDS 2012, vol. 7197, pp. 458–467, 19–21 Mar 2012. Kaohsiung, Taiwan, LNCS (2012)
16. Penhaker, M., Darebnikova, M., Cerny, M.: Sensor network for measurement and analysis on medical devices quality control. In: International Conference on e-Technologies and Networks for Development, ICeND 2011, vol. 171, pp. 182–196. Springer CCIS, Dar Es Salaam (2011)
17. Penhaker, M., Kasik, V., Stankus, M., Kijonka, J.: User adaptive system for data management in home care maintenance systems. In: Yeungnam, U. (ed.) 3rd International Conference on Intelligent Information and Database Systems, ACIIDS 2011, Wroclaw University, University of Information, vol. 6592, pp. 492–501. LNAI, Daegu (2011)
18. Horalek, J., Matyska, J., Sobeslav, V.: Communication protocols in substation automation and IEC 61850 based proposal. In: Proceedings of CINTI 2013—14th IEEE International Symposium on Computational Intelligence and Informatics, Art. no. 6705214, pp. 321–326 (2013)
19. Cimler, R., Matyska, J., Balík, L., Horalek, J., Sobeslav, V.: Security issues of mobile application using cloud computing. Adv. Intell. Sys. Comput. **334**, 347–357 (2015)
20. Pies, M., Hajovsky, R., Latocha, M., Ozana, S.: Radio telemetry unit for online monitoring system at mining dumps. Appl. Mech. Mater. **548–549**, 736–743 (2014)
21. Ozana S., Pies M., Horacek, O., Zidek J.: Concept of automated transportation center and its PIL model. In: World Congress on Engineering and Computer Science (WCECS'13)—International Conference on Modeling, Simulation and Control (ICMSC'13), vol. 2, pp. 992–996. ISBN: 978-988-19253-1-2 (2013)
22. Vanus, J., Novak, T., Koziorek, J., Konecny, J., Hrbac, R.: The proposal model of energy savings of lighting systems in the smart home care. In: Proceedings of IFAC, vol. 12 (Part 1), pp. 411–415 (2013)
23. Vanus, J., Koziorek, J., Hercik, R.: Design of a smart building control with view to the senior citizens' needs. In: Proceedings of IFAC (IFAC-Papers Online), vol. 12 (Part 1), pp. 422–427. ISSN: 14746670, ISBN: 9783902823533 (2013)
24. Jancikova, Z., Kostial, P., Bakosova, D., Ruziak, I., Frydrysek, K., Valicek, J., Farkasova, M., Puchky, R.: The study of electrical transport in rubber blends filled by single wall carbon nanotubes. J. Nano Res. **21**(16), 1–6. ISSN: 16625250 (2013)
25. Jonsta, Z., Kostial, P., Ruziak, I., Jonsta, P., Jurciova, J., Jancikova, Z., David, J., Kopal, I.: Advanced fillers enhancing thermal and mechanical properties of rubber blends. J. Nano Res. **13**, 27–32. ISSN: 1662-5250 (2011)
26. Machacek, Z., Slaby, R., Hercik, R., Koziorek, J.: Advanced system for consumption meters with recognition of video camera signal. Elektronika Ir Elektrotechnika. **18**(10), 57–60. ISSN: 1392-1215 (2012)

27. Kriz, P.: Automated WiFi-based localization and visualization of wireless network. In: 13th International Conference on Intelligent Software Methodologies, Tools, and Techniques (SoMeT), Langkawi, Malaysia, 22–24 September 2014. Frontiers in Artificial Intelligence and Applications, vol. 265, pp. 432–445 (2014)
28. Machaj, J., Brida, P.: Optimization of rank based fingerprinting localization algorithm. In: 3rd International Conference on Indoor Positioning and Indoor Navigation, pp. 1–7, Sydney, Australia (2012)

Three Dimensional (3D) Cost-Downtime Model for Hostel Facilities Maintenance

Yuseni Ab Wahab, Abd Samad Hasan Basari, Burairah Hussin, Kyrina Airin Abu Samah and Noramalina Mohd Sabri

Abstract This paper is part of an on-going research on the development of maintenance cost and down time model for Higher Education Institution Hostel facility maintenance in Malaysia where the case study is conducted at Kolej Universiti Islam Melaka (KUIM). The model is developed to assist maintenance staff based on calculated total cost and downtime from the maintenance data. The decision that has to be made is focused on the replacement action where all taking account on the frequency basis. This is due to the item is not subject to failure but consider the operating cost and downtime with use. The advantage of the graph of total cost downtime per unit time curve is with the optimal value, it consider the total cost around the optimum. If the curve is fairly flat around the optimum, it shows that no extensive plan for the replacements exactly at the optimum. However, if there is uncertainty regarding the value of the stated parameter required during the analysis, then the optimal replacement is doubtful. In addition, by giving variety set of values of the total cost and downtime, the curve could in consequence affect the optimal solution and graphically shown in 3 Dimensional (3D).

Keywords Hostel facilities maintenance · Replacement cost and downtime model · Three dimensional cost-downtime model

Y. Ab Wahab (✉)
Faculty of Business Innovation and Accounting, Kolej Universiti Islam Melaka (KUIM), Kuala Sungai Baru, 78200 Masjid Tanah, Melaka, Malaysia
e-mail: yuseni@kuim.edu.my

A.S.H. Basari · B. Hussin · K.A.A. Samah · N.M. Sabri
Faculty of Information and Communication Technology, Centre for Advanced Computing Technology (C-ACT), UTeM, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia
e-mail: abdsamad@utem.edu.my

B. Hussin
e-mail: burairah@utem.edu.my

K.A.A. Samah
e-mail: rinarinza@gmail.com

N.M. Sabri
e-mail: noramalinamohdsabri@gmail.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_55

1 Introduction

Some equipment operates with excellent efficiency when it is new. But as it ages, the performance deteriorates. An example is the door components in Kolej Universiti Islam Melaka (KUIM) hostel facilities maintenance. When new, it is considered as the equipment is in good condition. However if there has a small crack, it will affect the quality of the equipment. In general, replacements will cost money in terms of component and a balance is required between the money spent for replacements and savings obtained to reduce the operating cost [1, 2]. The optimal replacement policy can be determined by minimizing the sum of operating and replacement costs and downtime per unit time [3, 4]. The developed model can be used to optimize component replacement decision. The interest in this decision area is initiated by attempting to increase the reliability of the system or the building [5, 6].

Most equipment has being inspected by following the preventive replacement according to the schedule and requirements. Thus it is necessary to identify which component should require preventive replacement during maintenance and which should be run until breakdown. If the component is selected for preventive replacement during maintenance, then the related question to be answered is: What is the best time to perform maintenance? Based on the fact, primary goal addressed in this study is to make maintenance more reliable through preventive replacement [6, 7].

Replacement in maintenance can be classified as either deterministic or probabilistic (stochastic). Deterministic problem are those in which the timing and outcome of the replacement action are assumed to be known with certainty [8, 9]. For example, the component that is not subject to failure but whose operating cost increases with use [10]. In order to reduce this operating cost, a replacement can be performed. After the replacement, the trend in operation cost downtime is decreased. Meanwhile, probabilistic problem are those where the timing and outcome of the replacement actions are depended on chance where the equipment may be described as being good or breakdown. The probability law defining changes from good to fail is described by the distribution of time where completion failure is a random variable. The distribution then is named as the equipment's failure distribution [11, 8].

2 Literature Review

Related works regarding the cost and downtime models for hostel facilities maintenance is discussed in this section.

2.1 Optimal Replacement Time for Component

Normally, most component or equipment operates with outstanding efficiency when new. As it ages the performance will deteriorates. Generally a balanced replacement cost and downtime is required between the money spent on replacement and saving obtained by reducing the operating cost. Thus, in order to determine an optimal replacement policy, it is necessary to minimize the sum of operation and replacement cost per unit time [11, 8].

To deal with the optimization problem, in general, the aim is to optimize some measure of performance over a long period of time. This approaches is simpler mathematically when compare to developing a model for optimizing a measure of performance over a finite horizon [12].

Cost is usually conflicted and associated with optimization problem. This class of problem can be termed as short term deterministic since the magnitude of the interval between replacements is weeks or months. If the interval between replacements was measured in years, then the money changes in value over time would need to be taken into account in the analysis [12, 13].

2.2 Stochastic Preventive Replacement

In normal conditions, preventive replacement action is taken before equipment reaches a failed state. This requires two necessary conditions:

1. The total cost of the replacement must be higher when failure occurs during operation rather than during scheduled maintenance. This may be caused by a higher loss of production since replacement after failure is unplanned or failure of one piece of plant may cause damage to other equipment [13].
2. The hazard rate of the equipment is kept increasing. The failures occur according to the negative exponential distribution or equivalent with the Weibull distribution, where the shape parameter $\beta = 1.0$. When this is the case, replacement before failure does not affect the probability that the equipment will fail in the next operation, given the current condition is good. Consequently, if preventive replacement is applied to equipment that fails according to the negative exponential distribution, the maintenance is inefficient. Obviously, the hazard rate is decreasing when equipment fails according to the hyper exponential distribution or the Weibull whose β value is less than 1.0 and again preventive replacement should not be applied [13, 11].

2.3 Optimal Preventive Replacement

Failure is always unexpected and it is not unreasonable to assume that failure replacement is more costly than a preventive replacement. However, a balance is required between the preventive replacement cost and their resulting benefits that can reduce failure replacements [8]. The replacement policy is considered as a preventive replacement which occurs at fixed intervals of time. Meanwhile failure replacement occurs whenever necessary. The optimal interval between the preventive replacement can be determined to minimize the total expected cost downtime of replacing the equipment per unit time [12, 13].

2.3.1 Cost Model

The model construction is as follows.

Construction of the Model

1. *c(t)* is the operating cost per unit time at time *t* after replacement
2. *C*, is the total cost of a replacement.
3. The replacement policy/to perform replacements at interval length
4. The objective is to determine the optimal interval between replacements to minimize the total cost of operation and replacement per time

   The total cost per unit *C(t)* for replacement at time *t*, is
   C(t) = total cost in interval (0,t) length of interval

To use the equation $c(t_r) = C(t_r)$, it requires that the trend in operating costs is consider as an increasing function. Any costs, such as production losses incurred due to the duration of the replacement needs to be incorporated into the cost of the replacement action. Otherwise, a numerical solution is required as in Eq. (1):

$$C(t_r) = \frac{\int_0^{t_r} c(t)dt + C_r}{t_{r+T_r}} \tag{1}$$

2.3.2 Downtime Model

The purpose of downtime model is to minimize the total downtime per unit time. In some cases, the difficulties in costing or the desire to get maximum throughput or utilization of equipment require the replacement policy that minimizes the total downtime per unit time or, equivalently, maximizes availability [14]. The problem is to determine the optimal times at which replacements should occur to minimize total

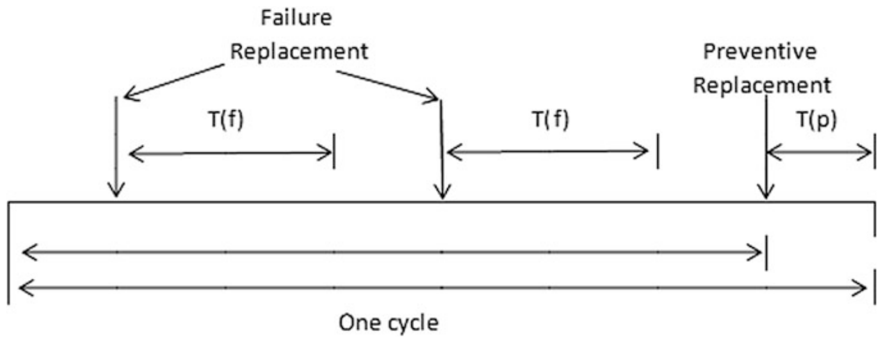


Fig. 1 Downtime minimization: optimal interval

downtime per unit time. The basic conflict are that as the preventive replacement frequency increase, there is an increase in downtime due to these replacement, but a consequence of this is a reduction of downtime due to failure replacements, and the aim is to get the best balance between them.

The model is developed to determine the optimal replacement interval between the preventive replacements in order to minimize the total downtime per unit time. The policy is illustrated in Fig. 1.

The total downtime per unit time, for preventive replacement at time t_p, denoted as $D(t_p)$ is in Eq. (2).

$$D(t_p) = \frac{\text{Expected downtime due to failures } + \text{ downtime due to preventive replacemene}}{\text{Cyclelength}} \quad (2)$$

Downtime due to failures = number of failures in interval $(0, t_p)$ × Time required to make a failure replacement = $H(t_p) \times T_f$.

Downtime due to preventive replacement = T_p.

Therefore Eq. (3)

$$D(t_p) = \frac{H(t_p)T_f + T_p}{t_p + T_p} \quad (3)$$

This is a model of the problem relating replacement interval t_p to total downtime $D(t_p)$.

3 Case Study

The hostel building maintenance data is gathered from ICYM for a period of time.

Table 1 Replacement cost for door

Month	MYR
1	17.99
2	17.81
3	17.71
4	17.64
5	17.54
6	17.38
7	17.35
8	17.27
9	17.23
10	17.19
11	17.20
12	17.25

MYR Malaysia ringgit

3.1 Cost Model

By referring to Eq. (1), Table 1 is obtained where the optimal replacement age is 10 months, and the associated cost per month is MYR 17.14. Table 1 also shows the deterioration trend from month 1 to 12 and increase again from 10 to 11. The associated graph of cost per month versus time is provided in Fig. 2, which includes the calculation of the optimizing criterion $c(t) = C(f_r)$ when the trend in operating cost is discretized. Therefore, the replacement is at the end of month 10, since the next period of operations and maintenance cost, $c(t = 10)$, is higher than the average cost to date (MYR 17.19).

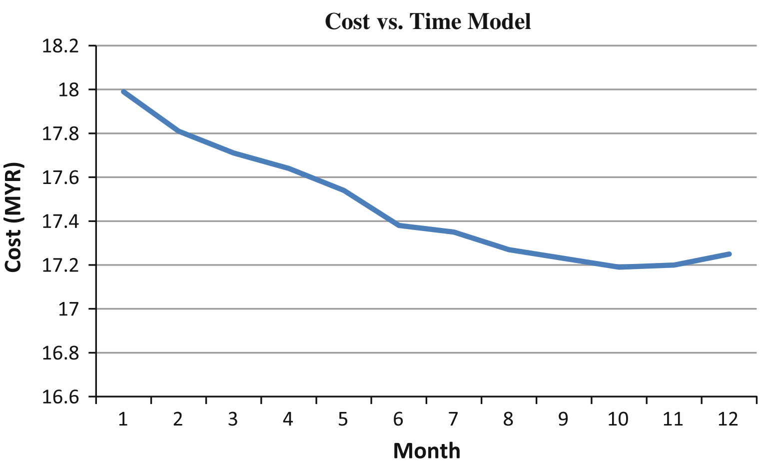


Fig. 2 Replacement cost for door

Sample numerical Calculation
The simplified after referring to Eq. (1) is:

$$C(tr) = (1/t) * (((600.44/30) * t + ((22172.94/30) * EXP(-X * (t))/X) - ((600.44/30)/X) + (200)/30))$$

where t = 24 h (1 Day) and X = Exponent/days

$$\begin{aligned} X &= 5357/30\,\text{days} = 0.178567 \\ &= (1/24) * (((600.44/30) * 24 + ((22172.94/30) * EXP(-X * (24))/X) \\ &\quad - ((600.44/30)/X) + (200)/30)) \\ &= 17.99 \end{aligned}$$

3.2 Downtime Model

The method applied from the downtime model, the corresponding curve of $D(T)$ in Table 2 for the preventive replacement and presented graphically in Fig. 3. The remarks that can be concluded are that the assumption verified by the curve that the preventive replacement plotted above the best time to the replacement of component hostel facility maintenance. It also shows that when the 18 value increased the curve will go nearer to perfect replacement from 6.6 value downtime, if the quality

Table 2 Replacement downtime for door

t (month)	Tf	Tp	lambda	f(t); if exponent	H(t); if exponent	Downtime (h)
	7	0.035	5.309	lambda * EXP(−lambda * t)	lambda * t	
1	25	0	5.309	0.026262939	5.309	132.725
2	34	3	5.309	0.000129919	10.618	72.8024
3	18	5	5.309	6.42694E-07	15.927	36.46075
4	14	12	5.309	3.17933E-09	21.236	19.3315
5	5	7	5.309	1.57277E-11	26.545	11.64375
6	3	7	5.309	7.7803E-14	31.854	7.889385
7	4	18	5.309	3.84881E-16	37.163	*6.66608*
8	11	60	5.309	1.90396E-18	42.472	7.752824
9	15	85	5.309	9.41864E-21	47.781	8.528883
10	31	120	5.309	4.65928E-23	53.09	13.583
11	40	90	5.309	2.30489E-25	58.399	24.01941
12	65	90	5.309	1.1402E-27	63.708	41.48059

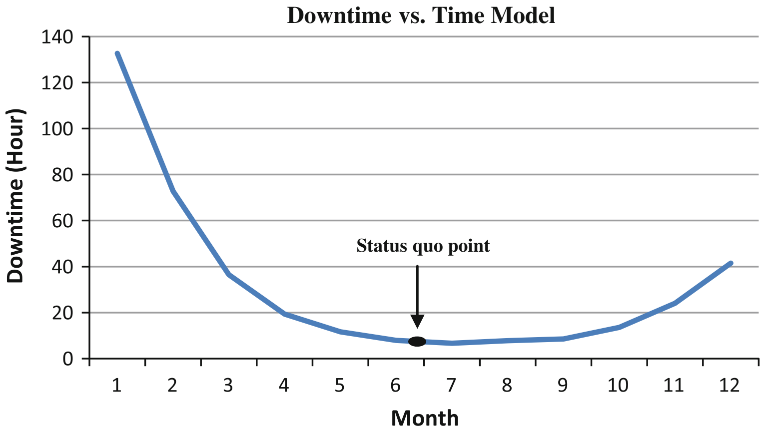


Fig. 3 Replacement downtime for door

of preventive replacement downtime, means that the more downtime detected, the downtime will reduce due to fewer breakdowns occurred during operations [15]. Details of the percentage of the expected downtime to fit the status quo point are also shown in Table 2 and Fig. 3.

3.3 *Three Dimensional (3D) Graph: Cost-Downtime Replacement Model*

The three dimensions can be labelled by a combination of cost, downtime and time. Any three directions can be chosen, provided that they do not all platters in the same plane. In physics and mathematics, a sequence of n numbers can be understood as a location in n-dimensional space. When $n = 3$, it can be considered as the three-dimensional of cost, downtime and time replacement space. It is represented by Table 3 and Figs. 3, 4.

The cost and downtime model, the corresponding curve of $D(T)$ in Table 3 for the cost and downtime replacement the graphically shown in Figs. 4 and 5. The remarks that can be concluded are that the assumption verified by the 3D curve that the best time to replacement cost and downtime plotted above the perfect replacement. Details of the percentage of the expected downtime to fit the status quo point are shown in Table 3.

Table 3 Three dimensional (3D) graphs: cost-downtime replacement model

Time (month)	Cost (RM)	%	Downtime (h)	%
1	17.99	8.58	132.725	34.7
2	17.81	8.5	72.8024	19.01
3	17.71	8.45	36.46075	9.53
4	*17.64*	*8.42*	*19.3315*	*5.04*
5	17.54	8.36	11.64375	3.04
6	17.38	8.29	7.889385	2.05
7	17.35	8.27	6.66608	1.75
8	17.27	8.24	7.752824	2.03
9	17.23	8.22	8.528883	2.23
10	*17.19*	*8.16*	*13.583*	*3.55*
11	17.2	8.28	24.01941	6.27
12	17.25	8.23	41.48059	10.8
Total	209.56	100	382.8836	100

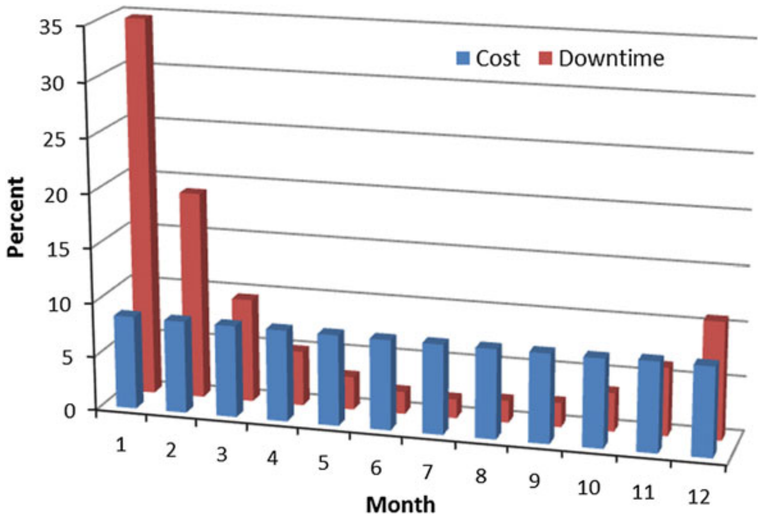


Fig. 4 Three dimensional (3D) graphs: cost-downtime replacement model

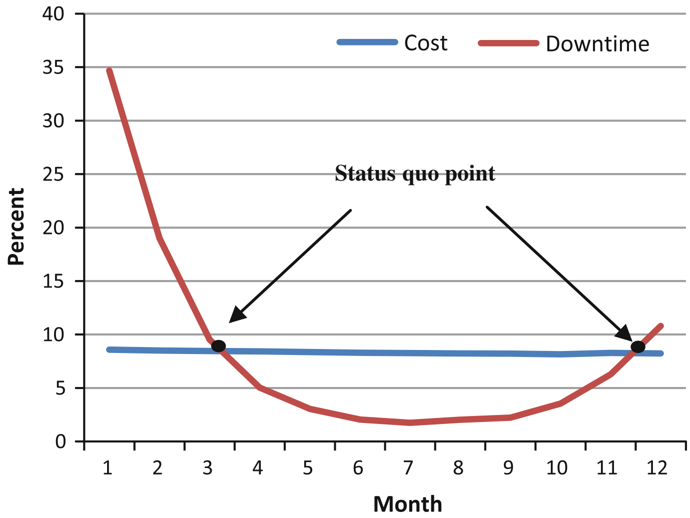


Fig. 5 Three dimensional graphs (3D): cost-downtime replacement model

4 Conclusion

The hostel facilities maintenance model shows that the total cost downtime curve has shown its optimal values. It is beneficial to plot the total cost downtime per unit time curve to assist maintenance staff in making appropriate decision. The advantage of the curve is by giving the optimal value of t, the total cost downtime is plotted around the optimum value. If the curve is fairly flat around the optimum, there is no urgency for the maintenance staff to prepare for the replacements to achieve the optimum value, thus giving some flexibility in scheduling the work. When necessary the replacement duration can be embedded into the replacement model, which the aim is to minimize the total downtime or equivalent and the maximization of item availability. This research has presented a model that can be used to establish the optimal time based which discard decision if the goal is to identify the interval of preventive replacement policy for future improvement, the model can be hybrid that with Artificial Intelligence (AI) and data mining technique to increase its accuracy.

Acknowledgments This research is supported by Kolej Universiti Islam Melaka (KUIM) and is part of the Degree of Doctor of Philosophy (Ph.D.) in the Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM).

References

1. Basari, A.S.H.: Maintenance modeling tools with special to incomplete data. PhD Thesis, Universiti Teknikal Malaysia Melaka (UTeM), Malaysia (2009)
2. Abinova, A.Y. (Zakerya): Innovations in the sphere of higher professional education: experience and future. Middle-East J. Sci. Res. **17**(1), 91–95 (2013)
3. Bamadi, M.: A review on education of dentistry. Middle-East J. Sci. Res. **18**(7), 878–886 (2013)
4. Taubaeva, S., Bulatbaeva, A., Bogatyrev, Y., Dementyev, A., Barsay, B.: Study of the educational space of a higher military educational establishment. Middle-East J. Sci. Res. **21** (1), 51–55 (2014)
5. Udefi, A.: The trouble with nigerian universities, the menace of bogus policy and speculative ideology. Middle-East J. Sci. Res. **21**(5), 855–869 (2014)
6. Jones, B., Jenkinson, I., Wang, J.: Methodology of using delay time analysis for a manufacturing industry. Reliab. Eng. Syst. Saf. **94**(1), 111–124 (2009)
7. Wang, W.B., Banjevic, D., Pecht, M.: A multi-component and multi-failure mode inspection model based on the delay time concept. Reliab, Eng. Syst. Saf. **95**(8), 912–920 (2010)
8. Ab Wahab, Y., Basari, A.S.H.: Best parameter of the hostel building component maintenance. Int. J. Comput. Appl. **110**(15), 17–20, ©IJCA (2015)
9. Ab Wahab, Y., Basari, A.S.H.: Hostel facility maintenance preliminary finding of higher education institution in Malaysia. Int. J. Sci. Res. **5**(5). ISSN 2250-3153 (2015)
10. Wang, W.B.: Delay time modeling. In: Murthy, D.N.P., Kobbacy, A.K.S. (eds.) Complex System Maintenance Handbook, pp. 345–370. Springer, Amsterdam (2008)
11. Ab Wahab, Y., Basari, A.S.H.: Identifying the best parameter distribution for university hostel building maintenance. Middle-East J. Sci. Res. **22**(8), 1145–1149. ISSN 1990-9233, ©IDOSI Publications (2014)
12. Ab Wahab, Y., Basari, A.S.H.: Replacement model for higher education institution hostel building maintenance in malaysia. Int. J. Trade Econ. Finance IJTEF **5**(5), 449–453. ISSN: 2010-023X (2014)
13. Ab Wahab, Y., Basari, A.S.H., Hussin, B.: Replacement model for hostel building case study: ICYM. Middle-East J. Sci. Res. **21**(11), 1977–1981. ISSN 1990-9233, ©IDOSI Publications (2014)
14. Ab Wahab, Y., Basari, A.S.H.: Analysis of down time and reliability estimation in hostel building maintenance—a case study. Middle-East J. Sci. Res. **17**(9), 1260–1268 (2013)
15. Ab Wahab, Y., Basari, A.S.H.: Building maintenance management preliminary finding of a case study in ICYM. Middle-East J. Sci. Res. **17**(9), 1260–1268 (2013)

A New Model of Crypt Edge Detection Using PSO and Bi-cubic Interpolation for Iris Recognition

NurulAkmal Hashim, Zaheera ZainalAbidin, AbdulSamad Shibghatullah, Zuraida AbalAbas and Norzihani Yusof

Abstract Several attempts have been made to improve the iris recognition system from first into second generation which is proficient to recognize unique iris features such as crypts. However, the always changing iris features create difficulties in comparison phase to determine the genuineness. Therefore, to determine genuineness, this study proposes a new model of iris recognition using combinational approach of particle swarm optimization (PSO) and Bi-cubic interpolation techniques in selecting the best crypt among unique iris features template. The particles in PSO searches the most optimal crypt features in iris texture meanwhile, Bi-cubic interpolation technique create sharp and refined crypt images. The results indicate an improvement of PSNR rates which are 22.6836 dB for CASIA and 22.3312 dB for UBIRIS database. Moreover, this method has indicates that the output from two databases are within tolerate noise rate. The implication of this study contributes to a new method of feature extraction using bio-inspired, which enhanced the ability of detection in iris recognition.

N. Hashim (✉) · Z. ZainalAbidin · A. Shibghatullah · Z. AbalAbas · N. Yusof
Optimization, Modelling, Analysis, Simulation and Scheduling (OptiMASS) Research Group, Fakulti Teknologi Maklumat & Komunikasi, Universiti Teknikal Malaysia Melaka, 76100 Hang Tuah Jaya, Durian Tunggal, Melaka, Malaysia
e-mail: akmal112@hotmail.com

Z. ZainalAbidin
e-mail: zaheera@utem.edu.my

A. Shibghatullah
e-mail: samad@utem.edu.my

Z. AbalAbas
e-mail: zuraida@utem.edu.my

N. Yusof
e-mail: norzihani@utem.edu.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_56

1 Introduction

Iris recognition is one of the systems in biometric identification. It is an autonomous system that uses mathematical pattern recognition techniques for measuring the human iris. It has evolved more than a decade since it is introduced by John Daugman. Each of the iris has a unique pattern that cannot be replaced by others even there are a twins [1]. These iris patterns determine the one of the kind information and only a specific human's iris contains a unique iris features biometric data [2].

However, these always changing iris features create difficulties in comparison phase to determine either genuine or impostor. It is shown that a failure was detected in 21 % of intra-class comparisons cases, taken at both three and six month intervals [3]. Therefore, in order to reduce the re enrollment or failure matching, the unique iris features such as crypts, radial furrows, concentric furrows, freckles and blotches are easy to identify and been recognize. The example of crypts and furrow are show in Fig. 1.

This unique iris features are depending on the micro characteristics of iris features [4]. However, the almost stable iris features are only sustained for a certain period of time [1]. Nevertheless, the change in iris feature shows that it can be sustainable up to six years [5].

Furthermore, due to the change in iris features, the number of not genuine is arising in iris recognition system proportionally. The non-genuine user is categorized into two types; the first type is someone who is trying to penetrate the systems and pretending to be the original person. Meanwhile, the second type is when the system rejects the genuine user who wants to access the system.

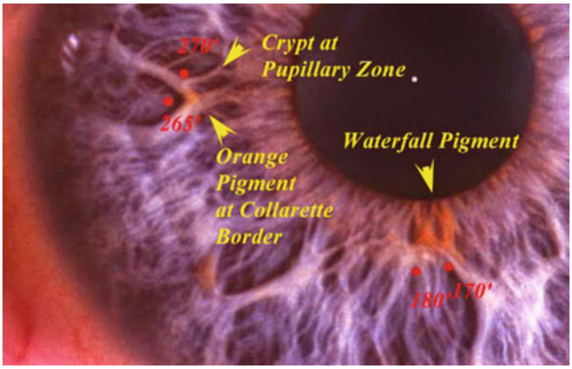


Fig. 1 Human iris [20]

2 Related Work

Continuous research proposal of iris recognition method exist with the same main goal of addressing the rejected genuine issue [6]. Example of methods which has been used by other researcher such as deblurring [7, 8], white noise insertion [9, 10], image enhancement [11], multiple biometric modality analysis [12–14], compression [15, 16] and the selection of unique iris features [17]. Among all, the selection of unique iris features research has gain latest attention. There are techniques in selecting unique iris features such as PSO [15], ACO [18], ant-CBIR [18] and Bee Optimization [19]. However, all of these techniques still produce high noise rate images and consume processing power. Therefore, a combination of PSO and Bi-cubic interpolation has been proposed to overcome the difficulties to determine either genuine or vice versa.

The Particle Swarm Optimization (PSO) algorithm is used to identify the significant and unique iris features, which is according to the bird flocking or fish schooling behavior, meanwhile Bi-cubic interpolation re-sampling the image to raise the number of pixels in a digital image.

In the proposed method as in Sect. 3, the existing technique of PSO has been under studied and evaluated. PSO is one of the techniques that inspired by the behavior of bird flying or fish schooling which introduced by Kennedy and Eberhart [12]. Moreover, PSO is one of the new heuristic algorithms that acts as randomly generated population and utilizes a fitness value to evaluate the population. In fact, PSO updates the population and searches for the optimum value with random techniques. In addition, PSO has more effective memory capability since each of the particles remembers its own previous and neighborhood best value [15]. Besides, PSO is easier to be implemented and have fewer parameters to be adjusted. Other than that, PSO only select the ‘best’ particle that shares the information to others. It is a one-way information sharing mechanism; the evolution is determined to provide for the best solution.

3 Propose Method

In the proposed research methods, techniques of PSO and Bi-cubic interpolation have been under studied and evaluated.

3.1 Particle Swarm Optimization

The performance of each particle is measured using a fitness function which depends on the optimization problem [15, 16, 21]. Each particle *i* flies through the n-dimensional search space R^n and maintain the following information:

- X_i, the current position of the particle i(x-vector)
- P_i, The personal best position of the particle i(p-vector).
- V_i, the current velocity of the particle i(v-vector).

The personal best position associated with a particle *i* is the best position that the particle has visited so far. If *f* denotes the fitness function, then the personal best of particle *i* at a time step *t* is updated as:

$$\mathrm{Pi}(t+1) = \begin{cases} \mathrm{Pi}(t) & \text{if } f(xi(t\ +\ 1)) \geq f(\mathrm{Pi}(t)) \\ x_i(t+1) & \text{if } f(xi(t\ +\ 1)) < f(\mathrm{Pi}(t)) \end{cases} \tag{1}$$

In order to locate the position of yielding the lowest error among all the P*i* is called the global best position and is denoted as gbest. It is shown in the Eq. 2.

$$\begin{aligned} \text{gbest} &\in \{P0(t), P1(t), \ldots, Pm(t)\} \\ &= \min\{f(P0(t)), f(P1(t)), \ldots, f(Pm(t))\} \end{aligned} \tag{2}$$

The velocity updates are calculated as a linear combination of position and velocity vectors. Thus, the velocity of particle I is updated using Eq. 3 and the position of particle I is updated using Eq. 4:

$$x_i(t+1) = x_i(t) + V_i(t+1) \tag{3}$$

$$V_i(t+1) = wv_i(t) + c_1 r_1 (p_1)(t) - x_i(t) + c_2 r_2\,\text{crypt} - x_i(t) \tag{4}$$

where,

v_i must be in a predefined range [V_{min}, V_{max}], where
if $v_i > V_{max}$ then $v_i = V_{max}$, and if $v_i < V_{min}$ then $v_i = V_{max}$.

In the formula, w is the inertia weight, c_1 and c_2 are the acceleration constants and r_1 and r_2 are random numbers in the range [0, 1]. Moreover, p*i* is the personal position of the particle *i* and crypt is the updated crypt (Fig. 2).

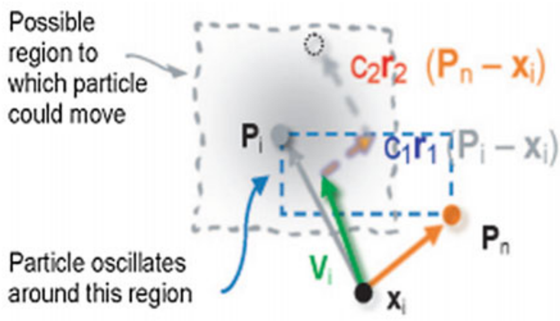


Fig. 2 A 2D illustration is shows, on how the velocity moves and updates the equation and regions [16]

3.2 Bi-cubic Interpolation of Iris Feature Template

Bi-cubic interpolation is a technique that interpolates a 2D image into new size of crypt image. In this method, the nearest 16 pixels are used to create an intermediate pixel $F(p', q')$. Therefore the output image quality is increased. In Fig. 3, an intermediate pixel $F(p', q')$ [Near to F(p, q)] is created by interpolating nearest [4 × 4] pixels from F(p − 1, q − 1) to F(p + 2, q + 2), then interpolating the results in the horizontal direction.

The following equation [17] is used to interpolate nearest 16 pixels. Number 2 on the sigma represent number of loops in the shaded code.

$$F(p', q') = \sum_{m=-1}^{2} \sum_{n=-1}^{2} F(p+m, q+n) R_c\{(m-a)\} R_c\{-(n-b)\} \tag{5}$$

where F(p + m, q + n) indicates pixel data at location (p + m, q + n). Rc() denotes a Bi-cubic interpolation function such as a BSpline, Traingular and Bell cubic interpolation function.

4 Experiment Result

An experiment had been set up to identify the crypts in iris image template using combination of PSO and Bi-cubic interpolation.

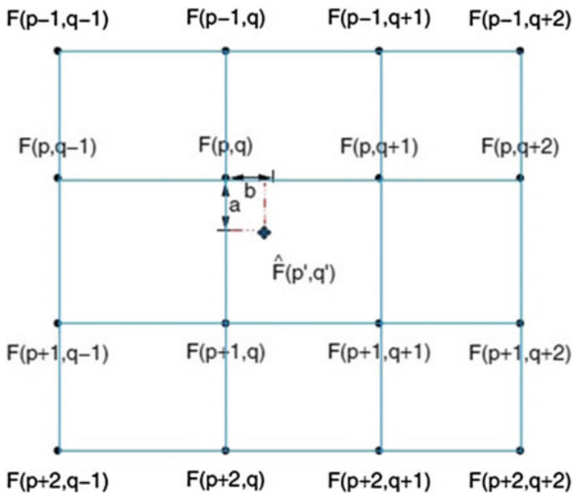


Fig. 3 A 2D illustration is shows, on how the Bi-cubic interpolation moves [17]

4.1 Test Image Set

In these experiments, a set of iris template images were used from the CASIA and UBIRIS database which example of images is shown in Figs. 4, 5, 6 and 7. Figure 4 shows a normalization image of iris template size [20 × 240] is taken from CASIA database and split it into four block size [10 × 120]. Then the first blocks is taken and apply PSO to find the best edge of the crypt and crop it out. This cropped image then will be apply on Bi-cubic interpolation for zoom out. Figure 5 shows a transformation of crypt using different type of method which PSO and Bi-cubic interpolation. Figure 6 shows example of UBIRIS database image with noise that contains diverse of grey levels (we name this a *noisy* image). Some of the objects' intensity is very similar to the background of the image and some shapes are more irregular. Compared with the simple shapes, the edge detection problem in this set of images is clearly more difficult. In Fig. 7, crypt image of size [29 × 21] have been interpolated into size [128 × 128] using bilinear, nearest neighbor and bi-cubic technique. The result shows the image quality improvements of Bi-cubic interpolation method as compared to others mentioned. Moreover, the result shows image that employing the enhancement can help in identification of iris image pattern matching techniques.

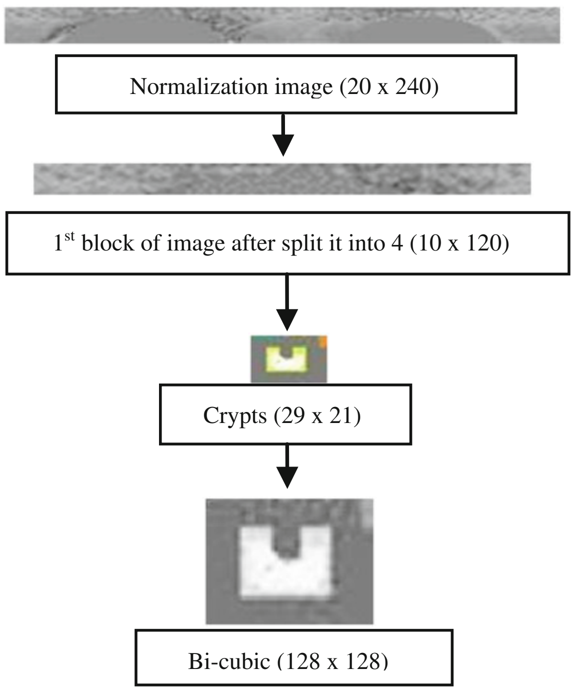


Fig. 4 Transformation of iris template and crypt (CASIA)

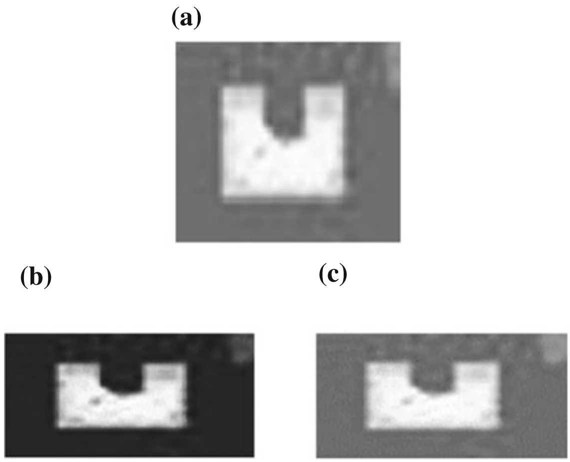


Fig. 5 Transformation of crypt image (CASIA). **a** PSO and bi-cubic, **b** PSO and nearest neighbor and **c** PSO and bilinear

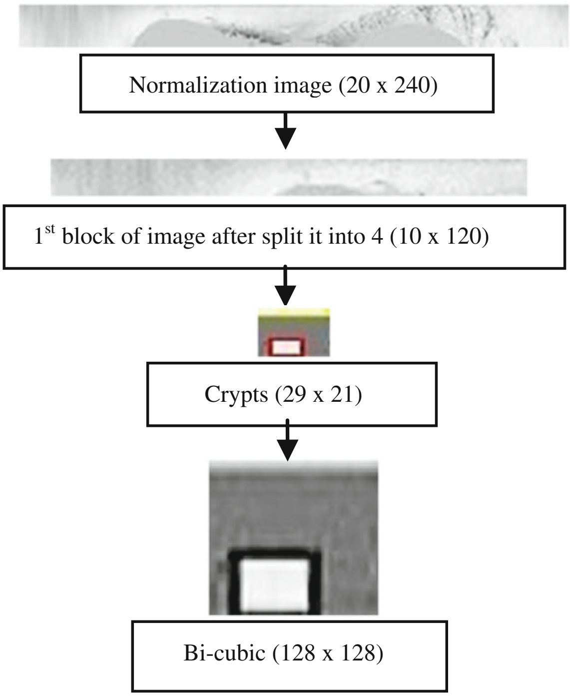


Fig. 6 Transformation of iris template and crypt (UBIRIS)

Fig. 7 Transformation of crypt image (UBIRIS). **a** PSO and bi-cubic, **b** PSO and nearest neighbor and **c** PSO and bilinear

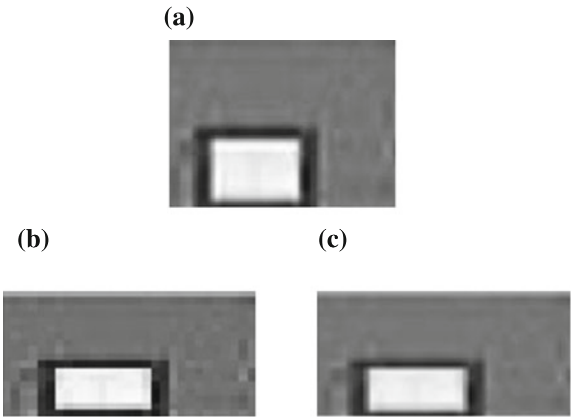


Table 1 Setting for PSO parameter

Population	50
Iteration	100
C1	2.05
C2	2.05

4.2 PSO Parameter Setting

The parameter for PSO setting is based on Table 1. All the values were chosen based on the common setting and an observed search from initial experiments.

4.3 Result

The experiment results have been grouped into high quality (CASIA) and low quality iris images (UBIRIS). Section 4.3.1 shows the high quality data meanwhile Sect. 4.3.2 gives the low quality iris.

4.3.1 CASIA for High Quality Image

The 'PSNR' in Table 2 is the peak signal to noise ratio of the crypt-image. The PSNR value for each method is calculated which are 22.6836 dB for PSO and Bi-cubic, 21.0286 dB for PSO and Nearest Neighbor method, 22.3851 dB for PSO and Bilinear method. Figure 8 shows the comparison in column graph that PSNR performance of the proposed method is better as compare to conventional methods for high quality iris image.

Table 2 PSNR values using different method of enhancement image for CASIA database

Method	PSNR value
PSO and bi-cubic	22.6836
PSO and nearest neighbor	21.0286
PSO and bilinear	22.3851

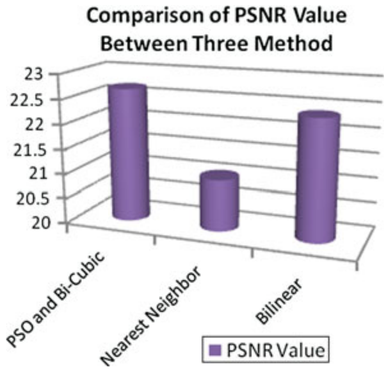


Fig. 8 Comparison of PSNR value between three methods (CASIA)

4.3.2 UBIRIS for High Noise Image

The PSNR values of different methods have been summarized in Table 3. The PSNR value for each method is calculated which are 22.3312 dB for PSO and Bi-cubic interpolation, 21.9385 dB for PSO and Nearest Neighbor method, 22.2764 dB for PSO and Bilinear method. Figure 9 shows the comparison in column graph that PSNR performance of proposed method is better as compare to conventional methods for high noise iris image.

Table 3 PSNR values using different method of enhancement image for UBIRIS database

Method	PSNR value
PSO and bi-cubic	22.3312
PSO and nearest neighbor	21.9385
PSO and bilinear	22.2764

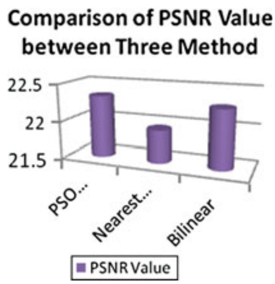


Fig. 9 Comparison of PSNR value between three methods (UBIRIS)

5 Conclusions

The proposed method is a combinational approach for identifying the crypts in iris images template in order to aid the recognition process of iris recognition system. The algorithm employs new covariance based interpolation for edges and iterative curve based interpolation for smooth areas. The results showed PSNR value has improved in terms of performance and suggest quality of visual enhanced crypts images with the propose method as compare to conventional methods. The new method is excellent since it can robustly detect the crypt image even though the image is high in noise rate. Finally, the effectiveness of the crypt image can be viewed in Figs. 8 and 9.

References

1. Muroň, A., Pospišil, J.: The human iris structure and its usage. Physica **39**, 87–95 (2000)
2. Wayman, J.L.: Technical testing and evaluation of biometric identification devices. In: Jain, A., Bolle, R., Pankanti, S. (eds.) Bio-metrics: Personal Identification in Networked Society, pp. 345–368. Kluwer, Dordrecht (1999)
3. Fenker, S.P., Bowyer, K.W., Dame, N., In, N.D: Analysis of Template Aging in Iris Biometrics, pp. 1–7 (2012)
4. Han, F., Li, J., Qi, M., Sheng, M.: An approach of iris recognition based on partical swarm optimization. In: 2010 Fifth International Conference on Frontier of Computer Science and Technology, pp. 541–545. doi:10.1109/FCST.2010.62 (2010)
5. Zhang, D., Ma, L., Li, N., Cai, Y., Zuo, W., Wang, K.: Iris-based medical analysis by geometric deformation features. IEEE J. Biomed. Heal. Inform. **17**(1), 223–231 (2012)
6. Militello, C., Conti, V., Vitabile, S., Sorbello, F.: An embedded module for iris micro-characteristics extraction. In: 2009 International Conference on Complex, Intelligent and Software Intensive System, pp. 223–230 (2009)
7. Proenca, H., Alexandre, L.A.: UBIRIS: a noisy iris image database. Covilha, Portugal (2005)
8. Alkoot, F.M., Box, H.P.O.: A review on advances in iris recognition methods. Int. J. Comput. Eng. Res. **3**(1), 1–9 (2012)
9. Hollingsworth, K.P.: Increased Use of Available Image Data Decreases Errors in Iris Biometrics. Notre Dame (2010)
10. Daugman, J.: Probing the uniqueness and randomness of iriscodes: results from 200 billion iris pair comparisons. IEEE Proc. **94**(11), 1927–1935 (2006)
11. Baker, S.E., Bowyer, K.W., Flynn, P.J., Phillips, P.J.: Template Aging in Iris Biometrics: Evidence of Increased False Reject Rate in ICE 2006, pp. 1–15 (2012)
12. Kennedy, J., Eberhart, R.: Particle Swarm Optimization, p. 20212. Purdue School of Engineering and Technology, Washington, DC (1995)
13. Militello, C., Conti, V., Vitabile, S., Sorbello, F. An embedded module for iris micro-characteristics extraction. In: 2009 International Conference on Complex, Intelligent and Software Intensive Systems, pp. 223–230. doi:10.1109/CISIS.2009.117 (2009)
14. Young, M.: The Technical Writer's Handbook. University Science, Mill Valley, CA (1989)
15. Bai, Q.: Analysis of particle swarm optimization algorithm. Comput. Inf. Sci. **3**(1), 180–184 (2010). doi:10.5539/cis.v3n1P180
16. Baker, S.E., Bowyer, K.W., Flynn, P.J., Phillips, P.J.: Template aging in iris biometrics : evidence of increased false reject rate in ICE 2006, pp. 1–15 (2012)

17. Gao, S., Gruev, V.: Bilinear and bicubic interpolation methods for division of focal plane polarimeters. Opt. Express **19**(27), 26161 (2011). doi:10.1364/OE.19.026161
18. Abidin, Z.Z., Manaf, M., Shibghatullah, A.S. ant-CBIR: a new method for radial furrow extraction in iris biometric. In: The International Conference on Information Security and Cyber Forensics (InfoSec2014). The Society of Digital Information and Wireless Communication, pp. 20–25 (2014)
19. Krishna, N.M., Reddy, P.C.S.: Iris Feature Extraction and Recognition using Unbalanced Haar Wavelets & Modified Multi Texton Histogram (2014)
20. http://sivasanta.blogspot.com/2014_05_01_archive.html
21. Setayesh, M., Zhang, M., Johnston, M.: A new homogeneity-based approach to edge detection using PSO. In: 2009 24th International Conference Image and Vision Computing New Zealand, IVCNZ 2009—Conference Proceedings, (Ivcnz), pp. 231–236. doi:10.1109/IVCNZ.2009.5378404 (2009)

8 Colour Quantization of Colour Construct Code in *CIELAB* Colour Space Using K-Means Clustering and Hungarian Assignment

Kam Meng Goh and Zubairy bin Ismaail

Abstract Color Construct Code (CCC) has been introduced to store data with larger size by applying colour feature as compared to traditional 2D QR code. To decode CCC, colour quantization and colour correction process in pre-decoding process are essentially important in order to obtain data. Unfortunately, colour feature in CCC is always easily influenced by illumination. Meanwhile, colour space i.e. *RGB* and *HSV* is also device-dependent. Our work addresses these two issues and introduces a method to quantize and correct the 8 colour of CCC based on its *CIELAB* colour space. We combined K-mean clustering for the colour reduction and Hungarian assignment for the colour correction purpose. The results showed that our method does not only manage to correct the colour under varying illumination, but it is also independent to the device.

1 Introduction

The usage of one dimensional barcode (1D) can be found everywhere because of its capability to store digital information based on the spacing and width of the parallel lines. However, the data amount that can be stored in 1D barcode is very limited. The capacity of 1D barcode can be increased by using multiple 1D barcode or enlarging the barcode size, and these actions may increase the printing costs. Therefore, plenty of researches have been conducted to address this issue. Barcode is then further upgraded into 2D barcode, where data is stored in vertical and horizontal dimension. Furthermore, different geometric shapes/patterns have been

K.M. Goh (✉) · Z.b. Ismaail
TM Innovation Centre, Telekom Malaysia Research and Development Sdn. Bhd., Lingkaran Teknorat Timur, 63000 Cyberjaya, Malaysia
e-mail: gkmeng@tmrnd.com.my

Z.b. Ismaail
e-mail: zubairy@tmrnd.com.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_57

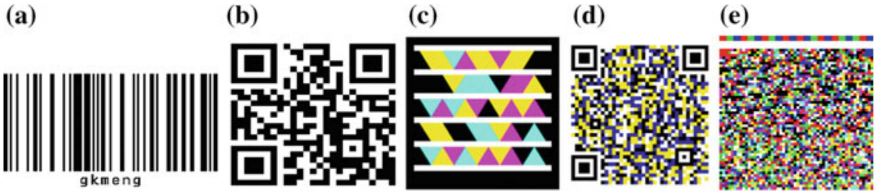


Fig. 1 **a** 1D barcode, **b** 2D barcode, **c** HCCB, **d** HCC2D, **e** color construct code

integrated into 2D barcode for storage enhancement. Figure 1 shows the example of 1D and 2D barcode.

There are several type of 2D codes developed recently. In 1994, Denso Wave has invented a well-known quick respond (QR) code that can store about 7000 numerals [1]. There are other types of 2D barcodes developed in the market such as Aztec Code [2], Data Matrix [3], PDF417 [4], Maxicode [5], and so on. Most of the time, 2D code is used for storing text data i.e. URL. Due to the extensive use of smartphone nowadays and the increasing demand of the internet, 2D code scanning technology is now applicable in the smartphones. The emerging demands of barcodes with higher capacity urge the development of 2D colour codes, where colours are integrating with 2D code. In 1996, Han et al. invented a first colour code, named as ColorCode to store characters and symbols [6]. ImageID has marketed another type of colour code in 2006 [7]. Xerox Prac has invented DataGlyphs in 2001 [8]. Microsoft Research teams introduced High Capacity Colour Barcode (HCCB) in 2007 [9], where the coloured triangle is used to replace traditional square cell in the 2D codes. However, the detection and alignment algorithms of HCCB are fragile and may cause the HCCB decoding failure. In 2010, High Capacity Colour 2D barcode (HCC2D) is designed by adding colour into QR code [10]. HCC2D has similar robustness as QR code, but the capacity is slightly lower than that of HCCB code. Colour Construct Code (CCC) was invented by Color Code Technologies Co. Ltd. that aims to store multimedia file such as video, audio or image in the 2D colour code [11]. CCC contains a row of header that store relevant properties information about the multimedia data type embedded in this code. The rest of the CCC is embedded with the binary data of the multimedia file, where the binary data is represented in terms of colour. The colour is arranged in rectangular cells, while the size of the CCC depends on the size of the multimedia file. Figure 1c–e shows the examples of 2D colour codes.

During pre-decoding process, quantization process plays an important role for correcting the colour. However, the dramatic colour changes may cause the quantization process to turn out as a failure. Colour value is easily distorted by some environmental factor i.e. illumination changes. Uneven illuminations on the 2D colour codes may cause the quantization to be fragile. Besides, every displaying device or printing device has its own displaying/printing characteristics. One colour may be presented or printed in different colours depending on the device. Furthermore, colour space i.e. *RGB* colour space or *HSV* colour space is device-dependent. Therefore, colour code decoding process becomes more challenging.

In this paper, we addressed the colour changes issue and we introduced a novelty of correcting the colour based on *CIELAB* colour space using K-mean clustering and Hungarian assignment. We used CCC in our research since this type of colour code fulfils our needs. Our method not only managed to correct the colour affect by the illumination changes, but at the same time we also managed to handle the distorted colour caused by printer characteristics. In Sect. 2, we discuss about the related previous works of quantization. We present our method in Sect. 3 and results in Sect. 4. Lastly, the conclusion is discussed in Sect. 5.

2 Related Work

The CCC decoder basically involves several stages consisting of capturing process, CCC cropping, perspective transformation, colour quantization, grid detection, and decoding process. The pre-decoding process will involve the mechanism from cropping until grid detection. This paper will focus on pre-decoding process only. Figure 3 illustrates the flow of pre-decoding process.

Basically, quantization methods can be classified into two categories. The first category uses splitting algorithms to quantize the image. This category is computational-efficient but will tend to alter the colour value during the process. Median cut algorithm [12], variance based algorithm [13], and Wu Quant [14] can be classified under this category. Since colour feature is very important to store data in colour code, therefore this category is only applicable if CCC is undisturbed by noise and illumination changes. Second category contains the quantization algorithm based on statistical clustering method. For example, K-Means clustering [15], Fuzzy C Mean [16], and Self-organizing Map [17]. This category is able to produce results without altering the colour, however it is computational-expensive and needs a very good initialization. This approach is able to produce the result as good as Wu-Quant but poorer than that of clustering methods.

All 2D colour code mentioned in the Sect. 1 assumed that there are no illumination changes or colour changes during capturing process, except [15]. Bagherinia and Manduchi [18] invented a method to quantize colour without referring to any reference colour list. Their algorithm tends to decode groups of colour bars at once, where they claimed that decoding colour bars is equivalent to finding the nearest reference colour. However, this method focuses on illumination changes in *RGB* spaces only. Querini and Italiano [19] made a comparison among a few of colour classifiers on HCC2D, which are SVM, Euclidean distance, LMT, Naive Bayes, and K-mean, in order to identify which classifiers will give the best results on correcting colour. The result showed that K-mean displayed the best results among all classifiers. Since colour carries important information.

Besides quantization methods, we also notice that colour space plays a important role in the decoding process. Colour spaces can be categorized into two classes, which are device-dependent and device-independent class [20]. A device-dependent colour space is the colour space where the colour delivered is based on the device properties. For example, the *RGB* colour value printed on the paper is definitely different from the *RGB* colour value displayed in monitor. Some well-known colour spaces are device-dependent, such as *RGB*, *HSV*, *YCbCr*, and *CMYK*. The second class is colour space which is device-independent, such as *CIELAB* and *CIELuv*. These colour space are based on the way human vision interprets colours, but not the way printer ink or monitor deliver colours. Therefore, all these colour spaces will deliver the same parameters or same value regardless of the device being used.

The previous study showed that K-mean will display the best results among all classifiers. Besides, *CIELAB* seems to be the most suitable colour space for our research purpose, since we are looking for the colour space that will not be affected by printer or illumination changes. Therefore, an approach using K-mean clustering on *CIELAB* colour space and Hungarian assignment to quantize CCC colour is proposed in the following section. Our work focuses on quantizing 8 colours, which are red, blue, green, cyan, magenta, yellow, black and white colour. For validation process, we compare our results with the other methods in *RGB* spaces as well as *CIELAB* colour space. Besides that, we also compare the quantization result in difference colour space such as *HSV* space and *RGB* space.

3 Proposed Quantization Process

In this section, we introduce our quantization scheme, which tries to cluster the colour in *CIELAB* colour space using K-means clustering. We use *CIELAB* colour space because it is a perceptual uniform colour space and device-independent. Figure 2 shows the general flow of our proposed method.

From Fig. 2, we first converted the image from *RGB* colour space into *LAB* colour space because of its robustness towards colour change. Then, we applied K-means clustering on converted image to categorize every pixel into relevant clusters. After the clustering process, we will be producing seven layers, where every layer represents one colour except black and white colour (BW) in the same layer. Six colour layers are assigned to reference ideal colour using Hungarian assignment. Lastly, we applied K-means clustering on BW layer again to correct black and white colour. In this paper, some pre-processing such as auto-cropping, affine transformation and perspective transformation are out of the scope, and input

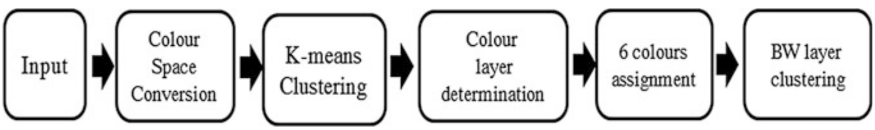


Fig. 2 General flow of proposed quantization process

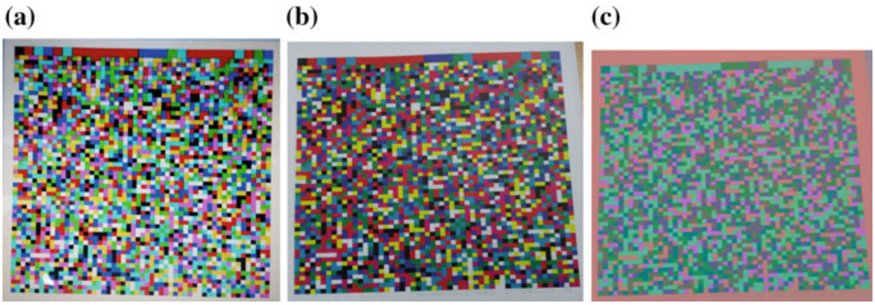


Fig. 3 **a** Image taken from displaying device, **b** image taken from printed paper, **c** *Lab* image of (**b**)

images are cropped manually before processing. All input images are taken from cameras. Figure 3 shows examples of the input images, where Fig. 3a is the image taken from the displayed monitor, while Fig. 3b is the same colour code taken from printed image.

The input taken is in *RGB* colour space. According to Fig. 2, we first converted *RGB* colour space to *CIELAB* colour space, where *L* represents luma (lightness or illumination information) while *ab* represents chroma information. *CIELAB* is a mathematical derivative of *CIE*XYZ*. Therefore, *RGB* colour space is first converted into *CIE*XYZ* colour space, then only it is being transformed into *CIELAB* colour space. The example of *Lab* image of Fig. 3b is shown in Fig. 3c.

After the conversion, we applied K-means clustering to classify every pixel of the input image to the corresponding cluster/group, where every cluster represents one colour. This was done by minimizing the sum of square between cluster centroid and the input data. In this paper, the input data will be *ab* component of *Lab* images. We abandoned *L* element since it represents lightness information and this element is easily affected. Pairs of *ab* elements of each pixel will be the feature for clustering. Figure 4 shows the clustering process.

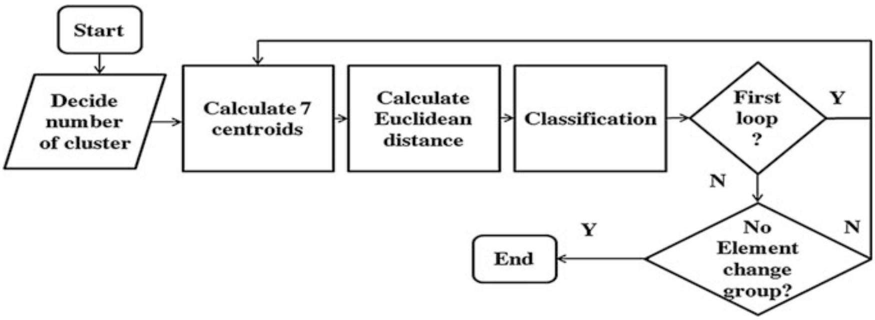


Fig. 4 Flow chart of *ab* elements K-means clustering

The clustering process is the iteration process of updating centroid of cluster and distance between feature and centroid. Referring to Fig. 4, first we assume that there are 7 clusters instead of 8. This is because value of *a* element is always equals to *b* element in black colour as well as white colour. Therefore, *ab* elements of black and white colour should be classified under the same cluster. The clustering process starts with first initial guessing of 7 centroids based on all available *ab* features. Then, Euclidean distance of every feature and every centroid of cluster is calculated, and every feature is assigned to be a member of the nearest cluster based on the minimum Euclidean distance. After the assignment, the algorithm will check if the current iteration is the first iteration. If it is the first iteration, then the algorithm will re-calculate centroids of clusters based on the mean value of all *ab* members under the corresponding clusters. The process will be repeated to find the nearest cluster of every *ab* element again until there are no changes of cluster anymore, as illustrated in the second decision in Fig. 4. The second decision in Fig. 4 shows that if there is no first iteration and there are still changes of members (elements) among the group, the algorithm will be repeated until it reaches its stability. After the clustering, we assigned every *ab* elements to the corresponding *RGB* value from the original image.

Then, the Hungarian assignment is applied to do the colour correction on these original *RGB* values of elements in the clusters. A cost matrix is needed for Hungarian assignment process. Before doing this, mode values of R, G, and B layers of all elements in every cluster are computed for the purpose of finding colour trend of each cluster. Note that this process is applied to all clusters except *BW* cluster. To exclude BW cluster, this cluster is selected from all clusters based on the cluster with maximum number of elements, since black and white colour are under the same cluster. Then, we compute Euclidean distances between 6 computed mode *RGB* values of clusters with reference ideal colours from the lookup reference table, in which this table contains ideal *RGB* value of 6 colours (*RGBCMY*). A cost matrix, *T* is generated as shown in (1):

$$T = \begin{bmatrix} E_{11} & \cdots & E_{61} \\ \vdots & \ddots & \vdots \\ E_{16} & \cdots & E_{66} \end{bmatrix}. \tag{1}$$

where E_{mn} is the Euclidean distance between ideal and mode *RGB* values of every colours generated using (2):

$$E_{mn} = \sqrt{(R_m - R_n)^2 + (G_m - G_n)^2 + (B_m - B_n)^2}. \tag{2}$$

m and n in (2) represents the index number of colour from 1 to 6. After this process, we applied Hungarian Assignment to find the optimal pairs of colours. Hungarian assignment is an algorithm to obtain the best or optimal assignment with minimum cost based on the given cost matrix, *T*. In this paper, the optimal assignment is about assigning the colours from the input image to the corresponding ideal reference colours. By having optimal assignment pairs, we replaced every input *RGB* values of these six colours with the corresponding assigned ideal *RGB* values. However, black and white colours are yet to be corrected. To separate these two colours, we applied K-means clustering again on BW layer into 2 clusters, where the input data are *RGB* values of original black and white colour. After the separation, we need to decide which cluster belongs to which colour. To do this, we summed up *RGB* value of every element/member in these 2 clusters and compared the maximum summation value from these clusters. The summation of *RGB* value from white cluster should be always larger than that of black colour. After that we replaced the input black and white colour with ideal values. Lastly, we combined all layers together to obtain the image with corrected colour as shown in Fig. 5c.

4 Experimental Results

In order to test the robustness of our proposed algorithm, we implemented the proposed method to different scenarios. We applied our method to 4 kinds of image, which are illustrated in Figs. 3a, b and 5a, b.

Figure 3a is the image captured from the displaying device (monitor), where there are some noises caused by the screen shadow. Figure 3b is the image captured from the printed paper, where the colour code is not affected by any environmental factor. However, the printed *RGB* colour value deviated from ideal *RGB* colour value. Figure 5a, b is printed images with uneven illumination changes.

These 4 samples were chosen to test the robustness of our algorithm towards colour changes and illumination changes. We tested these 4 samples with other algorithm, which is Wu-Quant, and median cut algorithm using Color Inspector 3D software. We figured out that adjusting the colour saturation of image to maximum value manually in Color Inspector 3D may achieve better results in certain circumstances. Therefore, we compared our method with these 4 methods, which are Wu-Quant, Wu-Quant S (saturation), median cut, median cut S (saturation). Note

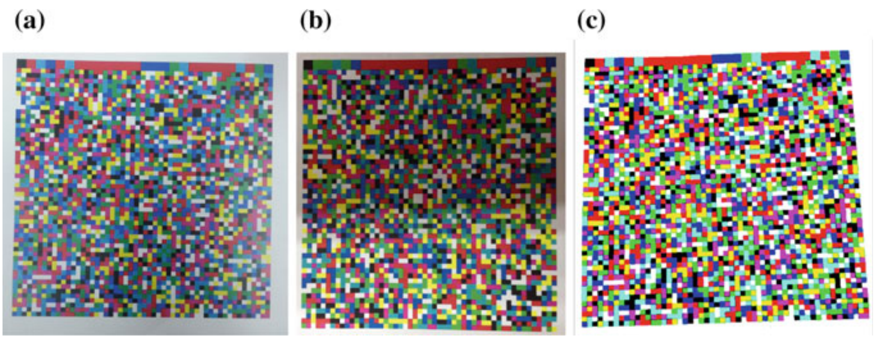


Fig. 5 **a** Test sample 3, **b** test sample 4, **c** corrected colour image from Fig. 4b

that the other 4 methods are applied in *RGB* space. The result is calculated based on the percentages of cells with correct colours as follows:

$$p = \frac{C_{correct}}{C_{total}} * 100\,\%. \tag{3}$$

where p is denoted as correct cell percentage, $C_{correct}$ is the number of cells with correct colour and C_{total} is the total number of cells from colour code. Table 1 shows the results of 4 scenarios.

Our proposed method produces the best and convincing results regardless of any situation. In Table 1, Wu-Quant and median cut only provide better results if the colour saturation is changed. This is because illumination changes and colour distortion factor is removed by adjusting the colour saturation value. However, the level of saturation of every image is different. Therefore, the level of saturation can only be adjusted experimentally and hence is inconvenient. Thus, changing saturation level is not a good option for us.

Figure 3a is the image directly captured from screen, where *RGB* value is deviated less from the ideal value. Therefore, all methods are able to quantize the colour nicely. On the other hand, the results produced by Wu-Quant and median cut in Fig. 3b is slightly lower compared to that of Fig. 3a, due to the colour value deviation since printers have its own properties to create colour. Wu-Quant and median cut is less robust to the printed colour. In this case, our proposed method shows the strong robustness towards colour changes compared to that of other four methods.

Figure 5a, b are printed CCC captured under uneven illumination. Illumination and colour distortion brought huge impact to the printed CCC, meanwhile also affecting the data stored inside CCC. Wu-Quant and median cut may misinterpret the colour and allocate the colour to the wrong category if these two methods are applied on RGB colour space, as shown in Table 1. Even if the colour saturation is adjusted, the correction is not significant since there is too much deviation in these two samples. Generally, our proposed method is strongly robust towards illumination changes and colour changes based on our experimental testing in Table 1. Besides that, we also investigate if *Lab* colour space is applicable on other quantization algorithm by replacing K-means clustering process with other quantization algorithms to compute the result, as displayed in Table 2.

Table 1 Comparison of accuracy between proposed method with other methods

Sample	Proposed method (*Lab*) (%)	Wu-Quant (*RGB*) (%)	Wu-Quant S(*RGB*) (%)	Median cut (*RGB*) (%)	Median cut S(*RGB*) (%)
Figure 4a	100	98.71	99.17	96.57	99.25
Figure 4b	100	96.84	100	92.01	90.15
Figure 5a	99.80	86.95	99.32	90.42	85.46
Figure 5b	99.85	78.68	81.90	77.03	77.10

Table 2 Comparison of lab colour space quantization using different quantization algorithm

Sample	Proposed method (%)	Wu Quant (%)	Median cut (%)	Fuzzy C-mean (%)	Self-organizing map (%)
Figure 4a	100	100	79.27	99.97	100
Figure 4b	100	100	89.77	100	100
Figure 5a	99.80	88.79	99.83	99.80	89.91
Figure 5b	99.85	100	97.22	100	100

Besides the methods used in Table 1, we also compared our results with fuzzy c-means (*FCM*) and self-organizing maps (*SOM*). *FCM* algorithm is very similar to the theory of K-means clustering, but it will run slower than K-means since FCM performs more operations. *SOM* is a type of neural network that also introduced for clustering process. It involves training and mapping process. As a result, *SOM* takes more processing time compared to that of K-means and *FCM*.

From Table 2, Wu-Quant produce better results in *Lab* Colour space since *Lab* is a promising colour space. However, the results of Wu-Quant are not as consistent as that of our proposed method. Result of Wu-Quant in Fig. 5a tells that it may quantize colour wrongly despite in *Lab* colour space. In this sample, all cyan colours in CCC are quantized as blue colours by Wu-Quant. *FCM* results trend is very similar to our proposed method, since *FCM* and K-means are very much alike. *SOM* can achieve good results except Fig. 5a, where 1 of 8 colours is wrongly quantized. Self-organizing maps is not suitable for our application because of its high computational cost and expensive algorithm, even though the result is good. The results of median-cut quantization are the poorest amongst all results, since it quantized colour wrongly. This significantly shows that median cut cannot produce good results even in *Lab* colour space. Generally, the combination of K-means or *FCM* and *Lab* colour space will give more promising results compared to other algorithms. Furthermore, we also tested K-mean quantization in *RGB* colour space and *HSV* colour space to find out which colour space is the most suitable to be implemented on CCC quantization. The results are shown in Table 3.

In Table 3, implementation of our method in *Lab* colour space provided more stable results in comparison to the other 2 colour space. *RGB* and *HSV* are device-dependent and will be easily affected by illumination changes, especially *RGB* colour space. Therefore, the results were lower than that of *Lab* colour space. *HSV* results were better than the *RGB* colour space results, even if there are uneven illumination changes. Overall in Table 3, our methods will give the most

Table 3 Results of implementation of our methods in different colour space

Sample	*Lab* (%)	*RGB* (%)	*HSV* (%)
Figure 4a	100	97.28	88.41
Figure 4b	100	99.29	89.13
Figure 5a	99.80	83.76	100
Figure 5b	99.85	57.18	95.86

convincing results when applied in *Lab* colour space under environmental changes. From all results, we concluded that colour space and quantization algorithm is equally important to do the colour reduction on our CCC code. The performance of our proposed method showed that our method can give more consistent and stable results compared to the other methods.

5 Conclusion

In this paper we have proposed 8 bit colour code quantization method which aims to correct colour of the code, while at the same time promotes robustness towards colour change and illumination changes. Our results showed that the novel method in this article leads to a more consistent and higher accuracy of correct colour cell compared to the other methods. For future work, we will improve the efficiency and computational cost of our method.

References

1. http://www.qrcode.com/en/history/
2. Longacre, A., Hussey, R.: Two dimensional data encoding structure and symbology for use with optical readers. Google Patents (US5591956) (1997)
3. GS1: Two-dimensional (2D) barcodes
4. Wang, Y.P.: System for encoding and decoding data in machine readable graphic form. Google Patents (US5243655) (1993)
5. Chandler, D.G., Batterman, E.P., Shah, G.: Hexagonal, information encoding article, process and system. Google Patents (US4874936) (1989)
6. Philip, B.: System for encoding and decoding data that utilizes a combination of colors and shapes. Google Patents (WO1997037321 A1) (1997)
7. Sali, E., Lax, D.M.: Color bar code system. Google Patents (US7210631) (2007)
8. Hecht, D.L.: Printed embedded data graphical user interfaces. Computer **34**, 47–55 (2001)
9. http://research.microsoft.com/enus/projects/hccb/
10. Querini, M., Grillo, A., Lentini, A., Italiano, G.F.: 2D color barcodes for mobile phones (2013). Int. J. Comput. Sci. Appl. (IJCSA) **8**(1), 136–155 (2011)
11. Takura, Y., Ichinose, K., Miwa, T.: Method for creating and method for decoding two-dimensional colour code. Google Patents (EP2544128 A1) (2013)
12. Heckbert, P.: Color image quantization for frame buffer display. ACM (1982)
13. Wan, S., Prusinkiewicz, P., Wong, S.: Variance-based color image quantization for frame buffer display. Color Res. Appl. **15**, 52–58 (1990)
14. Wu, X.: Color quantization by dynamic programming and principal analysis. ACM Trans. Graph. (TOG) **11**, 348–372 (1992)
15. Celebi, M.E.: Improving the performance of k-means for color quantization. Image Vis. Comput. **29**, 260–271 (2011)
16. Wen, Q., Celebi, M.E.: Hard versus fuzzy c-means clustering for color quantization. EURASIP J. Adv. Signal Process. **2011**, 1–12 (2011)
17. Papamarkos, N., Atsalakis, A.E., Strouthopoulos, C.P.: Adaptive color reduction. IEEE Trans. Syst. Man Cybern. Part B Cybern. **32**, 44–56 (2002)

18. Bagherinia, H., Manduchi, R.: A theory of color barcodes. In: IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 806–813. IEEE (2011)
19. Querini, M., Italiano, G.F.: Color classifiers for 2D color barcodes. In: IEEE Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 611–618 (2013)
20. Ford, A., Roberts, A.: Colour space conversions (1998)

Egress Cloud Computing with Big Data Attribution

Thepparit Banditwattanawong and Masawee Masdisornchote

Abstract To share big data stored in cloud in a distributed manner can cause so huge data transfer out of cloud that affects cloud service responsiveness, public cloud data-out monetary costs and network bandwidth consumption. Our previous works showed that deploying cloud cache with i-Cloud eviction scheme is an effective solution. Through trace-driven simulations, this paper aims to analyze algorithmic factors influencing the performance of i-Cloud. A main finding is that TTL attribute creates the greatest contribution in i-Cloud efficiency in distributed big data sharing. In addition, this paper presents a guideline for content providers to allow the distributed sharing of their big data in an effective manner.

Keywords Cloud cache · Cache replacement policy · Time-to-live · Data-out charge · Rate · Window size · Factor impact evaluation

1 Introduction

Big data including the archives of high-definition images and videos gathered via ubiquitous information-sensing devices have been increasingly stored in an off-premise cloud and distributed shared among users through various cloud computing services. This can cause downstream network bandwidth saturation, cloud service nonresponsiveness and costly cloud data-out charge. The sharing of big data can be achieved in a more efficient manner by means of client-side cloud cache located in user locality [1, 2].

A cloud cache aims for serving any repeated data requests by fetching the valid data copies from local cache repository instead of retrieving the data from remote

T. Banditwattanawong (✉) · M. Masdisornchote
School of Information Technology, Sripathum University, Bangkok, Thailand
e-mail: thepparit.ba@spu.ac.th

M. Masdisornchote
e-mail: masawee.ma@spu.ac.th

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_58

cloud. However, the cloud cache has been designed with different performance goal from traditional web cache [3] to truly suit big data era. As a cloud cache enabling mechanism, a newly established cache eviction algorithm called i-Cloud has been proposed along with full-fledged evaluation [4]. The main benefits of i-Cloud are its superior performance in terms of cost-saving, delay-saving, byte-hit and hit ratios simultaneously among other well-known algorithms in the realm of traditional web caching that are LRU, GDSF and LFU-DA. This performance characteristic became the motivation of further research in this paper. We extended our previous research by curiously conducting the analysis of i-Cloud algorithm to determine which of its parameter makes the most performance impact. The analysis result has been found absolutely valuable not only for big data providers but also cloud computing service providers and consumers.

To accommodate the understanding of this paper's contribution, the following organization is provided. Section 2 introduces metrics used for performance impact measurement during experiment and as a basis to comprehend subsequent sections. Section 3 summarizes i-Cloud algorithm including its all parameters. Section 4 describes the impact analysis methods of i-Cloud's parameters. Section 5 presents the measurement results and findings of the impact analyses. As our main contribution, the results are then translated into a real-world practice useful for actors in digital content ecosystems in Sect. 6 before concluding the whole paper.

2 Performance Analysis Metrics

The factor impacts were analyzed in four metrics. The first three were conventional caching performance metrics [5], defined as follows. For an object i, $\textit{byte-hit ratio} = \sum_{i=1}^{n} s_i h_i / \sum_{i=1}^{n} s_i r_i$, $\textit{delay-saving ratio} = \sum_{i=1}^{n} l_i h_i / \sum_{i=1}^{n} l_i r_i$ and $\textit{hit rate} = \sum_{i=1}^{n} h_i / \sum_{i=1}^{n} r_i$ where s_i is the size of i, h_i is how many times a valid copy of i is fetched from cache, r_i is the total number of requests to i, and l_i is the loading latency of i from cloud. The other metric was cost-saving ratio [3] used to capture economical performance impact. The metric measures how much money can be saved by serving the valid copies of requested objects from cache. Given a data object i, $\textit{cost-saving ratio} = \sum_{i=1}^{n} c_i s_i h_i / \sum_{i=1}^{n} c_i s_i r_i$ where c_i is the data-out charge rate or monetary cost for loading i from cloud.

3 i-Cloud Algorithm

This section gives the overview of i-Cloud algorithm and its major parameters. Based on the principle of contemporaneous proximity [6], i-Cloud has performed relatively well in optimizing the byte-hit, delay-saving, cost-saving and hit ratios. Figure 1 demonstrates the conceptual framework of i-Cloud. It consists of two main

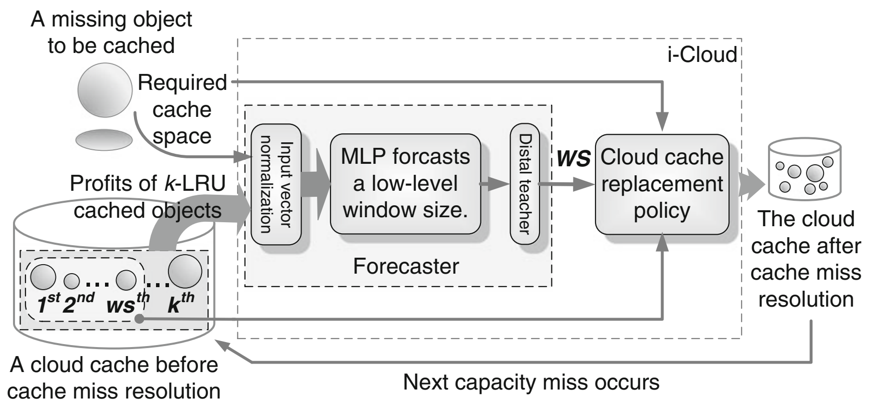


Fig. 1 The conceptual framework of i-Cloud

processing modules, Cloud cache replacement policy [3, 7–10] and a forecaster. Table 1 presents Cloud algorithm, which is explained as follows. When cache eviction is needed, Cloud is invoked. It first formulates a cluster of in-cache lru objects as many as instructed either by a *window size* parameter or a required cache space depending on which one is larger, it will be used. Once the cluster of lru objects has been formed, Cloud quantifies a profit associated with each object inside the cluster as follows: given an object i, $profit_i = s_i \cdot c_i \cdot l_i \cdot f_i \cdot TTL_i$ where s_i is the size of i, c_i is data-out charge rate for loading i, l_i is latency for loading i, f_i is the access frequency of i, and TTL_i is the remaining lifespan of i. An object with least profit is evicted first from cache. This eviction process is repeated on the next least profitable objects in the cluster until gaining enough cache room. The input parameter *ws* in Table 1 has its value supplied by a multilayer perceptron (MLP), which automatically learns both the latest cache state and a required cache space to forecast a near-optimal window size. As shown in Fig. 1, the forecaster basically comprises an input vector normalization, a feedforward propagation and an output denormalization (i.e., a distal teacher) parts, respectively. The MLP implementing the forecaster has the structure of 10,002/2/2/1.

The forecaster learned input patterns inherent in caching state history by means of supervised learning with a distal teacher and back-propagation. Each input pattern represented a cache miss state and was organized into a training vector of the form $<1, p_1, p_2, \ldots, p_{10,000}, rs>$ where constant activation 1 is bias, $\mathrm{p}_{\mathrm{i}\ =1-10{,}000}$ is the profit of ith lru object and rs is a required cache space. Each pattern was generated every time cache capacity miss occurred during the i-Cloud simulation of a certain trace, cache size and respective static optimal window size (determined for optimal byte-hit ratios based on a nonuniform cost model where two cloud subscriptions were assumed with different data-out charge rates of 0.0829 USD/GB as of Amazon S3 and 0.1535 USD/GB as of Google Cloud Storage). The result of each complete simulation session was a sequence of input

Table 1 Cloud algorithm

Algorithm: cloud
Input variables:
cd/*cache database (recency-keyed min-priority queue)*/;
ws/*window size*/;
rs/*required cache space*/;
Local variables:
ecd/*empty cache database (recency-keyed min-priority queue) */;
oc/*an object cluster of least-recently-used (LRU) objects (profit-keyed min-priority queue of evictable objects)*/;
co/*a candidate object to be included in a cluster*/;
$ts \leftarrow 0$/*total size of *ws* objects initialized to zero*/;
eo/*an evicted object*/;
$c \leftarrow 0$/*counter for objects in a cluster initialized to zero*/;
begin
if $cd.getTotalNumberOfObjects() < ws$
then $ws \leftarrow cd.getTotalNumberOfObjects()$;
$ecd \leftarrow cd$;
do
$co \leftarrow ecd.removeLeastRecentlyUsedObject()$;
$ts \leftarrow ts + co.getSize()$;
$oc.addObject(co)$;
$c \leftarrow c + 1$;
$while\ (c < ws) \vee (ts < rs)$;
do
$eo \leftarrow oc.removeMinProfitObject()$;
$cd.evict(eo)$;
while $cd.getFreeSpace() < rs$;
return $cd.getFreeSpace()$;

patterns, contained inside a single training data set. We generated two training data sets by the separate simulation sessions of two 15-day traces collected from different user communities located in Boulder (BO) and New York (NY) using the same 10 % cache size (i.e., percent of the maximum bytes of total unique objects referenced in each 15-day trace). We normalized every element value within each input vector except the bias constant by max-min linear scaling. Because the MLP outputted a low-level window size of a domain [0.0, 1.0] via sigmoid function, to achieve desired window sizes requires a distal teacher to denormalize the low-level window size to obtain a practical one of [0, 10,000]. Two target window sizes of the training data sets were the static optimal window sizes aforementioned. The learning phase terminated when mean squared error stabilized with a decreasing rate smaller than 1 per epoch.

Therefore, we had two forecasters, one trained with the 15-day BO trace and 10 % cache size and the other with the 15-day NY trace and 10 % cache size. Both of the forecaster have been validated in our previous work [11] to be good enough for generating performance baselines described in the following section.

4 Factor Impact Measurement

The focus of the impact analysis lies in three i-Cloud's parameters, which have not much been studied in other related algorithms [5], that are the window size parameter (*ws*), data-out charge rate (c_i) and object's remaining lifespan (TTL_i). The impacts are measured and reported in terms of the four performance metrics described in Sect. 2. To achieve this, it is important to figure out the performance baselines, then excluding the studied parameters one by one to observe how much four performance aspects are effected as compared to the baselines.

The performance baselines were gained by placing i-Cloud with the two forecasters to operate as normal based on the nonuniform cost model against each trace listed in Table 2 and each cache size (10, 20 and 30 %). The traces contained HTTP requests collected within 1 month period. The cache sizes were provisioned in the percents of the maximum bytes of total unique objects of the 1-month traces. Since there were two forecasters, two distinct evaluation traces, three simulated cache sizes and four metrics to consider, there were totally 48 performance baselines in a set. Fortunately, a lesson learned from our previous work [1] suggested that we used the forecaster learning the 15-day BO trace for simulation against the 1-month BO trace, and using the 15-day NY trace learning forecaster when simulating against the 1-month NY trace. Therefore, we conducted only 24 baseline measurements.

The contributions in the overall performance measurements of i-Cloud's parameters were simulated and measured according to the following three impact analysis configurations before comparing with the baselines.

- Configuration 1: To reveal the impact of window size, we simulated i-Cloud by disabling the forecaster and instead engaged the worst window sizes of the 1-month BO and NY traces.
- Configuration 2: The impact of data-out charge rate was rendered by hiding the parameter c_i from the object profit equation during simulations.

Table 2 Characteristics of evaluation traces

Feature	31-day BO	31-day NY
Total requests	639,187	1,311,880
Total requested bytes	4,149,211,314	17,067,821,671
Total unique objects	323,979	593,365
Max bytes of total unique objects	2,262,144,480	10,801,010,237

– Configuration 3: Similarly, the parameter TTL_i was omitted during simulations to see the impact of object remaining lifespan.

Each configuration was simulated with the pair of forecasters, the two separate traces, the three distinct cache sizes and the nonuniform costs. The simulation results were measured in the four metrics and compared against the baselines. Any performance variations from the baselines (for example, subtracting a cost-saving ratio with a removed factor from a baseline cost-saving ratio) were translated as the impacts of the considering parameters and reported in the next section.

5 Results and Discussions

Once the performance variations were determined, we selected to report only the maximum ones based on two steps. Firstly, between the 1-month BO and NY traces, only the greater impact of the same impact analysis configuration, performance metric and simulated cache size were considered in the next step. Secondly,

Fig. 2 The comparative cost-saving performance impacts yielded by the different impact analysis configurations

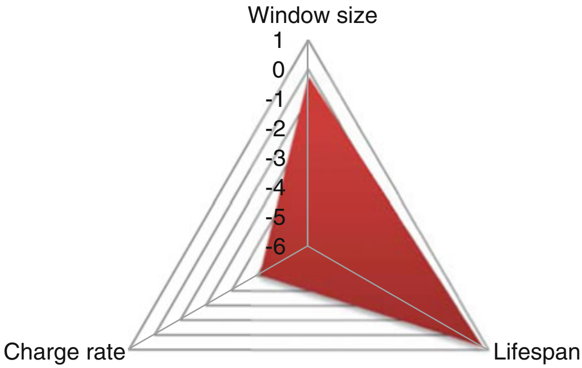
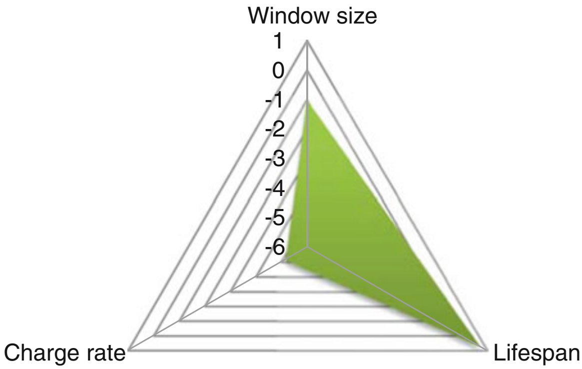


Fig. 3 The comparative delay-saving performance impacts yielded by the different impact analysis configurations



among the different cache sizes of identical performance analysis metric and identical impact analysis configuration, the impacts from the previous step have been finally reported as portrayed in Figs. 2, 3, 4 and 5 using logarithmic-scale radar charts. The three chart axes represent the performance impacts contributed by the window size, data-out charge rate and remaining lifespan derived from the impact analysis configurations 1, 2 and 3, respectively (Fig. 3).

The charts are described as follows.

- In terms of cost-saving ratio, ignoring object remaining lifespan has had the most impact whereas omitting data-out charge rate has had the least effect as illustrated in Fig. 2.
- In terms of delay-saving ratio, considering remaining lifespan has made i-Cloud reduce much more delay than skipping it whereas neglecting data-out charge rate has marginally degraded the delay saving as showed in Fig. 3.

Fig. 4 The comparative byte-hit performance impacts yielded by the different impact analysis configurations

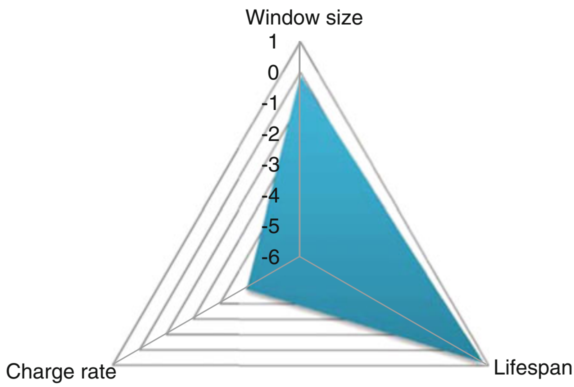
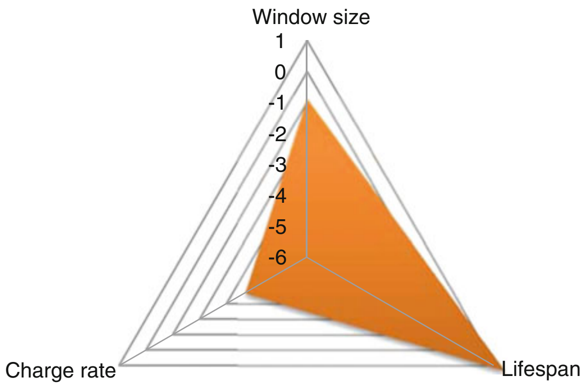


Fig. 5 The comparative hit performance impacts yielded by the different impact analysis configurations



- In terms of byte-hit ratio, network bandwidth has been saved substantially when regarding remaining lifespan; removing data-out charge rate has affected the byte-hit performance slightly as depicted in Fig. 4.
- Finally, hit rate has also been improved largely when taking remaining lifespan into account as can be seen in Fig. 5.

To recap, the trace-driven simulation has shown that the cost-saving, delay-saving, byte-hit and hit performances have declined most when i-Cloud ignored data remaining lifespan, moderately when not using optimal window size, and least when omitting data-out charge rate. As a consequence, a major observation can be inferred that is the data object remaining lifespan is the most significant factor to the performance of i-Cloud-based cloud caching.

This phenomenon can be discussed as follows. Expired objects had absolutely non chance to be hit any more an thus evicting them from cache storage redeem cache space from pollution with any false negative resulting in remarkable caching performance improvement. On the other hand, cached objects associated with lower data-out charge rates or staying inside window sizes might incur false positives when cache misses occur on them.

6 Recommendation

The previous section gives an inference that serves as a valuable guideline to cloud content creators and providers to gain i-Cloud's benefits including fast data delivery, reduced data accessing workloads on their virtualized servers and inexpensive data downloading monetary costs by their customers. The initial description of the guideline is as follows.

- First, it is always a good idea for cloud content providers to set their data expiration attributes rather than letting the data stay valid for unlimited time and become cache pollution at last. There have been supporting mechanisms to achieve this such as the Expires or the Max-age field inside HTTP protocol header [12] as long as the cloud data is delivered over HTTP.
- Second, the data expiration attributes should be set in a precise manner for user majority to avoid both false positive and false negative in order to maximize the advantages of i-Cloud's. We realize that this suggestion may be a complicated task for most content providers. As far as we know, there have been some related work that analyzed the lifetimes of traditional web documents such as [13]. Nevertheless, we believe that this research issue will have to be re-studied since cloud data especially in big data era has its common characteristic relatively different from that of traditional WWW [14].

7 Conclusion

This paper presents the performance impact analysis of TTL, window size and data-out charge rate that are factorized by i-Cloud cache eviction algorithm. The breakthrough finding is that cloud caching performance could be significantly improved when data remaining lifetimes are provided for i-Cloud processing. This discovery serves as a data attribution guideline to be adhered by cloud content providers so that the distributed sharings of big data could continue efficiently in terms of network scalability, responsiveness and economy.

Acknowledgment This research is financially supported by Sripatum University. The authors also thank Duane Wessels, National Science Foundation (grants NCR-9616602 and NCR-9521745) and the National Laboratory for Applied Network Research for the trace data used in this study.

References

1. Banditwattanawong, T., Masdisornchote, M., Uthayopas, P.: Economical and efficient big data sharing with i-cloud. In: International Conference on Big Data and Smart Computing, pp. 105–110 (2014)
2. Banditwattanawong, T., Masdisornchote, M., Uthayopas, P.: The smart distribution of social media contents. In: The 2014 International Electrical Engineering Congress (iEECON2014) (2014)
3. Banditwattanawong, T.: From web cache to cloud cache. In: Li, R., Cao, J., Bourgeois, J. (eds.) Advances in Grid and Pervasive Computing. Lecture Notes in Computer Science, vol. 7296, pp. 1–15. Springer, Berlin (2012)
4. Banditwattanawong, T., Masdisornchote, M., Uthayopas, P.: Hybrid cloud computing: economy, scalability and responsiveness optimization. Chiang Mai J. Sci (page in press) (2015)
5. Podlipnig, S., Böszörmenyi, L.: A survey of web cache replacement strategies. ACM Comput. Surv. **35**(4), 374–398 (2003)
6. Banditwattanawong, T., Hidaka, S., Washizaki, H., Maruyama, K.: Optimization of program loading by object class clustering. IEEJ Trans. Electr. Electron. Eng. **1**(4), 397–407 (2006)
7. Banditwattanawong, T., Uthayopas, P.: A client-side cloud cache replacement policy. ECTI Trans. Comput. Inf. Technol. **8**(2), 113–121 (2014)
8. Banditwattanawong, T., Uthayopas, P.: Cloud cache replacement policy: new performances and findings. In: 1st International Conference on Annual PSU Phuket 2012. PSU PIC 2012 (2013)
9. Banditwattanawong, T., Uthayopas, P.: Improving cloud scalability, economy and responsiveness with client-side cloud cache. In: 10th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2013. ECTI-CON 2013 (2013)
10. Banditwattanawong, T., Uthayopas, P.: Cloud-enabling technique for dissimilar charges. In: The 2014 International Industrial Information Systems Conference (IIISC2014), pp. 119–120 (2014)
11. Banditwattanawong, T., Uthayopas, P.: An intelligent cloud cache replacement scheme. In: Papasratorn, B., Charoenkitkarn, N., Vanijja, V., Chongsuphajaisiddhi, V. (eds.) Advances in

Information Technology. Communications in Computer and Information Science, vol. 409, pp. 23–34. Springer, Berlin (2013)
12. Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., Berners-Lee, T.: Hypertext transfer protocol–HTTP/1.1 (2012)
13. Chen, X., Mohapatra, P.: Lifetime behavior and its impact on web caching. In: IEEE Workshop on Internet Applications, pp. 54–61, Aug 1999
14. Liu, S., Huang, X., Fu, H., Yang, G.: Understanding data characteristics and access patterns in a cloud storage system. In: 13th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), pp. 327–334 (2013)

Intelligent Displaying of Notes Based on Current Position and Time

Jan Fogl, Jan Dvorak, Ondrej Krejcar and Kamil Kuca

Abstract This work comes as a reaction to continually growing availability of mobile devices throughout human population. Nowadays the mobile devices are being enhanced with plenty of sensors, modules and computing functions, which can be used in real life to save people's labour. This work concretely deals with utilization of GPS module and Google Services for smart filtering notes, which have been created by the user in provided application. The application has been primarily created for android system, because it's the most favourite and widespread operating system today. The essential problem of this work is an algorithm for searching notes that are as close as possible to the user's position by the location of their creation. For this functionality it was selected service geo fencing from Google Services API. This service enables to gather notes that are close to each other by their location and afterwards show these notes to the user. The goal of the whole work is to create functional user friendly application.

Keywords Android · Locating · Location based applications · Intelligent application · Intelligent filtering

J. Fogl · J. Dvorak · O. Krejcar (✉) · K. Kuca
Faculty of Informatics and Management, Center for Basic and Applied Research,
University of Hradec Kralove, Rokitanskeho 62, Hradec Kralove 500 03, Czech Republic
e-mail: ondrej@krejcar.org

J. Fogl
e-mail: Jan.fogl@uhk.cz

J. Dvorak
e-mail: jan.dvorak2@uhk.cz

K. Kuca
e-mail: kamil.kuca@uhk.cz

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_59

1 Introduction

These days the world of mobile devices develops rapidly. The mobile devices get progressively to the whole wide society and are able to provide people with entertainment or to help in all sorts of industries. Along with the mobile devices development are coming various applications as well based on receiving, processing and interpreting data from different sensors that become a standard equipment of all mobile devices today. An example of applications using these systems can be various navigation systems utilizing both GPS and internet connection, then systems recognizing gestures or systems that can save people's time and energy.

Plenty of such applications occur all around us, however, some of them are useless or the range of their usability is too low. Therefore it's necessary to think about concepts of these applications and to adapt them maximally to demands and needs of the widest spectrum of users.

As the most logical and the most frequent utilization of a mobile device it turns out the use of an actual location of the user. A good example of well worked out idea is a reminder application [1] that allows setting a reminder to a specific location and while getting close to that location it displays that reminder to the user.

The main goal of this work is realization of note application that will aspire to simplify user's work utilizing GPS sensor and internet connection. While saving the notes the application will register also user's actual location using GPS and WiFi and consequently while viewing these notes it will suggest to the user the notes, which respond the most to the actual user's position. Application reflects the idea that the notes created for example at the work will be likely viewed by the user at the work again, i.e. in the similar location.

Obtaining the position will be executed primarily through GPS module, but to save energy in the places, where it's possible, it will be used WiFi and Google API, which enables to obtain the last reachable user's position. In the text [1] it's described the option how to gain this position. There are plenty of approaches to algorithms for sequencing individual notes. One direct way is to localize the user and afterwards calculate his distance from the individual saved notes. The algorithms to calculate the location can be found in the article [2]. Another method is using "geo fencing". Geo fence is an area formed by a specific range around given point. Android system enables to monitor all the formed geo fences and to notify about entries/outputs from these areas. The principle of geo fencing is well explained in specialized work [3].

The principal of geo fencing will be used in this work. For higher efficiency it will be extended with a time dimension, which means that sequencing of elements will depend not only on current user's position, but as well on the time when the note was created. This dimension of time shall allow that the user—without self-effort—gains the notations he/she most probably needs.

2 Problem Definition

As it was already said the objective of this work is to design an intelligent system at the platform of android operating system for displaying user's notes in relation to his actual position, time, when the note was created and time, when we want to display the individual record. The android system was chosen according to its popularity and extensiveness throughout the whole population. Therefore it will not be any problem to assure distribution of the application to a wide range of people.

The main vision of the whole application is to simplify control and search for the created notes. This functionality should be user friendly, which means it should simplify user's work with the application and save plenty of time that the user would have spent with a manual notes search.

This issue is a frequent target for many developers. Application stores are full of different intelligent applications, which display specific data on the base of GPS coordinates or time priorities. This whole theme led to creation of a large number of specialized works and also of many patents that deal with both position systems and target marketing, which means displaying the data that the user will most probably need. In the following part of the work will be specified several possible approaches to locate the position, to filter the records and to display the data by users.

In the work [1] the authors concern with noting down the position for application managing reminders and consequent displaying the reminders in the relation to the current position. In this work it is possible to find out the principal of saving the actual position, while the application is divided into inner and outer part. Whereas in the outer part it's used the GPS module to get the information and in the inner part the location is gained by WiFi connection that is strong enough and by the Google API service is located the current user's position. This mean of position localization is definitely good and usable way, that's why we can take inspiration in this localization system, but retrieving the available data on the basis of distance calculation seems slightly difficult to me and certainly it will be preferable to suggest a different easier and smarter way [4–6].

In the patent publication [7] there is published very interesting approach to locate the user's position. This principal is based on transmitting radio frequencies. The wanted area is divided into several subareas covered by the radio signal. These radio signals periodically transmit information about actual subarea position. The mobile device receives these information and gains its own current position out of them. This approach is very interesting and innovative, but it isn't useful for wide utilization, because it is not possible to cover the whole world with this signal enough. Thereby this way matches only with use in smaller spaces, such as interiors. In this case we would have to use the principal of distance calculation again, which is slightly difficult as it was said already.

The patent publication [8] deals with complicacy of displaying the data according to user's current position. In fact the work concerns with target marketing —according to the user's position it shows advertisement and offers that the user could be interested in. The displaying principal is based on "geo fencing". The areas are divided into geo fences, which help us to detect, where the user is situated. The application enquires by its position if, or in which geo fence it is located and accordingly it provides displaying the advertisement. This way we found optimal because with this approach it is not necessary to make such complex calculations of distances between actual position and the notes' position. Unfortunately it lacks the time dimension, which enriches the created application.

The last method to gain current user's coordinates that will be mentioned here is WSM, which means Weighted Screening Method. It is a method of obtaining the positions purely by WiFi signal [9]. This method is based on a database of positions of individual WiFi networks, which means, that there must exist a database, where the positions of all the known WiFi networks will be available. The mobile device will try to find out its own position by enquiring the server with the database and announcing currently available WiFi connection to the server [10–16].

In this chapter there were presented several approaches that can be followed. None of them is absolutely perfect; therefore to develop this application it will be the most appropriate to choose a combination of described approaches. As the best variant it seems a combination of obtaining the current location by GPS module or by Google API [1], and consequently using these coordinates enquiring, if the user is situated in any geo fence and eventually displaying to the user all the notes loaded in this geo fence filtered according to time.

3 New Solution

In the previous chapter there were described several possible approaches to create the application. In this chapter it will be described its creation process.

As a solution for this application it was chosen a combination of approaches already specified in the previous chapter with an implemented secure and safe option of solution [17, 18]. The application can be divided into two main parts—for obtaining current user's coordinates it will be used a principal described in specialized work [1], which uses GPS module and in the places with access to internet connection it uses Google Services API. The next phase, as it was already mentioned, is creating and consequent receiving and sequencing available notes, for which it will be used geo fencing principal that is well explained in the specialized article [3].

Therefore the fundamental principal of the application will be this: while creating a new note it will be activated locating current user's position. In the application's inner structure it will be verified, if there exists any active geo fence nearby, to which the note could be allocated. In the case this geo fence doesn't exist, it will be created a new geo fence with a midpoint in actual position and with a set radius. While listing available notes the application locates the current user's position, verifies again, whether it doesn't belong to already existing geo fence. If there's an existing geo fence around the current position, the application displays all the notes contained in this geo fence sequenced by the time of their creation. But if any geo fence doesn't exist at the place, it means, that nearby this position there wasn't created any note yet, so the application gives back all the notes sequenced by the time of their creation.

As it was already mentioned, for locating the position it will be used GPS system in combination with WiFi. Android system has got integrated access to GPS module, which saves a lot of work with creating a communication with GPS module alone. Therefore the access to GPS module is accomplished by Location manager that is able to transmit communication between the application and GPS module alone. For the Location manager it is suitable to create "Location Listener", which enables us to scan current coordinates by several methods. This way we gain current user's coordinates that can be sent forward for further processing by another part of application. Other way of obtaining the coordinates is using internet connection. There exist a few possibilities to gain the current coordinates by WiFi connection—using internet connection—which have been already described in the previous chapter. For this application it was chosen the method using Google Services [4]. Google Services are made by Google Company and enable to note down the current user's position into the cloud in periodic intervals. But a specific user can have disabled access to this service; therefore the information from Google Services are not always a reliable source. When it's needed to obtain the coordinates we simply make sure the service is active. In the case it is not active, the method to gain the coordinates will be as described before—activating GPS module. If the service is active, we gain the last known position by the interface and we consider it to be the current one. Then we forward the coordinates to proceed.

Android operation system contains geo fencing service in the Google Services pack. This pack allows creating, deleting, detecting and—in general—controlling everything connected with geo fencing; it means that the Google Services interface can provide the whole second part of this notes application. As it was said already, while creating a new note the application gains current user's coordinates and the Google Services interface generates an enquiry, if there exists any geo fence in this locality yet. If yes, the application conjugates the note with the specific geo fence. But in the other case the application creates a new geo fence with a set radius parameter using the Google Services interface again. While enlisting the notes that are the closest to the user, the application by the interface again detects the geo fence, where the user is situated, and by an intern conjugation it displays all the notes that belong to the geo fence.

The process of creating the notes and their consequent displaying is shown in the flow diagrams (Fig. 1a, b). The last part of the application is the upgrade by the time dimension. It means that the user will most likely want to display the notes created in the similar time of the day as the current displaying time. Because of this functionality the application comes through the filtered notes once again and sequences them according to a set time level, which means according to an absolute time distance from the time of their display.

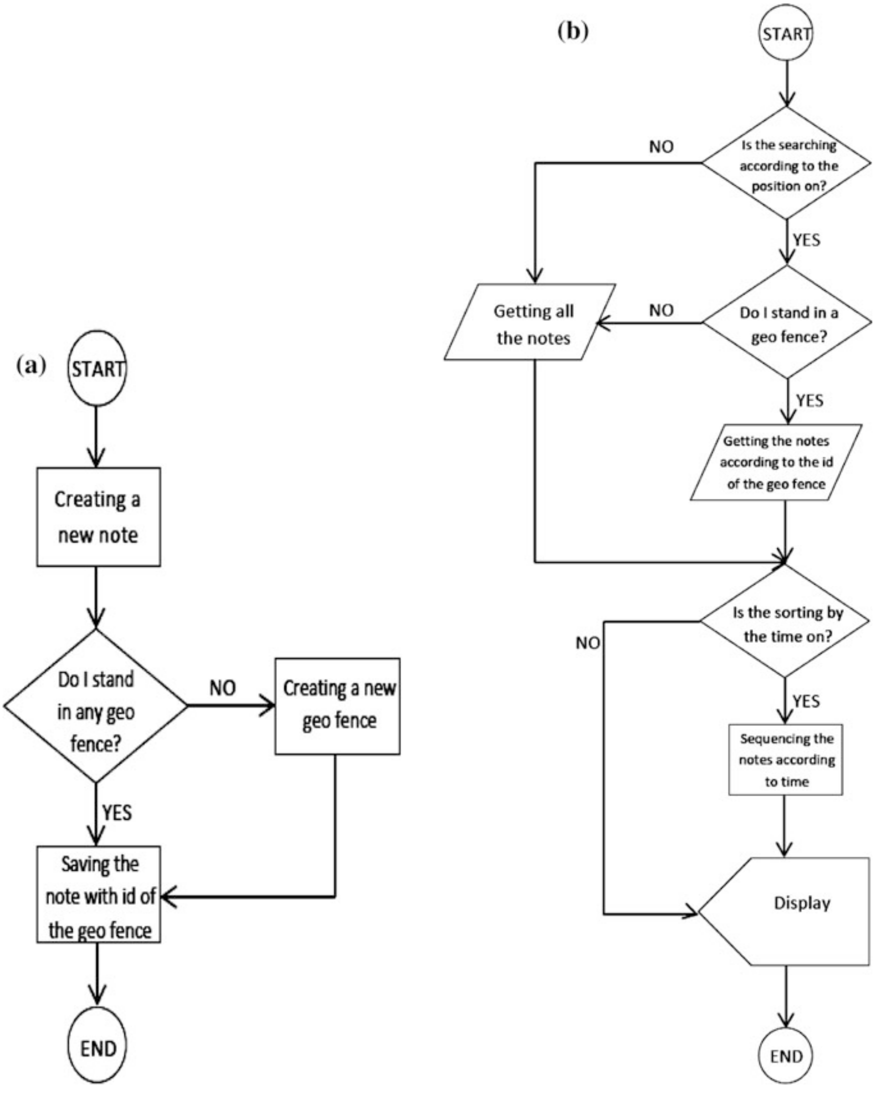


Fig. 1 **a** *Left*—flow diagram of creating the note; **b** *Right*—flow diagram—user has initiated scanning the notes

4 Implementation

As it was said before, for implementation of the application it was chosen the android platform. For sure there are more reasons to use this platform, but the main reason is common availability and accessibility of this system. Concretely it was selected version of android 4.0.3—the reason is its compatibility even with older devices.

First of all it was necessary to project the application structure and systems that the application will apply. For the first sight it is obvious that this application has to contain a database. For access to this database it was chosen ORMLite tool, which simplifies the work with database a lot and allows concerning more important parts of the application to a developer. Another technology or extern system is utilization of Google Services that are used for obtaining position loaded in the cloud by the android system. The application structure was selected the easiest possible and the most user friendly one. The application consists of one main page, which can be used to create a new note, browse the existing notes, display a calendar of user who is logged in the android system, and not the least to make a setting where it's possible to turn on or off the systems for sorting the notes.

Now it's appropriate to look over the individual parts of the application. The application can be primarily divided into three general parts—the model part that contains model objects needed for communication with the database, the view part that contains all activities serving to the interaction with the user and the intelligent part that concerns with the sorting notes according to set criteria itself.

The model part consists of categories Note and Person and their access objects. The category Note is an object applied for storing the information about the note and the category Person is an object keeping information about the users who are connected to the note. To these categories obviously belong their access objects so-called DAO objects, which means access object data. The access category for the Note object is formed by following methods: method "save", which serves to save the note into the database, method "delete", which serves to delete the note out of the database, next there are methods "getAll, getById, getByName and getByReqId". These methods are used to obtain the notes from the database. The last method is "edit", which is made to edit already existing note. The access category for the "Person" object consists of these methods: method "save" and "delete", which are used for manipulation with the record and method "getAll" and "getById", which serve to read the data from the database (Figs. 2 and 3).

In the view part there are created following categories: NewNoteActivity, ListNotesActivity, DetailNoteActivity, SettingsActivity and CalendarActivity. All of these activities are standard activities of android system without any customization. The category NewNoteActivity includes a text field, where it's possible to enter the note name, text field for entering the note itself, and a button for adding an event into the calendar. Next activity serves to display the detail of individual notes, which contains just non-editable text fields with the displayed note values. The SettingsActivity contains only two alternation switches, which allow turning on and off the sorting according to position and time. The last activity serves to display a list of already existing notes. The activity consists of a standard ListView component with notes and one text field, which is used for searching the notes according to a requested chain.

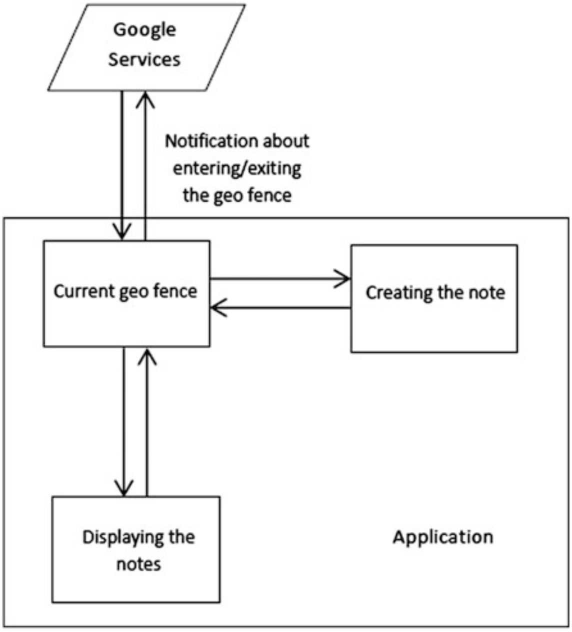


Fig. 2 Obtaining information about geo fence

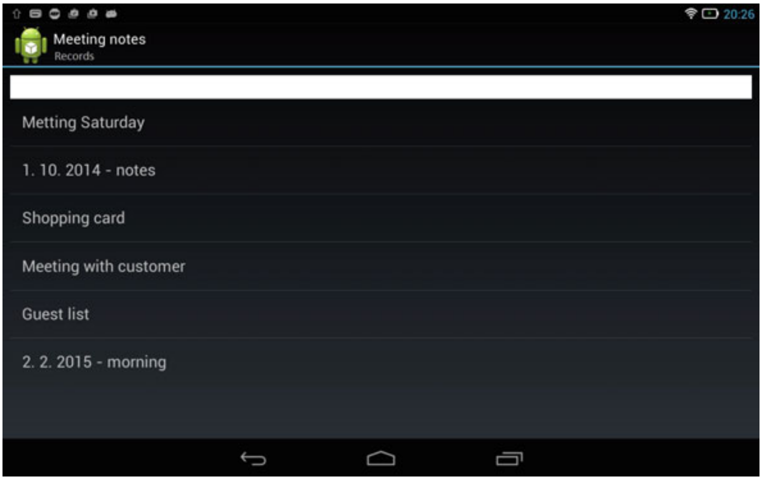


Fig. 3 Browsing the notes

The next few paragraphs will describe practical implementation of sorting the notes. Sorting according to time is a simple enhancement of common displaying notes. This sorting in the application is provided by category TimeStrategy. This category contains a method sortByTime, which sequences inserted.

The sequencing on the base of current user's position is slightly more difficult yet and furthermore it requires internet connection. The principal of displaying the data is this: In relation to the current position the application maintains either not any or specific identical number of geo fence, depending whether it is situated in one or not.

Sorting according to the position requires a reliance on GPS module and internet connection as well. An implementation using GPS module is obvious, but practically it's not used a lot, because it's quite difficult and it shortens battery endurance. If the user is satisfied with less precise position, it's evidently better to use internet connection to gain the last available position from Google Services.

5 Testing of Developed Solution

Focus of this chapter is absolutely clear—it's necessary to determine test scenarios that will enable to assess the new created application subsequently [19–22].

Testing this application can be conceived by several points of view. The first possible approach can be to test the speed of locating the position by GPS in comparison with the speed of locating the position by internet. The second point of view is to test the speed of displaying individual notes with the sorting on or off. The third and last approach is to detect, if this sorting has a meaning at all and if it is valuable for the user.

Testing the application with regard to the speed of locating position—testing this functionality will be realized as follows: Into the source code itself it will be inserted several time snapshots that will enable us to detect the time, when the application started to locate the position and the time, when the application finished the locating. Gathering these snapshots will be obviously realized more times and consequently the test will be clearly compiled and analysed.

The second test—if it's possible to see a difference in displaying the notes while the intelligent sorting is on or off—will be realized similarly as in the previous test, by reason that in this case it's not possible to use manual measurement, because the time diversion would be very small.

The third test has a long-period character and it's necessary to widen this application among the users. At the same time this test is very subjective, because every user evaluates the application himself according to a fact, whether the application helps him in the common life or not. As it was already said, to realize this test it's necessary to widen the application among the users and at the same time to make a questionnaire that will enable the users to evaluate the application certain time later. Because of the long-period character the results of this test are not presented in this work (Tables 1, 2, 3 and 4).

Out of the test results we can deduce following conclusions. While testing the time that the application needs to find its position by GPS method it was reached 12,142 time units on the average, on the opposite of that while locating the position by Google Services it was reached 0.8 time units on the average. Out of these

Table 1 Testing time for locating the position—GPS module

Number	Start (ms)	Stop (ms)	Difference (ms)
1	1,421,693,036,435	1,421,693,041,160	4480
2	1,421,693,202,072	1,421,693,221,500	19,428
3	1,421,693,223,857	1,421,693,241,568	17,711
4	1,421,693,386,526	1,421,693,401,915	15,389
5	1,421,693,451,736	1,421,693,462,090	10,354
6	1,421,693,503,735	1,421,693,522,153	18,418
7	1,421,693,583,503	1,421,693,602,353	18,850
8	1,421,693,640,107	1,421,693,642,251	2144
9	1,421,693,692,179	1,421,693,702,424	10,254
10	1,421,693,738,083	1,421,693,742,474	4391
Average			12,142

Table 2 Testing time for locating the position—Google Services

Number	Start (ms)	Stop (ms)	Difference (ms)
1	1,421,694,174,506	1,421,694,174,507	1
2	1,421,694,232,542	1,421,694,232,543	1
3	1,421,694,294,023	1,421,694,294,023	0
4	1,421,694,351,540	1,421,694,351,541	1
5	1,421,694,402,857	1,421,694,402,858	1
6	1,421,694,493,421	1,421,694,493,422	1
7	1,421,694,554,328	1,421,694,554,329	1
8	1,421,694,658,954	1,421,694,658,954	0
9	1,421,694,663,251	1,421,694,663,252	1
10	1,421,694,712,489	1,421,694,712,490	1
Average			0.8

Table 3 Testing time of displaying items—without sorting

Number	Start (ms)	Stop (ms)	Difference (ms)
1	1,421,695,251,911	1,421,695,252,091	180
2	1,421,695,310,783	1,421,695,310,896	113
3	1,421,695,351,541	1,421,695,351,736	195
4	1,421,695,391,629	1,421,695,391,810	181
5	1,421,695,429,307	1,421,695,429,484	177
6	1,421,695,461,761	1,421,695,461,947	186
7	1,421,695,507,477	1,421,695,507,670	193
8	1,421,695,537,080	1,421,695,537,264	184
9	1,421,695,579,996	1,421,695,580,173	177
10	1,421,695,611,804	1,421,695,611,986	182
Average			177

Table 4 Testing time of displaying items—sorted

Number	Start (ms)	Stop (ms)	Difference (ms)
1	1,421,695,729,034	1,421,695,729,214	180
2	1,421,695,782,231	1,421,695,782,389	158
3	1,421,695,817,292	1,421,695,817,484	192
4	1,421,695,854,215	1,421,695,854,392	177
5	1,421,695,887,153	1,421,695,887,339	186
6	1,421,695,921,178	1,421,695,921,354	176
7	1,421,695,954,818	1,421,695,954,990	172
8	1,421,695,991,528	1,421,695,991,707	179
9	1,421,696,025,112	1,421,696,025,292	180
10	1,421,696,056,612	1,421,696,056,788	176
Average			177

results it arises that use of locating the position by Google Services is markedly faster. The second test concerned the time of displaying the created items. By this test it was found, that while sorting the items it wasn't reached longer times than while not sorting, out of this test we can estimate, that sorting the notes doesn't have basic influence to slow the application down. Both of these tests are obviously of an orientation type, because they were realized on a very little sample.

6 Conclusions

The goal of this work was to create an application for creating and administration of user's notes. This application was intended to be enhanced with an intelligent searching of already existing items. As the most suitable solution it was chosen geo fencing for locating according to current position and sequencing the notes according to the current searching time. The application was programmed and it was tested by two basic kinds of tests. The first test intended to find out the differences between locating the position by GPS method and inquiring the Google Services. Out of the test it's obvious, that Google Services can locate the position a lot faster than GPS, but of course less precisely. The second test proved time of displaying the created items with and without sorting them. This test didn't approve distinctive difference between both approaches. This application can certainly achieve to find its favourers, who will be satisfied with its functionality. By the time with growing number of users it will be possible to realize the test of satisfaction and its benefits into common life e.g. in biomedical area [23–26].

Acknowledgement This work and the contribution were supported by project "SP-103-2015—Smart Solutions for Ubiquitous Computing Environments" Faculty of Informatics and Management, University of Hradec Kralove, Czech Republic. Last but not least, we acknowledge the technical language assistance provided by Pavlina Simkova.

References

1. Lin, C.-Y., Hung, M.-T.: A location-based personal task reminder for mobile users. Pers. Ubiquit. Comput. **18**, 303–314 (2014)
2. Chizari, H., Poston, T., Razak, S.A., Abdullah, A.H., Salleh, S.: Local coverage measurement algorithm in GPS-free wireless sensor networks. Ad Hoc Netw. **23**, 1–17 (2014)
3. Snider, L., Arbor, A.: Systems and methods for sharing of geo-tagged file segments based on location of a mobile device. Pub. No.: US 2014/0248912 A1. Pub. Date: 4 Sept 2014
4. Maresova, P.: Decision making criteria for cloud computing deployment. In: Proceedings of Advanced Multimedia and Ubiquitous Engineering, Lecture Notes in Electrical Engineering, vol. 352, pp. 93–98. Future Information Technology, Hanoi, Vietnam, 18–19 May 2015, Springer, Berlin. ISBN 978-3-662-47486-0
5. Maresova, P., Klimova, B.: Investment evaluation of cloud computing in the European business sector. Appl. Econ. **47**(36), 3907–3920 (2015)
6. Maresova, P., Halek, V.: Deployment of cloud computing in small and medium sized enterprises in the Czech Republic. E & M Ekonomie a Management **17**(4), 159–174 (2014)
7. Joshi, A., Bhargava, S., Suryanarayanan, R.: System and method of locating users indoors. Pub. No.: US 2013/0249672 A1. Pub. Date: 26 Sept 2013
8. Zeto III, M.J., Rippetoe, D., Shaw, D., Mercer, R., Gaxiola, G. Jr., Williams, R.T., Anders, E., Johansson, O.: System and methods for delivering targeted marketing content to mobile device users based on geolocation. Pub. No.: US 2013/0268353 A1. Pub. Date: 10 Oct 2013
9. Liu, H.-H., Lo, W.-H., Tsen, C.-C., Shin, H.-Y.: A WiFi-based weighted screening method for indoor positioning systems. Wirel. Pers. Commun. **79**, 611–627 (2014). doi:10.1007/s11277-014-1876-y
10. Behan, M., Krejcar, O.: Modern smart device-based concept of sensoric networks. EURASIP J. Wirel. Commun. Netw. (1), 155 2013. doi:10.1186/1687-1499-2013-155
11. Behan, M., Krejcar, O.: Adaptive graphical user interface solution for modern user devices. Lecture Notes Comput. Sci. LNCS **6592**, 411–420 (2012)
12. Krejcar, O.: Threading possibilities of smart devices platforms for future user adaptive systems. In: Lecture Notes in Computer Science, 4th Asian Conference on Intelligent Information and Database Systems, ACIIDS 2012, vol. 7197, pp. 458–467, 19–21 Mar 2012, Kaohsiung, Taiwan, LNCS (2012)
13. Mlynka, M., Brida, P., Machaj, J.: Modular positioning system for intelligent transport. 5th International Conference on Computational Collective Intelligence Technologies and Applications, Craiova, Romania, Studies in Computational Intelligence, vol. 513, pp. 115–124 (2014)
14. Gantulga, E., Krejcar, O.: Smart access to big data storage—android multi-language offline dictionary application. Lecture Notes Comput. Sci., LNCS **7653**, 375–284 (2012)
15. Obermeyer, L., Oliver, A., Oliver, K.W., Shetty, P., Rustgi, A.: Geo-fence entry and exit notification systém. Pub. No.: US 2013/0045753 A1. Pub. Date: 21 Feb 2013
16. Benikovsky, J., Brida, P., Machaj, J.: Proposal of user adaptive modular localization system for ubiquitous positioning. In: 4th Asian Conference on Intelligent Information and Database Systems, ACIIDS 2012, Kaohsiung, Taiwan, pp. 391–400. ISBN 978-3-642-28489-2
17. Horalek, J., Matyska, J., Sobeslav, V.: Communication protocols in substation automation and IEC 61850 based proposal (2013) CINTI 2013—Proceedings of 14th IEEE International Symposium on Computational Intelligence and Informatics, Art. no. 6705214, pp. 321–326
18. Cimler, R., Matyska, J., Balík, L., Horalek, J., Sobeslav, V.: Security issues of mobile application using cloud computing. Adv. Intell. Syst. Comput. **334**, 347–357 (2015)
19. Machacek, Z., Slaby, R., Hercik, R., Koziorek, J.: Advanced system for consumption meters with recognition of video camera signal. Elektronika Ir Elektrotechnika **18**(10), 57–60 (2012). ISSN 1392-1215

20. David, J., Svec, P., Frischer, R., Stranavova, M.: Usage of RFID wireless identification technology to support decision making in steel works. In: 21st International Conference on Metallurgy and Materials, pp. 1734–1738 (2012). ISBN 978-80-87294-31-4
21. Pies, M., Hajovsky, R., Latocha, M., Ozana, S.: Radio telemetry unit for online monitoring system at mining dumps. Appl. Mech. Mater. **548–549**, 736–743 (2014)
22. Ozana, S., Pies, M., Horacek, O., Zidek, J.: Concept of automated transportation center and its PIL model. In: World Congress on Engineering and Computer Science (WCECS'13)—International Conference on Modeling, Simulation and Control (ICMSC'13), vol. 2, pp. 992–996 (2013). ISBN 978-988-19253-1-2
23. Cerny, M., Penhaker, M.: The circadian cycle monitoring. Paper Presented at the 5th International Workshop on Wearable and Implantable Body Sensor Networks, BSN 2008. In Conjunction with the 5th International Summer School and Symposium on Medical Devices and Biosensors, ISSS-MDBS 2008, Hong Kong (2008)
24. Cerny, M., Penhaker, M.: Wireless body sensor network in health maintenance systems. Elektronika Ir Elektrotechnika **9**, 113–116 (2011)
25. Penhaker, M., Darebnikova, M., Cerny, M.: Sensor network for measurement and analysis on medical devices quality control. In: 1st International Conference on e-Technologies and Networks for Development, ICeND 2011, CCIS, vol. 171, pp. 182–196. Dar Es Salaam. Springer, Berlin
26. Penhaker, M., Krawiec, J., Krejcar, O., Novak, V., Bridzik, R., Society, I.C.: Web system for electrophysiological data management. In: Proceedings of Second International Conference on Computer Engineering and Applications, Iccea 2010, vol 1, pp. 404–407. doi:10.1109/iccea.2010.85

Artificial Immune System Based MAC Layer Misbehavior Detection in MANET

Shailesh Tiwari, K.K. Mishra, Nitin Saxena, Nitin Singh and A.K. Misra

Abstract MAC layer misbehavior drastically degrades the network efficiency even in the presence of secure ad hoc routing protocols. Even small number of malicious nodes may cause network partitioning or lead to failure of whole network. Simple attacks such as jamming or disruption on the 802.11 MAC, protocol if not taken care properly, propagated to the network layer. Detecting misbehaving node and punishing them is the only way for network survival. This paper introduces a Misbehavior Detection System (MDS) for MANET based on Artificial Immune System (AIS). Negative Selection technique is used for generating the detectors for identifying deviation from normal behavior. The proposed system detects malicious and selfish nodes performing misbehavior at MAC layer with the ability of learning and detecting new misbehavior. The system performance is evaluated using network simulator NS2 for MANET MAC layer 802.11 protocols over two on demand routing protocols AODV and DSR. Detection rate, false positive rate and Packet delivery rate are used as metrics for evaluation.

Keywords MANET · Artificial immune system · AODV · MAC layer · Negative selection

K.K. Mishra · N. Saxena · N. Singh · A.K. Misra
Department of Computer Science and Engineering,
Motilal Nehru National Institute of Technology Allahabad, Allahabad, India
e-mail: kkm@mnnit.ac.in

N. Saxena
e-mail: nitinsaxena777@gmail.com

N. Singh
e-mail: nitins@mnnit.ac.in

A.K. Misra
e-mail: akm@mnnit.ac.in

S. Tiwari (✉)
CSED, I.T.S Engineering College, Greater Noida, India
e-mail: shail.tiwari@yahoo.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_60

1 Introduction

Mobile Ad Hoc Networks (MANETs) are gaining popularity due to its fast and easy deployment even in intricate surroundings and have been employed in broad range of applications such as emergency services, pollution monitoring and vehicular networks [1]. MANET is a wireless network consisting of mobile nodes with no existing infrastructure support such as gateway or access points. Mobile nodes can directly communicate (in one hop) to each other in their transmission range, while multi-hop wireless links are used for those nodes that are not in their transmission range. Each node has to play the role of traffic forwarder (router) for multi-hop away mobile nodes, apart from traffic generator (Source) and traffic consumer (destination) in operating network. In MANET it is assumed that each node follow the protocols specification ideally employed at different layers such as AODV, DSR etc. at network layer and 802.11 DCF at MAC layer. Due to open and unsupervised model of MANET it is easier for misbehaving nodes to introduce them in the network and compromise the performance of network. These misbehaving nodes apply modifications in the default definition of protocols. These misbehaving nodes can be classified as malicious or selfish nodes. Malicious nodes try to disturb the normal functioning of network by dropping/introducing true/false protocol events, while selfish nodes aim to achieve the given goal by utilizing the network resources and save own resources. Malicious/selfish nodes can misbehave at physical, MAC or network layer. Selfish nodes may not forward data/control packets for saving battery power while malicious nodes may send false control/data packet to waste the available network bandwidth. At MAC layer in 802.11 DCF protocol, selfish node may not respond to RTS by its corresponding CTS packet even though route is possible through that selfish node. In other situation malicious node may apply jamming attack at MAC layer by placing fake RTS packet at regular interval for request of large timeslot for transferring data. Well behaving nodes must identify these malicious/selfish nodes for achieving maximum efficiency of the network. The problem of identifying these malicious nodes becomes more difficult due to nodes mobility. Nodes in MANET can move arbitrarily and result in frequent change of nodes connectivity. Distinguishing nodes that perform misbehavior deliberately from genuine one is not straight forward.

Based on the form of audit data adopted, Intrusion Detection System (IDS) can be classified as Host based IDS (HIDS) and Network based IDS (NIDS). A network-based IDS captures and analyzes packets from network traffic while a host-based IDS uses operating system or application logs in its analysis [2]. Other classification includes detection technique parameter for classifying IDS as anomaly or misuse detection system [3]. The misuse detection system stores patterns (or signatures) of known attacks and uses them to compare with the captured data. Any matched pattern is treated as an intrusion. The drawback of misuse detection system is that it cannot detect new kinds of attacks.

The misuse detection system defined in [3] use predefined misbehavior pattern as signature represented in the form of specific sequence of events. The system requires knowledge of potential misbehaviors and works efficiently only if when known attacks are performed. Existing approaches are not suitable for learning new misbehavior and adapt accordingly. In open and distributive environment of MANET, it is difficult to predict and define complete list of all possible misbehaviors. This motivates the employment of Artificial Immune System (AIS) technique along with string pattern matching in the proposed system for detecting misbehavior. Instead of storing the activity pattern of well behaving nodes Negative Selection Algorithm is used to generate pattern detectors randomly that encompass possible misbehavior pattern.

In this paper we focus on identifying misbehavior at MAC layer i.e. the problem of detecting malicious/selfish nodes that do not follow the standard definition of IEEE 802.11 DCF protocol. The formation of detectors includes the protocol events (defined in Table 3) of both MAC and Network layer. AODV and DSR on demand routing protocol is assumed with constraint of no misbehavior at network layer. Once well behaved node confirm about the presence of malicious/selfish node in the network, counter measures can be applied to minimize the effect of those malicious/selfish nodes. Possible reactions of misbehavior can range from avoid responding/using malicious/selfish node to exclude them from any participation in the network. This work encompasses the detection of malicious/selfish nodes and does not discuss possible reactions after detection. The network simulator *ns2–2.35* is used for evaluating the performance of proposed system. The parameters used for evaluation include detection rate, false positive rate and packet delivery rate.

We define the framework and module structure for proposed Misbehavior Detection System (MDS). The correspondence of Natural Immune system (NIS) and MANET characteristics is shown in Table 1 which conform the path for using AIS in misbehavior detection. A list of *Imprints* focusing on key protocol events and their order is defined and used in the formation of detectors. Finally the performance of proposed system is measured over multiple parameters by implementing the system in network simulator ns2.

1.1 Contribution and Paper Organization

We define the framework and module structure for proposed Misbehavior Detection System (MDS). The correspondence of Natural Immune system (NIS) and MANET characteristics is shown in Table 1 which conform the path for using AIS in misbehavior detection. A list of *Imprints* focusing on key protocol events and their order is defined and used in the formation of detectors. Finally the performance of proposed system is measured over multiple parameters by implementing the system in network simulator ns2.

The rest of the paper is organized as follows: Sect. 2 presents brief description of MAC and Network layer protocols for MANET as well as potential threats of IEEE 802.11 protocol as background knowledge; Sect. 3 covers the literature survey; Sect. 4 discuss the Natural Immune System (NIS) and AIS techniques. Section 5

Table 1 Correspondence of NIS and MANET characteristics

Characteristic	NIS	MANET
System	Body	Mobile ad hoc network
Elements	Molecules, cells and organs	Mobile nodes
Controlling	No centralized control	Autonomous nodes
Objective	Identification of malfunctioning cells and elements	Identification of selfish and malicious nodes

covers the details of proposed Misbehavior Detection System (MDS); Sect. 6 describes the simulation model and presents the simulation results. Section 7 finally concludes the paper.

2 Background Details

2.1 MAC Layer in MANET

The IEEE 802.11 standard for wireless networks is responsible for fair access to the medium for all users in the network. The IEEE 802.11 defines two basic access modes for wireless devices: (1) Distributed Coordinated Function (DCF); (2) Point Coordinated Function (PCF). PCF is a centralized mechanism and requires access points while DCF a distributed mechanism. 802.11 DCF uses Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) and random backoff time. It is designed to reduce collisions when multiple nodes access the shared medium. The 802.11 DCF provides two communication options: (1) four-way handshaking, i.e. RTS-CTS-DATA-ACKA, suitable for long frame data transmission; (2) two-way handshaking, i.e. DATA-ACK, suitable for short frame data transmission. The proposed system makes use of four-way handshaking in 802.11 DCF.

A node that wants to transmit packets first senses the medium for any transmission. If the medium is idle for at least Distributed Inter-frame space (DIFS) duration, it requests the channel by sending a short control frame Request to Send (RTS) to the receiver node. If the receiver correctly receives the RTS, it will reply with a short control frame Clear to Send (CTS). On reception of CTS at transmitter, it starts transferring DATA. After receiving DATA successfully, the receiver sends an ACK to the transmitter. The RTS/CTS scheme used prior to the actual data transmission reduces the high collision probability of large data frames since RTS and CTS frames exchanged by sending and receiving nodes contain information about the length of the packet to be transmitted. The RTS/CTS contains a duration field indicating the time (in microseconds) after the end of present frame transmission that the channel will be reserved to complete the data or management frame transmission. Any node within the transmission range of either sending node or receiving node hears the RTS/CTS will learn the medium reservation and adjust its Network Allocation Vector (NAV), which indicates the amount of time that the node should defer. The collision will mostly happen when the current node completes its transmission and multiple nodes are waiting for contending for the channel. Thus each node with data to transmit will generate a random backoff number from range [0, CW] for an additional deferring time, where CW is the contention window size maintained by each node. While the channel is idle, the backoff number will decrease by one after one time slot, which is defined in the IEEE 802.11 standard and stop decrementing when the channel is busy. Once the backoff counter reaches zero, the sending node will reserve the channel by exchanging RTS/CTS as described above.

If a node sends RTS but does not receive CTS within certain time, the node will defer by doubling its CW size and choosing a random value from the new range and retransmit RTS with limited times. If the RTS retry time is more than the station short retry count (SSRC) the sending node will drop the DATA packet and inform the network layer of a link breakage. Alternatively, if the ACK is not received within certain time, the sending node will retransmit the DATA packet for limited time (i.e. SSRC for a short frame DATA, or station long retry limit SLRC for a long frame DATA).

2.2 Network Layer in MANET

Several standard routing protocols such as AODV, DSR and DSDV etc. are available for providing the functionality of network layer in MANET. The primary job of any routing protocol is to find the shortest path between source and destination. Even all routing protocols share the same objective of establishing route from source to destination, their operational part may be extensively different. One possible classification divide routing protocols in proactive and reactive categories based on when route discovery process is scheduled. Proactive routing protocols keep the updated route (if it exists) from each node to all other nodes in the network. Contrary to proactive protocols, reactive (also known as on-demand) routing protocols discover route (only from source to destination) when source node have some data to transfer to destination node. In this paper we use Ad hoc On-demand Distance Vector (AODV) and Dynamic Source Routing (DSR) reactive routing protocols.

2.3 IEEE 802.11 DCF Potential Threats

With the implementation of MAC protocol in software rather than hardware or firmware in network access cards, it is easy to modify the protocol by a selfish or malicious node [3]. A malicious node may apply flooding to almost prevent the well behaved nodes in its vicinity from transmitting valid data and iterates negatively effects the network performance that could lead to Denial of Service (DoS). A more smart malicious node apply the similar attack with out wasting its own energy in transmitting false data packets to generate flood by just sending RTS packet intelligently, containing request of time slice required for sufficient large data packets. Alternatively malicious node may target one or more key nodes of the network such that victim nodes are burdened to receive or forward large number of packets, with the intention of draining out the energy of targeted nodes.

A selfish node may choose not to cooperate with other nodes of the network i.e. do not involve in routing activity to save its energy. In this attack the selfish node drop all or selective RTS/DATA packets. It is difficult for sender to judge that absence of CTS/ACK is due to collision/interference of control packets or misbehavior of some selfish nodes. Further this selfish node modifies its backoff mechanism to gain access of channel more frequently than others.

3 Related Work

Artificial Immune System (AIS) techniques are being applied for detecting intrusion in both wired and wireless networks. Hofmeyer [4], Hofmeyer and Forrest [5] have given intrusion detection system for wired LAN using AIS. Kim and Bentley [6] map artificial immune system concepts to computer network elements efficiently. Le Boudec and Sarafijanovic [7] adapt the structure of self/nonself and matching function of [6] to their framework, followed by mapping and construction of AIS for MANET elements. For wireless ad hoc networks most of the work focus on detecting attacks on network layer. In [7] antigens and antibodies, the key AIS concepts, are defined on the basis of DSR routing protocol events, managed to handle only routing misbehavior vulnerabilities of DSR. Similarly Balachandran et al. [8] use negative selection and structured genetic algorithm (sGA) as hybrid negative selection algorithm (HNSA) for detecting routing misbehavior. Behavior based Anomaly Intrusion Detection System (BAIDS) [9] detects wormhole attack in MANET and use hybrid negative selection algorithm (HNSA). Although BAIDS includes 802.11 MAC layer protocol event sequences along with routing protocol events in their genes formation but give attention to routing misbehavior and evaluated the system for AODV, DSR and DSDV routing protocols. Our Misbehavior Detection System (MDS) uses negative selection algorithm for detecting misbehavior such as jamming and disruption in 802.11 MAC protocols of MANET [10].

4 Artificial Immune System (AIS)

Artificial Immune System (AIS) techniques have inspired from immune system of body i.e. Natural Immune System (NIS) to build computational tools for the vast range of domain areas like pattern recognition [9], intrusion detection etc.

4.1 Natural Immune System (NIS)

All living species have certain level of immunity to defend external attacks. The vertebrate's immune system commonly referred as Natural Immune System (NIS) consists of variety of molecules, cells and organs spread all over the body [9]. The functioning of NIS is not controlled by any central organ. The immune system is responsible for identifying (1) harmful body cells (cancer or tumor cells) i.e. playing negative role in the body (2) external elements **causing** diseases (viruses and bacteria). Every element that can be recognized by the NIS is called *antigen* (Ag). Immune system labels these elements as *self-antigen* or *self* (harmless cells of the body) and *nonself antigen* or *nonself* (disease causing elements) [4].

Immune system recognizes antigens (self or nonself) with the help of immune cells categorized as B-cells and T-cells. Immune cells contain receptor molecules on their surface capable of recognizing vast range of antigens. B-cells are capable of recognizing antigens free in solution while T-cells require antigens to be presented by the accessory cells as illustrated in Fig. 1a and b [11]. Once immune system recognizes the presence of *nonself antigens* in the body with certain *affinity* named *affinity threshold*, immune response is mounted.

4.2 Artificial Immune System (AIS)

An Artificial Immune System (AIS) is a computational model of Natural Immune System (NIS) and incorporates properties of NIS including diversity, distributed computation, error tolerance, dynamic learning and adaptation and self-monitoring. AIS identify Negative Selection, Clonal Selection and Immune Network Theory as key techniques for mapping NIS process. Negative selection algorithm is used to solve anomaly detection problems [6].

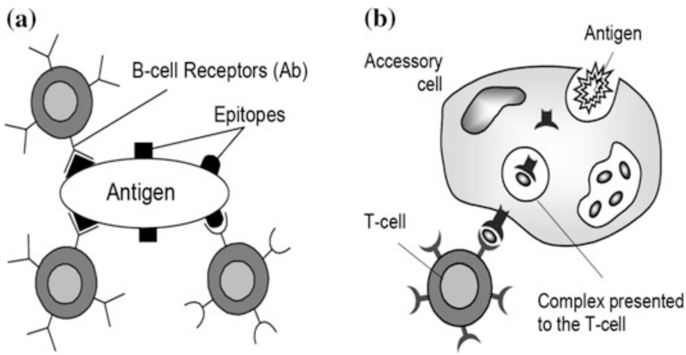


Fig. 1 **a** B-cell recognizing an antigen free in solution. **b** T-cell recognizing an antigen presented by accessory cell

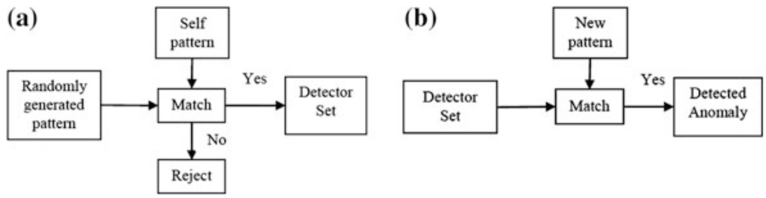


Fig. 2 **a** Generation of detector set. **b** Anomaly detection using detector set using negative selection algorithm

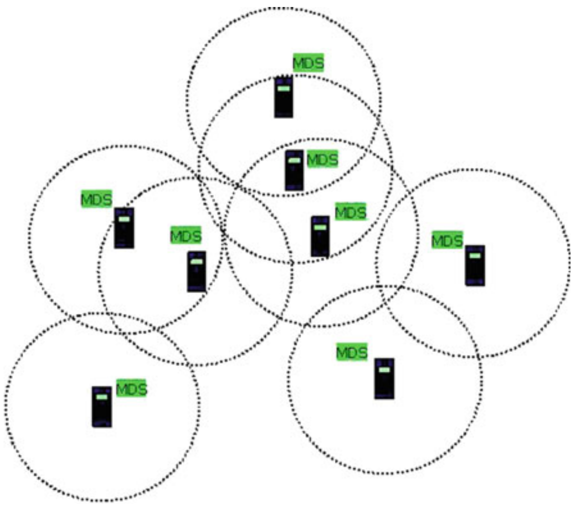
Negative Selection technique described in [3] defines the pattern of normal behaving nodes as *self*.

The process of negative selection algorithm to generate detector set and anomaly detection [3] is shown in Fig. 2a and b.

5 Proposed Work

The proposed Misbehavior Detection System (MDS) as shown in Fig. 3 is collaborative and distributed in nature suitable for MANET. Each node detects malicious and selfish activities by monitoring other nodes in its broadcast range. Subsequently, nodes share their experience to other nodes for collaborative update. Every node is running MDS module along with network protocol stack. The MDS functionality is carried out in two folds: (1) Independent local detection, (2) Global or System wide detection. Local detection uses monitored node traces for misbehavior detection. System wide detection incorporates received detection parameters from other nodes along with its own outcome.

Fig. 3 MDS implementation in MANET



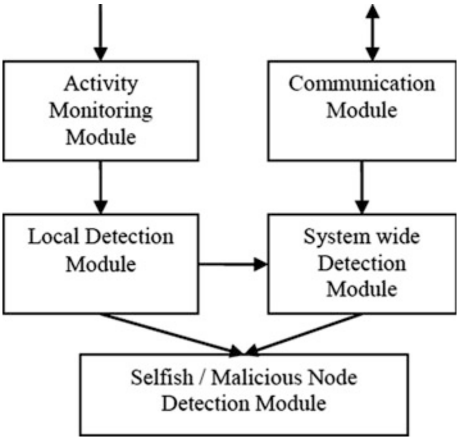


Fig. 4 Module structure of MDS

Misbehavior Detection System (MDS) running at each node can be divided into modules as shown in Fig. 4. Activity Monitoring Module is responsible for collecting event traces of each neighbor. Local Detection Module take misbehavior decision using this collected event traces. System wide detection module combines detection parameters received by neighbor nodes using Communication Module and local detection for formation of global observation. Final decision about nodes is carried out in Selfish/Malicious Node Detection Module which takes input from both local and system wide detection module.

5.1 System Methodology

The proposed MDS is an anomaly detection system where ideal or normal behavior of nodes/network are stored and compared with captured audit data. Deviation (larger than some predefined threshold) from normal behavior is counted as misbehavior.

Due to unique characteristics of MANET, it is difficult to compose a complete set of node/network profile [9]. Alternatively we are using the specification based approach for maintaining network profile. Detectors trained under ideal environment, are employed for predicting anomaly in audit data. Identification is being done by the system in two phases i.e. (1) Learning phase (2) Detection phase.

5.2 Mapping of NIS and MANET

The activities in MANET closely resemble the functioning of Natural Immune System and this motivates the application of AIS techniques for detecting malicious

Table 2 Mapping of NIS elements to MANET elements

NIS elements	MANET elements
Body	The entire mobile ad hoc network
Self cells	Ideal nodes
Nonself cells	Malicious nodes
Genes	Correlated protocol event subsequences (imprints)
Antigen	Pattern of protocol event sequence used by self cells
Antibody	Pattern of protocol event sequences used by nonself cells
Chemical binding	Mapping of antibody to antigens
Negative selection	Negative selection block for antibodies generation

nodes in MANET. The correspondence of NIS and MANET characteristics are shown in Table 1.

Applying AIS for misbehavior detection in MANET requires mapping of NIS elements to MANET Elements [2] as shown in Table 2.

5.3 *Protocol Events, Imprints Representations and Antigen/Antibody Formation*

In MANET nodes use routing algorithms such as AODV, DSR, DSDV etc. for network layer functioning, 802.11 DCF at MAC layer which subsequently employ radio links at physical layer to transmit data over the wireless channel. The key protocol events at MAC and Network layer [9], mentioned in Table 3 forms the basis of antigens for the proposed system.

Each monitoring node collects one protocol event trace per monitored node in their neighbor. MANET operation time is divided into non overlapping time intervals of duration Δt (e.g. $\Delta t = 20$ s). Each protocol event trace is collection of data sets (up to maximum of N_s events in each data set) collected one each in Δt

Table 3 Observed protocol events in MANET

Layer	Packets transmitted		Packets received	
	Label	*Events*	*Label*	*Event*
MAC	X1	RTS	X2	RTS
	Y1	CTS	Y2	CTS
	Z1	ACK	Z2	ACK
Network	A	RREQ	E	RREQ
	B	RREP	F	RREP
	C	RERR	G	RERR
	D	Data sent and IP address not of monitored node	H	Data sent and IP address not of monitored node

time interval. Protocol event in each data set is converted to corresponding label (refer Table 3) e.g.

E_1 = (EAFBX2Y1HX2Y1HX2EDX1Y2DX1Y2, …)

Specific protocol events are correlated i.e. occurrences of one event have pre-requisite of other event to be occurred. Example sending of CTS (Y1) event requires receiving of RTS (X2) event. An *Imprint* is an atomic pattern for matching occurrences of correlated protocol events. Representation of *Imprints* provides more meaning to raw protocol events sequence. The System is using following *Imprints* for formation of antigens.

Imprint 1: **count** (X2)
Imprint 2: **count** (X2*Y1)
Imprint 3: **count** (X1)
Imprint 4: **count** (X1*Y2*)
Imprint 5: **count** (X1*Y2*D)

where **count** (sub-pattern) is representing frequency of *Imprint* in a data set and * mapped to any label with zero or one occurrence in sub-pattern. Data sets of protocol event traces are mapped to *Imprint* values in that data set. For example E_1 is converted to E_2 where,

E_2 = (3 2 2 2 1)

Each imprint is encoded on N bits (e.g. N = 10) and the last step of antigen formation has been adopted as given in [12]. The imprint value range below some threshold is divided by N, partitions the range in N intervals. Intervals are then assigned numbers sequentially starting from 1. An imprint whose value is less than the threshold belongs to one interval. Set the interval number bit position in the N-bit pattern to 1. Imprint values larger than the threshold are encoded as if they belonged to the last interval. In this way E_2 is converted to E_3 for N = 10:

E_3 = (0000000010, 0000000010, 0000000010, 0000000001)

Each node generates one antigen in every Δt time interval for each monitored (neighboring) node. Every bit in this representation is “nucleotide” of genes in NIS.

Both antibodies and antigens are having the same format and length other than the occurrences of ‘1’ per imprint. In antigen each imprint will include exactly one ‘1’ per imprint while antibody may have any number (0 to N) of ‘1’ per imprint. Matching of antibody to an antigen is defined in [13]. Match is true if the antibody has a 1 in every position where the antigen has a 1. This allows detection system to have good coverage of a large set of possible nonself antigens with a relatively small number of antibodies [5].

Negative Selection technique of AIS is employed while accepting antibodies for the system. Initially antibodies are generated randomly distributed uniformly over the set of possible antibodies. Any self-antigen matches to generated antibody is discarded [14].

5.4 Learning and Detection Phase

All nodes behave ideally in the learning phase of system that leads to the generation of true network traffic profile. The learning phase is long enough so that system able to capture ideal behavior correctly. Antigens are formed on the basis of applied traffic that reflects the ideal behavior of nodes. Subsequently antibodies are generated randomly that encompass nearly all possible misbehaviors. Antibodies work as detectors for identifying deviation from normal behavior in detection phase.

In the detection phase certain number of nodes are set to misbehave that disturb the normal working of network. Detectors generated during the learning phase are applied to antigens of currently running network traffic. Match of any detector to antigen represent the potential misbehavior.

6 Performance Evaluation

The proposed Misbehavior Detection System (MDS) is implemented in network simulator *NS2 2.35* for analyzing the effectiveness of the system in detecting selfish or malicious behavior for 802.11 MAC protocol in MANET. MDS extends the functionality of 802.11 by adding misbehavior detection capabilities. Simulation results are based on detection rate, false positive rate and packet delivery rate metrics.

6.1 Simulation Description

Mobile Ad hoc Network consist of 40 nodes in a rectangular area of dimensions 800 m × 800 m is simulated for 60 min. The initial positions of nodes are chosen randomly defined in input file. Nodes move randomly with fixed speed of 1 m/s. Among 40 nodes 7 nodes are sources of data, generate Constant Bit Rate (CBR) traffic of packet size 512 bytes sent 5 packets/s. The simulation parameters and their default values are summarized in Table 4.

Misbehavior is introduced by modifying the standard specification of 802.11 protocol for certain number of nodes marked as misbehaving nodes. The other nodes of the network following the default definition of protocol are well behaving nodes. In our simulation we incorporate two misbehaviors: (1) Not replying with CTS packet corresponding to RTS packet received by selfish nodes. (2) Sending of fake RTS packet at regular interval by malicious nodes.

Table 4 Simulation parameters

Simulation environment parameters	
Parameter	Default value
MAC protocol	802.11 DCF
Routing protocol	AODV, DSR
Simulation time	60 min
Simulation area in meters	800 × 800
Number of nodes	40
Radio range	250 m
Propagation path loss model	Two-ray
Mobility model	Random way point
Mobility speed (no pauses)	1 m/s
% misbehaving nodes	5, 10, 15, 20, 25, 30, 35, 40
Traffic type	CBR
Connection type	UDP
Packet size	512 bytes
Traffic generation rate	5 packets/s

6.2 Performance Metrics

The performance of system is measured using the following metrics: (1) Detection Rate, (2) False Positive Rate, (3) Packet Delivery Rate (PDR).

- Detection Rate is the percentage of malicious nodes identified correctly.
- False Positive Rate is percentage of well behaving nodes identified as malicious.
- Packet Delivery Rate is percentage of packets delivered at destination.

6.3 Simulation Results and Analysis

Based on the performance metrics we have analyzed the simulated results as given below:

(1) Detection Rate: Simulation result represented in Fig. 5 shows that detection rate of MDS decreases for 802.11 MAC protocol under both routing protocol DSR and AODV. The system's detection rate is more than 70 % even in the presence of 40 % misbehaving nodes.

The detection rate of MDS increases as we increase the number of detectors we employ in detection phase as shown in Fig. 6. Antibodies generated in learning phase work as detectors in detection phase. The MDS able to detect more than 55 % of misbehaving nodes using only 100 detectors while it reaches close to 98 % when number of detectors are increased to 500 for both DSR and AODV routing protocols running over 802.11 MAC protocol.

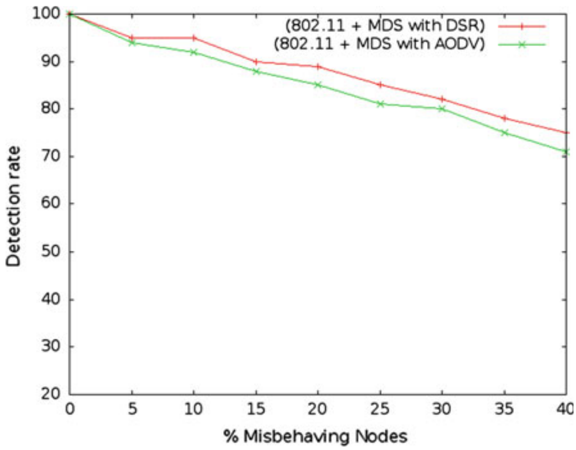


Fig. 5 MDS detection rate with varying number of misbehaving nodes

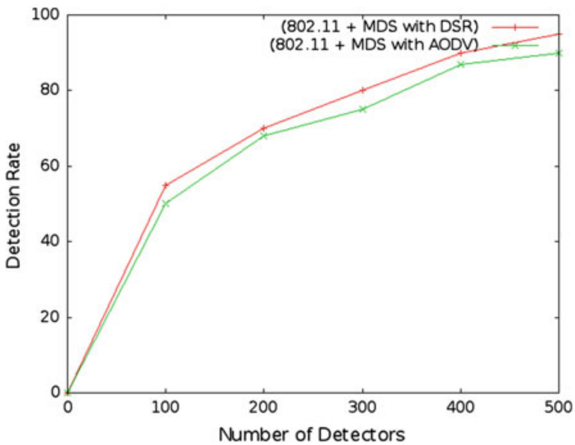


Fig. 6 MDS detection rate with varying number of detectors

(1) False Positive Rate: False positive rate of proposed system MDS ranges from 3 to 10 % for 802.11 under both routing protocols. The false positive rate is maximum when network contains 40 % misbehaving nodes (Figs. 7 and 8).

(2) Packet Delivery Rate: Packet delivery rate of underlying network decreases as we increase the number of selfish and malicious nodes. This is same when 802.11 are working alone without any detection system. It is reasonable because our proposed system is performing misbehavior detection only and does not incorporate any possible reactions.

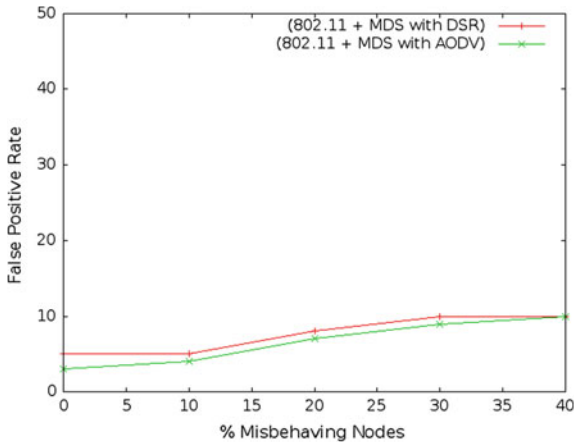


Fig. 7 MDS false positive rate with varying number of misbehaving nodes

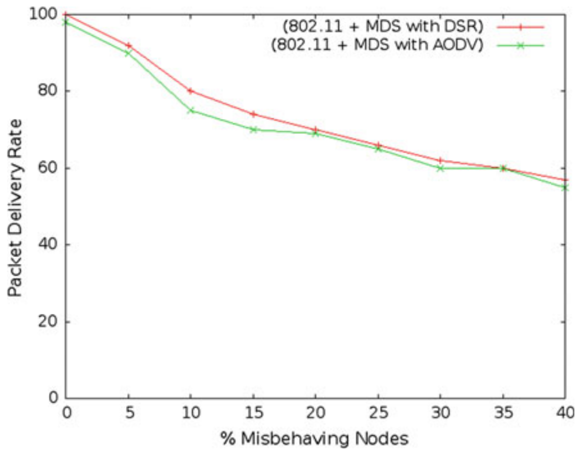


Fig. 8 Packet delivery rate with varying number of misbehaving nodes

7 Conclusions

We have proposed a Misbehavior Detection System (MDS) for detecting nodes indulging in selfish or malicious activities in 802.11 MAC layer protocol of MANET environment. MDS use negative selection for generating antibodies in learning phase that later work as misbehavior detectors in detection phase. The proposed system requires learning in ideal environment so that it can detect nodes anomaly behavior by computing the deviation from standard behavior. The system is also capable of learning and identifying new misbehavior. The detection system is tested for two on demand routing protocols DSR and AODV. Misbehavior is incorporated by modifying the behavior of nodes as selfish or malicious. Selfish nodes do not respond to RTS packet by CTS packet while malicious nodes send fake RTS packet in regular time interval to reduce the chance of well-behaved nodes for sending genuine data.

Simulation results confirm good detection rate even in the presence of large percentage of misbehaving nodes. The proposed system is able to identify nearly all misbehaving nodes for sufficiently large number of detectors with very small false positive rate. The advantage of the proposed system is that it can identify 802.11 MAC layer misbehaviors efficiently. On the hand limitation of system is it assumes that no misbehavior in routing protocols.

References

1. Issariyakul, T., Hossain, E.: Introduction to Network Simulator NS2. Springer, Berlin (2012)
2. Anantvalee, T., Wu, J.: A survey on intrusion detection in mobile ad hoc networks. In: Wireless Network Security, pp. 159–180 (2007)
3. Forrest, S., Perelson, A.S., Allen, L., Cherukuri, R.: Self-nonself discrimination in a computer. In: Proceedings of the IEEE Symposium on Research in Security and Privacy. IEEE Computer Society Press, pp. 202–212 (1994)
4. Hofmeyr, S.A.: An Immunological Model of Distributed Detection and it's Application to Computer Security. Ph.D. thesis, Department of Computer Sciences, University of New Mexico, April 1999
5. Hofmeyr, S.A., Forrest, S.: Architecture for an artificial immune system. Evol. Comput. **7**(1), 45–68 (2000)
6. Kim, J., Bentley, P.: An evaluation of negative selection in an artificial immune system for network intrusion detection. In: Proceedings of the Genetic and Evolutionary Computation Conference, Morgan Kaufmann, pp. 1330–1337 (2001)
7. Le Boudec, J.-Y., Sarajanovic, S.: An Artificial Immune System Approach to Misbehavior Detection in Mobile Ad-Hoc Networks. Technical Report IC (2003)
8. Balachandran, S., Dasgupta, D., Wang, L.: A hybrid approach for misbehavior detection in wireless ad-hoc networks. Published in Symposium on Information Assurance (2006)
9. Sundararajan, T.V.P., Ramesh, S.M., Maheswar, R., Deepak, K.R.: Biologically inspired artificial intrusion detection system for detecting wormhole attack in MANET. Wirel. Netw. J. **20**, 563–578 (2013)
10. Ansari, M.S.A., Inamullah, M.: Misbehavior detection in mobile ad hoc networks using artificial immune system approach, pp. 1–6 (2011)
11. de Castro, L.N., Timmis, J.: Artificial immune systems: a novel paradigm to pattern recognition. In: Artificial Neural Networks in Pattern Recognition SOCO 2002, pp. 67–84 (2002)
12. Gonzalez, F.A., Dasgupta, D.: Anomaly detection using real-valued negative selection. J. Genet. Prog. Evolvable Mach. **4**, 383–403 (2003)
13. Dasgupta, D., Gonzalez, F.: An immunity based technique to characterize intrusions in computer networks. IEEE Trans. Evolut. Comput. **6**, 281–291 (2002)
14. Hofmeyr, S.A., Forrest, S.: Immunity by Design: An Artificial Immune System. GECCO 1999

Wideband Monopole Antenna for WWAN Services

Mohamad Amir Imran Mohd Hasli, Ahmad Rashidy Razali, Aslina Abu Bakar, Mohd Aminudin Murad and M. Feroze Akbar J. Khan

Abstract An approach to design a compact monopole antenna for use in portable RF transceivers that feature wideband performance and miniaturized dimension is presented. The antenna structure is initially designed as a microstrip-fed planar monopole quarter elliptical shape supported by a 60 × 90 mm^2 ground plane bolstered on FR-4 substrate. A parasitic microstrip stub and a ground plane cut are introduced to achieve wideband characteristic. The proposed approach modifies the antenna's structure by introducing two miniaturization steps on both the radiator and ground plane. The modification that partially removes the radiator and ground plane to enact a slot is the first step in the proposed miniaturization. In the second step, the main radiator is carefully corrugated near the tapered end of the elliptical radiator to further reduce antenna's size without undermining the wideband performance. The simulation and experimental results are compared and discussed. The designed antenna operates at all wireless wide area network (WWAN) services from 850 MHz to 3 GHz at 6 dB return loss reference.

M.A.I. Mohd Hasli (✉) · A.R. Razali · A. Abu Bakar · M.A. Murad · M.F.A. J. Khan
Faculty of Electrical Engineering, Universiti Teknologi MARA (UiTM),
13500 Permatang Pauh, Pulau Pinang, Malaysia
e-mail: amir.imran07@yahoo.com

A.R. Razali
e-mail: ahmad073@ppinang.uitm.edu.my

A. Abu Bakar
e-mail: aslina060@ppinang.uitm.edu.my

M.A. Murad
e-mail: aminudinmurad82@gmail.com

M.F.A. J. Khan
e-mail: muhammad.feroze@yahoo.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_61

1 Introduction

With the rapid and growing interest in wireless communication system and its application, the compact multiband and wideband antenna are highly demanded. Such antennas are required to handle various wireless services over the 850 MHz to 3 GHz frequency range. These include AMPS, GSM, GPS, DCS, PCS, UMTS, WLAN and WiMAX service bands. Many research that have been done show that the smaller the antenna size and dimension, the lower is the bandwidth [1]. This challenge always becomes the drawbacks to antenna designer in order to meet the compact criteria came from the industry. To overcome the impact, designers introduce lots of technique which manipulate the current distribution on the antenna radiator. Techniques such as truncated ground plane slot [2, 3], parasitic element introduction [4, 5] and radiator slot [6–8] have been explored in order to achieved compact size, broadband characteristics and desired band of interest.

This paper present a new approach of the antenna designed in [9] with reduction of the antenna size using a combination of a few method that have been experimentally proved. The wideband characteristics are achieved by introducing a symmetrical corrugated slot along the radiator end which improved overall return loss. Then, a slot is applied in the middle of the radiator to further enhance overall return loss. Finally a symmetrical ground slot near the feed line is attached to tune the resonance to cover most of significant services within this frequency range.

This paper is organized as follows. Section 2 describes the geometry and the design process of the proposed antenna. Comparison between original antenna in [9] and proposed antenna is described in full electromagnetic (EM) wave simulation by comparing the return loss performance of both antennas. A few parametric studies have been presented to show the effect of modification to achieve optimum design parameters. Next, in Sect. 3, the experimental antenna results are described. This includes a comparison and discussion about simulation and experimental return loss, antenna gain, and radiation pattern at the band of interest. In the end, the conclusion is resolved.

2 Antenna Geometry

2.1 Initial Antenna

The initial structure of quarter elliptical monopole antenna is as shown in Fig. 1. It is initially proposed by the work in [9] that provide a wide band antenna for wireless service band below 6 GHz. The main radiator is designed using the combination of quarter ellipse with major axis radius, a = 50 mm, minor axis b = 30 mm and a 13 mm radius circular structure as the curvature edge. The radiator

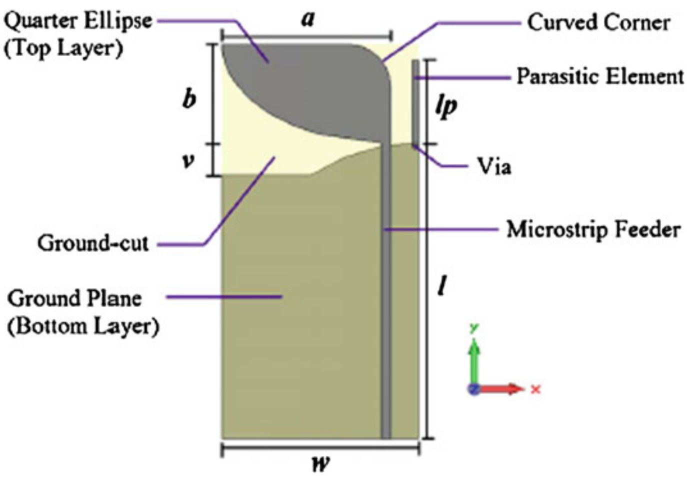


Fig. 1 Configuration of initial antenna [9]

is located at the top layer of 1.6 mm FR4 substrate with dielectric permittivity of 4.3. The radiator is fed by a microstrip line of width $w_f = 3$ mm to achieve the 50 Ω characteristic impedance feedline. The initial design positioned the feed point of the microstrip line towards the right end of the substrate. The antenna radiator is supported with w = 60 mm × l = 90 mm ground plane at the bottom layer of the substrate. An open circuit microstrip parasitic element is positioned parallel to the right edge of the main radiator with dimension of $l_p = 24$ mm and $w_p = 2$ mm. The lower end is connected to the ground with a shorting pin. The ground plane designed begin with a cut, v of 10 mm combined with a 10 mm radius circular structure extended to the left end of the substrate width as shown in Fig. 1.

2.2 *Proposed Antenna*

The focus of the proposed antenna modification is to improve the bandwidth of the resonance and to achieve miniaturization. First, corrugated slots are introduced to the original quarter ellipse radiator in the manner shown in Fig. 2. The introduction of these corrugated slots reduces the radiator size while maintaining the resonance performance as discussed in [10, 11]. The slot depth dimension is set to be less than a quarter of the effective wavelength at the lowest operating frequency. The corrugated slots present an inductive reactance, as well as lengthened the current path on the radiator. The initial width of the corrugations follows the guideline from [10]. The final dimensions are obtained from parametric optimization in CST Microwave studio. The second modification introduced in this proposed antenna is

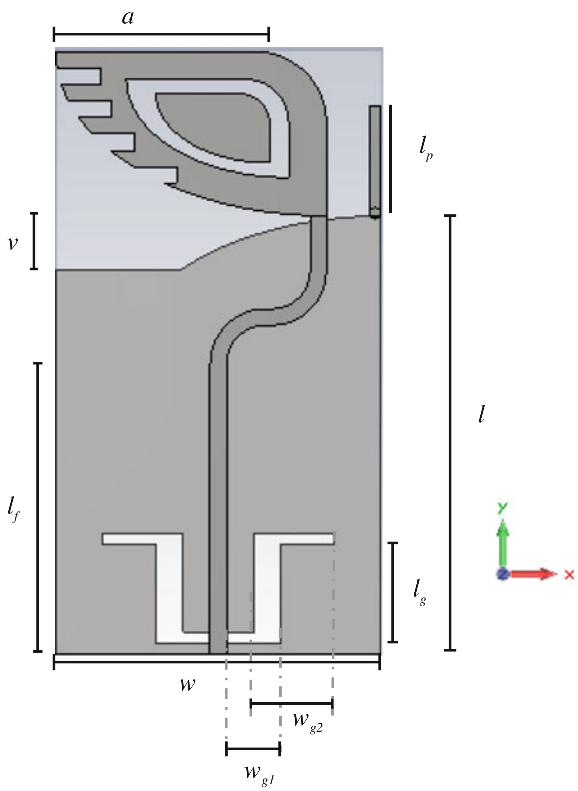


Fig. 2 Configuration of the proposed antenna design

the inclusion of a slot in the middle of the radiator. It is shown in [7] that an introduction of slot at the radiating element excites a new frequency resonant. The effect of this slot will be discussed in the later section. In order to further enhance the impedance bandwidth of the proposed antenna, a symmetrical stair-shaped ground slot is introduced as shown in Fig. 2. To apply this, the microstrip feedline of the radiator is initially bent into S shape with the aid of two 10 mm radius quarter circulars. With this modification, the feed point is moved to the middle of the substrate lower end. This technique are explained in [12], where the author discussed the possibilities of bandwidth enhancement using ground slots, while in [13], the coupling of open ended ground slot with feedline is capable to generate a new frequency resonant. Therefore, this proposed slot design exploits the ground slots promising advantage to enhance the bandwidth. On top of that, the proposed antenna also achieves length reduction by 10 mm. From the return loss analysis, this reduction has a very minimal effect at lower resonant frequency. Therefore, the changes in losses and efficiency of antenna due to this reduction can be ignored. The final dimension of antenna structure is summarized in Table 1.

Table 1 Proposed antenna parameter

Parameter	Value (mm)
a	37
w	60
l	80
v	10
l_p	20
l_g	20
w_{g1}	10
w_{g2}	10

2.3 Parametric Analysis

Using CST Microwave Studio, the analyses done to the antenna performance are based on ideal and loss free condition. The first analysis performed is the effect of corrugated slot to antenna return loss. Researched done by [10, 11] concluded that corrugated radiator reduce surface area and maintained antenna characteristics. As shown from the results in Fig. 3, the return loss and bandwidth are slightly better near the high frequency range after the corrugated slot is adapted to the antenna radiator.

Next, analysis is performed by studying the behavior of the slot in the middle of the radiator. In literatures, most antenna designs with different types of slot in the radiator show improvement on the impedance bandwidth. For example, according to [7], the application of U-Slot and Y-Slot achieve a dual band results. Furthermore, [8] shows that the slot excites a new resonant frequency as it provide a new current path. Hence, by using this technique, the overall antenna's return loss performance throughout the frequency range can be enhanced significantly. Figure 4 shows the effect of a slot to the original return loss. The slot cut at main

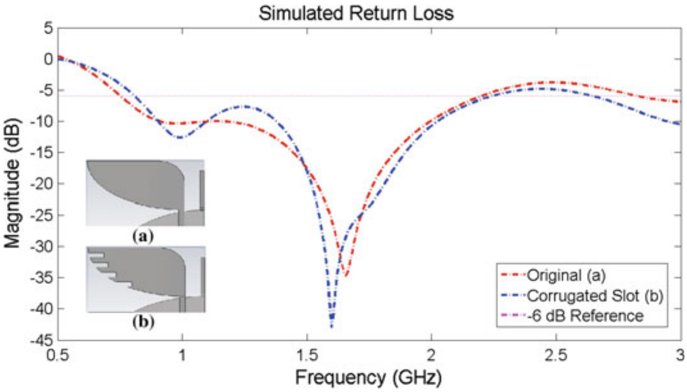


Fig. 3 Simulated return loss for effect of corrugated slot on radiator

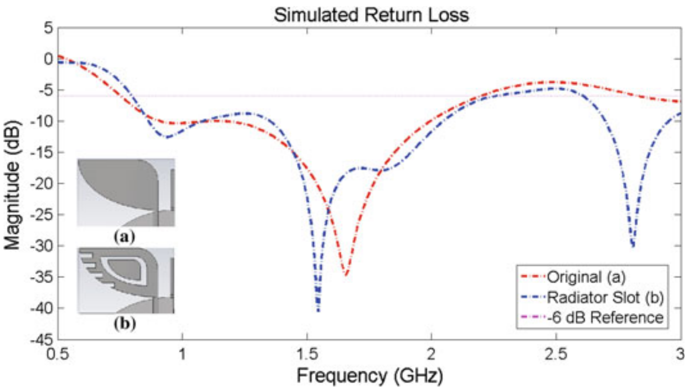


Fig. 4 Simulated return loss for effect of slot on radiator

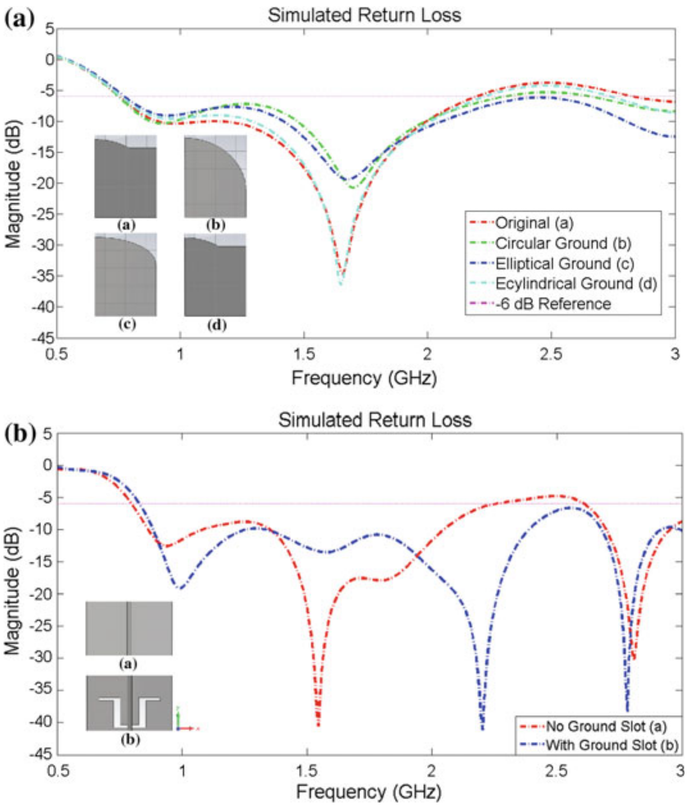


Fig. 5 **a** Simulated return loss for different types of ground plane shape. **b** Simulated return loss for effects of ground slot

radiator improves the resonant frequency at the lower band and exhibit a new resonant frequency at the higher bandwidth.

Another analysis that has been done is the ground plane modification that improves overall bandwidth and return loss of the proposed antenna.

As conclude in [12–14], there are some considerable effects to antenna performance with modification of its ground plane. In this case, Fig. 5 shows the simulated return loss due to effects of ground plane modification. Figure 5a compares different ground plane shape while Fig. 5b shows the significant of ground slot effect. The elliptical + bricks ground shows a progress response and hence has been chosen as the ground plane of the proposed antenna. Adding a symmetrical slot at the ground plane improves the resonant at WLAN and WiMAX bands. The final bandwidth at 6 dB reference improved from 2.3 to 2.6 GHz.

3 Results and Discussion

The prototype of the proposed antenna was built to verify the performance of the CST MWS simulation results. The fabricated prototype of the antenna is shown in Fig. 6. The comparison between original antenna, simulated results, and fabricated

Fig. 6 Fabricated prototype of the proposed antenna

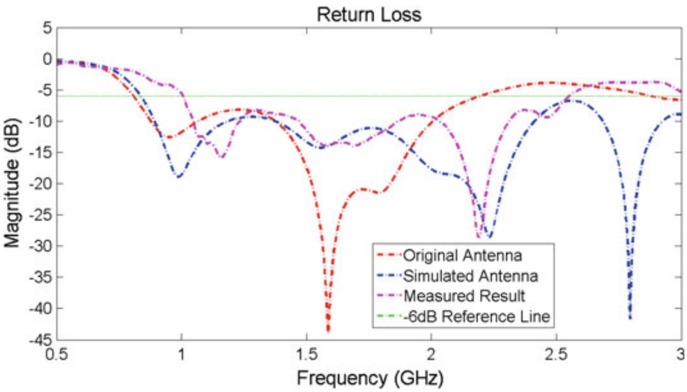


Fig. 7 Comparison between measured and simulated return loss of the proposed antenna with original antenna

prototype antenna in terms of their return loss response are plot together in Fig. 7. The 6 dB return loss reference was used prior to the toleration made in manufacturer's specification [15–17]. The presented return loss response shows that the simulated antenna is workable for most of the WWAN services includes AMPS, GSM, GPS, DCS, PCS, UMTS, WLAN and WiMAX. However for the fabricated antenna, measured results shows a small disagreement at the lower frequency and above 2.5 GHz range. This discrepancy could be caused by the imperfection of the prototype antenna fabrication. Also, the simulation results obtained are based on ideal and loss free case. Measurement condition and surrounding effect could also affect the end results. The far-field radiation pattern of the proposed antenna is shown in Fig. 8. It shows pattern at popular wireless bands including 900 MHz, 1.57, 1.8, 2.0, and 2.45 GHz services. The simulated gain of the antenna is shown in Fig. 9. The graph shows the antenna gain for proposed antenna ranges from 1.3 to 5 dBi.

At low frequency, the antenna exhibit nearly omnidirectional characteristics while at higher frequency, the antenna radiate at particular direction especially on H-plane. This is shown in radiation pattern in Fig. 8.

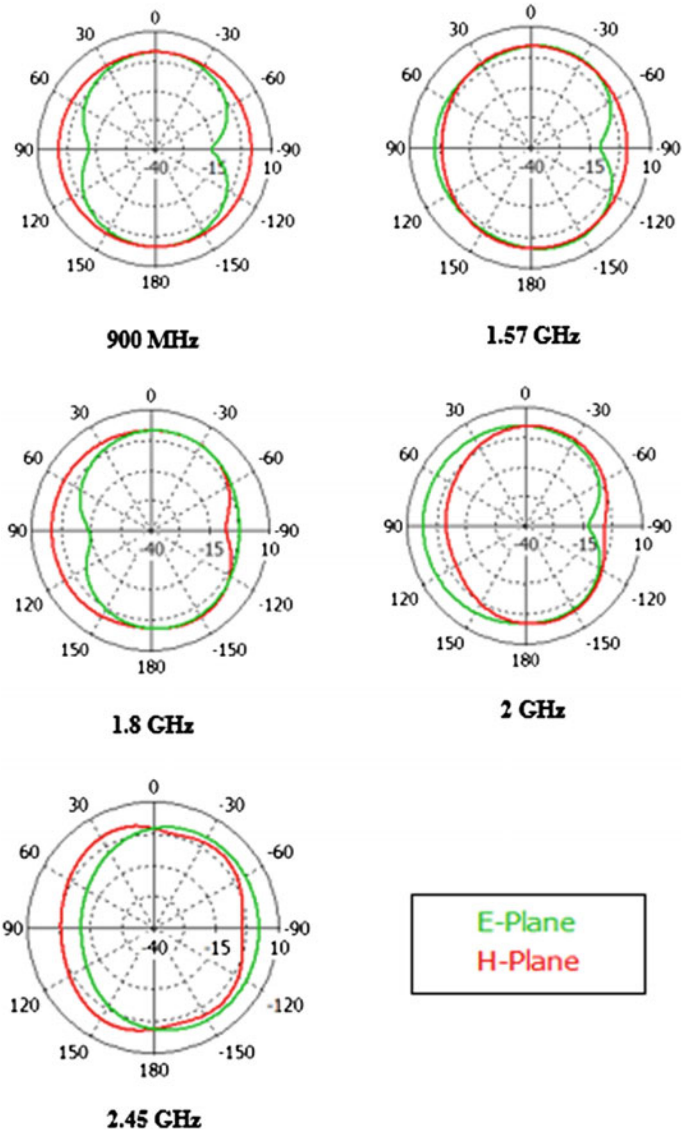


Fig. 8 Simulated 2D radiation pattern of the proposed wideband antenna at several frequencies used by modern wireless services

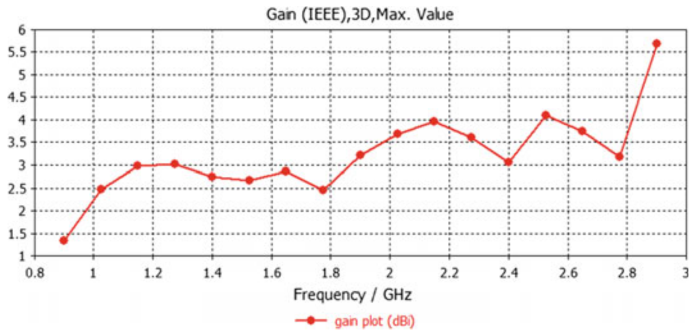


Fig. 9 Simulated antenna gain (dBi)

4 Conclusion

The design of wideband monopole antenna in the form of quarter-elliptical radiator for use in WWAN services has been presented. The proposed antenna operates from 850 MHz to 3 GHz at 6-dB return loss reference. It produces almost omnidirectional radiation patterns and has the gain between 1 and 5.5 dBi. With its small, compact size and good bandwidth coverage, it can be a good candidate for use in portable RF transceiver to access modern WWAN services.

Acknowledgement The author would like to acknowledge the Ministry of Science and Technology (MOSTI), Malaysia through EScience Research Grant Project No. 04-01-01-SF0620 that support this work.

References

1. Moradhesari, A., Moosazadeh, M., Esmati, Z.: Band-notched UWB planar monopole antenna using slotted conductor-backed plane. Microwave Opt. Technol. Lett. **54**(10), 2237–2241 (2012)
2. Majidzadeh, M., Ghobadi, C., Nourinia, J., Poorahmadazar, J.: Small monopole antenna with modified slot ground plane for UWB applications. In: 20th Iranian Conference on Electrical Engineering, May 2012, pp. 1078–1082
3. Shetawy, S.A., Abdallah, P.E.A., Darwish, P.: Slotted ground plane of rectangular patch microstrip antenna with enhanced bandwidth and size reduction, pp. 286–292 (2008)
4. Karimian, R., Tadayon, H.: Multiband MIMO antenna system with parasitic elements for WLAN and WiMAX application. Int. J. Antennas Propag. **2013**, 1–7 (2013)
5. Tseng, C.-F., Huang, C.-L., Hsu, C.-H.: Microstrip-fed monopole antenna with a shorted parasitic element for wideband application. Prog. Electromagnet. Res. Lett. **7**, 115–125 (2009)
6. Cao, Y., Yuan, B., Wang, G.: A compact multiband open-ended slot antenna for mobile handsets. IEEE Antennas Wirel. Propag. Lett. **10**, 911–914 (2011)
7. Aj, U.P., Raj, S.M.G., Mathew, P.K.: Performance analysis of U-Slot and V-Slot patch antenna for wireless applications, pp. 6–9

8. Zhao, F., Tang, H., Zhao, C., Gao, X., Zhuo, P., Zhang, F.: Design of novel dual band-notched disk monopole antennas. In: Proceeding of 2012 International Workshop on Microwave Millimeter Wave Circuits System Technology (MMWCST 2012), pp. 226–233 (2012)
9. Bialkowski, M.E., Razali, A.R., Boldaji, A.: Design of an ultrawideband monopole antenna for portable radio transceiver. IEEE Antennas Wirel. Propag. Lett. **9**, 554–557 (2010)
10. Abbosh, A.M., Member, S.: Miniaturized microstrip-fed tapered-slot antenna with ultrawideband performance. IEEE Antennas Wirel. Propag. Lett. **8**, 690–692 (2009)
11. Abbosh, A.M., Member, S.: Miniaturization of planar ultrawideband antenna via corrugation. IEEE Antennas Wirel. Propag. Lett. **7**, 685–688 (2009)
12. Rafiee, M., Ain, M.F., Aftanasar, M.S.: A new ultra-wideband antenna with unique ground plane shape. Progr. Electromagnet. Res. Lett. **35**, 165–179 (2012)
13. Razali, A.R., Abbosh, A.M., Antoniades, M.A.: Compact Planar Multiband Antennas for Mobile Applications. INTECH Open Access Publisher, Croatia
14. Razali, A.R., Bialkowski, M.E., Boldaji, A., Tsai, F.-C.E.: Investigations on a multiband microstrip-fed inverted-F antenna with ground slots. Microwave Opt. Technol. Lett. **52**(8), 1713–1719 (2010)
15. Wong, K.L., Lee, L.C.: Multiband printed monopole slot antenna for WWAN operation in the laptop computer. IEEE Trans. Antennas Propag. **57**(2), 324–330 (2009)
16. Chi, Y.W., Wong, K.L.: Quarter-wavelength printed loop antenna with an internal printed matching circuit for GSM/DCS/PCS/UMTS operation in the mobile phone. IEEE Trans. Antennas Propag. **57**(September), 2541–2547 (2009)
17. Wong, K.-L.: Planar Antennas for Wireless Communications. Wiley, Hoboken (2003)

Smart Meters in Smart Cities: An Application of DLMS-COSEM on 169 MHz WM-Bus

Giuseppe Di Leo, Consolatina Liguori, Vincenzo Paciello, Antonio Pietrosanto and Paolo Sommella

Abstract Advanced Metering Infrastructures (AMI) are going to represent the backbone of all the Smart City projects where metering and other public services are supposed to be smart. The scenario proposed by OPEN METER project is taking shape in some urban areas where smart meters, concentrators and central access systems have been experiencing. In this paper some topics concerning the use of DLMS-COSEM (Device Language Message specification-Companion Specification for Energy Metering) protocol on a 169 MHz gateway network based on WM-bus for collecting and manage data coming from smart meters and other electronic devices are discussed by the authors. In particular, an innovative solution for water metering and the architecture of a data Central Access System, are described in detail.

1 Introduction

Urban areas are today a privileged ground of challenge for those who want to develop services aimed at improving the quality of life based on the use of innovative technologies. One of the aspects, which an important game is played on, is surely made by energy consumptions. Mode of reading for billing and possibility of monitoring offered to customers greatly influence the perception of the service quality that the utility companies can grant [1]. The perceived quality, however, adds further costs thus determining the final cost of the service. To date, the costs that the Authorities recognize the companies are often determined by assuming that readings are to be committed to staffs who go there to read the data displayed on the front panel of an analog mechanical meter. This modality inevitably causes difficulty in making regular readings, with consequent recourse to advance and balance

G. Di Leo · C. Liguori · V. Paciello · A. Pietrosanto (✉) · P. Sommella
Department of Industrial Engineering, University of Salerno, Salerno, Italy
e-mail: apietrosanto@unisa.it

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_62

in charging consumption, and low efficiency of the service itself. Also detachments of users from the grid for emergency or lawsuits are difficult or, at times, impossible [2, 3]. This is the reason why Smart Metering and Advanced Metering Infrastructures (AMI) for full-scale bi-directional data communication are being more and more widespread. During last years, a lot of Automatic Metering Infrastructures have been experiencing, based on different physical channels and available to different kind of utilities (electric power, gas, water) [4, 5].

As for water and gas metering the problem is twofold and concerns both meters and infrastructure. In the most of installations meters are still mechanical and then output cannot be transmitted. New electronic smart meters are going to be placed on the market but they are battery devices and data connection is wireless. Consequently, battery life and reachability represent the major problems of smart metering in these fields. Moreover, no dedicated network for measurements and commands exists, therefore ad hoc solutions for infrastructures have to be experienced. In the electric energy field the problem of automatic metering had been faced since some years. The replacement of electromechanical meters with numeric counters and the indispensable connection of the meter to the power grid allowed the development of a communication system that uses the last mile of the grid (PowerLineCommunication). As a result, since for power meters connected in part to a wired network no energy consumption problem and limited problems of reachability arose, this solution has been widespread. Particularly relevant in this field has been the OPEN Meter project [6], whose goal is to specify a comprehensive set of open and public standards for AMI. The 169 MHz wM-Bus is suggested for water, gas and electric power smart meters. Some other devices, “concentrators” should be spread in the urban area to interface 169 MHz smart meters with a central unit through the GSM-GPRS network. Among the protocols suggested for the communication between smart meters and central unit, the DLMS-COSEM [7] seems to be the solution accepted by the most of the stack holders. Based on data models and functional processes in an object oriented notion (OBject Identification System, OBIS), the DLMS-COSEM specifies data exchange procedures and access services for the smart devices at the application layer, by allowing interoperability between devices from different vendors. Thanks to the OPEN Meter also national regulations have accepted the recommendations about infrastructure, systems and communication protocols for data storage [8].

In the gas sector, there was a deep innovation of meters to turn them from volumetric flow meters for measuring mass flow [9]. This redesign has prompted the introduction of electronics and allowed easy integration of a wireless communication module with the metrological part. As for gas meter, manufacturers are already releasing models featured with 169 MHz radio modules implementing the DLMS-COSEM protocol, but very few network infrastructures at 169 MHz have been installed until now.

In water meter market, during last few years new principles have been coming out which are based on electromagnetic induction or ultrasounds and allow electronic smart meters to be made, but they are not available yet. Solutions instead that adopt traditional measurement principles featured with electronic readers (reed contacts) and transmitters are available for a long time, but transmitters exploit physical channels mainly designed for integration in Home Area Networks, thought to connect together smart meters and smart devices within the home premises, and not for the direct integration of a meter in AMI. This means that wireless technologies like ZigBee and wireless BACnet, or wired busses like HomePlug [10] are generally adopted. This article presents an AMI that fully implements all the specifications of the project OPEN Meter which have been transposed in the national regulations. The proposed solution is based on a 169 MHz network which allows the bi-directional connection of water, gas and power meters to a single Central Access System. In the work is described the organization of the network relatively to the problems of: (i) occupation of the channel by the various users; (ii) distribution of the gateways within the urban area; (iii) design and construction of the Central Access System (SAC); (iv) network management by the "third operator". In particular, an original solution for the realization of a water smart meter for 169 MHz networks fully compatible with the DLMS-COSEM protocol is described. Finally the former results are given of a pilot project that is being carried out in the town of Salerno thanks to funds given by the Italian Gas Authority to facilitate the experiments of 169 MHz infrastructures.

2 Advanced Metering Infrastructure in Smart City

AMI must be thought as a system that measures, collects and makes available to utility companies data coming from electricity meters, gas meters and water meters, through various communication media on request or on a pre-defined schedule. In this architecture, measurement data are forwarded to a central unit, featured with suitable web interfaces which allow eventual control strategy to be implemented in and/or pricing data and commands to be sent to meters by utility.

The communication architecture for the proposed AMI is based on a hierarchical topology, as shown in Fig. 1: sensor nodes (WSN nodes) are connected to master nodes, acting as data concentrator unit/(DCU/GW), which in turn send information to a centralized control and monitoring system, where data are stored and processed. More in detail, the leaf nodes in the hierarchical network should be able to perform short or medium range radio transmissions at a low power consumption, while master nodes are equipped with long-range transmission capacity. Each master node is responsible for the collection and organization of data generated from several peripheral nodes, that are to be delivered to the central monitoring unit. Here, data are processed in order to identify and locate possible faults in the distribution network, in a real time fashion.

The choice of the individual components, especially for the numerous remote nodes, must be made with the aim to reduce maintenance operations to a minimum. The architecture of the monitoring network will provide master nodes equipped with GSM/GPRS modem to communicate with the central unit, over typical distances of a few kilometers. To this aim, master nodes are usually powered by the

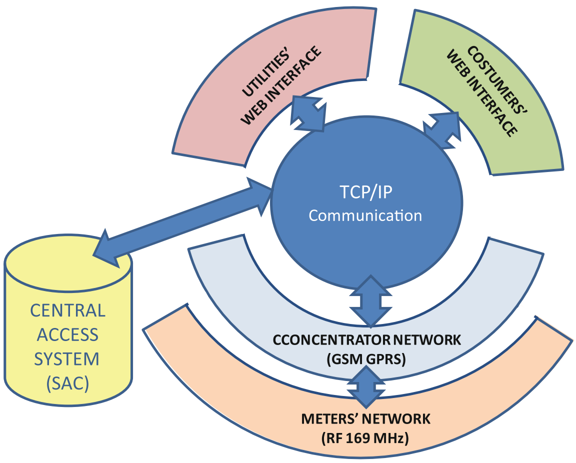


Fig. 1 Block scheme of the AMI

grid, or from solar cells. Short/medium range communications, as performed by the leaf sensor nodes, must require the minimum power consumption, usually provided by batteries. For these reasons, the wM-Bus standards is assumed as a good tradeoff among the power requirements due to radio transmission, and the coverage range that allows to increase the number of sensor nodes to locate on the multi-service grid.

3 The RF 169 MHz Smart Meter Network

The wireless M-Bus (EN 13757-4:2013) has emerged as the recommended standard for metering scenario in European market because it represents a good tradeoff among the power requirements due to radio transmission, and the coverage range that allows to increase the number of sensor nodes to locate on the multi-service grid. In details, the standard specifies the communication protocol (at both the physical and data link layer of the ISO/OSI model) between low-cost, batteries operated meters and a stationary data concentrator or mobile reader. wM-Bus transceivers require low energy thanks to a low-overhead protocol, transmission-only modes (which do not require an idle receive phase) and long-range sub-GHz transmission bands. While the first document EN 13757-4:2005 prescribed the adoption of 868 MHz ISM and 468 MHz bands (using FSK modulation with data rate ranging up to 100 kbit/s), the last version EN 13757-4:2013 added new transmission modes at 169 MHz (GFSK, GMSK and 4-GFSK modulations) with lower data rates (up to 19.2 kbit/s). The lower 169 MHz frequency band enables longer transmission range due to the inherently lower path losses, while the reduced data rates enable higher sensitivity for the receiver, allowing a reduction of the transmission power at the transmitter or a longer transmission range for the same transmission power [11].

For these reasons, only the N2a-f modes for stationary readout (169 MHz frequency band with 12.5 kHz channel separation and data rate ranging from 2.4 up to 4.8 kbit/s) have been included in the national specification UNI/TS 11291-11-4:2014 for gas metering and considered in the following. In these bidirectional communication modes, the meter starts the transmission to the concentrator, being the latter always in reception. The transmission session cannot be started by the concentrator since the meters mostly remain in sleep mode to save battery life and open the reception window only after the transmission of the first packet, waiting for the concentrator command or request. Then, the meter switches to the Frequent Access Cycle (FAC): it shall repeat the last message periodically with a FAC-Transmission delay t_{TxD} until the next request or command is received from the concentrator. The FAC lasts until the FAC-Timeout after the last successful reception of a command/request from the concentrator. The FAC can be earlier stopped by the concentrator, which sends a suitable message at the end of the communication.

The transmission mechanism assured by the wM-Bus mode N have led the authors to develop a pilot Smart Meter Network (including 2500 sensors/actuators), which is able to support the multiservice applications as specified in Table 1, where the corresponding communication requirements are reported, both on uplink (from the endpoint to the network) and on downlink (network to the sensor).

Further to the traditional gas, water and electricity metering services for billing purposes, the identified applications also include the parking management (based on suitable sensors to be installed on cars and used within the municipality by registered users) and the health alarming (through suitable indoor sensors activated by ISM band radio-buttons available to older/disable people).

For every application, the reported dataset includes the payload both for wM-Bus and DLMS-COSEM protocols. Suitable OBIS codes have been introduced for data information associated to water metering, parking management and health alarming. As previously introduced, a visual sensor node (shown in Fig. 2a) is

Table 1 Summary of application requirements

Application	End-points	Uplink			Downlink		
		Periodicity	Dataset (bytes)	Daily load (bytes)	Periodicity	Dataset (bytes)	Daily load (bytes)
Gas metering	1.000	3 times/day	150	450	1 time/month	50	–
Water metering	1.200	1 time/month	10k	330	1 time/month	4, 5k	150
Parking management	200	6 times/h	50	300	6 times/h	50	3
Health alarm	80	Random	50	–	Random	50	–
Electricity metering	20	3 times/day	150	450	1 time/month	50	–

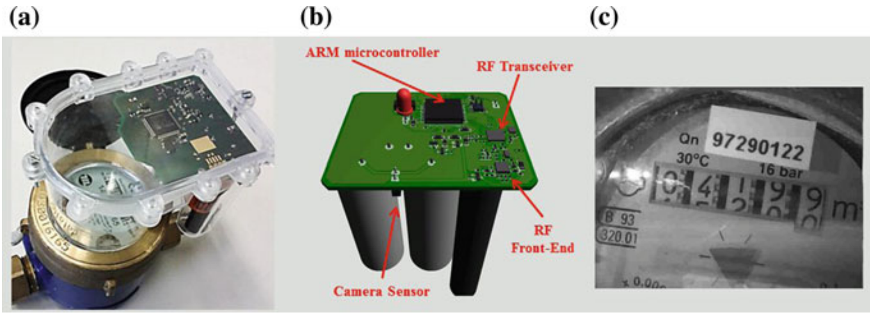


Fig. 2 Water smart meter with 169 MHz RF wM-bus communication module

adopted for water metering, which is monthly requested for the image acquisition and transmission on the basis of the last snapshot date and sensor status (daily dataset). About the parking management, suitable OBIS codes for the cumulated parking time (from the subscription date), the user code and the associated gateway are transmitted during the service requested by the car owner, whereas OBIS codes for user identification and SMS destination phone number are included in the transmitted dataset for the health alarming.

About the hardware implementation of the RF 169 MHz Smart Meter Network, except for the gas meters (manufactured by two different companies according to the national market specifications) multiple prototypes have been developed for the water and electricity metering as well as the parking management and the health alarming, which all share the common platform for radio transmission capabilities. More in details, the smart meters and sensors include: (i) the high performance ARM®Cortex™-M3 32-bit RISC core as Microcontroller; (ii) the sensing element; (iii) the Texas Instruments CC1120 chip as RF Transceiver, (iv) the SKY65367-11chip as RF front-end module. The selected family of microcontrollers pays great attention to power saving that's why there are three operation modes that are conceived to economize battery consumption these are: Sleep, Stop and Standby modes. Analogously the other modules work at very low power and low voltage so that they can match the requirement for low power operation of the single node.

As an example Fig. 2 shows the circuit schematics of the RF 169 node (including the micro-camera as sensing element) when applied to the traditional water meter for retrieving the image of the front analog panel.

A slight difference exists for the electricity metering: the prototype unit introduced in [12] is completed with an add-on module which makes the meter a wM-Bus Repeater according to the standard EN 13757-5:2011: it is able to collect the measurement data from the water, gas meter and other end-points by adopting the N2a-f channels (both at 2.4 and 4.8 kbps) and extend the transmission range of the network by using the prescribed N2g mode (at 19.2 kbps).

4 The DCU Network

Data concentrator units are essential devices in the AMI. The principal function is to gather data from several smart meters, compress gathered data and then pass on to the central system. This reduces the number of application association establishment among smart meters and the central system. If the AMI has no DCUs, millions of smart meters have to communicate with the central system directly. This is practically impossible due to the number of transactions, the long distances and cost incurred. The data collected from smart meters involve energy usages, important events and so on. The control system analyses the collected data for billing and quality of service. Conversely, the central system sends commands, firmware upgrades and billing methods to smart meters via DCUs.

A prototype of DCU for multi-service AMI has been developed and is reported in Fig. 3. The hardware architecture includes: (i) a BeagleBone Black board able to processing information at application level (DCU functionalities) and (ii) as many ARM microcontrollers as the different services (gas, electricity, water, parking and health assistance) are intended to be simultaneously managed by the concentrator. More in details, at DLMS-COSEM level, DCU behaviors both as a client (with respect to the smart meters) and server (with respect to the central system), also providing a transparent end-to-end communication. These concentrator functionalities are implemented at the meter side by the same hardware devices as previously introduced for the water meter, whereas, at the other side, the DLMS-COSEM communication profile over TCP/IP (developed for GPRS transmission within a Public Network) is implemented by the BeagleBone Black board. Considering 10 concentrators for the whole system and considering the total amount of the meters (as reported in Table 1). The mean number of meters per each concentrator is 250. Considering the characteristic of the 169 MHz radio-network and to avoid the

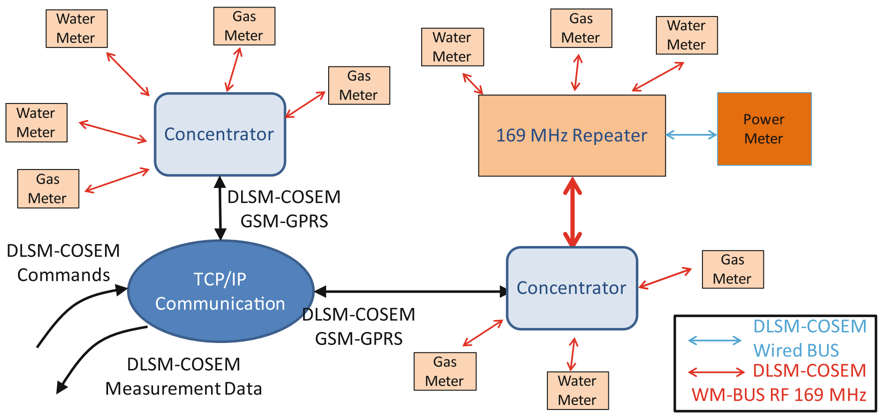


Fig. 3 Concentrator and repeater collect measurement data from meter

packet collision (multiple channel access), a time division scheme has to be considered. Each meter has the RTC clock and it is possible to program an awake time for different each meter.

5 The Central Access System (SAC)

The function of the Central Access System in the remote reading system is the networking management: all utilities that need to communicate with the meters must do so exclusively through the central access system that is the only component of the infrastructure to delegate that function. The implemented Central Access System (depicted in the scheme of Fig. 4) consists of three software modules: (i) a module developed in JAVA that implements communication with concentrators and/or meters using the protocol DLMS-COSEM and the mobile (GPRS/UMTS) network; (ii) a web application developed in PHP that allows users and utilities to send commands to the meters and access the stored data; (iii) a MySQL relational database that records all data uploaded by the meters. More in detail, the Business Logic of the web application is responsible both for the automatic networking processes (such as clock synchronization, meter enrollment, regulation of the power transmission for battery optimization, periodic data recording) and the services requested by the utilities and end-users through the Web Server. The meta-commands coming from the utility (such as orders for Tariff Plan changing, new metering point associations, meter valve closing) are translated in the corresponding OBIS codes and transmitted through the DLMS-COSEM messaging. Analogously, the end customers may access the measurement data (request for consume profiles) by adopting a user-friendly web interface, which makes the underlying communication protocols and objects transparent. About the security features, the Central Access System implements mechanisms for authentication to allow access to the concentrators and

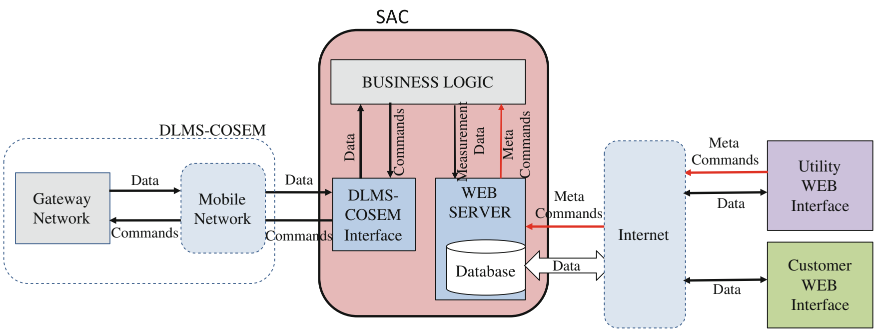


Fig. 4 Central access system and utilities' web interface

utilities. The data that is exchanged between head-end system and concentrators or utilities are encrypted with AES-GCM 128 bit; it is ideal for protecting packetized data, because it has minimum latency and minimum operation overhead.

6 Experimental Results

In order to evaluate the performance of the proposed multiservice smart metering scenario, a preliminary investigation has been carried out about the diverse energy needs for the visual sensor node at the different operation modes. Figure 5 reports the average current absorption figures for the most energy demanding prototype (visual node equipped with the STM32F207 ARM microcontroller, TI CC1120 transceiver and camera sensor). Task energy estimates are based on measured values for the microcontroller operating at 3.6 V and the CC1120 transceiver operating in wM-Bus *N2a* mode at maximum transfer power. The expected life-time of the visual node (battery discharge) reveals to be greater than 13 years, when the smart meter is configured as follows: (i) 3 daily random push sessions (data rate = 4.8 kbs and wM-Bus frame length = 180 byte); (ii) a monthly image acquisition and transmission (FAC session); and (iii) 2 lithium batteries for a total capacity equal to 7000 mAh.

About the communication capabilities of the visual nodes, a comparison has been carried out between simulated and measured data taken into account the RSSI received by the DCU/GW when adopted in urban scenario with medium inhabitant density. A popular wireless sensor network simulator (Castalia, which exploits the facilities offered by Omnet++) has been exploited to model both the electrical characteristics of the sensor node (transmission power, antenna size and pattern, functional states of RF Transceiver) and the physical channel of the radio communication. More in details, two well-known models (LogNormal and Okumura-Hata) have been included in the analysis by varying the corresponding parameters in the Path Loss formulas (such as the path loss exponent in dependence of the propagation environment and the fading factor). A summary of the results is reported in Fig. 6, which highlights the inefficiency of the models proposed in literature (which tend to underestimates the path loss) and the stringent need for a

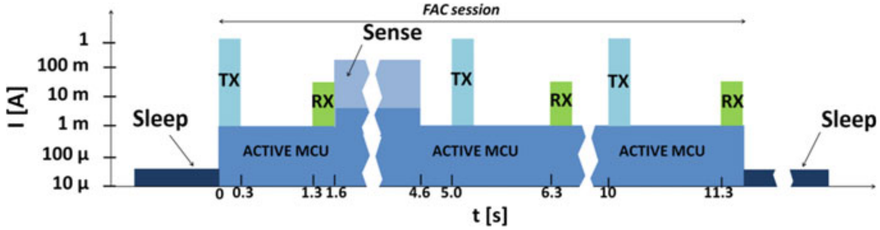


Fig. 5 Electrical current absorption of the visual water meter

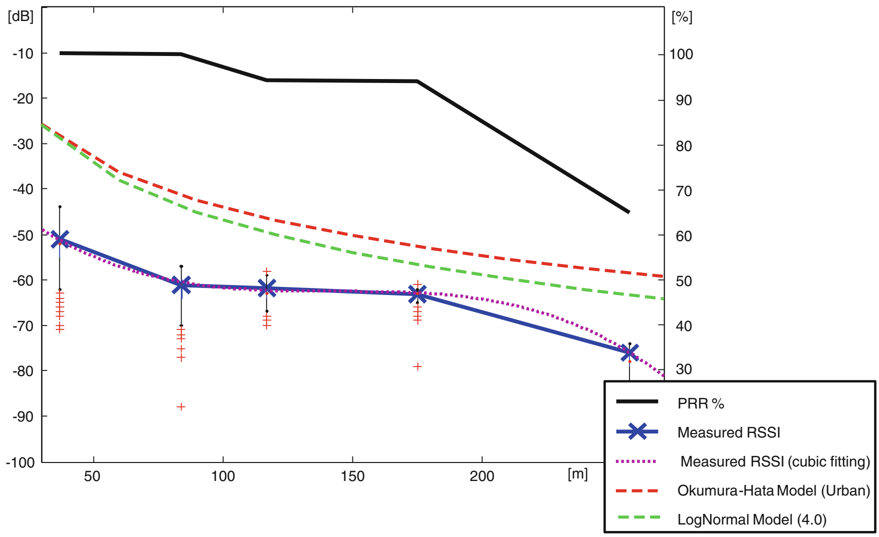


Fig. 6 Communication distance and quality of the wM-bus network in urban scenario

synthesis and experimental calibration of ad hoc model about the wM-Bus communication. In Fig. 6 is also reported the observed Packet Received Ratio (PRR%) as varying as the measured RSSI, when a wM-Bus frame length equal to 125 bytes is transmitted by the visual sensor node at the maximum allowed power (the typical

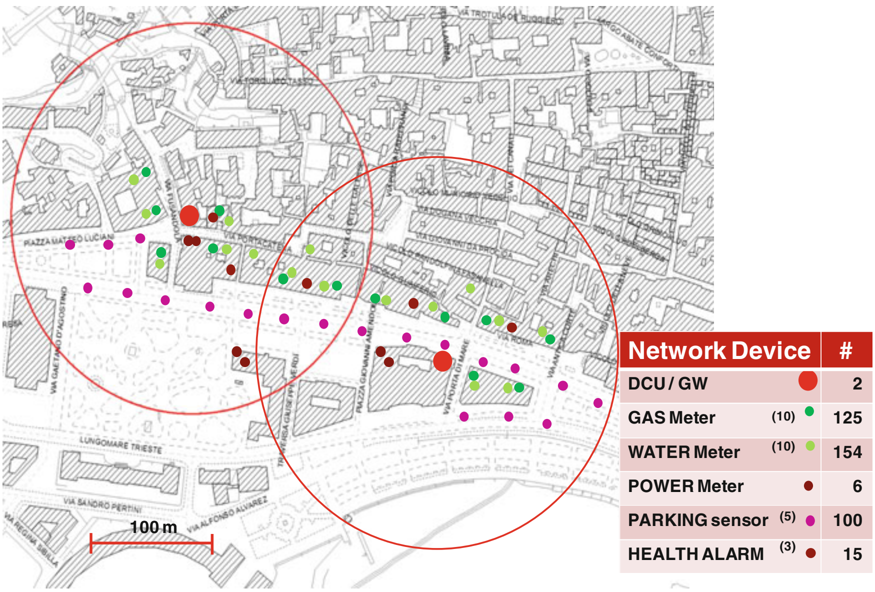


Fig. 7 Installation plan of the smart network in urban scenario (Salerno downtown)

DLMS-COSEM payload for the proposed image block is taken into account). Lead by the measurement results, in order to assure a reliable communication for the most demanding application (by maintaining the need for packet re-transmission as low as possible and optimizing the life-time of the visual nodes) the selection of the installation sites of both the remote sensors and the DCU/GW for the Pilot project has been carried out by considering the maximum communication distance equal to 150 m. As an example, in Fig. 7 is reported a scheme of the installation plan for an urban area within the historical downtown in the city of Salerno, where two DCUs are supposed to assure the data retrieve by 400 remote sensors.

7 Conclusions

The AMI presented in the paper has demonstrated that the DLMS-COSEM protocol leaves very few points of ambiguity if its implementation is rigorous and complete, even though the number of OBIS implemented to grant the maximum interoperability among smart meters is probably exorbitant respect to the actual needs of utilities. Former results concerning the in situ optimal siting of gateways confirm the complexity of the problem and its substantial dependence on environmental characteristics of urban areas more than on R.F. modules and their antenna performance. Moreover, some indication have been given on the distribution of the several services using the gateway network among channels and during the day time. The use of images in water smart metering forced the authors to think a proprietary solution to allow the implementation of DLMS-COSEM paradigms that showed good reliability in data transmission, even if the occupation of the channel is quite heavy. During the first part of the pilot project which has been carrying out in the city of Salerno the SAC designed by the authors demonstrated good functionality and ease of access. Future developments will be aimed at further improvements of the gateway nodes in the area through the use of mobile devices of investigation of the area characteristics.

References

1. Smart meters and smart meter systems: A metering industry perspective, Mar 2011. Edison Electric Institute, Washington, DC (2011)
2. Stribac, G., Gan, C.K., et al.: Benefits of advanced smart metering for demand response based control of distribution networks. In: Centre for Sustainable Electricity and Distributed Generation Imperial College, London, Apr 2010
3. Di Leo, G., Landi, M., Paciello, V., Pietrosanto, A. : Smart metering for demand side management. In: Proceedings of the IEEE Instrumentation and Measurement Technology Conference, I2MTC 2012, Graz, Austria, May 2012, pp. 1798–1803 (2012)
4. Mohassel, R.R., Fung, A., Mohammadi, F., Raahemifar, K.: A survey on advanced metering infrastructure. Electr. Power Energy Syst. **63**, pp. 473–484 (2014)

5. Rafiei1, M., Eftekhari, S.M.: A practical smart metering using combination of power line communication (PLC) and WiFi protocols. In: Proceedings of 17th Conference on Electrical Power Distribution Networks (2012)
6. OMS Group: Open Metering System Specification. OMS Group (2011)
7. DLMS User Association: DLMS-COSEM Architecture and Protocols, 7th edn. (2009)
8. Feuerhahn, S., Zillgith, M., Wittwer, C., Wietfeld, C.: Comparison of the communication protocols DLMS/COSEM, SML and IEC 61850 for smart metering applications. In: IEEE International Conference on Smart Grid Communications, SmartGridComm 2011, art. no. 6102357, pp. 410–415 (2011)
9. Henry, M.P., Clarke, D.W., Archer, N., Bowles, J., Leahy, M.J., Liu, R.P., Vignos, J., Zhou, FBa: A self-validating digital Coriolis mass-flow meter: an overview. Control Eng Pract. **8**, pp. 487–506 (2000)
10. Homeplag powerline alliance, HomePlug AV specification, Ver. 1.0.05, Oct 2006
11. Communication systems for meters and remote reading of meters. Part 1: Data exchange; English version EN13757-1 (2002)
12. Ferrigno, L., Paciello, V., Pietrosanto, A.: Visual sensors for remote metering in public networks. In: Proceedings of the IEEE Instrumentation and Measurement Technology Conference, I2MTC 2011, pp. 1006–1011 (2011)

Empirical Estimation of Multilayer Perceptron for Stock Market Indexes

J. Chandra, M. Nachamai and Anitha S. Pillai

Abstract The return on investment of stock market index is used to estimate the effectiveness of an investment in different savings schemes. To calculate Return on Investment, profit of an investment is divided by the cost of investment. The purpose of the paper is to perform empirical evaluation of various multilayer perceptron neural networks that are used for obtaining high quality prediction for Return on Investment based on stock market indexes. Many researchers have already implemented different methods to forecast stock prices, but accuracy of the stock prices are a major concern. The multilayer perceptron feed forward neural network model is implemented and compared against multilayer perceptron back propagation neural network models on various stock market indexes. The estimated values are checked against the original values of next business day to measure the actual accuracy. The uniqueness of the research is to achieve maximum accuracy in the Indian stock market indexes. The comparative analysis is done with the help of data set NSEindia historical data for Indian share market. Based on the comparative analysis, the multilayer perceptron feed forward neural network performs better prediction with higher accuracy than multilayer perceptron back propagation. A number of variations have been found by this comparative experiment to analyze the future values of the stock prices. With the experimental comparison, the multilayer perceptron feed forward neural network is able to forecast quality decision on return on investment on stock indexes with average accuracy rate as 95 % which is higher than back propagation neural network. So the results obtained by the multilayer perceptron feed forward neural networks are more satisfactory when compared to multilayer perceptron back propagation neural network.

J. Chandra (✉) · M. Nachamai
Christ University, Bangalore, India
e-mail: chandra.j@christuniversity.in

M. Nachamai
e-mail: nachamai.m@christuniversity.in

A.S. Pillai
Hindustan University, Chennai, India
e-mail: anithasp@hindustanuniv.ac.in

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_63

Keywords Return on investment (ROI) · Multi layer perceptron (MLP) · Neural network (NN) · Feed forward (FF) · Adaptive linear neuron (ADALINE) · Back propagation (BP) · Mean of magnitude relative error (MMRE)

1 Introduction

Stock market forecasting is an important task for an investor to make good returns on investment. The aim of the investor in investment society is to maximize the values on return on investment. The return on investment of stock market indexes keep changing from time to time. Therefore, it is very difficult to predict next high values of stock indexes [1]. The strategy used for selecting the stock index is one of the biggest challenges faced by the investor society. According to Indian stock market, the stock market investments are more volatile and highly perilous in the field of investment. Forecasting the stock market indexes in the investment field is considered as one of the major challenges in the investor society. A variety of statistical methods and systematic indicators have been used with different results but these methods have limitations and are not completely capable of providing accurate results. Many researchers have already implemented multilayer perceptron neural network methods to resolve the problems faced by the investor in the stock market but accuracy of the prediction is not standard [2, 3], so to set the benchmark for the prediction of stock indexes, the comparative analysis of multi layer feed forward and back propagation neural network is used. Even though many research works have happened with stock market prediction using neural network methods, when checked across the original values of the stock prices on next days, the performance rate was very low (i.e., the average accuracy rate for the MLP back propagation neural network was 63–85 %). The regression model is considered as one of the popular methods in predictive analysis. The performance of MLP neural network behaves better than regression models [4]. Many years, the MLP neural network has also been applied to different applications in financial services. Performing the stock prediction is not a simple job even for an expert, because the data set used for these predictions may be linear or nonlinear. Here, the focus is more on accurate prediction of stocks and commodity trading. In 1943, Mcculloch and Pitt developed neural network models on the basis of neurology.

The neural network model makes numerous assumptions and the working condition of neural network is based on simple neurons which are essential as binary devices. The neural network method is very flexible, an adoptive methodology that can be applied to problems as stream flow forecasting ground water, rainfall runoff modeling. Neural Network has gained a lot of interest in the last decades. The neural network methods can be found in numerous application areas such as stock forecasting, digital image processing, pattern recognition and optimization. Building MLP neural network was an attempt to understand investment behavior of human understanding, learning and cognitive process.

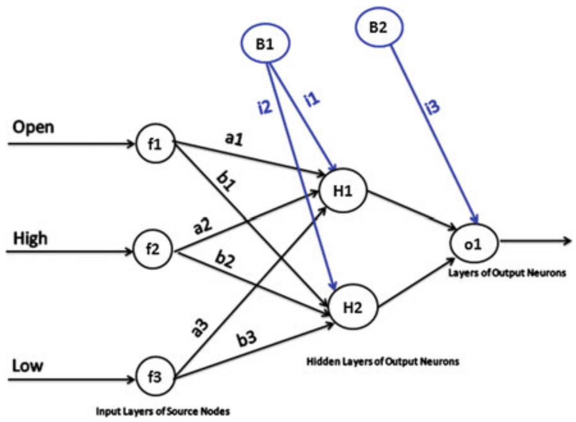


Fig. 1 MLP neural network architecture

The MLP neural network is the most widely used neural network in optimization and high quality prediction. The neural network can also be implemented in numerous application areas such as stock forecasting, digital image processing and pattern recognition. The objective of the research is to focus on building intelligent systems that are invented to replicate individual intelligent applications for forecasting. The MLP neural network is a collection of layers in elementary units called neurons, connected to one another by weighted connection. The MLP neural network structure consists of linked cells. The MLP neural network neurons obtain impulses from any input cells to other neurons. The MLP neural network is used to achieve various types of transformation of input and transmit the result of the transformation to other neurons and perform some kind of transformation to the output cells (see Fig. 1).

The MLP neural network is an intelligent method used to identify design patterns from any given data. Figure 1 shows MLP neural network architecture which contains three inputs such as f1, f2, and f3, two hidden layers as H1, H2 and one output layer O1 and B1 and B2 are intermediate nodes which are used to connect hidden nodes and output nodes. The input pattern is offered to the input layer of the neural network. The given inputs are propagated throughout the network until they achieve the output units. This forward pass produces the real or predicted output pattern.

2 Related Work

The Auto Regressive Integrated Moving Average model is one of the popular and very old methods which was implemented in SPSS, SAS to forecast the financial prices. Kohzadi et al. [5] compared ARIMA model with neural network for forecasting the commodity price. The comparison is checked with different products like wheat, cattle prices from 1956 to 1990 and the same research was repeated many times in three years. Through the comparative experiment, the author found that the neural network model performs better than ARIMA model due to absolute mean error. With experimental result, it was also found that the neural network methods perform extremely well with wheat as well with cattle but ARIMA model could do only with wheat product. According to the author's experimental work, it is concluded that neural network works better than ARIMA for price prediction.

Ravichandren et al. [6] has compared neural network and statistical methods for performance prediction of return on investment in Indian share market indexes, which has been considered as a dominant method in data analysis. The neural network result was compared with traditional statistical method for estimating the values of return on investment and it is found that the neural network method provides better accuracy than the traditional statistical method.

Kryzanoueski et al. [7] has compared the performance on back propagation neural network and multi linear regression model to forecast the stock prices. The comparison is evaluated with the help of mean squared error. Based on mean squared error, it is proved that the back propagation model performs better than multi regression model.

Saad and Wanseh [8] has compared the MLP back propagation neural network with recurrent neural network for stock market prediction based on current historical data. Tang et al. [9] has implemented the predictive analytics with time series data and checked the comparison with respect to positive, negative and neutral ROI on stock prices. Mizunu et al. [2] has predicted the Tokyo stock market using neural network and this method was able to predict with accuracy rate of 63 %. Sexton et al. [3] predicted Singapore stock market by integrating neural network and genetic algorithm which has the accuracy as 81 %. The current research focuses on improving the predictor accuracy by evaluating different MLP neural network. According to related works, the neural network model is compared with other existing methods, but the neural network methods perform extremely well in all cases. So the different MLP neural network models were taken into consideration for experimental comparison.

3 Methodology

The MLP neural network is the most popular method in the field of optimization. The MLP neural network is an intelligent method which is used to identify design patterns from any given data. In general, the MLP neural network models are in used neural network to build any complex process. The perceptron has three layers with the middle layer known as hidden layer. Another MLP ADALINE feed forward neural network was developed in 1960 by Widrow and Hoff in Stanford University [2].

3.1 MLP Back Propagation Neural Network

The MLP back propagation neural network method was proposed by Rosen Blatt in 1957. Several neural network methods were proposed for doing classification and prediction, one of them is MLP neural network which became useful with the introduction of the back propagation algorithm [10]. For any commodity trading and stock market prediction, there are a number of models in the domain such as daily, weekly and monthly investment based on performance measures. In this paper, the daily investment model is considered for the stock market. The MLP back propagation neural network is very simple and it is computationally efficient to involve its linear synaptic weights of neural network. A weakness of the MLP back propagation is, it does not always connect and can be noticeably slow [11]. The neural network results are subtracted from specified results and an error signal is produced. On the basis of the error signal produced by back propagation neural network, whereby the errors are passed back through the neural network by calculating the contribution of each hidden processing unit and deriving the corresponding adjustment needed to produce the correct output. The connection weights are then adjusted and the neural network learns from an experience. Figure 2 describes the neural network structure for back propagation neural network and feed forward neural network. The traditional back propagation network algorithm is widely used in solving many practical problems. A back propagation neural network consists of at least three layers of units: an input layer, at least one intermediate hidden layer, and an output layer [12]. Typically, units are connected in a feed-forward fashion with input units fully connected to units in the hidden layer and hidden units fully connected to units in the output layer. Because back propagation neural network is a supervised learning algorithm, the desired outputs are given as part of the training vector [13]. The actual network outputs are subtracted from the desired outputs and an error signal is produced.

This error signal is calculated on the basis of back propagation step, whereby errors are passed back through the neural network by computing the contribution of each hidden processing unit and deriving the corresponding adjustment needed to produce the correct output [14]. Once the network is trained, it will provide the desired output for any of the input patterns. From Fig. 2, the error rate as 32,842,283.932 and steps involved is 4909 and the other back propagation structure contains error rate as 57,171,881 and the steps involved for training the model is 170. From Fig. 2, it is understood that the increase of hidden layers will decrease the number of steps involved in training the back propagation networks and the different values are shown in Table 1.

The back propagation neural network method always changes the weights to cause more error. The weights represent the knowledge that the neural network contains about a specific training data. Their values will directly affect the output of the neural network. The back propagation neural network requires multiple steps to build predictive models.

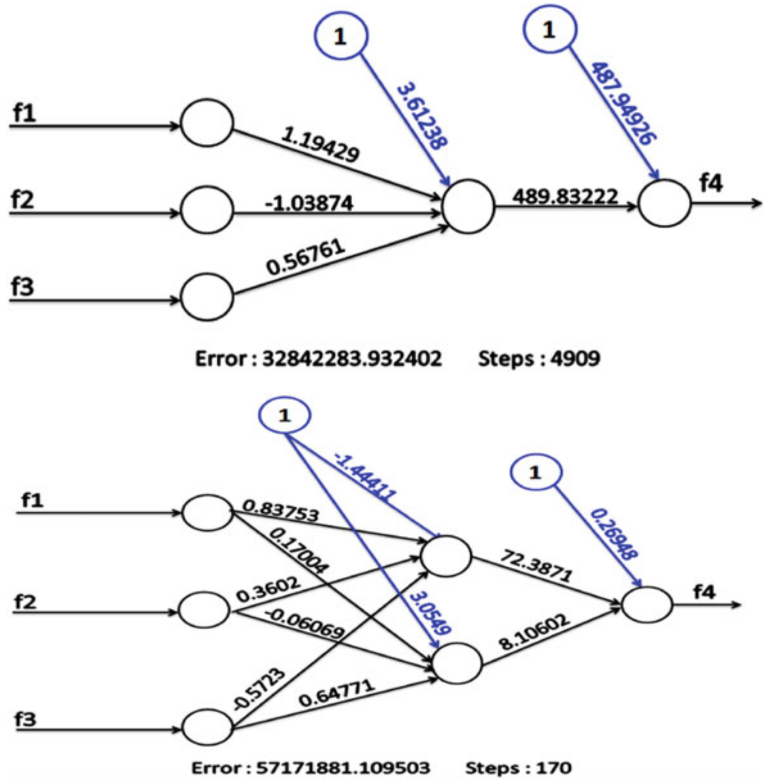


Fig. 2 MLP back propagation neural network on stock indexes

Table 1 Comparative result on different back propagation neural network

Sl no	Algorithm	Hidden layer	Error	Reached	Threshold	Steps	Error function
1	Default	1	1	3E+07	0.0048	4909	CE
2	Back propagation	5	1	6E+07	00	02	CE
3	Back Propagation	2	1	6E+07	0.0083	170	CE
4	rprop+	2	1	6E+07	0.0097	55	CE
5	Slr	2	1	6E+07	0.009	62	CE

3.2 MLP Feed Forward Neural Network

The MLP feed forward neural network, neurons are connected only in forward direction. Each layer of the neural network contains links (connections) to the next layer but there are no connections back. In general, network consists of a set of source nodes that constitute the input layer, one or more hidden layers of computation nodes and an output layer of computation nodes. Usually, most neural network has one hidden layer and it is unusual of a neural network to have more than two hidden layers. The input signals propagates through the network in a forward direction, on a layer by layer basis, these neural networks are referred as Multilayer perceptron.

The method can be used for supervised and unsupervised learning. The feed forward computes error signal as $E_i = D_i - Y_i$ where D_i is the preferred reply and Y_i is the real productivity produced by the network in response to the input X_i [15]. The current experiment investigates accuracy rate and the performance of high profit value while training MLP neural network for forecasting the stock market. The scope of the comparative experiment is to find better prediction on ROI values and also to minimize the risk in future investment. The neural network is robust enough to handle noisy data with errors and can be applied to any number of data sets. The aim of investigational end result is to realize different accessible variations of neural network model for investment predictions. The main feature of the ADALINE feed forward method is to increase the performance of predictive analytic process.

The MLP feed forward neural network consists of set of inputs, hidden and output layers of neurons. Every node is linked to every other node in the neighboring layers. The performance of the MLP feed forward neural network basically depends on the present input and weighted values. The feed forward neural network can use line active function as well as non linear processing [16]. The objective of training MLP neural network is used for producing an accumulated result whereas a collection of input is applied to the network.

Figure 3 shows the structure of ADALINE feed forward network. Here the input nodes are Open, Low, High values are considered as input nodes and predict-High

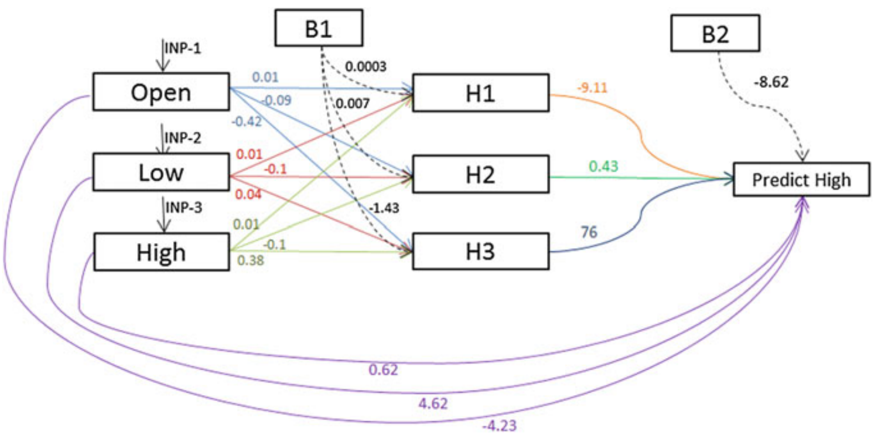


Fig. 3 MLP feed forward ADALINE neural network on stock indexes

is the predictor variable. The predictor variable is found with the help of the following equations in Sect. 3.5. The feed forward neural network model is considered as one of the best methods in MLP networks.

3.3 Data Set Description

The National Stock Exchange (NSE) is one of the popular stock market exchanges in India and it is considered as Asia's fastest exchange market. The historical data of NSE stock prices are considered for deploying the model [17]. Many stock prices like IT, CG TECH, FMCG, PSU, METAL, OILGAS and BANKEX etc., are considered for the experiment. The data set contains more than thirteen attributes such as item-no, item-name, expiry-month(dd-mm-yy), open(Rs), high(Rs), low (Rs), ltp(Rs), pcp(Rs), change(%), buyqty, sellqty, sellprice(Rs) etc., Except the name and expiry-month, the remaining attributes like 'item No', 'open', 'close' etc., are considered as numeric. The values are measured based on Indian rupees and the closing price of every stock is used for building the MLP classifier.

3.4 Feature Selection

The main objective of doing feature selection is to reduce the number of features in the given data set and the focus is to improve the relevancy and efficiency of the computation in predictor classification. The Pearson's correlation coefficient is considered as one of the efficient feature selection methods based on the classifier accuracy and it is mainly used in all data mining applications like machine learning, genetic algorithm and neural network. Through the correlated values, the attribute 'open', 'high', 'low ', 'sell prices 'values are considered as highly correlated which is taken as the input variables for training MLP neural network models.

3.5 Building MLP Neural Network

The various steps involved in building the MLP ADALINE feed forward neural network are shown below.

1. Load the optimized data set D. The data optimization is done using genetic algorithm.
2. Do the correlation based features selection.//Pearson's correlation is considered as efficient feature selection method.
3. Build the MLP NN, the linear regression formula used with the help of independent and dependent variable.

4. Compute $h_{1,}$ h_2 … h_n then calculate $\log(h_1)_,$ $\log(h_2)$ … $\log(h_n)$ where h_i is a hidden neuron.
5. Compute Predicted-high-value = $i_n + d_1h_1 + d_2h_2 + d_nh_{n.}$ Where i is the intermediate neuron and d is the output neuron.
6. Compute MRE and MMRE

$$\text{Compute Accuracy} = 1 - \text{MMRE}$$

7. Check the Accuracy; if the Accuracy is >90 % then accept the model for prediction otherwise repeat step-1 to step-6 until the accuracy value is >90 %

In MLP NN, generally the hidden layer neurons are calculated with the following equations

$$H_1 = i_1 + a_1f_1 + b_1f_2 + c_1f_3 \tag{1}$$

$$H_2 = i_2 + a_2f_1 + b_2f_2 + c_2f_3 \tag{2}$$

$$H_3 = i_3 + a_3f_1 + b_3f_2 + c_3f_3 \tag{3}$$

The equations to estimate predicted high value are as follows.

$$H_n = i_n + a_nf_1 + b_nf_2 + c_nf_3 \tag{4}$$

$$\text{Log}(h_n) = \exp(h_n)/(1 + \exp(h_n)) \tag{5}$$

$$\text{Predicted-high-value} = i_n + d_1h_1 + d_2h_2 \tag{6}$$

where f_1, f_2, f_3 are open, high, low values of input attributes from NSE data set and H_1, H_2, H_3 are hidden layers and i_1, i_2 and i_3 are intermediate neurons to hidden layers(d_n) and d_1, d_2 are intermediate neurons to output layer (O_n). The predictor variable high values of every share prices of the next business day is calculated using the Eq. (6) and the values are described in Table 2 in Sect. 4.

Table 2 Predictive result on stock indexes using MLP feed forward neural network

Item num ber	Current (Rs)	Open (Rs)	High (Rs)	Low (Rs)	Previous (Rs)	Predicted high-value	Actual high-value (5/02/2014)
1	1735.4	1731.6	1740.35	1720.25	1728.6	1757.34	1757.4498
2	8585.95	8516.55	8600.6	8494.6	8497.55	8615.22	8615.3284
3	5386.3	5371.85	5419.35	5353.7	5368.65	5449.17	5449.1929
4	17868.9	17969.9	17989.5	17816.8	17951.8	18053.5	18053.542
5	2700.65	2658.55	2720.7	2658.55	2655.7	2741.91	2741.9199

4 Experimental Result

Stock market prediction plays an important role in everyday life. Even though, many people have developed multiple methods of forecasting, enough work is still evolving around accurate prediction. So accuracy is the major concern for doing stock market forecasting using different MLP neural network. The MLP back propagation neural network and feed forward neural network methods are implemented to forecast the stock market indexes. The first model MLP back propagation neural network uses cross entropy (CE) to measure the performance of the MLP back propagation neural network. The MLP back propagation is implemented with CE error function as stopping criteria without exceeding the specified maximum number of cycles or using fixed number of epochs for learning MLP back propagation neural network. The reason for using cross entropy in MLP back propagation neural network is to accelerate the performance of back propagation easily within a short span of time. The second model uses feed forward neural network with generalized error function such as MRE and MMRE. Both neural networks were deployed and compared against the original stock prices. During the training phase, new patterns are discovered and the accuracy of the discovered new pattern is checked against the next high values of the original data. The performance of the method is also evaluated with the MRE, MMRE and accuracy rate.

Table 1 describes the training phase of back propagation neural network on stock market indices which give error rate as one in all kinds of back propagation algorithm with different hidden layers. The code for MLP feed neural network implemented using 'NNET' in R tool is given below:

```
library (nnet) proj1.nn <- nnet(High ~ Previous
+ Current + Open,data = newdata, size = 3, decay = 1e-3,
linout = T, skip = T,maxit = 1000, Hess = T) summary (proj1.nn)
proj1.nn$wts summary (proj1.nn)
```

The MLP back propagation neural network is implemented with the help of 'NeuralNet' package in R tool and the code is as follows.

```
library(neuralnet)
trading<-data.frame(trading data)
Neural<-neuralnet(High~open+current+previous,trading,
hidden=5,learningrate=0.01,algorithm="backprop",err.
fct="ce", linear.output = FALSE)
```

Through the experimental result, the predicted output values are calculated and checked with real time values. When it is checked with the real time high values of the next, day, the MLP feed forward neural network predicted values are very closer to actual values than the back propagation values, which are approximately close to the original high values. The performance of the methods are measured based on accuracy rate. The accuracy rates of MLP back propagation neural network values are between 80 and 85 % and the MLP feed forward neural network gives 99 % as the accuracy rate. The accuracy rate of feed forward ranges from 90 to 99.5. By this experimental work, it is found that MLP back propagation accuracy is less than the MLP feed forward neural network. The MLP feed forward neural network values are also checked with all the back propagation algorithms for final verification. So the MLP feed forward ADALINE neural network could be claimed as the best method for forecasting the stock market prediction.

Table 2 shows the input variables item number, Open, High, Low values of historical data of the day 4/02/2014 which were used to measure the predicted-high values. By this experiment, it was found that the predicted high values using FF NNs are very closer to the actual values of next day high values.

5 Conclusion

The ROI values of stock mark indexes are predicted based on historical data of every day closing price of NSEindia. The MLP feed forward neural network can be concluded as an efficient predictable method than MLP back propagation neural network. Although the MLP back propagation neural network is widely used in many applications, the performance and the accuracy rate is relatively bad when compared to feed forward neural network. The train-test implementation of the network was repeated on numerous cycles within a week-time model, with everyday historical stock index data. Through the corroborated experimental result, the accuracy rate for MLP feed forward gave the precision value as 99.05 %, therefore the MLP feed forward neural network can be taken as the most optimal method for making accurate prediction in stock market indexes.

References

1. Zekic, M.: Neural network applications in stock market predictions—a methodology analysis. J. Investing **2**(1), 59–68 (1993). www.efos.unios.hr/IMJ (2014)
2. Mizuno, H., Kosaka, M., Yajima, H., Komoda, N.: Application of neural network to technical analysis of stock market prediction. Stud. Inf. Control **7**(3), 111–120 (1998)
3. Sexton, R.S., Dorsey, R.E., Dohnson, J.D.: Toward global optimization of neural networks: A comparison of the genetic algorithm and back propagation. Decis. Support Syst. **22**, 171–185 (1998)
4. Refenes, A.N., Zapranis, A., Francis, G.: Stock performance modeling using neural networks: a comparative study with regression models. J. Neural Netw. **7**(2), 375–388 (1994)
5. Kohzadi, N., Boyd M.S., Kermanshahi, B., Kaastra, L.: A comparison of ANN and time series model for forecasting commodity prices. Neuro Computing **10**(2), 169–181 (1996)
6. Ravichandran, K.S., Thirunaukarasu, P., Nallu Swamy, R., Babu, R.: Estimation of return on investment in share market through ANN. J. Theor. Appl. Inf. Technol. (2005)
7. Kryzanoueski, L., Galler, M., Wright, D.W.: Artificial neural networks to pick stocks. Financ. Anal. 21–27 (1993)
8. Saad, E.W., Wanseh, Dr: Comparative study of stock trend prediction using time delay, recurrent and probabilities neural networks. Neural Netw. IEEE Trans. **9**(6), 1045–1047 (1998)

9. Tang, Z., de Aimeida, C., Fishwick, P.A.: Time series forecasting using neural networks vs. Box-Jenkins methodology. Simulation, 303–310 (1991)
10. Schoeneburg, E.: Stock price prediction using neural networks: a project report. Neuro Computing **2**, 17–27 (1990)
11. Naeini, M.P., Taremian, H., Hashemi, H.B.: Stock market value prediction using neural networks. In: International Conference on Computer Information Systems and Industrial Management Applications (CISIM), pp. 132–136 (2010)
12. Donate, J.P., Sanchez, G.G., de Miguel, A.S.: Time series forecasting: a comparative study between an evolving ANN system and statistical methods. Int. J. Artif. Intell. **21** (2012)
13. Morgan and scofield: Neural Networks and Speech Processing. Kluwer Academic Publisher, Berlin (1991)
14. Raman, J., Jamed, A.: Back propagation and recurrent neural network in financial analysis of multiple stock market returns. In: 29th Annual Hawai International Conference on System Sciences (1996)
15. Kimoto, T., Asakawa, K., Yoda, M., Takeoka, M.: Stock market prediction system with modular neural network. In: International Joint Conference on Neural Networks Proceedings, pp. 1–6 (1990)
16. Elmen: TLEARN-Simulator program for neural networks center for Research in J-aiwxe, c-008. University of California, San Diego, La Jolla, CA 92093, www.Jatit.org (1990)
17. http://www.Nseindia.com

Age Range Estimation Based on Facial Wrinkle Analysis Using Hessian Based Filter

Husniza Razalli, Rahmita Wirza O.K. Rahmat, Fatimah Khalid and Puteri Suhaiza Sulaiman

Abstract Aging is a normal process that has an effect on different parts of the human body under the influence of various biological and environmental aspects. The most prominent changes that occur on the face are the form of the skin wrinkles, which are the main objective of this research. Specific wrinkle detection is an important task in face textural analysis. Previously, some researchers have been proposed the age range estimation based on wrinkle analysis in literature, but poor localization limits the performance of the whole age estimation process. This is because, when less number of wrinkles are detected or extracted, it will consequently affect the process to estimate the correct age. Therefore, we address this issue to enhance age range estimation method using a new approach to extract correct facial wrinkles for further analysis. We propose a method to extract facial wrinkle in face image using Hessian based filter (HBF) for age estimation. In other word, this research focus on age range estimation method based on facial wrinkle analysis extracted from facial image obtained from FG-NET database using hessian based filter. The proposed filter is theoretically straightforward, however, it significantly increases the wrinkle analysis result compared to previous methods. The result shows that HBF successfully obtained higher accuracy with over 90 % estimation rate.

Keywords Age range estimation · Wrinkle analysis · Wrinkle density · Hessian based filter

H. Razalli (✉)
Signal Processing Lab, TM Research and Development, Cyberjaya, Selangor, Malaysia
e-mail: husniza@tmrnd.com.my

R.W.O.K. Rahmat · F. Khalid · P.S. Sulaiman
Computer Graphics, Vision and Visualization Research Group, Faculty of Computer Science and IT, Universiti Putra Malaysia, Serdang, Selangor, Malaysia
e-mail: rahmita@fsktm.upm.edu.my

F. Khalid
e-mail: fatimahk@fsktm.upm.edu.my

P.S. Sulaiman
e-mail: psuhaiza@fsktm.upm.edu.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_64

1 Introduction

The human face contains a lot of information related to personal characteristics, including identification, emotion, age, gender and race. This information has been used extensively in face recognition and verification systems which are capable of interpreting the facial information found in human-to human communication [1, 2, 3]. Currently, the research related to age range estimation using face images is more important than ever, because it has many applications, such as an internet access control, underage cigarette-vending machine use [1, 4], age-based retrieval of face images [4], age prediction systems for finding lost children and face recognition robust to age progression. In the human aging process, two biophysical change or growth happen on the face [5], which are;

- The craniofacial bony aging due to a relative change in bone expansion and loss most probably refers to face shape, and
- The change of face texture and elasticity of the skin.

During adult aging, from adulthood to old age, the most perceptible change becomes skin aging (texture change). The shape change still continues, but less dramatically, mostly due to typical patterns in skin and tissue [6]. Originally shown in [7] and also in [8], Fig. 1 shows six face sketches, each of which associated with age of 30, 40, 50, 60, 70 and 80 years old, respectively. Biologically [8], as the face matures and ages with loss of collagen beneath skin as well as gravity effects, the skin becomes thinner, darker, less elastic, and more leathery. A dynamic wrinkles and blemishes due to biologic aging gradually appear. Dynamic wrinkles and folds due to muscle motion become more distinct. In the areas of deeper attachment, such as cheeks, eyelids, chin, and nose, elasticity of muscles and soft tissues gets weak and fat continues depositing. In other areas, fat may atrophy or be absorbed. These changes lead to the downward descent or sagging of the skin, such as double chin, dropping cheek, and lower eyelid bags [9]. The bony framework underneath the skin may also deteriorate to accelerate the development of skin aging, such as wrinkles, creases, and droops. In addition, face aging during this age period may cause the loss of flexible control of facial muscles so that the facial movements and behaviors may also change unintentionally [5]. In other words, face aging throughout this age era suitably can be measure based on wrinkle analysis. The

Fig. 1 Face aging sketches from 30 to 80 years with 10 years per sketch, originally shown in [8], also in [7]

wrinkle and skin features are generally appeared by high frequency components on face images and easily distinguishable to human eyes, however it is a challenging task for computer vision systems to detect them automatically. Therefore, traditionally edge detector or high frequency images are used to extract local features.

2 Existing Method

Most of the researchers focus on edge detection in order to extract wrinkle in estimating the age of a person through wrinkle analysis process. Dehshibi et al. [10] and Shima et al. [11] estimates facial age using hybrid of geometric features extraction and wrinkle analysis. In their wrinkle analysis process, they used canny edge detector to extract wrinkles. Based on the extracted wrinkles, they calculated the density of the wrinkles. Meanwhile Jiang et al. [12] proposed the classification of the skin aging level at the lower eyelid regions. The wrinkles analysis process used in this work was also based on canny edge detector and used wrinkle density and intensity as measurements. Wrinkles are measured as line segment series. Wrinkle should not be confused with edge. Edge is the border between two areas while wrinkle is a line that is either darker or lighter than their neighbourhood. Therefore, edge detection methods such as Canny are not suitable for wrinkle detection because it will produce wrinkle boundaries, not the wrinkle. Based on this statement, wrinkle density and intensity obtained from canny edge detected probably give inaccurate wrinkle analysis results and leading to reduction of age estimation accuracy.

Batool and Chellappa [2] proposed a stochastic wrinkle detection method based on marked point process (MPP). They employed a second derivative linear filter to extract line structures from an image, and penalized an overlap of wrinkle segments. However, their solution strongly depends on the initial condition, and fails to detect complex patterns of wrinkles.

In this work, we propose a new method for extracting facial wrinkle from a face image using Hessian based filter (HBF) and use these extracted wrinkles to estimate the age range. We focused on the problem of classifying an adult age range (age around 19–70) since wrinkles happening during these period. And as a nature process, wrinkle only occur when a person reaching adulthood period. Furthermore, this research was based on the publicly available FG-NET (Face and Gesture Recognition Research Network) aging database which contained 1,002 color and grayscale images of persons from 0 to 69 years of age [13].

The rest of this paper is organized as follows. Section 2 provides step by step methodology of the proposed work. Section 3 demonstrates an experimental setup while, Sect. 4 describes experimental results. Finally, conclusions are drawn in Sect. 5.

3 Methodology

The proposed facial age range estimation method using HBF is briefly outlined in this section. The technique starts with the input of face image from FGNet database. This technique comprises four major processes which are; Wrinkle Region of Interest (ROI) Localization, Wrinkle Detection and Extraction, Wrinkle Density Calculation and Age Range Classification as illustrated in Fig. 2.

This proposed method will recognize person's age group based on their face image. The range between subsequent age groups is 10 years, which mentioned in [5] this age gap give maximum result to wrinkle progression during aging. The ranges of the age are grouped as below;

- R1—Range 1 age 19–30,
- R2—Range 2 age 31–40,
- R3—Range 3 age 41–50 and
- R4—Range 4 is for 51+

3.1 Wrinkle ROI Localization

During age progression, wrinkles on faces become more and more clear. The wrinkles on human face become more prominent when he or she getting older. Aged people often have clear wrinkles on the following areas of the face [14]:

- The forehead has horizontal furrows.
- The eye corners have crows' feet.
- The cheeks have obvious cheekbones, crescent shaped pouches, and deep lines between the cheeks and the upper lips.

Fig. 2 Block diagram for age range estimation process using Multi-SVM Classification

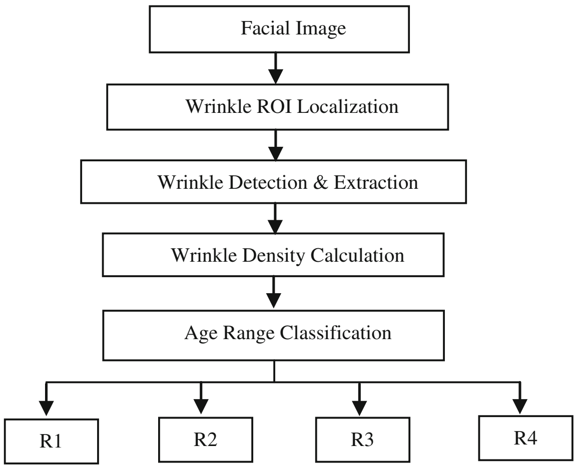


Fig. 3 Sample images in FG-NET database for every age range

However, according to [15] the upper ROI face wrinkles increase the performance of age estimation and as mentioned in [12], skin aging tendency can be indicated well using lower eyelid region since this area possesses important information for aging level identification. Based on these two statements, only selected ROI are used to detect and extract the wrinkle, which are forehead area and lower eyelids area. The measurement of the area is carried out based on region of the eye as describe in Fig. 3. We labeled the forehead box as A, lower eyelid for left eye as B and lower eyelid for the right eye as C. The height of all ROI areas is determined by the height of eye region, denoted by 'n'. The width for ROI area A and B is also the same with the width of the eye region. By contrast, the width of ROI area A is 3 times of that of the width of the eye region width. The definition of ROI areas is defined in Eq. (1) and (2):

$$Area\,B = Area\,C = Area\,of\,Eye \tag{1}$$

$$Area\,A = EyeHeight * (EyeWidth * 3) \tag{2}$$

The ROI is also selected based on consideration about the nature of face and circumstances that may occur during detection, such as if we choose the area around mouth, we should consider the gender of the sample. This is because we try to extract the wrinkle pixels, most of older male samples in our FG-NET database consist of the face with mustache and beard which will disturbed the result of the detection. Therefore, in order to reduce the number of error detection we only consider the above-mentioned ROI areas in our proposed algorithm.

3.2 Wrinkles Detection and Extraction

Wrinkles are considered as stochastic spatial arrangements of line segment sequences. Wrinkle should not be confused with edge which refers to the border between two areas. By contrast, wrinkle is a line that is either darker or lighter than their neighbourhood as shown in Fig. 1. Therefore, edge detection methods such as

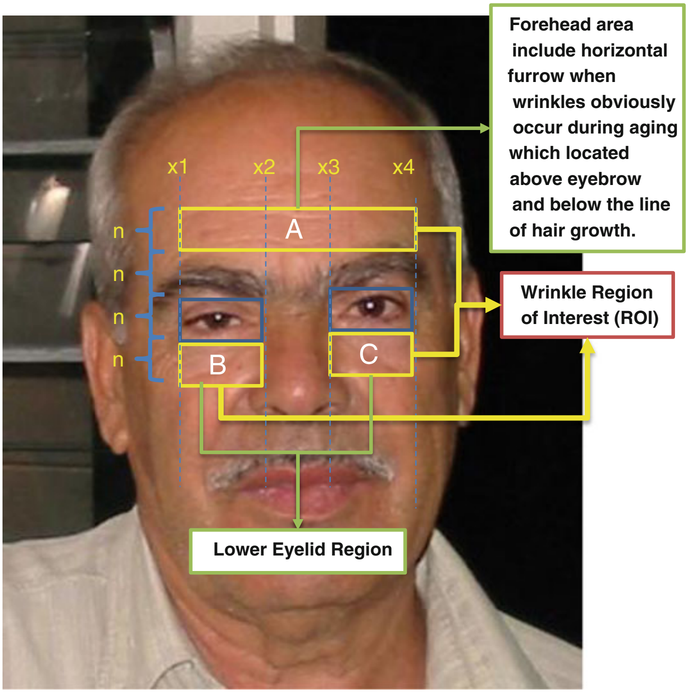


Fig. 4 The selected areas of facial wrinkles for HBF analysis

Canny and Sobel are not suitable for wrinkle detection because it will only produces wrinkle boundaries, not the wrinkle.

The HBF is applied in order to extract wrinkles from selected ROI of facial wrinkles as shown in Fig. 4 where the HBF with the scale range, scale ratio and correction constant parameter was used according to the selected facial, as cited from [15]. Here, the HBF with [1 1] scale range, and 1 scale ratio were chosen. This was done after some empirical analysis with the method, where these settings exhibited the best performance.

After extracting the wrinkle from selected ROI areas of the facial images, Wrinkle Extraction Process can be applied on facial images. In this stage, several steps are used to obtain exact wrinkles pixels as describe in Fig. 5. Firstly, the input of wrinkle ROI will be converted to grayscale image. Then, Hessian Based Filter [16] is applying to the grayscale image to get the wrinkle edge likeliness (Fig. 6).

After that, morphological operation was used to remove unwanted pixels which are considered as a noise. Finally, the wrinkle pixels result was passed to another process which is equivalent to measuring the density of wrinkles.

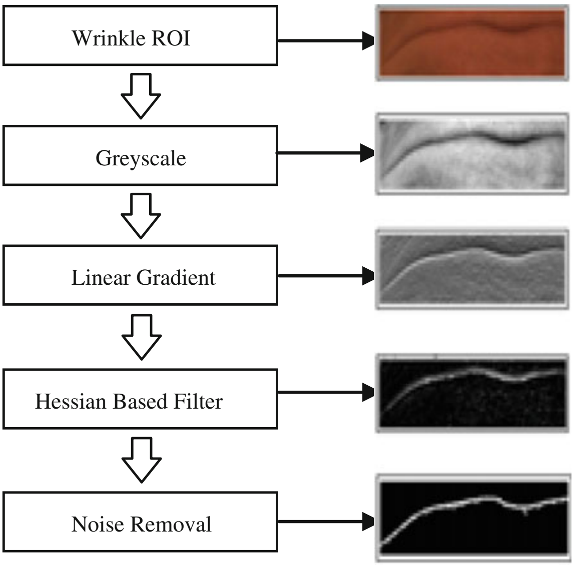


Fig. 5 Block diagram for wrinkle extraction process using HBF

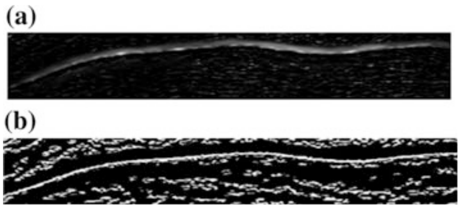


Fig. 6 **a** Wrinkle edge using Hessian Based Filter, **b** Wrinkle edge using Canny Edge Detector

3.3 Calculate Wrinkle Density

In our proposed method, the binary wrinkle image from ROI A image after noise removal process is taken as the input. The function returns 1 when it finds wrinkle pixels in the input image and returns 0 elsewhere. The wrinkle density $\mathrm{WD_A}$ in area A is defined as:

$$\mathrm{WD_A} = \frac{|WA|}{|PA|} * 100 \quad (3)$$

where WA stands for the total of all wrinkles pixels in area A, PA is the set of all pixels in area A.

Table 1 Average wrinkle densities for selected wrinkle ROI

Age range	Wrinkle ROI		
	Area A	Area B	Area C
Range 1	8.09	11.60	10.70
Range 2	9.55	12.80	12.40
Range 3	13.20	14.80	13.90
Range 4	15.40	18.20	16.40

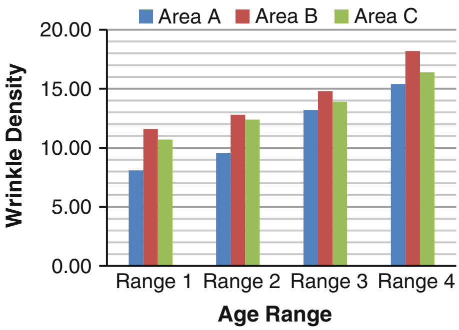


Fig. 7 Distribution of average wrinkle densities in selected ROI for every age range

In order to measure the effectiveness of our method, we firstly find the wrinkle density pattern from 100 adult face images with age around 19–60 years old obtained from FG-NET database. The wrinkle density then grouped according to its age range. After that, the average wrinkles density associated with every age range is calculated for all selected ROI. These values are listed in Table 1.

From Table 1 we distribute the wrinkle densities in a graph to clarify the pattern of the extracted wrinkle density during age progression using our proposed method as shown in Fig. 7.

From the figure, we can observe the pattern of the wrinkles progression during aging and we can conclude that those seleted ROI contributed to wrinkle analysis.

3.4 Age Range Classification

For the classification process, modified Multi Support Vector Machine (Multi-SVM) developed by Mishra [17] was chosen because it is designed for binary class problems which select the optimal linear decision hyper-plane. In SVM based classification, each data point in the dataset is represented by a k-dimensional vector with n-ratios. Assuming, each data point belongs to only one of two classes, the SVM Separate the dataset with a k-1 dimensional hyper-plane with maximum separation between the two classes. The same data point and process goes to Multi-SVM classifier. However, in Multi-SVM assumes, each data point belongs to one of more than two classes as shown in Fig. 7. This phase is used as an experimental setup to measure the accuracy of the proposed age range estimation method. Detail explanation about classification process is discussing in the next section.

4 Experimental Setup

The proposed age range estimation based on facial wrinkle analysis using hessian based filter is evaluated on the FG-NET aging database. Out of 1002 face images from 82 subjects with the age ranging between newborns up to 69 years, only 180 proper frontal images with no spectacles, no beard and moustache and no extreme facial expression were chosen, which are selected randomly from different ages range between 19 and 69 years old. We use the Multi-SVM classifier to estimate the age range of the facial image as describe in Fig. 8.

The n ratios are the densities for all ROI, and the data is split into training and testing sets with k rows. The ¼ k data as a training set which is used for generating the classification model and the ¾ k data as a testing set is used to test the classification performance of the classifier. As it is shows in Fig. 5, the classifiers categorize the images into four age groups as mentioned earlier.

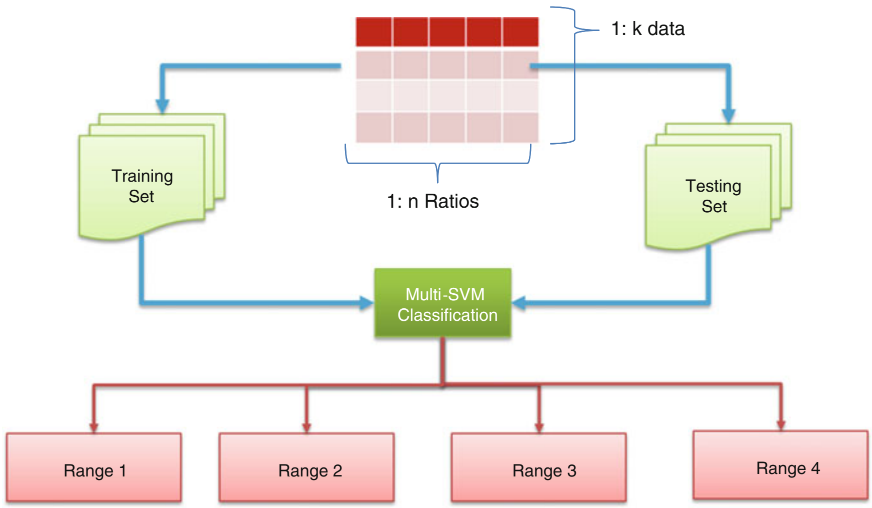


Fig. 8 Block diagram for age range estimation process using Multi-SVM Classification

Table 2 Estimation rate for the propose age range estimation method

Ground truth	Total image	Detected image	Accuracy (%)
Range 1	89	83	93.25
Range 2	48	42	87.50
Range 3	25	23	92.00
Range 4	18	17	94.44

Table 3 Performance comparison of Hessian Based Filter (HBF) with the work of other researchers

Ground truth	Accuracy (%)
Canny Edge Detector proposed by [10, 11, 12]	82.356
Proposed HBF	91.798

5 Result and Discussion

This section presents the experimental results for the age range estimation method based on facial wrinkle analysis extracted from facial image using HBF. Wrinkle analysis proposed in this work was compared with the method proposed in [10, 11]. The results are listed in Tables 2 and 3.

From the results, this method successfully classifies the face images acquired from FG-NET database into preferred ranges using Multi-SVM classifier with higher accuracy of estimation rate (%). The average accuracy of our proposed method also demonstrates a promising result compare to others.

6 Conclusion

In this work we extract the wrinkles on upper face region using HBF and then calculate the wrinkle density to measure the age range of a person's face. We experimented with different wrinkle analysis method proposed by previous work to compare accuracy of the age range estimation method. Although higher estimation rate achieved by using our proposed estimation method, this approach can be beneficial in hybrid features extraction method since it can reduce the need of unwanted features and decrease time execution. However, the drawback is the image have to be frontal (without tilt) with no facial accessory, radial lighting condition and consideration of image quality.

References

1. Geng, X., Zhou, Z. H., Zhang, Y., Li, G., Dai, H.: Learning from facial aging patterns for automatic age estimation. In: Proceedings of the 14th annual ACM international conference on Multimedia, pp. 307–316 (2006)
2. Batool, N., Chellappa, R.: Modeling and detection of wrinkles in aging human faces using marked point processes. In: Computer Vision–ECCV 2012. Workshops and Demonstrations. Springer, Berlin, Heidelberg (2012)
3. Razalli, H., Rahmad, R.W.O., Mahmod, R.: Real Time Face-Tracking and Iris Localization: The Design of a Real Time Face-Tracking and Iris Localization System Using Image Processing and Circle Fitting Technique. LAP Lambert Academic Publishing, Germany (2012)
4. Lanitis, A., Draganova, C., Christodoulou, C.: Comparing different classifiers for automatic age estimation. IEEE Trans. Syst. Man Cybern. B Cybern. **34**(1), 621–628 (2004)
5. Fu, Y., Guo, G., Huang, T.S.: Age synthesis and estimation via faces: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **32**(11) (2010)
6. Geng, X., Zhou, Z.H., Miles, K.S.: Automatic age estimation based on facial aging patterns. Proc. IEEE Trans. Pattern Anal. Mach. Intell. **29**(12), 2234–2240 (2007)
7. Gonzalez-Ulloa, M., Flores, E.: Senility of the face: basic study to understand its causes and effects. Plast. Reconstr. Surg. **36**, 239–246 (1965)
8. Stone, A.: The aging process of the face & techniques of rejuvenation. http://www.aaronstonemd.com/Facial_Aging_Rejuvenation.shtm (2010)
9. Cula Gabriela, O., et al.: Assessing facial wrinkles: automatic detection and quantification. Skin Res. Technol. **19**(1), e243–e251 (2013)
10. Dehshibi, M.M., Bastanfard, A.: A new algorithm for age recognition from facial images. Sig. Process. **90**(8), 2431–2444 (2010)
11. Izadpanahi S., Toygar, O.: Human age classification with optimal geometric ratios and wrinkle analysis. Int. J. Pattern Recognit. Artif. Intell. **28**(02) (2014)
12. Jiang, X., Liao, Q.: Skin wrinkles analysis based on digital image processing. In: Wavelet Active Media Technology and Information Processing (ICWAMTIP), 2012 International Conference on Digital Object Identifier:10. 1109/ICWAMTIP. 2012. 6413437 Publication Year: 2012, pp. 50–53
13. FGNET.: Aging database. Available: http://www.fgnet.rsunit.com/ (2007). Accessed on 13 Feb 2013
14. Thomas, C.: Make-Up: The Dramatic Student's Approach. Theatre Arts Books, New York, U. S. A (1968)
15. Dahlan, H.A.: et al. Age estimation using specific gabor filter on upper facial area. In: Proceedings of the Fourth International Conference on Signal and Image Processing 2012 (ICSIP 2012). Springer India (2013)
16. Dirk-Jan Kroon: Enhancement of Vessel/ridge like structures in 2D/3D image using hessian eigen values. http://au.mathworks.com/matlabcentral/fileexchange/24409-hessian-based-frangi-vesselness-filter (2010). Accessed 1 June 2014
17. Mishra, A.: Multi class support vector machine. Matlab central. http://www.mathworks.com/matlabcentral/fileexchange/33170-multi-class-support-vector-machine (2011). Accessed 10 June 2014

Bending Effects on Wearable Antenna with Silver Nanowires and Polydimethylsiloxane

Mohd Aminudin Murad, Aslina Abu Bakar, Ahmad Rashidy Razali, M. Amir Imran M. Hasli and Muhammad Feroze Akbar J. Khan

Abstract This paper presents an investigation of an inset-fed microstrip patch antenna designed with silver nanowires (AgNWs) radiating patch and polydimethylsiloxane (PDMS) substrate. Highly conductive and stretchable radiating element AgNWs is embedded in flexible and durable PDMS elastomer. The antenna is designed to be wearable and flexible for body-centric wireless communications (BCWCs) applications, specifically in the 2.45 GHz Industrial, Scientific and Medical (ISM) band. The radiation characteristics are simulated and analyzed when the antenna is under flat and bent conditions. The proposed antenna demonstrates good performance for H-plane and E-plane bending orientations. The antenna offers great mechanical flexibility and robustness, which indicates its good potential for BCWCs applications.

Keywords Wearable antenna · Polydimethylsiloxane · Microstrip patch · Silver nanowires · PDMS · Body centric

1 Introduction

In recent years, research on wearable antennas are becoming more important especially for body-centric wireless communications (BCWCs). Body centric communication systems is applicable for specific applications such as medicals, fire fighters, military and public safety, tracking and navigation, which demand for robust and flexible antennas. Generally, wearable antenna for modern applications requires light weight construction, low cost, almost maintenance-free and easy

M.A. Murad (✉) · A.A. Bakar · A.R. Razali · M.A.I.M. Hasli · M.F.A.J. Khan
Faculty of Electrical Engineering, Universiti Teknologi MARA Pulau Pinang, 13500 Permatang Pauh, Pulau Pinang, Malaysia
e-mail: aminudinmurad82@gmail.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_65

installation. Specific requirements for the design of wearable antennas are: planar structure; flexible conductive materials in the patch and ground plane; and flexible dielectric materials [1]. Various materials and construction methods have been used to develop wearable and flexible antennas.

Wearable antennas require the ability to function under various modes of structural deformations while being worn. Conventional antennas are fabricated by printing or etching conductor patterns on rigid substrates, which can fail to operate properly when subjected to mechanical deformation such as bending, twisting and stretching. One of the dominant research in antennas for BCWCs applications is wearable, fabric-based antennas. These wearable fabric antennas are known to be sensitive to environmental conditions such as humidity and temperature. Their performance may be affected significantly or even fail to operate under certain circumstances.

Recently, more flexible substrates have been reported such as polymer, micro fluids/liquid metals, paper and plastic to be used in flexible antenna development. The polymer-based antennas specifically, are becoming popular topic in the field on flexible antennas. These flexible polymer-based antennas, which incorporates highly conductive radiating element have demonstrated better performance when subjected to mechanical deformations and environmental challenges as well as having other unique qualities to be integrated in BCWCs applications. Latest advancement in flexible antennas includes thin film metal deposited on polydimethylsiloxane (PDMS) elastomer [2–6]. PDMS is a silicon-based elastomer, which is chemically inert, thermally stable, permeable to gases, easy to manipulate, and exhibits isotropic and homogeneous properties. The initial liquid state of PDMS allows versatile fabrication process [7]. The advantages of using such polymers as flexible substrate is that they are inexpensive and able to withstand mechanical strains.

In order to achieve fully flexible wearable antenna, the radiating element of the antenna must be able to handle mechanical deformation such as bending. A flexible antenna with limited mechanical flexibility was developed by [4] using PDMS and copper. A more recently used flexible metals such as silver nanowires (AgNWs) [2, 3], eutectic gallium indium (EGaIn) [5, 6] and graphene [8] facilitate better mechanical flexibility. The AgNWs used in the proposed antenna has also been previously incorporated in various electronics and circuits applications due to its transparent, flexible and stretchable characteristics [2]. In this paper, we report a development of a flexible wearable antenna by exploiting the flexibility and good dielectric properties of AgNWs and PDMS in the antenna design. Simulations performed on antenna bent in H-plane and E-plane orientations are conducted to determine the radiation characteristics and performance.

2 Antenna Design and Simulations

2.1 Antenna Design

In this paper, a wearable rectangular microstrip patch antenna operating at 2.45 GHz Industrial, Scientific and Medical (ISM) band was designed for BCWCs applications. The proposed antenna integrates a highly conductive AgNWs as the radiating patch element and PDMS elastomer as the dielectric substrate. The dielectric substrate separates the ground plane from the transmission line and patch antenna. The PDMS elastomer appeared as encasing the patch radiator and ground plane. This designed is required to hold the antenna layers together due to the low-adhesive characteristics of PDMS as reported by [7].

The proposed wearable antenna structure shown in Fig. 1 utilizes the transmission-line model of the rectangular microstrip patch antenna. This model is represented by two slots, separated by a transmission line. Fringing fields are generated at the edges of the patch when the microstrip transmission line received a signal, which resulted to radiation of electromagnetic waves. The amount of fringing fields generated are influenced by the dimension of the patch and the height of the substrate [9].

The resonant frequency f_{res} of the antenna is a function of the width W, the length L of the patch element and the thickness of the substrate h. The resonant frequency f_{res} can be calculated using

$$f_{res} = \frac{c}{2L\sqrt{\varepsilon_{reff}}} \tag{1}$$

where

c free-space velocity of light

L length of microstrip patch

ε_{reff} effective relative permittivity of microstrip patch

The targeted resonant frequency f_{res} is 2.45 GHz. Existing literatures on PDMS dielectric characteristics reported the relative permittivity ε_r of 2.76–3.00 and lossy

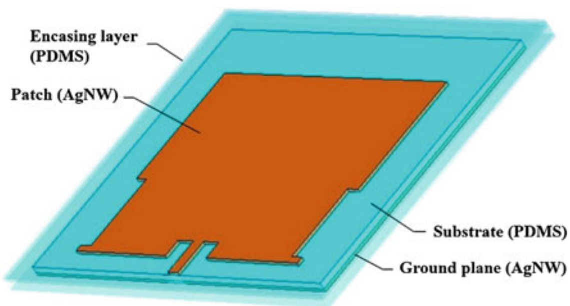


Fig. 1 Wearable AgNWs microstrip patch antenna on PDMS substrate

tangent tanδ of 0.01–0.05 over an operating frequency of 0.2–5.0 GHz [6, 7]. The proposed antenna was modelled with relative permittivity of ε_r = 2.76 and lossy tangent tanδ = 0.02 accordingly. The AgNWs/PDMS patch has a conductivity of ~8,130 S/cm [10].

A practical patch width, W for efficient radiator to achieve a good radiation efficiency can be determined using

$$W = \frac{c}{2f_{res}}\sqrt{\frac{2}{\varepsilon_r + 1}} \tag{2}$$

where
W width of microstrip patch
f_{res} resonant frequency
ε_r relative permittivity of microstrip patch

and the actual length, L of the patch antenna is determined using

$$L = \frac{c}{2f_{res}\sqrt{\varepsilon_{reff}}} - 2\Delta L \tag{3}$$

where
ΔL extended length of microstrip patch
f_{res} resonant frequency
ε_{reff} effective relative permittivity of microstrip patch

The effective relative permittivity or effective dielectric constant ε_{reff} and the extended length of microstrip patch ΔL are given by

$$\varepsilon_{reff} = \frac{\varepsilon_r + 1}{2} + \frac{\varepsilon_r - 1}{2}\left[1 + 12\frac{h}{W}\right]^{-1/2} \tag{4}$$

$$\Delta L = 0.412h\frac{\left(\varepsilon_{reff} + 0.3\right)\left(\frac{W}{h} + 0.264\right)}{\left(\varepsilon_{reff} - 0.258\right)\left(\frac{W}{h} + 0.8\right)} \tag{5}$$

where h is the height of the substrate. Equation (5) gives a practical approximation of normalized length extension ΔL, which is a function of effective dielectric constant and the width-to-height ratio (W/h). The patch of the microstrip antenna appears to be larger electrically than its physical dimension due to the fringing fields at the patch edges.

Using Eqs. (4) and (5), the calculated length L and width W of the patch antenna are 36.5 and 44.7 mm, respectively. The thickness of the AgNWs patch layer and ground plane is 0.5 mm. The radiating patch element is separated from the ground plane by a PDMS substrate layer of 1.0 mm thickness. Cutouts were introduced on both sides of radiating patch element to achieve the 2.45 GHz resonant frequency, while suppressing unwanted resonant frequencies as well as increasing the

Table 1 Optimized antenna dimensions

Parameters	Optimized dimension (mm)
Patch length, L	35
Patch width, W	43
Ground length, L_g	47
Ground width, W_g	56
Inset feed length, y_0	3
Inset feed width, x_0	3
Feed length, L_f	3
Feed width, W_f	2
Patch Cutout Width, W_c	1
Patch Cutout Length, L_c	14

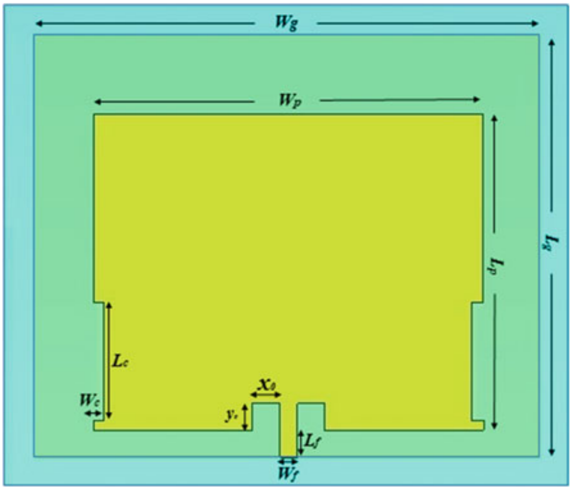


Fig. 2 The geometry of AgNWs microstrip patch antenna on PDMS substrate with optimized dimensions

bandwidth. Effective matching of the patch antenna with 50 Ω microstrip feed line can be achieved using an inset feed of recessed distance y_0. The cutout inset slots dimensions and the width of the microstrip feedline W_f were optimized using CST Microwave Studio (CST MWS) simulation software from CST Computer Simulation Technology AG to achieve optimal performance.

The optimized antenna design dimensions and geometry are described in Table 1 and Fig. 2.

2.2 Antenna Bending Simulations

All modeling and simulation work of the antenna were performed using CST MWS software. The proposed antenna was bent along the H- and E-planes of the antenna

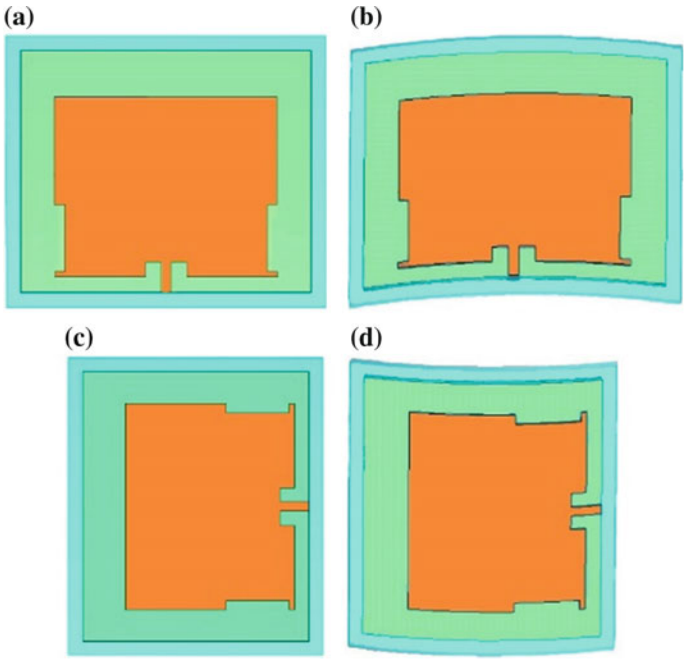


Fig. 3 Antenna bending configurations: **a** H-plane flat; **b** H-plane bending; **c** E-plane flat; **d** E-plane bending

structure to investigate the effects of bending on antenna performance as shown in Fig. 3. The antenna was bent over 17 radii between radius, r = 30 mm to r = 170 mm, along both planes, which resembles different curves and shapes of human body parts. An example of antenna being located and bent on human arm is shown in Fig. 4.

3 Results and Discussion

The following discussion is focused on the response of the proposed wearable antenna under flat (i.e., no bending) and E-plane and H-plane bending conditions. Figure 5 shows the return loss, S_{11} of the simulated antenna under flat and bent conditions for both planes. A reliable wearable antenna is expected to be subjected to physical structure deformation such as bending without significant performance degradation when incorporated in BCWCs applications. Therefore, it is critical for the proposed antenna to be simulated under flat and bent conditions. The smaller radius of bending simulation resembles the antenna positioned on smaller body parts such as wrist, while the larger bending radius describe the performance when the antenna is positioned on a more relaxed conditions on body parts such as on human chest, which is more flat.

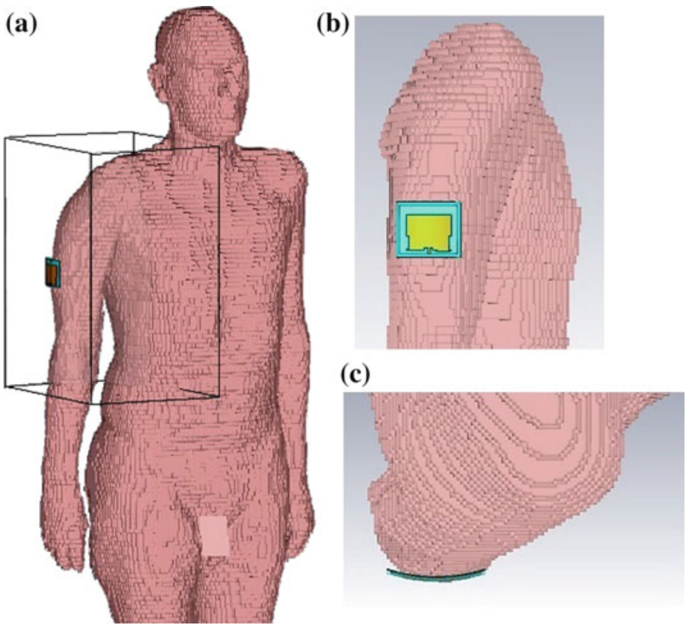


Fig. 4 An example of an antenna located on human arm (r = 100 mm): **a** front view; **b** side view; **c** top view

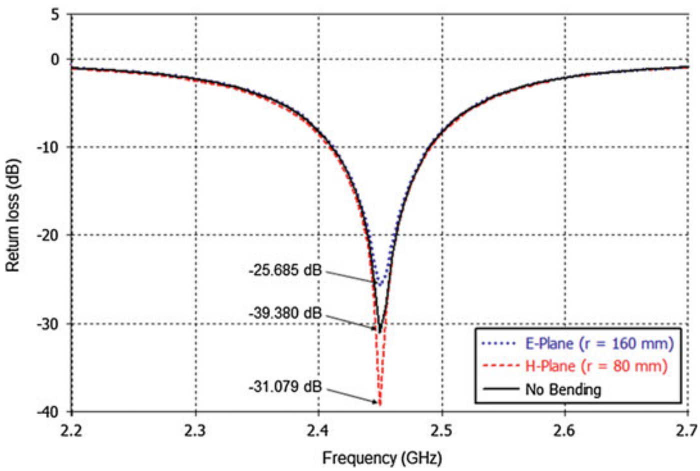


Fig. 5 Return loss of the proposed antenna

The best return loss performance achieved for H-plane and E-plane is shown in Fig. 5. At 2.45 GHz resonant frequency, the best return loss was demonstrated when the antenna was bent over radius, r = 80 mm and 160 mm for H- and E-planes, respectively. A sample 3D far field radiation pattern for the simulated

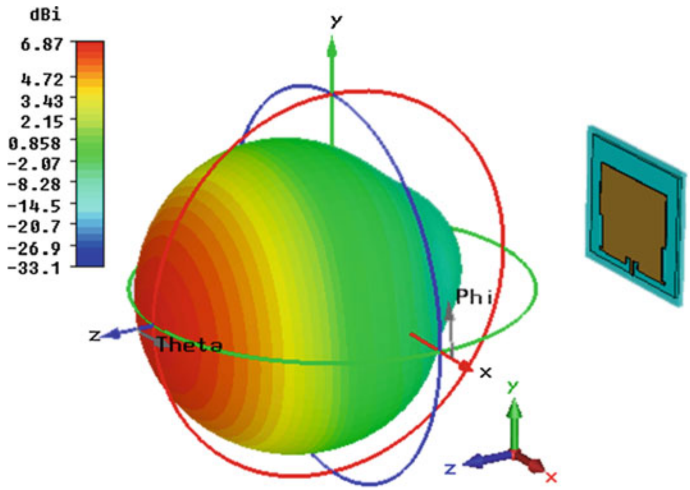


Fig. 6 Simulated 3D radiation pattern of the proposed antenna

antenna is shown in Fig. 6. The directive radiation pattern and the maximum gain, which is normal to the patch is compatible for BCWCs, specifically for off-body communication for sending signals from human body to a base station or remote terminal [11, 12]. The simulated far field radiation patterns of the antenna for flat and bent conditions are shown in Fig. 7. The simulation results demonstrated that the antenna was able to maintain a directive radiation with minimal back lobes for all bending conditions.

Table 2 summarizes the radiation characteristics of the proposed antenna for H-plane and E-plane bending conditions. The antenna still covers the 2400–2483.5 MHz ISM band for all bending conditions with relatively insignificant frequency shift. The radiation performance of the antenna bent along H-plane exhibited minimal variation in gain, bandwidth, directivity and voltage standing wave ratio (VSWR). The antenna bending along E-plane demonstrated moderate changes to the antenna performance. The gain and bandwidth of the antenna bent along E-plane decreased gradually as the degree of bending increased. The antenna gain was reduced by almost 69 % when bent over radius, $r = 30$ mm compared to $r = 170$ mm. The antenna also has 16.5 MHz less bandwidth when it was bent over radius, $r = 30$ mm compared to $r = 170$ mm on E-plane. The antenna bent along H-plane over the same radius experienced minimal bandwidth changes. However, the directivity of the antenna increased marginally when the antenna was subjected to higher bending degree along E-plane. The more substantial antenna performance degradation observed for E-plane bending was resulted from the more significant changes in the antenna structure and layers. The antenna structure and layers alignment are more symmetrical along the H-plane and were not significantly affected by the bending procedure, thus producing more consistent performance (Table 3).

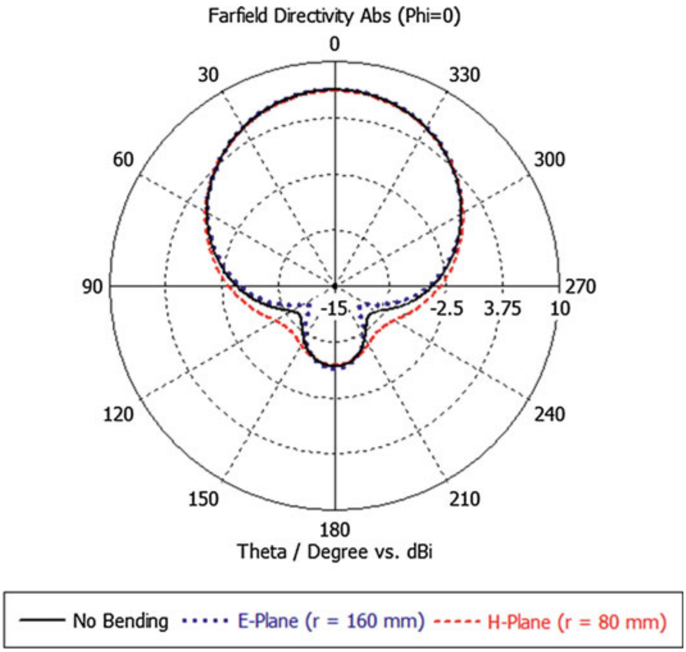


Fig. 7 Simulated far field radiation patterns of the antenna for flat and H- and E-planes bending

Table 2 H-plane bending antenna radiation characteristics

Bending radius, *r* (mm)	Parameters					
	Resonant frequency (GHz)	Return loss, S_{11} (dB)	Gain (dB)	Bandwidth (−10 dB) (MHz)	Directivity (dBi)	VSWR
30	2.440	−19.958	2.230	79.6	6.250	1.223
40	2.435	−23.900	2.235	81.8	6.471	1.136
50	2.440	−28.230	2.297	82.9	6.584	1.081
60	2.430	−30.670	2.193	81.2	6.661	1.060
70	2.445	−32.601	2.279	81.7	6.691	1.048
80	2.450	−39.380	2.279	82.8	6.715	1.022
90	2.435	−35.390	2.121	81.1	6.749	1.035
100	2.450	−34.491	2.230	81.7	6.736	1.038
110	2.460	−42.409	2.276	82.2	6.738	1.015
120	2.440	−35.407	2.087	80.1	6.873	1.035
130	2.440	−39.450	2.109	79.9	6.789	1.021
140	2.442	−33.762	2.066	80.4	6.806	1.042
150	2.445	−45.891	2.063	80.6	6.805	1.010
160	2.450	−36.099	2.148	79.6	6.795	1.032
170	2.440	−34.449	2.055	79.5	6.822	1.039
0 (Flat)	2.450	−31.079	1.803	79.5	6.827	1.057

Table 3 E-plane bending antenna radiation characteristics

Bending radius, r (mm)	Parameters					
	Resonant frequency (GHz)	Return loss, S_{11} (dB)	Gain (dB)	Bandwidth (−10 dB) (MHz)	Directivity (dBi)	VSWR
30	2.479	−19.048	0.585	63.0	7.146	1.251
40	2.479	−26.426	0.907	71.7	7.087	1.100
50	2.435	−20.777	0.910	63.8	7.127	1.201
60	2.430	−41.797	0.957	73.5	7.098	1.016
70	2.430	−31.494	1.163	70.9	7.074	1.055
80	2.445	−35.380	1.389	71.7	7.045	1.035
90	2.440	−29.780	1.315	76.2	7.019	1.067
100	2.455	−22.090	1.619	77.9	6.988	1.171
110	2.430	−22.783	1.451	76.8	7.007	1.157
120	2.426	−21.776	1.465	76.7	7.007	1.177
130	2.421	−24.159	1.406	77.4	6.996	1.132
140	2.450	−24.066	1.653	78.4	6.946	1.134
150	2.440	−24.900	1.607	77.9	6.952	1.121
160	2.450	−25.685	1.721	78.1	6.931	1.110
170	2.450	−24.588	1.750	78.4	6.925	1.125
0 (Flat)	2.450	−31.079	1.878	79.5	6.827	1.057

4 Conclusion

A wearable 2.45 GHz inset-fed rectangular patch antenna incorporating AgNWs radiating patch and PDMS elastomer substrate is reported in this paper. The proposed antenna has demonstrated good radiation performance when subjected to structural bending. The proposed antenna has been verified by simulations to work within the 2400–2483.5 MHz ISM band as stipulated by the Federal Communications Commission (FCC) when subjected to bending.

Acknowledgement The authors would like to acknowledge the Ministry of Science and Technology (MOSTI), Malaysia through EScience Research Grant Project No. 04-01-01-SF0620 that supports this work.

References

1. Rais, N.H.M., Soh, P.J., Malek, F., Ahmad, S., Hashim, N.B.M., Hall, P.S.: A review of wearable antenna. In: Antennas & Propagation Conference, 2009. LAPC 2009. Loughborough, pp. 225–228 (2009)
2. Song, L., Myers, A.C., Adams, J.J., Zhu, Y.: Stretchable and reversibly deformable radio frequency antennas based on silver nanowires. ACS Appl. Mater. Interfaces. **6**, 4248–4253 (2014)

3. Rai, T., Dantes, P., Bahreyni, B., Kim, W.S.: A stretchable RF antenna with silver nanowires. IEEE Electron Device Lett. **34**, 544–546 (2013)
4. Lin, C.P., Chang, C.H., Cheng, Y.T., Jou, C.F.: Development of a flexible SU-8/PDMS-based antenna. IEEE Antennas Wirel. Propag. Lett. **10**, 1108–1111 (2011)
5. So, J.H., Thelen, J., Qusba, A., Hayes, G.J., Lazzi, G., Dickey, M.D.: Reversibly deformable and mechanically tunable fluidic antennas. Adv. Funct. Mater. **19**, 3632–3637 (2009)
6. Hayes, G.J., So, J.-H., Qusba, A., Dickey, M.D., Lazzi, G.: Flexible liquid metal alloy (EGaIn) microstrip patch antenna. TAP_IEEE Trans. Antennas Propag. **60**, 2151–2156 (2012)
7. Trajkovikj, J., Zurcher, J.F., Skrivervik, A.K.: PDMS, a robust casing for flexible W-BAN antennas [EurAAP corner]. IEEE Antennas Propag. Mag. **55**, 287–297 (2013)
8. Mansor, M.M., Rahim, S.K.A., Hashim, U.: A 2.45 GHz wearable antenna using conductive graphene and polymer substrate. In: 2014 International Symposium on Technology Management and Emerging Technologies (ISTMET). IEEE, 2014, pp. 29–32. Bandung, Indonesia (2014)
9. Balanis, C.A.: Antenna Theory Analysis and Design, 3rd edn. Wiley, New Jersey (2005)
10. Xu, F., Zhu, Y.: Highly conductive and stretchable silver nanowire conductors. Adv. Mater. **24**, 5117–5122 (2012)
11. Hall, P.S., Hao, Y. (eds.): Antennas And Propagation for Body Centric Wireless Communications. Artech House, London (2012)
12. Conway, G.A., Scanlon, W.G.: Antennas for over-body-surface at 2.45 GHz. Trans. Antennas Propag. **57**, 844–855 (2009)

Real Time Signal Detection and Computer Visualization of the Patient Respiration

Jan Grepl, Marek Penhaker, Jan Kubicek, Ales Liberda and Reza Mashinchi

Abstract The article deals with the realization of the measurement chain for monitoring patient respiratory activity during radiotherapy. The primary requirement is a real time process monitoring of patient respiration during radiotherapy. The main part of the measuring chain is sensing element which converts the patient's breathing movements into electrical signals. Different possibilities of obtaining these data are described. An essential part of a comprehensive monitoring system is a software environment for the analysis and testing of the acquired data. The output signals which correspond to the ventilation state of the patient are graphically visualized in the user environment, which is created in LabVIEW.

1 Introduction

Current medicine is literally overwhelmed with the number of devices, which in many cases has been providing quality services to the end users. In the field of radiotherapy it is certainly doubly true. Just because of the use of ionizing radiation

J. Grepl (✉)
Faculty of Mechanical Engineering, VSB—TU Ostrava, Dep. 330/IT4I, Ostrava, Czech Republic
e-mail: jan.grepl@vsb.cz

M. Penhaker · J. Kubicek · A. Liberda
FEECS/Department of Cybernetics and Biomedical Engineering, VSB—TU Ostrava, Ostrava, Czech Republic
e-mail: marek.penhaker@vsb.cz

J. Kubicek
e-mail: jan.kubicek@vsb.cz

R. Mashinchi
UTM-IRDA Digital Media Centre, Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
e-mail: r_mashinchi@yahoo.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_66

for the treatment of cancer are demands for quality performance and safety in use of these devices much higher than in other areas of medical instrumentation, which is of course reflected in the price of such devices [1].

Nowadays, however, in many cases manufacturers of radiotherapy devices adheres trends when comes to additional functions in radiotherapy apparatus, and therefore some problems in most of today's devices are unresolved or resolved only partially. One of such problems of the radiotherapy apparatus is monitoring of respiratory activity, which undoubtedly affects the position of tissue during radiotherapy [2]. This complicates the surgery, especially if the patient does not cooperate with medical personnel perfectly. In such situations patient's body movement monitoring, when breathing, related to automatic tracking of application point of ionizing radiation by device itself.

2 The Proposed Solution

The first requirement for the monitoring of respiratory activity and breath hold detection is to use a suitable analog-to-digital converter to digitize two or three analog signals. These signals are obtained from the sensing elements when patient is breathing during radiotherapy.

The second requirement is to transfer thus digitized signals from output of the converter to a computer placed in the irradiation control room and to present them in real-time using software interface to medical personnel. Additional design requirements of the measuring device are:

- must be safe for the patient,
- have to minimize restriction of patient comfort,
- must be reliable and easy to use and maintain,
- it's implementation should be the least expensive.

The most significant breathing symptoms are temperature changes in the input areas of the airways (nose and mouth), respiratory muscle movements i.e. changes of the volume of the chest and belly and the influence of respiration on the chest tissue thus influence on the ECG signal [3, 4] (Fig. 1).

2.1 Thermal Biological Respiration Signal

Assuming the ambient room temperature about 25 °C breath cycles can be monitored due to differences in temperature between ambient and exhaled air. When inhale through the nose goes cooler ambient air while during exhale air is heated to body temperature i.e. approximately 37 °C. Unfortunately worse detection is while holding the breath, after inhale or after exhale. In both cases the temperature at the boundary nose-surroundings approximately equalize to ambient air temperature,

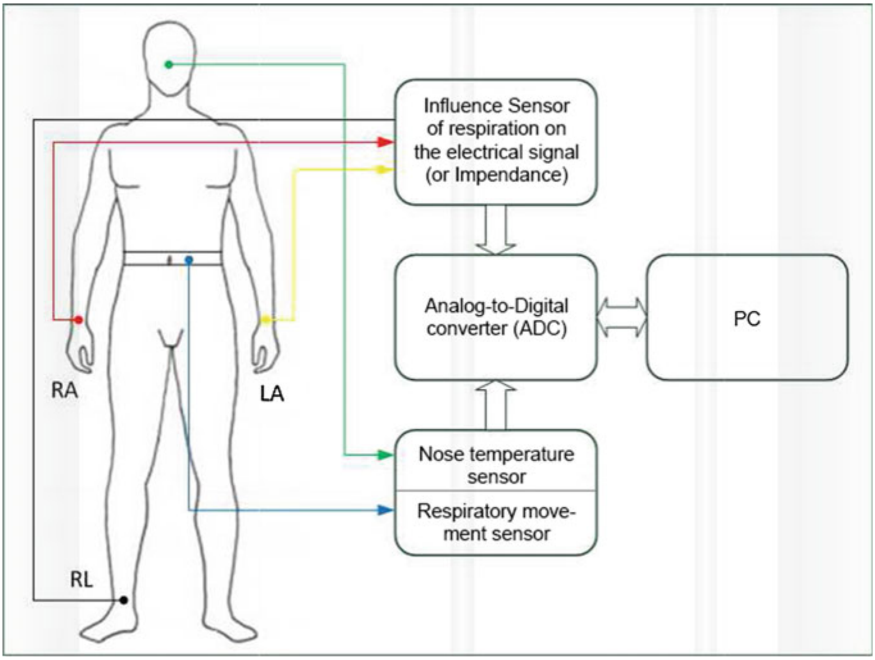


Fig. 1 Scheme of the proposed measurement chain

which could be mistaken tes for ambient air temperature (which can be mistaken for slow non-intensive inhalation or exhalation) [5].

A sensor for measuring such a signal have to have dimensions only in couple of millimeters, to not restrict a patient by its size. Its response to temperature changes must be fast and the sensitivity must be high at relatively small range of temperatures (about 20–50 °C). Additional requirements are high durability, low cost and good availability.

2.2 *Respiratory Movements Signal*

This biomechanical signal is significant symptom of respiratory activity. Especially during inspiration, which is always active process of respiratory muscles and therefore is well detectable even visually. Because expiration is not always active process of respiratory muscles detection of the signal is more complicated [6]. However usually during passive expiration at least relaxation of respiratory muscles, which leads to volume change of chest and belly, is detected.

The signal can be obtained by various methods, but in most cases these methods have one thing in common. Thus the sensor itself is placed on elastic belt or elastic cloth, which is surrounding human body and its respiratory movement. Sensor in this position usually doesn't limit the patient much, however the requirement of patient comfort has to be satisfied. If there are only "markers" on the elastic belt, then their location is usually determined optically. This system is due to high technical and financial difficulty unsuitable. More convenient is to use elastic belt with motion or volume-change sensors already built in. The main requirements of these sensors are high sensitivity, fast response, linearity and small dimensions. The price and availability is of course important too.

2.3 Description of the Measurement Chain

Several sensors were chosen for the respiration monitoring. First sensor is for electrical signal (or impedance) change measurement, second is temperature sensor for temperature difference measurement in the inlet (nose) area of the respiratory system and third is elastic belt measuring mechanical changes caused by breathing muscles.

All of these sensors are providing analog (continuous) signal, therefore it's necessary to digitize signals from each sensor for further processing. For this reason, a digital ADC converter is needed. This will represent the interface between the analog and the digital part of measurement chain.

It's important to realize that everything from the measurement chain will be placed in the room together the patient and the irradiator, thus in the shielded room and acquired data will be displayed outside of this room [7]. Most of the irradiation room has a connector panel which represents a data bridge to the control room. Therefore it's necessary to provide at least 10 m long cable which will lead from the ADC module to the connector panel.

3 Realization

The most important thing, in the realization of the patient respiration monitoring, is selection of suitable components, based on the proposed solution. Next step is interconnection all selected components to final prototype of measurement chain which could be tested and eventually used in practice.

3.1 Realization of the Nasal Temperature Signal Sensor

After examining of all possibilities, a thermistor as the sensor was chosen. Thermistor meets temperature range requirement as well as sufficient sensitivity and response speed. In a small temperature range is almost linear, it's very small and affordable in different designs.

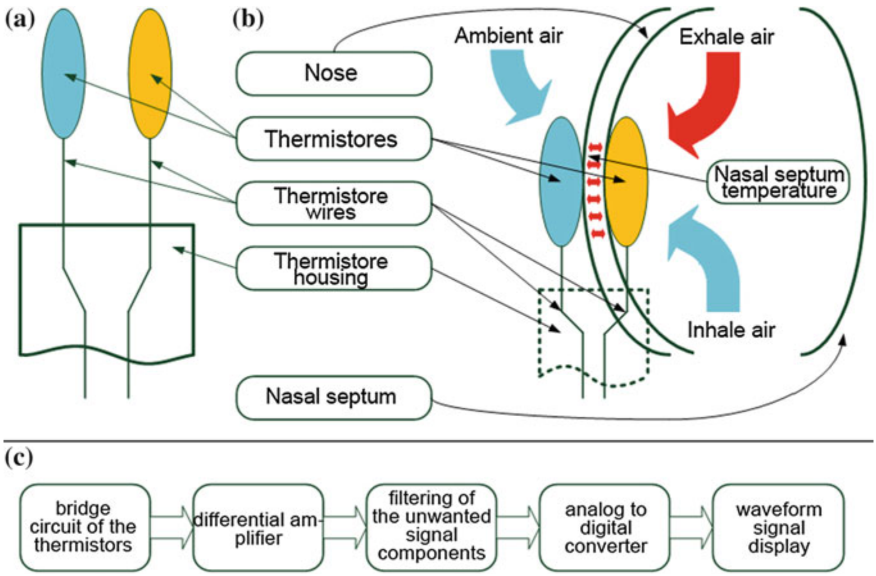


Fig. 2 Implementation location and nasal block diagram of the temperature sensor

The basic idea is to use two thermistors of the same electrical resistance (at 25 °C) in a bridge circuit. Thus, each thermistor is connected to the voltage divider and the voltage difference between the divider (in the bridge) represents temperature variation caused by respiration of the patient. One thermistor is located on the outer side of the nose and the other on the inner side (Fig. 2).

The temperature difference is caused by the flow of inhaled and exhaled air and its influence on the thermistor located on the inner side of the nose. Temperature differences between thermistors obviously cause voltage differences between their voltage dividers and those are then amplified and appropriately placed in the desired voltage range by differential amplifier. The output analog signal from the differential amplifier, before entering the analog-to-digital converter, passes through a passive lowpass filter of the first order, in order to remove possible undesired components. After that the signal is digitized by the analog-to-digital converter and appropriately displayed on a computer. (The circuit is shown on Fig. 2).

The differential voltage is amplified by the corresponding gain setting, and to it a reference voltage (signal shift on the Y axis) is added, which is also adjustable. The last part is a passive analog lowpass filter with critical frequency 16 Hz to eliminate unwanted high frequency signal components (Fig. 3).

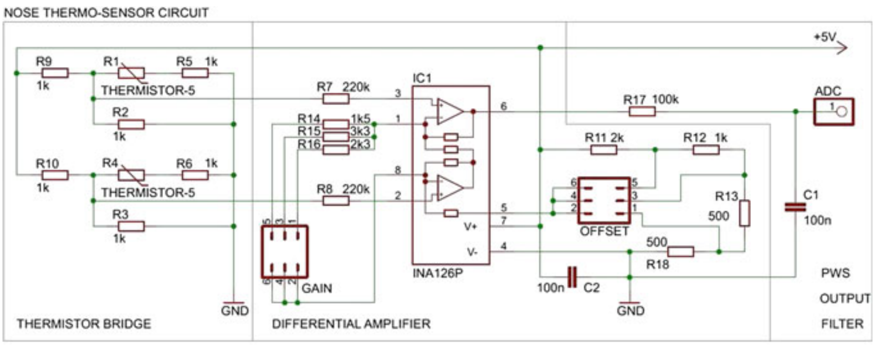


Fig. 3 Nose thermo-sensor circuit

3.2 Realization of the Respiratory Movement Sensors

Sensor selection of respiratory movements, like the selection of the temperature sensor, is based on sufficient fulfillment of the requirements for respiration monitoring equipment. The main element of such a sensor is an elastic strip serves as a support medium for the sensor itself. The mechanical tensions is measured between the two belt ends reproducing the patient’s body.

In this device a piezoelectric sensor is used and placed between the two ends of the elastic strip. Reasons are excellent properties of this kind of sensors such as the sensitivity, response speed, small size, linearity and a relatively high resistance to external influences. It is already functional piezoelectric strip from ADInstruments (MLT1132/D Piezo Respiratory Belt Transducer). The piezoelectric strip has an output range of 20–400 mV, the sensitivity of 4.5 ± 1 mV/mm and high input resistance. The length of the elastic belt is 300 mm (total length is adjustable) and can be stretched about 100 mm, width is 45 mm.

This part deals mainly with compatible wiring the above mentioned selected piezoelectric strip from ADInstruments with the ADC module and a suitable processing of the output signal of the band.

The basis for wiring the piezoelectric strip (ADinstruments) is the correct wiring connector compatible type DIN8, which is achieved through appropriate documentation for piezoelectric strip (Fig. 4). Like the nose temperature sensor is used connections passive lowpass filter of the first order with a critical frequency of 16 Hz to remove unwanted high frequency signal components, but in this case the filter is before the amplifier. Another part of the circuit is a non-inverting amplifier connections and the possibility of setting two amplification. Finally, as with nasal temperature sensor enters the amplified and filtered signal to an analog-to-digital converter ADC where it is digitized and then transferred to a computer for display (Fig. 5).

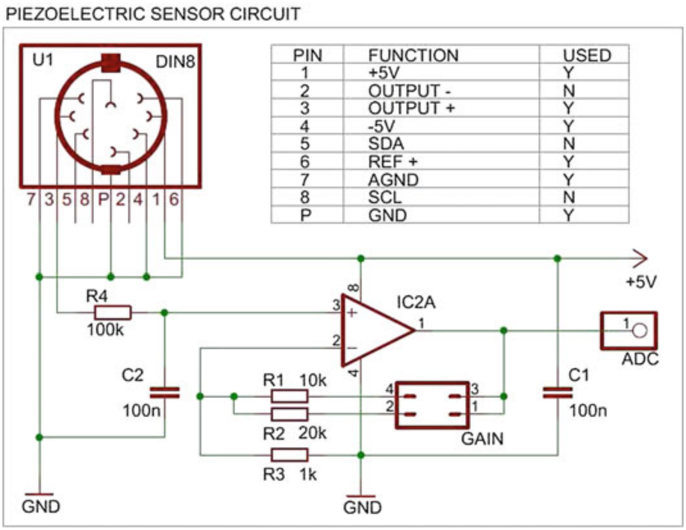


Fig. 4 Electrical connection of the piezoelectric strip, respectively. functional and compatible wiring connector for connecting DIN8 piezoelectric strip

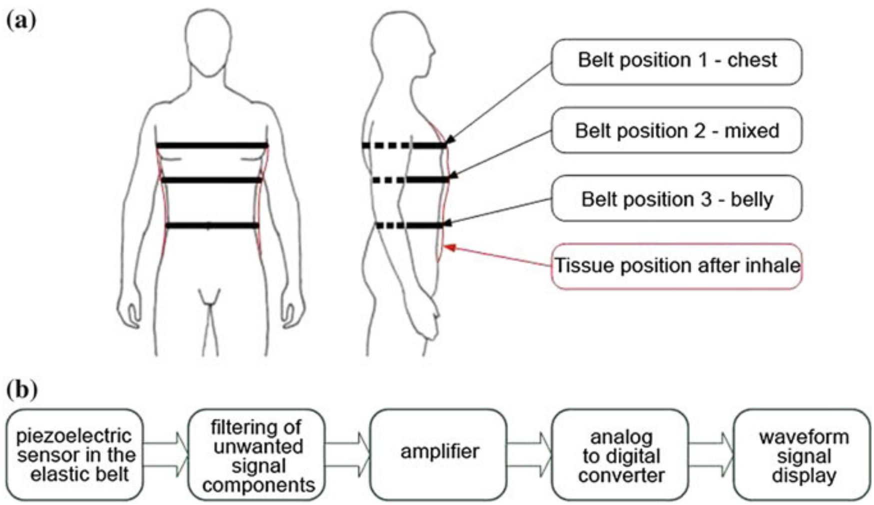


Fig. 5 Placement options (**a**) and block diagram (**b**) of the piezoelectric strip

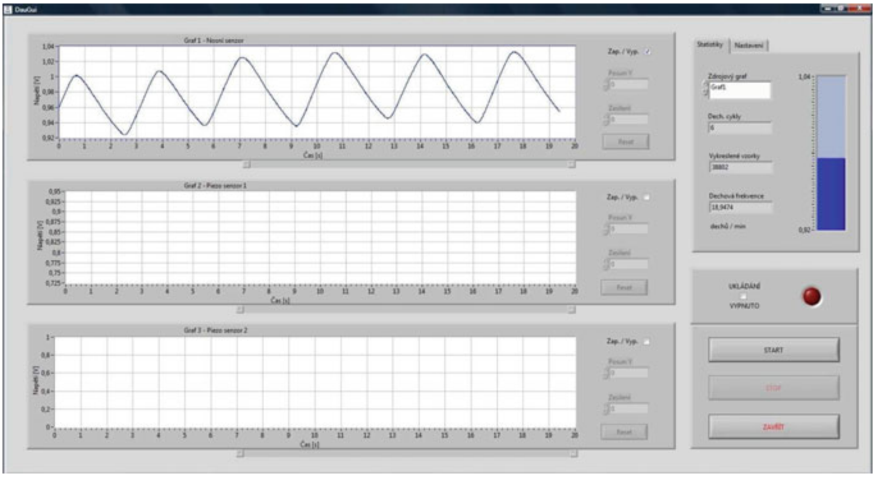


Fig. 6 Software interface

3.3 Software for the Analysis of the Measured Data

The software interface (Fig. 6) is realized by LabView. The software is subject to the following requirements:

- Simplicity, clarity and maximum speed for the signals.
- View the respiratory rate.
- Viewing trends signal amplitudes.
- Independence software on other systems.

4 Testing

All parts made by measuring chain have been tested mainly in the very course of their development, i.e. at the design stage and then mainly in the realization phase, when they were gradually debugging and testing. In this chapter the latest progress of the measurement chain and software is shown.

Parts of the test measuring chain are a module containing a circuit for piezoelectric strip and nasal temperature sensor, ADC module, piezoelectric strip, nasal temperature sensor and extension data cable simulating the distance between the output of the ADC module and a computer with installed software, which is the last link to the actual interpretation of the measured data. Additional module for sensing changes in impedance at the time of testing was not available, and therefore will not be included in this section.

Conditions during testing were virtually identical to those that would have been provided for testing in the laboratory, where a measured object was a test person.

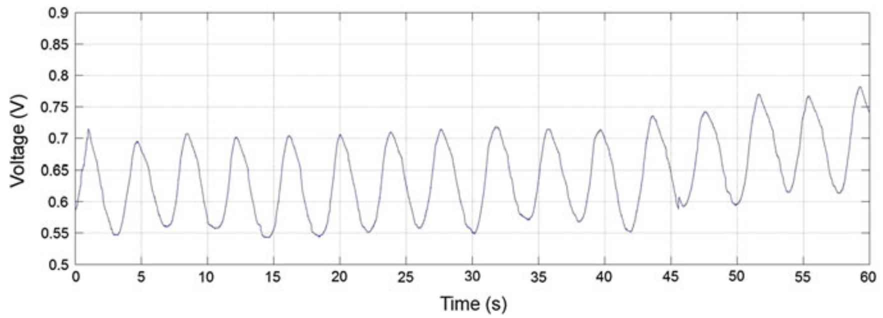


Fig. 7 Signal piezoelectric strip at a real gain 11

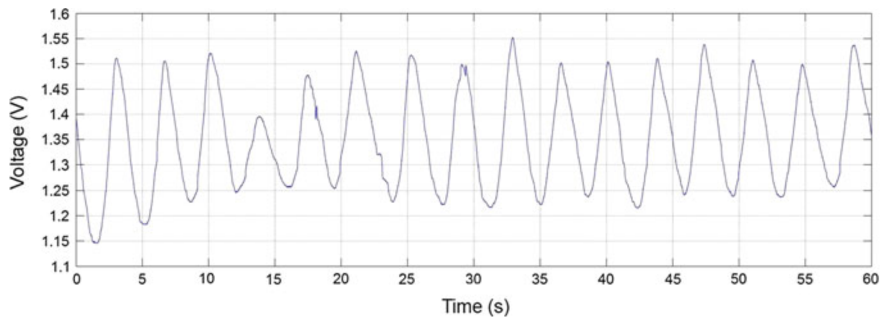


Fig. 8 Signal piezoelectric strip at a real gain 21

When selecting the gain 11 (Fig. 7) the signal is in the range of about 0.7 V, whereas when selecting the gain 21 (Fig. 8) the signal is approximately 1.5 V. In the measured progressions is also possible to observe small “vibrations” or “jumps” of the signal, which are probably caused by non-linear movement of the tissue while breathing and by structure of the belt itself. Results of measuring by the nasal temperature sensor can be seen on Figs. 9 and 10.

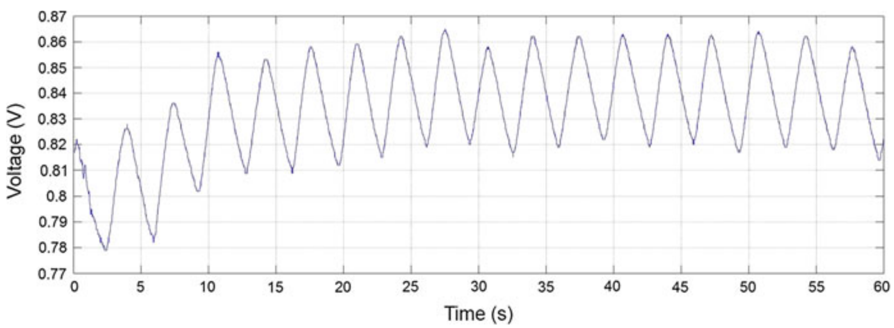


Fig. 9 Signal of nasal temperature sensor at 0.625 V offset and gain 30

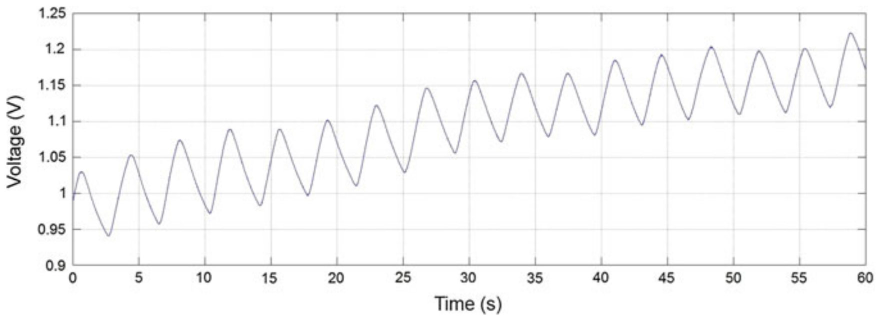


Fig. 10 Signal of nasal temperature sensor at 0.625 V offset and gain 60

5 Conclusion

The article deals with complex design of hardware and software solutions for monitoring of patient respiraton during radiotherapy. In terms of evaluation of the results of the work, namely evaluation of test results achieved by devices and software, it can be said that in order to monitor the state of respiration of the patient during the radiotherapy procedure is a combination of both the sensors desired result sufficient. Another part of the testing was also verify of some additional software features, such as detector breaths, which was further calculated to respiratory rate.

The last in a more indirect tested parameter was effect of data extension cable, the implementation of which there were concerns about the loss of intensity of the transmitted signal, because of the length management and conservation voltage levels for USB cables. However, these fears were refuted by testing, because without the extension cables, there was virtually no perceptible difference.

Since the proposed measurement chain is a unique solution and at this time there is no other similar instrument, it is not possible to compare the results with any other data. However the measurement chain is under continuous development and the results are still analyzed.

Acknowledgment The work and the contributions were supported by the project SP2015/179 'Biomedicínské inženýrské systémy XI', and this paper has been elaborated in the framework of the project "Support research and development in the Moravian-Silesian Region 2014 DT 1—Research Teams" (RRC/07/2014). Financed from the budget of the Moravian-Silesian Region. The paper was elaborated in the framework of the IT4Innovations Centre of Excellence project, reg. no. CZ.1.05/1.1.00/02.0070, supported by the Operational Programme Research and Development for Innovation,' funded by the Structural Funds of the European Union and the state budget of the Czech Republic.

References

1. Penhaker, M. Imramovský, M. Tiefenbach, P. Kobza, F.: Lékařské diagnostické přístroje: Učební texty. Ostrava: VŠB—Technická univerzita Ostrava (2004) 320 p. ISBN 80-248-0751-3
2. Spurný, V. Šlampa, P.: Moderní radioterapeutické metody: Základy radioterapie - VI. Brno: Institut pro další vzdělávání pracovníků ve zdravotnictví, 1999. 118 p. ISBN 80-7013-267-1
3. Trojan, S.: aj. Lékařská fyziologie. 4. vydání. Grada Publishing, Praha (2003), 772 p. ISBN 80-247-0512-5
4. Jirák, Z.: aj. Fyziologie pro bakalářské studium na FZS OU. Ostravská univerzita, Ostrava (2009), 250 p. ISBN 978-80-7368-234-7
5. Heřman, P.: Biosignály: z pohledu biofyziky. Praha: 2006. Available from: http://cs.wikisource.org/wiki/Biosign%C3%A1ly_z_pohledu_biofyziky#Biosign.C3.A1ly_v_organismu. ISBN 80-902899-7-5
6. Horáček, J.: Anatomie pro bakalářské studium se zdravotnickým zaměřením. VŠB – Technická univerzita Ostrava, Ostrava (2006), 113 p. ISBN 80-248-1203-7
7. Penhaker, M., Darebnikova, M., Cerny, M.: Sensor network for measurement and analysis on medical devices quality control. E-Technol. Netw. Dev. **171**, 182–196 (2011)

Intelligent Car Localization with the Use of Andruino Platform and Cloud Storage

Josef Horalek and Vladimir Sobeslav

Abstract Nowadays the possibility to localize a car is a frequently used method for completion of log-book, analysis of company car effectivity, or most importantly as a part of car security systems in the case of their theft. The proposed solution is designed on financially affordable components and platform-independent open-source software. Thanks to intelligent behaviour of the proposed solution which uses cloud storage and freely accessible Google Earth maps, the solution is competitive not only due to its low acquisition costs, but chiefly because of its operating costs and platform independence.

Keywords Arduino · GPS · GSM · Localization · Dropbox · Cloud · Google earth

1 Introduction

The issue of car localization has been the subject of research among many commercial as well as non-commercial subjects, as exemplified by [1]. The possibilities which are offered by modern data and localization technologies are to be found in [2]. The methods calculation car localizations based on GPS and GSM data are treated in [3], whose conclusions are used also in the present solution proposal. The subject of great potential for experimental development and prototype creation is location display in real time, where most significant results are presented in [4] and [5]. Thanks to the development of smart devices and their applications, it is relevant to adjust new solutions to be able to display the location on these smart devices,

J. Horalek · V. Sobeslav (✉)
Department of Information Technologies, Faculty of Informatics and Management,
University of Hradec Králové, Hradec Králové, Czech Republic
e-mail: vladimir.sobeslav@uhk.cz

J. Horalek
e-mail: josef.horalek@uhk.cz

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_67

be it phones of tablets, as manifest in [6]. The majority of available solutions that can be used have several drawbacks limiting their greater popularity. The first drawback is their platform dependency, which is nowadays a great disadvantage; therefore the proposed solution offers platform independency thanks to the use of Arduino microcontroller. [7–9], as well as the use of open-source software for the end-user. Another significant disadvantage of standard solutions is the use of commercially available map sources for car localization; our solution offers freely accessible Google Earth applications. The final significant innovation of our solution is its low price which is a result of the implementation of Arduino microcontroller and an edited GPS receiver, whose communication with the microcontroller via UART communication has been optimized.

2 New Solution Proposal

The actual innovative solution is composed of hardware and software parts, which will be introduced in further detail below.

2.1 Hardware Part

The hardware part represents the connection between individual electronic circuits and a controlling programme for the microprocessor. The connection of the proposed solution is composed of three independent components, described in Fig. 1. These are in particular:

- GPS receiver ensuring the reception of individual orbits location information and via NMEA sends the information in second intervals to the serial port.
- Microprocessor which receives on the serial port the data retrieved from GPS, checks their validity and through GSM communicator sends them as UDP packets to a remote server in set intervals.

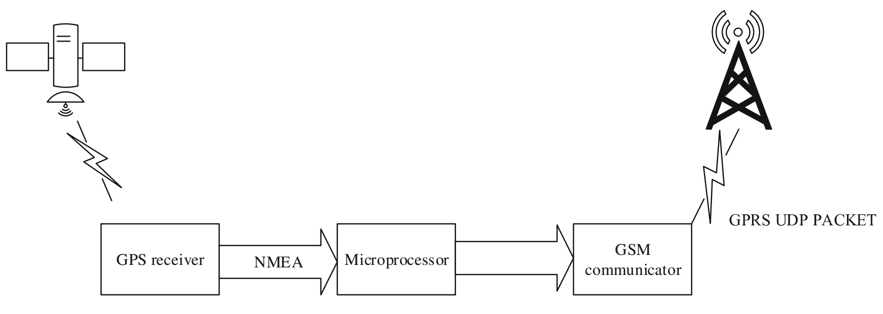


Fig. 1 Block scheme of hardware solutions

- GSM communicator then sends data obtained from the microprocessor through GRPS transmission.

2.1.1 UART Serial Link

The UART serial link (Universal Asynchronous Receiver and Transmitter) is used to transmit data between two devices. UART transmits on a pin labelled as TX (transmit) and receives on a pin RX (receive). The microprocessor ATMEGA 2560 planted on the Arduino Mega 2560 toolkit, which was used as a prototype for the proposed solution, uses ZV logic, which means that log. 0 equals the voltage from 0 to 0.8 V, log. 1 consequently 2–5 V. Although the GPS module uses the feed of mere 3.3 V, voltages for individual logic states are ensured. If there is no communication, the quiet mode of signal is log. 1. The transmission is initiated by the change of signal on log. 0 for the duration of one bit (called start-bit). The communication continues from the lowest data bit to the last, most significant data bit, which is followed by the stop bit, having the level of log. 1. After the stop-bit has been transmitted, the transmission on the new bit starts. Both communicating devices must have their communication speed set equally. Standard speeds are 1200, 2400, 4800, 9600, 19,200, 38,400, 57,600, 115,200 bps (bits per second).

2.1.2 Individual Proposal of System Component Interconnection

Complete control of the whole connection is provided by the microprocessor. The prototype uses the eight-bit microprocessor ATMEGA 2560 planted on Arduino Mega 2560 toolkit. The processor has four serial ports (UART), then 54 digital and 16 analogue inputs/outputs. The proposed solution uses two serial ports connected to a processor with GPS and PSM module. Considering the fact that all circuits are powered from the car battery, the solution is designed so as not to discharge the battery while the ignition is switched off. While the ignition is off, the powering of all circuits is terminated. After connecting the powering (car ignition in the ON position) circuits run automatically; only the GSM module remains in StandBy mode, therefore it must be initiated via a programme through the microprocessor by using one digital output, where log. 1 in the duration of 2.5 s is sent to the POWER GSM module input. The interconnection of individual components is exemplified in Fig. 2. To ensure functional operation, correct communication speeds must be set on serial channels. This is ensured by the processor programme, introduced below.

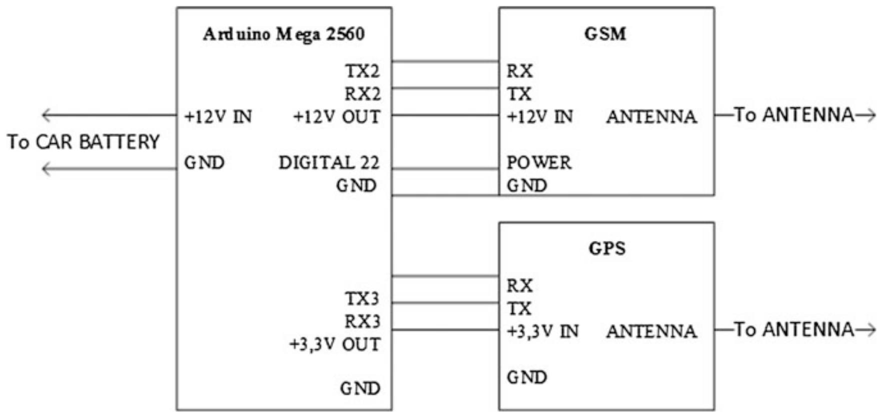


Fig. 2 Block scheme connections between components

2.1.3 GPS and GSM Module Operation

The GPS module prototype uses a custom-made GPS receiver planted with Fastrax IT300 module. The module was selected on a right price/performance relation as well as relatively simple programming and sufficient technical parameters. The GPS receiver contains voltage stabilization to required 3.3 V, aerial connector, external power pins, and a serial port. The most significant parameters of IT300 module are SiRFstarIII chipset GSC3e/LPX, low energy consumption (75 mW at 3.0 V), impressively high sensitivity of 159 dBm (Tracking), the support of NMEA a SiRF binary protocols. The operation of GPS module requires no special demands, only correct communication speed on serial processor need to be assured; in this case the speed is 4800 bps. When powering is connected, the GPS module initiates the transmission to a serial port, and individual NMEA and microprocessor provides data reception and processing.

The GSM module prototype uses developmental kit PGSM of Pandatron Company planted with M10 GSM module from Quectel. PGSM includes all basic supporting circuits for GSM module operation (SIM card slot, SMA aerial connector, Jack 3.5 mm connector for hands-free). The connection of the device is enabled via two standard pin ledges with 2.54 mm distance. The most significant of M10 planted module is the four-band version: 850/900/1800/1900 MHz, GPRS multi-slot: Class 12/10/8, GPRS mobile station Class B, power voltage range of 3.4–4.5 V (4.0 V type) and low-input operation 1.1 mA at DRX = 5 a 0.7 mA at DRX = 9.

The GSM module is controlled via AT commands transmitted to the module through the serial port, whose communication speed is 9600 bps. AT commands enable to set module parameters, initiate various actions (dial, send message, connect, send data etc.), or can request different module states (checking of network login, signal reception, unread messages or missed calls, etc.). Among the most significant and used AT commands belong:

- AT + CGREG—request of the registration state to GSM network. The response of the module is the following: +CGREG: <n>,<stat>[,<lac>,<ci>]. Stat is an important parameter which can range from 0 to 5 and signals the state of the GSM network login.
  - (0) SIM is not registered into any network and the module seeks no login operators,
  - (1) SIM je registered in domestic network,
  - (2) SIM is not registered in any network and the module is currently seeking login operators,
  - (3) Registration was denied,
  - (4) Unknown state,
  - (5) SIM is registered in a new network—roaming.
- AT+QIFGCNT—(initial) connection initialization, data carrier setting on CSD or GPRS.
- AT+QICSPG=1, "title APN"—Selects GPRS mode and sets the name of access point for internet connection.
- AT+QIMUX=0—restricts MUXIP functions—the connection will be initiated only with one server, not with more simultaneously.
- AT+QIMODE=0—sets non-transparent mode of data transmission, enables the protocol for error corrections.
- AT+QIREGAPP—registers TCP/IP architecture. The implementation of TCP/IP architecture.
- AT+QIACT—activates selected context.
- AT+QIOPEN= "connection type" IP address, port—initiates connection with a remote server, identifies connection type (TCP/UDP), IP address of the remote server and the port where the remote server "listens."
- AT+QISEND—sends data to the remote server.

2.1.4 Main Methods of Microprocessor Controlling Programme

This part introduces main methods of microprocessor controlling programme, which were written in Wiring programming language. The aim is to secure the reception of coordinates from the GPS module, checks the coordinates validity and consequently send the data to the remote server for further processing.

- *Setup()* method—the method executes all initial settings. The transmission speed of the serial port where the GSM module is connected is set to 9600 bps. The speed of the serial port where the GPS receiver is connected is set to 4800 bps. Furthermore, pin *PwrOn* serving to initiate the GMS module as an output is set. Also the methods of *Pwr()*, S*etConection() a SendStart()* are called.
- *Loop()* method—instructions within this method are executed repeatedly; it is in fact the main programme initiating other methods. If data available in the buffer of serial port 3 (they are always available unless the GPS module malfunctions),

the method checks every received sign saved in variable *inByte*, whether it is not "$" (sentence beginning). Then it reads other individual signs until "LF" (sentence ending), which it adds into variable *data*. As soon as the sentence is complete, it executes check-back to ensure it is not an RMC sentence. If so, then it checks coordinate validity and if the coordinate is valid, it adds one to RMC sentence meter (*count* variable) and then proceeds again from the beginning. Once variable *count* reaches the value of 15, it sends the last RMC sentence, stored in *data* variable, via the command of AT+QISEND to a remote server. The following represents a part of the code reading the whole sentence from "$" to "LF" and records variable data.

If the sentence starts with "$GPRMC" and includes valid data, it increments the meter and sends each fifteenth sentence to the server.

- *Pwr()*—the method is responsible for the initiation of the GSM module and the check of GSM network login. The method sets digital pin nr. 22 to log. 1 for the duration of 2.5 s. This pulse driven to POWER of the GSM module secures its initiation. Consequently the processor waits for 15 s for SIM logging into GSM network. Then it transmits command AT+CGREG? to the GSM module via serial port nr. 2 and reads the response form the GSM module. This is repeated in 150 ms intervals, until SIM card logs in home or foreign network (n = 1, or n = 5).
- *SetConection()*—the method establishes connection between GRPS and a remote server. After sending AT commands sequence, the method verifies if the connection was successful. If not, the commands are repeated until successful connection with the server is achieved. Once connected, the word "Start" is sent to the server, which signalizes the beginning of a new route (it was proceeded by switching the ignition in the OFF position).
- *Read()*—this method read responses from the GSM module. When data in serial port buffer are available, the method reads individual signs and connects them into resulting variable r*dData* and into variable 1, where the length of data read is stored.

2.2 Software Solution Proposal

This part of the article presents the software part of the proposed solution. The coordinates sent by the module placed in a car, as described in previous chapters, need to be processed and prepare for being displayed in Google Earth. Cloud storage Dropbox is used for storing data considering they are available from different sources [10]. The reception and processing of the data retrieved from the car is provided by a recently designed Java application which runs on a computer with a public IP address, functioning as a data-processing server. The programme consists of two parts. The first is graphic user interface (GUI); the second is the actual data-processing part. GUI serves for users to set parameters important for the application running. The second part secures the actual data reception, their conversion into *.kml format and their storage to Dropbox.

2.2.1 Graphic User Interface (GUI)

Graphic user interface is created via Swing library, because with its help it is possible to create windows, dialogues, buttons, frames, roll-up lists, etc. in Java script. Class hierarchy of Swing graphic components is based on AWT (Abstract Window Toolkit), which has been integrated in Java SE since its 1.2 version. The actual application uses JFrame class, which is a basic Swing container for the creation of titled application windows, then Jpanel, where other components are placed. BorderLayout is used to organize the components within the container into five areas. The application uses NORTH area to display information concerning the author; WEST area for path setting to saved files; CENTER area for FTP server setting parameters; SOUTH area contains the button for server activation. Each area contains a JPanel, where other components like JLabel are placed. JLabel serves to display static texts in window and complete texts to unlabelled components. JTextField creates a textual component serving as a single-line input field. JCheckBox stands for ticking boxes and JButton for a button which is set as so-called listener, which after clicking activates the method. The path to the file with current coordinates can be set either manually or by using JFileChooser, after clicking the Browse button. JFileChooser enables GUI to move in the file system and subsequently to select a file or a folder. After the application is run/activated/started, it is necessary to set the path to the file nmea.kml, which contains the information of the current location of the car. While saving the file to FTP server it is necessary to tick the corresponding checkbox and complete the connection parameters. The application is subsequently activated by clicking the Start button.

2.2.2 Data Reception and Processing

The actual programme is divided into several methods securing individual elements: UDP packer reception, retrieved data conversion into *.kml file and its saving, editing well as saving the file with route information, saving onto FTP if allowed.

- *udpComn()*—method ensures reception of UDP packets sent from the GSM module in a car. The method uses DatagramChannel class, which can send, receive and listen to datagrams on a certain UDP port. After their reception, the data are stored in buffer and then read by signs and saved into gps variable.
- *gpsConv()*—data received via the previous method creates NMEA theorem where individual information pieces are divided by commas. The method loads

data between separators and edits them for storing in nmea.kml files. The method further recalculates location format from “Deg° Min” to “Degrees,” speed from knots to kilometres per hour, and provides separators with date and time information.

- *nmeaRewrite()*—the method corrects data in nmea.kml file after their conversion by the gpsConv() method. The content of nmea.kml file is progressively loaded via FileInputStream class and saved into the variable of String type.

Subsequently the position of tags is located <coordinates> </coordinates>, among which the current GPS coordinates of the car are saved. The information concerning date and time of retrieved coordinate and current speed of the car is saved between tags <name></name>.

After the correction of method file content, FileOutputStream class rewrites the content of nmea.kml file.

- *roadRewrite()*—the application records both the current car location and its route. While activating the ignition, the GSM module in the car sends the word “Start.” If the runServer() method locates presence of the data in buffer, it establishes new file whose name comprises of current date and time. Additionally, its first coordinate is provided with the label Start to designate where the route started. The file then records each incoming coordinate until another “Start” word arrives. Similarly to nemeaRewrite(), the method edits road.kml by progressively recording route coordinates of the car.
- *sendFtp()*—if it is required to save the file nmea.kml to FTP server, the method initiates the connection with corresponding server and saves the file nmea.kml via FTPClient class.
- *runServer()*—the application records both the current car location and its route. While activating the ignition, the GSM module in the car sends the word “Start”. If the method locates presence of the data in buffer, it establishes new file whose name comprises of the current date and time. Until the arrival of another “Start,” the method saves each received coordinate in the file by calling the above-mentioned methods.

3 Cloud Service, Dropbox and Google Earth Use

One of the basic requirements for our solution was availability of the coordinate file from various platforms and from more devices at the same time. Therefore cloud storage was selected. GoogleDrive, Copy.com and Dropbox storages were tested. In the long-term Dropbox was a solution that offered highest synchronization speed and is also platform independent, therefore was eventually selected for our solution. The system is programmed to enable all abovementioned solutions based on each user's preferences. Dropbox client enables to load any file into a synchronized folder; thereby the file is saved into the web storage that is available to all synchronised devices. Users can also send data manually—directly via web interface or web browser [11, 12].

To display the current location of the car the Google Earth software is used. This programme was selected due to its complete worldwide map coverage and its freeware version availability. The programme handles online processing of coordinate files as well as automatized location update based on the content change of the monitored file. Another feature is the ability to display individual routes saved in the *.kml format. Google Earth is also platform independent. For the full functioning of the proposed solution it is necessary to set regular check of the file content and move over a map on the current position during its change. The delays between content update can be set based don users' requirements. The testing proved the most ideal delay value was of 15 s. In this setting the map automatically moves to the current location after applying changes in the nmea.kml by the Java application.

4 Application Testing

For testing purposes of the HW application part the Arduino toolkit was connected to a computer serial port, and via nTerminal programme communication between microprocessor and individual modules was monitored. While tuning the programme for microprocessor, a code capable of sending variables to serial ports was implemented in the programme. The code was displayed via the abovementioned nTerminal programme. During testing of the SW application part, HW module was substituted by programmes Packet Sender for PC and UDP sender for Android platform. Both programmes enable to send UDP packets to a particular IP address and port. While tuning of the application they simulated sending UDP packets by the GSM module. For the application testing purposes the unit with the GPS receiver and the GSM module was placed in four cars and for a period of five months was monitored their location and their routes recorded. The location and route display in a mobile phone is exemplified in Fig. 3.

When data was collected, blackouts were determined; they were caused by server connection failures resulting from weak or missing signal connection on GSM network. Based on these findings the controlling programme of microprocessor was adjusted to re-establish the connection once GSM signal is retrieved after a blackout. Blackouts stopped appearing after these adjustments had been implemented.

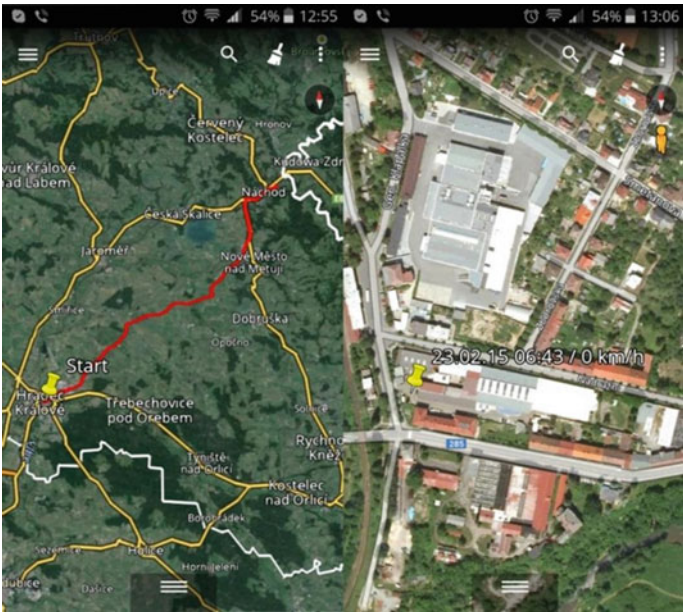


Fig. 3 View the route and the current position in the mobile app

5 Conclusions

The aim of the proposed solution was to create a platform-independent and cheap solution of car localization using a GPS module, data transmission via GPS and securing multiple accesses to data and car location using cloud solution. The proposed innovative solution is designed to function independently on expensive map basis, but uses freely accessible application Google Earth, which provides sufficiently accurate worldwide map coverage. Thanks to the Dropbox platform saved data are accessible anywhere and anytime, and enable further processing possibility by other software to create e.g. detailed route catalogues for company car use efficiency calculations. Furthermore, the proposed solution can be used for car localization in the case of its theft.

Acknowledgement Support of the Specific Research Project of FIM UHK “SCM, Control of Markets and Production in Agent-based Computational Economics” is gratefully acknowledged.

References

1. Ishikawa, K., Amano, Y., Hashizume, T., Takiguchi, J.I.: A study of precise road feature localization using mobile mapping system. In: IEEE/ASME International Conference on Advanced Intelligent Mechatronics. IEEE, pp. 1–6 (2007). doi:10.1109/AIM.2007.4412541
2. Luettel, T., Himmelsbach, M., Wuensche, H.J., Takiguchi, J.I.: Autonomous ground vehicles—concepts and a path to the future. In: Proceedings of the IEEE 2012, Special Centennial Issue, vol. 100, pp. 1831–1839 (2012). doi:10.1109/JPROC.2012.2189803
3. Najjar, E.L.M.E., Bonnifait, P., Philippe, H.J., Wuenschea, H.J., Takiguchi, J.I.: A road-matching method for precise vehicle localization using belief theory and kalman filtering. Auton. Robots IEEE **19**(2), 173–191 (2005). doi:10.1007/s10514-005-0609-1
4. Wang, J., Cho, J., Lee, S., Ma, T.: Real time services for future cloud computing enabled vehicle networks. In: International Conference on Wireless Communications and Signal Processing (WCSP). IEEE, 2011, vol. 19, issue 2, pp. 1–5 (2011). doi:10.1109/WCSP.2011.6096957
5. Liao, L., Patterson, D.J., Fox, D., Kautz, A.H.: Building personal maps from GPS data. Ann N. Y. Acad. Sci. IEEE **1093**(1), 249–265 (2006). doi:10.1196/annals.1382.017
6. Lee, SJ, Tewolde, G., Kwon, J., Kautzm H.: Design and implementation of vehicle tracking system using GPS/GSM/GPRS technology and smartphone application. In: IEEE World Forum on Internet of Things (WF-IoT). IEEE, 2014, vol. 1093, issue 1, pp. 353–358 (2014). doi:10.1109/WF-IoT.2014.6803187
7. Macdougall, J., Tewolde, G.S., Kwon, J., Kautz, H.: Tour guide robot using wireless based localization. In: IEEE International Conference on Electro-Information Technology, EIT 2013. IEEE, 2013, vol. 1093, issue 1, pp. 1–6 (2013). doi:10.1109/EIT.2013.6632690
8. Perez, M.S., Carrera, E.V., Kwon, J., Kautz, H.: Acoustic event localization on an Arduino-based wireless sensor network. In: IEEE Latin-America Conference on Communications (LATINCOM). IEEE, 2014, vol. 1093, issue 1, pp. 1–6 (2014). doi:10.1109/LATINCOM.2014.7041861
9. Loore, C., Carrera, E.V., Kwon, J., Kautz, H.: Ultraviolet observations of the be star and X-ray binary 4U 1145–61 (= HD 102567 = Hen 715) obtained with the IUE. IEEE Latin-America Conference on Communications (LATINCOM). IEEE, 2014, vol. 1093, issue 1, pp. 1–6 (2014). doi:10.1145/2536853.2536898
10. Maresova, P., Klimova, B.: Investment evaluation of cloud computing in the European business sector. Appl. Econ. **47**(36), 3907–3920 (2015)
11. Benikovsky, J., Brida, P., Machaj, J.: Proposal of user adaptive modular localization system for ubiquitous positioning. Lect. Notes Artif. Intell. **7197**, 391–400 (2012)
12. Mlynka, M., Brida, P., Machaj, J.: Modular positioning system for intelligent transport. In: 5th International Conference on Computational Collective Intelligence Technologies and Applications, Craiova, Romania, Studies in Computational Intelligence 513, vol. 513, pp. 115–124 (2014)

Intelligent Heating Regulation

Josef Horalek and Vladimir Sobeslav

Abstract Heating through own boiler is very widespread and popular. However, most of heating systems do not solve or solve inefficiently the heating regulation. Mostly, a residential unit contains one central thermostat controlling the boiler but cannot set different temperatures in individual rooms. Possible solution is to measure the temperature in every room and control valves on individual radiators. These solutions use proprietary communication protocols and they are unreasonably expensive. Controlling algorithm is fixed without a possibility of an adjustment. By the design and realization of own regulation we gain not only financial savings, but we mainly improve client's convenience. This article represents possibilities of smart heating solution using Arduino platform that provides enough performance and also supports effective programming and open source solution utilization.

Keywords Arduino · Temperature regulation · Home automation · Smart metering

1 Introduction

Currently the central heating is one of the most widespread heating means. Its principle is simple: every flat in a house has its own boiler. It can be for example coal boiler, but mostly it is a gas boiler. The heating system in flat consists of a central boiler and radiator system distributed throughout the flat. Every residential room has at least one radiator; the maximum number of radiators is not limited in

J. Horalek · V. Sobeslav (✉)
Department of Information Technologies, Faculty of Informatics and Management, University of Hradec Králové, Hradec Králové, Czech Republic
e-mail: vladimir.sobeslav@uhk.cz

J. Horalek
e-mail: josef.horalek@uhk.cz

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_68

theory, but usually there are not more than three in one flat [1]. To control the whole system, one central thermostat is mostly used. Its connection to the boiler (considering gas boiler) is realized through two wires. Earlier boilers were using alternating voltage of 230 V, then there was one more, protective conductor in the wire. Modern boilers work with direct voltage of 12 V. Additionally, in the case of cheap boilers, there is no use of any smart communication bus. If the thermostat processes the need to heat through its sensors, it just connects both wires leading to the boiler. When the thermostat evaluates enough temperature height, it disconnects wires and so it shuts down the boiler. Probably all boilers not older than ten years make evaluation of water temperature returning from radiator system to the boiler through their own built-in microprocessor in control unit. If it is high above temperature set on the boiler, they stop heating automatically. If there is permanent wire connection, nothing but overheating happens.

The ideal solution would be to use equithermic regulation same as on agents level introduced by [2] and [3]. Within this type of regulation the temperature of water in the system is changed. The lower temperature is required for boiler to heat water, the lower energy it consumes. The regulation is given by equithermic curves that describe mutual dependence of outside temperature and heating water temperature that is required to reach required temperature in rooms. For equithermic regulation it is necessary to have a possibility of electronic control of water temperature on the boiler. Cheap boilers have only a mechanical potentiometer. Although it would be possible to exchange the mechanical potentiometer on a boiler control board for a digital one, but the boiler would not be approved by a revision check and could be dangerous then. If we cannot ensure different temperature in every room through the regulation, a client will not feel comfortable and it will mean high energy costs. Existing accessible solutions does not solve this problem at all, or they rely on open doors through all rooms, so one heating unit, or they are very expensive and usually based on commercial manner.

If an automation system is not based on free software and solutions and it uses proprietary protocols, it is not possible to make a system change easily. We expose ourselves to a risk that a company stops selling the system after some time and so technical support ends. If any element breaks itself then, it will not be possible to buy new one instead of it and we will have to change whole regulation system. This entails construction works and also it might be necessary to change valve elements on all radiators. It is necessary to shut down the whole heating system.

Therefore, a target should be scanning temperature individually in every room [4] and through own control unit consistently evaluate if the room has reached required temperature. If it does, the unit has to be able to shut down the heating in this room. It is possible practically through only one way—by closing an access of heating medium to heating element [5].

2 Problem Definition

The most widely used system is models with central boiler, central thermostat and heating elements system. This centralized system is from its principle bad solution, because whole heating unit (flat and house) is controlled by the temperature in one room. If we have model scheme of heating unit with two rooms, it generally comes a situation, when in room A we want to have 20 °C and in room B 22 °C. While using mechanically settable radiator valve it is possible through testing to define right position for keeping temperature in empty rooms. Once there are one or more people in any room, the whole regulation stops working. It is known that a human delivers heat corresponding to 100 W bulb. If there are five people and some home appliance turned on in the room with the thermostat, they will produce a relatively large amount of heat. This will cause the boiler to shut down and in other rooms will not be heated regardless to the required temperature [6].

Two solutions are possible. The first solution lies in opening doors through all rooms and so the temperature equalizes. It is clear that this solution is very uneconomical, if there are unoccupied rooms (for example a hall) unnecessarily heated. Furthermore, client's convenience is important when the target state is to have the same temperature in the whole apartment. The second solution lies in the possibility of individual control of every room and every radiator. To be able to close and open the radiator again, there has to be an actuator with an electronically controlled valve on it [7]. Electronically controlled valves are differed to two basic types according to the way of closing the radiator valve. The first one is the valve powered by electricity and thermoelectric valves.

Electricity-powered valves contain small electromotor and small transmission through which the valve piston is squeezed. As their disadvantage there remains a limited service life of transfers that are usually from plastic. While changing the state they produce noise that can bother the client, especially in a case of bedroom. An advantage is a power consumption that is necessary only for the state change. Valves then hold the state without permanent power consumption.

Autonomous valves are powered by an electromotor. They scan the temperature directly in valve element that means several centimetres from radiator that makes them inaccurate. They have a display and buttons through which the client can limitedly set required temperature and timetable. The power is provided by batteries. Thermoelectric valves use same principle as classical mechanical thermostatic valves. They have an element reacting on heat inside. Once the element is heated, it expands itself and so squeeze valve piston. Whole process is absolutely without a noise. The disadvantage is electric power consumption for holding the state. There has to flow power through the heating element constantly. It is much smaller than during/in the case of state change, but it still exists. Consumption is usually under 2 W per one valve, it differs according to the manufacturer. Another important selecting criterion is valve operating voltage. There are two types produced: 230 V AC and 24 V AC/DC. 230 V valves require mains voltage. Therefore their installation can be performed only by someone with particular certification and there is dangerous voltage wiring in the flat. An advantage is simplicity and presence of 230 V in every room [8]. 24 V version is incomparably safer than 230 V version. Valves from most manufacturers can usually work with both, alternating voltage and direct voltage. Control voltage is often transferred directly from the control unit, thou there occurs relatively great loss along such low voltage. After bringing the voltage the valve can open or close the radiator. They are so-called NC (Normal Close) and NO (Normal Open) types. As it is clear from the name, NC type is closed without the voltage. NO is opened without the voltage. Most widely used are NC models, because radiator opens himself only in the case it has to heat. Otherwise it does not consume any energy for the valve. Very important parameter when selecting the valve is its connection. Currently it is most widespread the connection through metric thread M30 × 1.5. Usually it is necessary to use valves supported by a manufacturer, or a reduction. There are many existing solutions, but each has some weakness. Mostly they are a high price or commercial (unfree) solution. Wireless regulation system HomeControl works on a principle of wireless communication of individual valves and a central unit. System uses autonomous valves with measurement straight on the radiator and so it is more inaccurate. A set for four rooms can be bought for a price of approximately $175.

Regulation from the Czech company Jablotron requires besides individual valves central, the control unit for a price of around $240, which is very expensive. Wireless intelligent regulation system Fibaro uses for communication between individual elements wireless bus Z-Wave. The cheapest central control unit ranges around $320. It can communicate with a great number of other sensors in whole apartment, but if we want only heating regulation, its price is unbearably expensive. Wireless regulation system PocketHome has the cheapest central control unit, but the problem is in expensive control valves.

3 New Solution

New solution uses individual control of all heating elements in a flat and scanning the temperature in every room. Equitherm regulation is not possible, because no unprofessional gas boiler intervention is possible. After the analysis of control valves we concluded that the most suitable valve is thermoelectric one that does not contain any mechanical elements. Suitable voltage is 24 V, because then it is possible to transfer also a power from the central unit. Used wire will be UTP type, so it can be powered by passive PoE (Power Over Ethernet). From the logical and technical point of view it is not a problem to use 230 V AC valves and 24 V AC/DC valves of both, NC type and NO type. It is enough to modify firmware in the central control unit.

To enable evaluation for closing radiators, central unit has to know exact temperatures in rooms. For a higher accuracy there can be greater number of thermometers in every room from which central unit can count the average value. For temperature measurement there can be used more types of sensors. The cheapest temperature sensor is thermistor that changes its resistance according to the surrounding temperature. There is an NTC version based on a principle of lowering the value of electrical resistance while lowering temperature and PTC version [9, 10] while increasing temperature its resistance rises as well. For these needs it is always necessary to know precise characteristic of used thermistor. Then, the biggest problem is operating with analogue value. Once we use even of few centimetres longer or shorter power cord, we immediately loose required accuracy. Therefore it is necessary to have thermistor including the cable calibrated [11]. The second variant is the temperature sensor with analogue output, for example LM35. It gives the temperature recalculated to voltage. 1 °C equalls 10 mV. For measuring then it serves well general voltmeter, when we measure 20 °C as 200 mV. The disadvantage is again a requirement of calibrated cable. The advantage is direct temperature reading along which we do not care about inner characteristic of the temperature sensor. The best solution is temperature sensor communicating on digital bus that brings several advantages. Most importantly, we do not have to make cable calibration from central unit to temperature sensor. Most buses' communication is protected by a checksum mechanism on transferred frame that ensures the detection of wrongly transferred values. Due to the bus characteristics, it is possible to connect on one cable (bus) greater number of temperature sensors. The most well-known is digital temperature sensor Dallas DS18B20.

The control unit acts as a brain of a whole heating system. It should be powerful enough for whole system running, trouble-free, cheap and easily programmable.

Because it is not necessary to log current data constantly, we will use control board with a single-chip microcontroller. It contains whole computer including RAM, ROM, processor and other supporting circuits in one chip. It would be possible to make distant logging and have an opportunity to change algorithm values for heating in real time, but then it would be necessary to use board on classical ×86, or ARM architecture. But that would make the whole solution much more expensive and complicated. A failure rate would arise a lot—single-chip microcomputer with unchanging program running in one single thread has generally lesser chance of breakage compared with common ×86 PC on Intel platform and common operating system Microsoft Windows, or Linux. The control unit will use a simple algorithm. If the temperature measured in a room is smaller than the client required, it opens the radiator. If boiler does not run, it turns it on. It controls all rooms like this. If at least one radiator is opened, it always turns on boiler. Reversed process is the same. If all radiators are closed already, the unit turns of boiler. We solve central unit performance part through a relay. Every valve on every radiator is switched through its own relay. The boiler has also own relay through which comes up a possibility of using control valves various models. We do not care about required 230, 24 V voltages, because they are switched by a relay. The cable distribution in every room of the heated unit is necessary to be prepared. Individual power lines have to lead to every heating unit and control the valve. It would be possible to use the common ground, but because of the performance there are used power wires for every heating element in our design. Digital thermometers communicate on bus and for its operation they need only two power wires. They are all connected on this bus and due to their communication protocol they are able to communicate with the central control unit. It is necessary to pay attention to the cross-section of the cable leading to the valve on the radiator. There must be no large voltage drop. Thermometer cable is much more complicated. Every bus has defined requirements on the maximal length, capacity, cable type and the maximum number of simultaneously connected devices directly in its own standard. For example, within bus I2C the bus maximal length is limited by a cable that cannot exceed capacity of 400 pF. Branching is not arbitrary, some buses enable star topology (Ethernet), others, on the other hand, bus topology (RS-485).

4 Solution Implementation

For the model flat with two rooms, each containing a heating element realization, we needed two electrically controlled valves, central control unit and connecting cables. Because our testing rooms were small and we did not require extreme accuracy, we placed to each of the rooms in flat only one temperature sensor. If greater accuracy is needed, it is possible to add more temperature sensors additionally and adjust firmware in the central control unit. Because valves powered by electromotor are noisy and do not have such service life, valves powered by thermoelectricity were used. Controlling was carried out through stretched element heating. From the security reason valves were connected to 24 V DC. Used type was NC (Normal Close, so closed without voltage). Due to this fact it was necessary to keep valve powered only for the time needed to heat the room to required temperature. To every radiator there was led individual 24 V cable. These cables were connected to relay board within the central control unit. To the third relay we connect boiler instead of original thermostat. Boiler will be galvanically separated from control unit, so we will seamlessly meet security standards. As temperature sensors it is theoretically possible to use various types. Thermistors were rejected for testing, because it is necessary to know their characteristics and it is necessary to count cable resistance. The first sensor LM35 we used has accuracy of ±0.75 °C [12] that is for the purpose of heating regulation sufficient. The problem is reading values that we do through reading the voltage value Fig. 1. Again it is necessary to

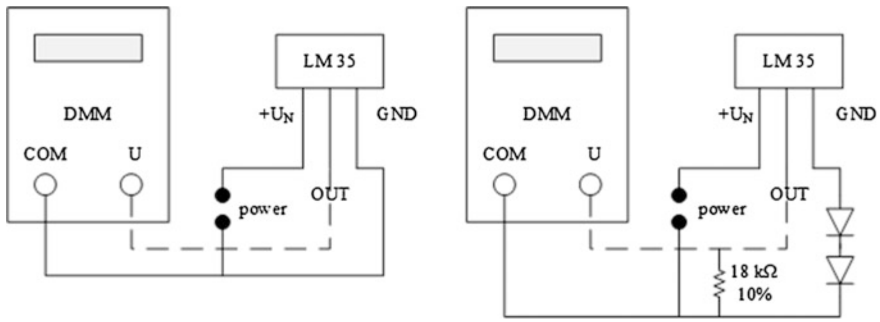


Fig. 1 Scheme of temperature measurement using a digital multimeter LM 35

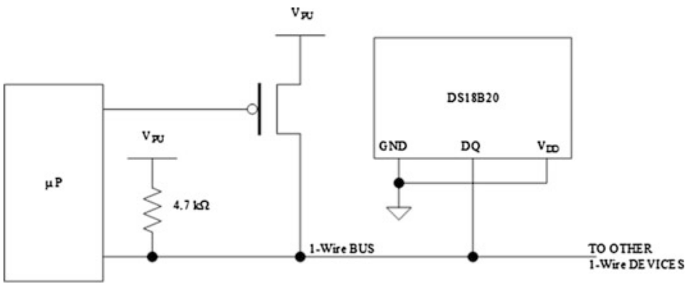


Fig. 2 Supplying the parasite-powered DS18B20 during temperature conversions

make sensors calibration in every room and count cable to the measurement. To every sensor it is necessary to have attached an individual data conductor. Despite easy measuring we rejected LM35 sensors after long-term testing from reasons of its problematic data reading.

As the best affordable solution within the implementation appears itself temperature sensor DS18B20. Sensors DS18B20 communicate with central control unit through 1-Wire bus, as seen in Fig. 2. Due to the bus it is possible to have only one cable distributed through the flat and connect sensors in parallel to it. 1-Wire bus contains always one master device and from one to n slave devices.

Communication is always started by master and slave devices responds to it. Operation on bus is divided to timeslots. Bus is much slower in comparison to I2C and it is not suitable for great data transfers. However, it is perfectly suitable for reading values from various sensors (Fig. 3).

For unambiguous identification every 1-Wire device contains worldwide unique 64 bit serial number (ID). Because DS18B20 has very small power consumption, it is possible to perform it in so-called “parasitic mode.” We attach only ground and data conductor to sensors. Power (+) is transferred through the data conductor. All sensors have built-in miniature condenser that charges itself in a case of logical 1 on

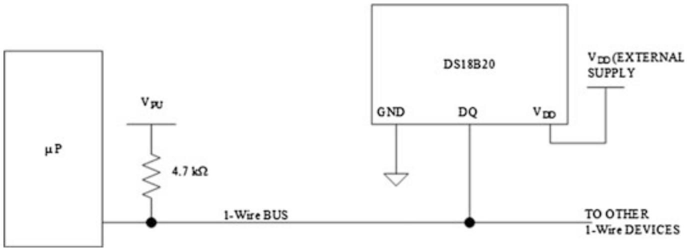


Fig. 3 Powering the DS18B20 with an external supply

a bus and from it transmissions are then realized. Within greater distances it is possible to use three-wire connection plus, ground and data.

As a central control unit we chose Arduino, concretely Arduino Nano board. It is fully compatible with general Arduino Uno and Arduino Duemilanove, but dimensionally smaller and enables direct connecting to the printed circuit board. Several different microcomputers were tested, mainly AVR, PIC and STM32. STM32 are 32 bit ARM processors on approximately 84 MHz. For stated regulation they are unnecessarily powerful and they have complicated programming. They are not produced in DIL (Dual Inline Package) solution, but only in SMD (Surface Mount Device), so they are not suitable for manual installation. The family of 8 bit PIC microcomputers appears itself to be low powered, problem is mainly in small RAM. Usually, they have 256 bytes RAM. Considering future possibility extension of Ethernet communication it is unusable low. As the most advantageous show themselves microcomputers from AVR family from Atmel company, particularly ATmega328. It has 2 kb RAM, 1 kb EEPROM and it works up to 16 MHz.

Considering simple programming we used programming language Arduino. It is modified C++ with easy learning curve. For Arduino there are a great number of libraries available and it has excellent community support. Arduino Nano has 14 input output pins. Setting of required temperatures is saved in EEPROM in integrated 1 kb EEPROM. If user wants to change settings, they have to connect control unit to a computer and rewrite EEPROM memory. For communication it is used USB with integrated converter to standard serial port RS-232. Implemented protocol is textual and with a device it is communicated via ASCII signs. For basic testing, tool PuTTY was used, because control application for easier setting and value check was not programmed in the testing phase. There is an illustration of the part of the control code as we solved it below.

#!/usr/bin/env python

```
import serial
ser = serial.Serial(  # port='/dev/ttyUSB0',
                      port='/dev/ttyUSB0',
                      baudrate=9600,  #writeTimeout = 2,
                      parity=serial.PARITY_NONE,
                      stopbits=serial.STOPBITS_ONE,
                      bytesize=serial.EIGHTBITS )

try:  # //If you already have the port open, try to close
it
    if ser.isOpen():
        ser.close()
except Exception, e:
    print "error open serial port: " + str(e)
    exit()

try:
    ser.open()
except Exception, e:
    print "error open serial port: " + str(e)
    exit()

if ser.isOpen():
    ser.flushInput()  # flush input buffer, discarding
all its contents
    ser.flushOutput()  # flush output buffer, aborting
current output

    t1 = float( raw_input("Enter Temperature 2 ") )
    t2 = float( raw_input("Enter Temperature 2 ") )
    print " Temperature 1=", t1
    print " Temperature 2=", t2

    t1a = int(t1 / 256)
    t1b = int(t1 % 256)

    t2a = int(t2 / 256)
    t2b = int(t2 % 256)

    temperature = str(t1a) + "," + str(t1b) + "," +
str(t2a) + "," + str(t2b)
    send = '<EEPROM,0,100,200,2,3,4,5,' + teploty + '>'
    print "SEND: " + poslat

    ser.write(poslat)
    #ser.write("<EEPROM,0,100,200,2,3,4,5,
0,190,0,231>")
    ser.flushOutput()

else:
    print "cannot open serial port "
```

5 Conclusions

The aim of the article was to introduce cheap and innovative solutions of intelligent heat regulation with a use of simple control unit for intelligent flat heating regulation. Primary requirement was the lowest possible price and use of free solution, therefore the method of current commercial automation systems modifications was rejected at the very beginning. As the control unit Arduino platform was chosen, for its low price, reliability and popularity. All the regulation is done solely in a control unit that is autonomous among the whole system. Thereby the situation when the internet connection drops/fails and heating stops working cannot occur. The design was implemented and tested in a simple model flat. From the testing it is clear that in the next solution phase it is necessary to create graphical application for controlling of the whole system [13] and its prospective connection to cloud for remote access option and system parameterization without a necessity of physical access [14].

Acknowledgement Support of the Specific Research Project of FIM UHK "SCM, Control of Markets and Production in Agent-based Computational Economics" is gratefully acknowledged.

References

1. Geth, F., Verveckken, J., Leemput, N., Beerten, J.V.N., Tielens, P., De Smedt, V., Iacovella, S., Hunyadi, B., Koolen, N., De Clercq, H., Gielen, G., Puers, R., Van Huffel, S., Belmans, R., Deconinck, G., Dehaene, W., Driesen, J.: Development of an open-source smart energy house for K-12 education. In: General Meeting of the IEEE-Power-and-Energy-Society (PES) (2013)
2. Sharples, S., Callaghan, V., Clarke, G.: A multi-agent architecture for intelligent building sensing and control. Sens. Rev. **19**(2), 135–140 (1999). doi:10.1108/02602289910266278
3. Morel, N., Bauer, M., EL-Khoury, M., Krauss, J.: Neurobat, a predictive and adaptive heating control system using artificial neural networks. Int. J. Solar Energy. 2007-03-23, **21**(2–3), 161–201 (2007). doi:10.1080/01425910108914370
4. Sun, S.H, Jin, Y.S, Zhang, W.J.: Design and simulation of remote temperature monitor and control system based on embedded web server, In: International Conference on Instrumentation, Measurement, Circuits and Systems (ICIMCS 2011) (2011)
5. Fuentes, M., Vivar, M., Burgos, J.M., Aguilera, J., Vacas, J.A.: Design of an accurate, low-cost autonomous data logger for PV system monitoring using Arduino (TM) that complies with IEC standards. Sol. Energy Mater. Sol. Cells **130**, 529–543 (2014). doi:10.1016/j.solmat.2014.08.008
6. Minebea Co Ltd (MINW-C). Thermoelectric energy converter for, e.g. sensor has thermal reservoirs in which one thermal reservoir serves as heat source and other serves as heat sink while second is formed by housing. DE102013204166-A1, 2014-R02161. Uděleno 16.02.2014 (2014)
7. Toyota, J.K.K (Toyt-C).: Thermoelectric-generation apparatus for engine of vehicle, has electronic control unit to diagnose failure of valve based on electrical property of thermoelectric conversion module, when transmitting opening or close signal to actuator. JP2013147974-A, 2013-M02201. Uděleno 20.02.2013 (2013)
8. Honeywell Technologies Sarl(Hone-C).: Thermoelectric actuator for use in room-temperature controller for regulation of e.g. radiator, has Peltier element that heats or cools expandable and

contractible element based on electrical current flow through Peltier element. EP2559898-A1, 2013-C52829. Uděleno 17.8.2011 (2011)
9. Zeng, Y., Lu, G.X., Wang, H., Du, J.H., Ying, Z., Liu, C.: Positive temperature coefficient thermistors based on carbon nanotube/polymer composites: In: Nature Publishing Group, Macmillan Building, 4 Crinan St, London N1 9XW, England Journal, vol. 4 (2014). doi:10.1038/srep06684Bhadrakumari
10. Rabiya, A.S.: Electrical properties of composite temperature: Dependent resistors. In: Taylor & Francis Ltd, 4 Park Square, Milton Park, Abingdon OX14 4RN, Oxon, England Journal, vol. 21, Issue 2, pp. 173–177 (2014). doi:10.1080/15685543.2014.856196
11. Instruments, Texas.: LM35 Precision Centigrade Temperature Sensors. LM35 datasheet, Aug (1999)
12. Penhaker, M., Darebnikova, M., Cerny, M.: Sensor network for measurement and analysis on medical devices quality control. Commun. Comput. Inf. Sci. **171**, 182–196 (2011)
13. Behan, M., Krejcar, O.: Modern smart device-based concept of sensoric networks. EURASIP J. Wireless Commun. Networking. 1(1155) (2013). doi:10.1186/1687-1499-2013-155. Received 9 October 2012; Accepted 15 May 2013; Published 6 June 2013. ISSN 1687–1499
14. Machacek, Z., Slaby, R., Hercik, R., Koziorek, J.: Advanced system for consumption meters with recognition of video camera signal. Elektron. Ir Elektrotechnika. **18**(10), 57–60 (2012). ISSN: 1392-1215

Measuring of Electric Energy Consumption in Households by Means of Arduino Platform

Josef Horalek and Vladimir Sobeslav

Abstract The paper presents possibilities of the measuring of electric energy consumption in households. Current approaches and technological solutions are analysed and closely examined with respect to their positive and negative aspects. Based on the analysis a new solution eliminating detected disadvantages is introduced using Arduino platform.

Keywords Arduino · Android · Measuring of consumption · Wattmeter · Smart metering

1 Introduction

Every household has at least one wattmeter planted in the main distribution board of the house, which measures the consumption of the entire household and a provider of the electric energy regularly on a time specific basis subtracts of wattmeter readings and issues a clearing invoice according to the selected tariff [1]. The proportion between the consumed energy and the price is continual, of course. In present day the prices of electricity are stagnating, even falling, but long-term predictions declare that the prices will continually rise. From this reason a large number of households are trying to reduce the wattage of appliances [2, 3] and thus decrease the expenses by using electrical devices with higher efficiency. Because data provided by manufacturers are not always exact [3], it is appropriate to validate the energy consumption. In this way it is possible to find out which appliance has the greatest wattage and for example start considering its exchange. Wattmeter from

J. Horalek · V. Sobeslav (✉)
Department of Information Technologies, Faculty of Informatics and Management,
University of Hradec Králové, Hradec Králové, Czech Republic
e-mail: vladimir.sobeslav@uhk.cz

J. Horalek
e-mail: josef.horalek@uhk.cz

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_69

the energy provider is inappropriate for this purpose because it shows the sum of wattages of all electrical devices in the home. Easy consumption identification of particular devices and appliances is thus impossible.

As a consequence the whole range of wattmeters is being produced [4–6] which enable us to measure the consumption of individual appliances [15–17]. Here however comes the problem. Offered devices do not allow long-term logging (saving) of measured data and executing of their analysis. Furthermore, there is no wattmeter that would allow remote access from the web browser or from a mobile application running on a smartphone. Therefore it is not possible to comfortably execute analysis of appliances' wattage [7]. The following parts of the paper therefore propose the solution by which the thereinbefore drawbacks are successfully eliminated. The solution is then implemented, thoroughly tested and verified. The last part contains the summary of all findings and suggests areas for further improvements.

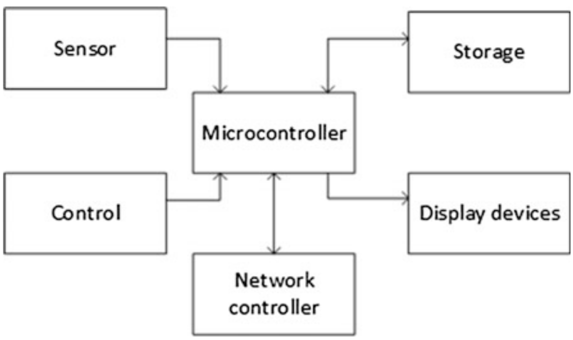
There is a wide range of wattmeters supply on the market [8]. Generally it is possible to divide the types of wattmeters into following categories: socket plug, industrial, pliers, and distribution board wattmeters. The socket plug wattmeter is attached into sockets in the house. Measured appliance is then plugged into the socket in the wattmeter. It is a cheap device which allows immediate consumption monitoring [21–24]. Selected models are also able to save the history of consumption in short-term interval (units of days). The industrial wattmeter is, accordingly to its name, used in the industrial section of the market. These meters can monitor and log measured data. It is possible to access the data via certain interface or bus (for example Ethernet, Modbus or Profibus through bus RS-485, etc.). However, their price is high and they require specialized technical service, thus their use in households is not suitable. The last type of wattmeters is the pliers wattmeters. These are used especially for measurement in field and in laboratories. Better equipped models have the option to store wattage values and examine the data in computer through USB. Yet their use is inappropriate for long-term monitoring on a single place. All the above mentioned solutions share a common drawback. It is not possible to realize long-term logging and storing of measured values in order to examine and analyse the data afterwards. The exceptions are industrial wattmeters, which can include Ethernet interface, and whose wattage can be monitored with special application. Naturally, their price is high and they can not be used in households. Manufactured wattmeters also lack advanced functions such as conversion of consumption to fee that needs to be settled according to a household tariff. Socket plug wattmeters are feasible for casual measurement, but if measurement of an appliance which does not have constant consumption (such as desktop computer or television) is necessary, it would be more desirable to use a wattmeter which is affordable and enables to browse and check measured values. Another useful feature can be the connection with smartphones, where it would be possible to work with measured wattages further. Such a wattmeter is not being produced and it is necessary to design it [9]. Requirements for a suggested solution thus arise from the lack of commonly available wattmeters with wattage measurement accurate enough, with simple operation and simple controls, with local storage of measured values, with network interface, with highest possible current load, with high level of reliability and safety. Another requirement is a mobile application that communicates with the wattmeter and allows detailed work and analysis of measured data. It is possible to generally summarize the requirements into few points. Support of Android, displaying of consumption in charts, conversion to price according to the household's tariff and notifications when the price or consumption surpasses the preset value.

2 New Solution Proposal

The proposed solution is depicted in Fig. 1, which in the form of a diagram describes innovative solutions of electric energy consumption measurement in the household.

The central element of the suggested solution is a microcontroller. The microcontroller is proposed because, unlike microprocessor, it contains not only the processor, but also integrated RAM, flash memory for a program, and it can also have an EEPROM. In order to be able to communicate with its surroundings, the microcontroller is equipped with a number of digital inputs and outputs, analog inputs and usually also containing bus lines. There is a fair amount of microcontrollers from various manufacturers and of different architectures. The most common is RISC architecture with 8, 16 or 32-bit instructions. The advantage of microcontrollers over microprocessors is considerably lower consumption. The primary task of a microcontroller is to read data from the sensor which measures the wattage. This operation will repeat cyclically with a period of five seconds. The data will be stored together with the time stamps on a central repository. A control element is attached to the microcontroller in order to enable initial setup of communication and for eventual reading of measured values on display. The last block of suggested wattmeter is a network controller that provides the communication with the surroundings through standard data and mobile data networks.

Fig. 1 Block diagram of functional solutions



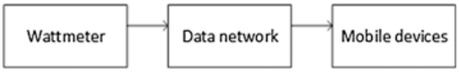


Fig. 2 Communication power meter with mobile devices

It is necessary to use the data network for the communication with mobile devices, as shown in Fig. 2. According to presently utilised standards, it is possible to use cable connection through Ethernet technology, wireless connection via WiFi or ZigBee technology.

2.1 The Technical Proposal of the Solution

The microcontroller model ATmega2560 from the company ATMEL was selected as a controlling component [10]. It is 8-bit microcontroller based on RISC architecture. It contains 256 kb of flash memory, 8 kb RAM and 4 kb EEPROM. It works on 16 MHz clock frequency and it allows the use of up to 86 inputs/outputs. Since the component in question is SMD (Surface Mount Device), which is characteristised by problematic development on the prototype basis, the development board base Arduino Mega 2560 was used for the purposes of the wattmeter. Arduino is a prototypic board that noticeably accelerates the development of electronical projects. A part of the Arduino project is also the Integrated Development Environment and the compiler. Its advantage is the existence of a large number of already debugged libraries supporting most of the commercially available parts. The basic version Arduino Uno [11] was not chosen because of its built-in microcontroller. The ATmega328, which contains only 32 kb of flash memory and 2 kb RAM. Those values are sufficient enough for simple projects, however, in the case of the wattmeter proposed by us, it would be very difficult to realize all the requested features. As a consequence the board Mega 2560 containing higher capacity of the memory was selected.

The sensor used for the measurement of the consumption was Hall Effect Current Sensor Module ACS712 [12]. It uses Hall's effect for the measurement of the flow of electric current. Thus the electric power is not measured directly, because such a wiring is very expensive and complicated [13, 14]. Effective voltage in powerline in the Czech Republic is 230 V with maximal permissible tolerance −10 and +6 %. As long as we measure only the electric current, which will then be multiplied by constant 230 we will get resulting value of consumption with maximal deviation 16 %. In this way we obtain values that are commonly reported by commercially available wattmeters in price range around 20$. The advantage of the sensor ACS712 is its maximal possible electric current 20A which can run through it without being damaged. Consequently it is possible to measure electric power up to the value of 4600 watts. In order to make the communication between wattmeter and a mobile platform possible, our solution design uses ethernet controller Wiznet

W5100 [15], or more precisely, extensive development board for Arduino on which the controller is used. The controller has hardware support of the set of protocols from the TCP/IP family. Thereby it is possible to use DHCP for automatic configuration of IP adress. The controller has also an implemented slot for MicroSD memory cards that supports FAT and FAT32 file system. The maximal supported capacity of the cards is thus limited on 16 GB. Thereby the storage large enough for measured values is guaranteed. NTP server is used for further accuracy of time stamps. The control of the wattmeter is solved by the choice of 4 × 4 matrix keypad which contains numbers and keys for A,B,C,D,* and #. Via the keypad it is possible to configure the settings of IP adress and to review the measured data. The two line display with twenty characters per line is used to show the information on the wattmeter. The display also enables to display the information about immediate consumption. Since the wattmeter uses cable Ethernet interface and most of the Android devices do not have the cable Ethernet, the solution of the wattmeter designed by us is connected by the cable into a wireless router. In this way the application for Android can make inquiries through the WiFi network with simple TCP packets. A part of the packet is also the opening and final time of measured data. The application can also have implemented a service class that would ensure the automatic periodic download of the data. Measured values can then be browsed either textually or in a form of a chart. Furthermore, there will also be the possibility to set up notifications when the tariff-price or the consumption over defined term surpass the preset value.

3 The Solution Implementation

The implementation of a prototype of the new wattmeter solution was created on the contact field. The Ethernet controller Wiznet W5100 was attached to the Arduino development board. The controller in question is realized as a ready-made model that attaches to the Arduino with the use of common pins. It also contains the controller for SD and SDHC memory cards. The SD card with a 1 GB capacity and FAT32 file system was used during the development. The schematic representation Fig. 3 illustrates the connections of other components to Arduino.

The pin 16, set up as an output, is connected through PNP transistor BC557C to a piezo speaker, which serves as a warning signal in case of surpassing of the consumption value set by the user. The transistor is used in order to not exceed the maximal allowed current drain on the pin. The pin 19 is connected to the keypad through second hardware UART (Universal Synchronous/Asynchronous Receiver and Trasmitter). Another component is Hall Effect Current Sensor Module ACS712, which is connected on first analogue input A0. This is a ready-made board with built-in protection and five pins. Two of them are conducted through a terminal plate and the measured current flows through them. The remaining three are supply input for +5 V, ground and output on which changes the voltage in range from 0 to 5 V in dependence on the electric current flow. The last module is two

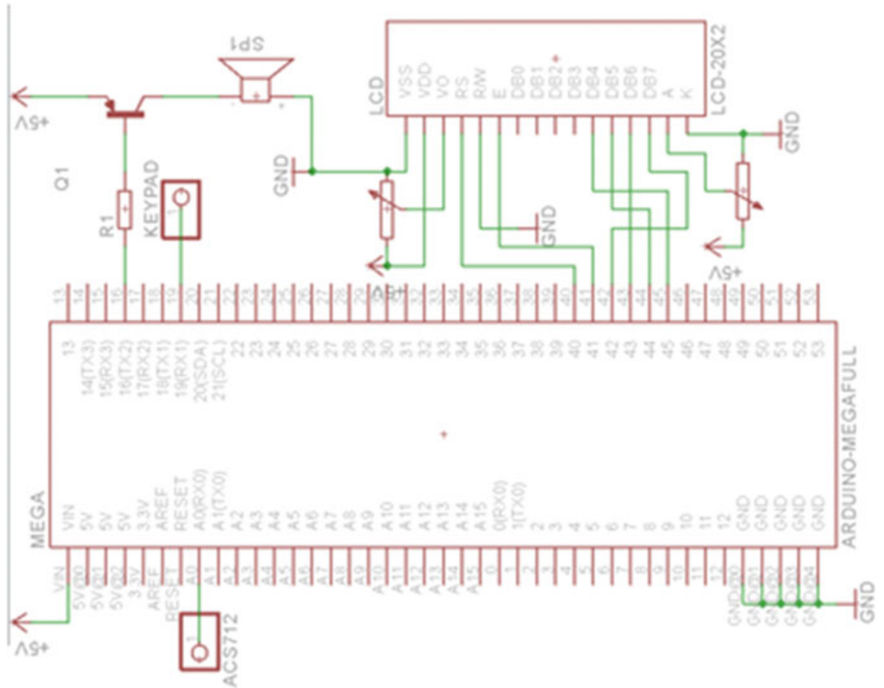


Fig. 3 Wiring diagram wattmeter

line LCD display with twenty characters per line with Hitachi HD44780 controller. A parallel interface is used for service of the display on the ground of low overhead costs for hardware. The display is attached on pins 40 to 45 as well as to two potentiometers. The first potentiometer is attached to pin Vo and serves for changes of the display contrast, the second one regulates the voltage between anode and cathode of the backlight [18–20].

Since the 4 × 4 matrix keypad was selected for the direct control of the wattmeter, it was necessary to implement secondary microcontroller. Handling of the matrix keypad has to run in infinite loop in which each line of the matrix is in sequence set on a logical value 1. Then the values are ensured in the columns. If there is also a logical value 1 on some of them, pressing of a key is recorded on the specific position. Considering that the Arduino has to communicate with the LCD display, Ethernet interface and SD card, it would not be able to record everytime a key is pressed. Thus the keypad is attached to a microcontroller PIC16F690, which has 18 input/output pins, 4 kb of flash memory and 256 bytes RAM. It has implemented simple code in programming language C, compiled in Integrated Development Environment MikroC, that runs the aforementioned algorithm and that sends a code through UART interface to Arduino when a key is pressed. Arduino contains hardware buffer, that is being regularly checked and if it contains a code of a key, corresponding action is realized. In case that Arduino is running an

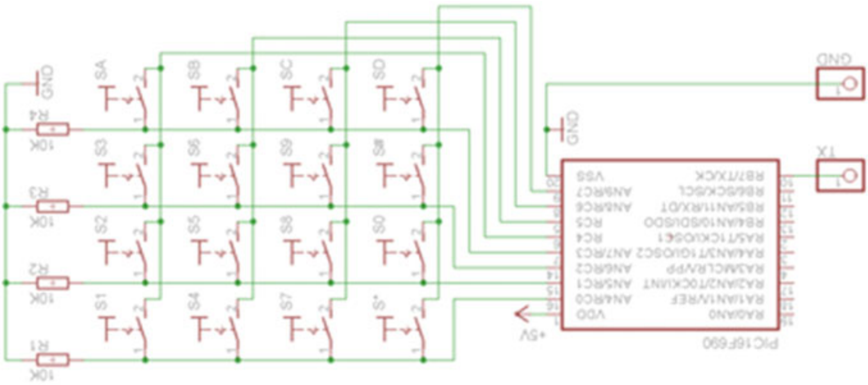


Fig. 4 Wiring diagram keyboard

uninterruptable code, there may be a delay inbetween the press of a key and a reaction. However, such delays are in the units of miliseconds and do not exhibit themselves in any distinctive way (Fig. 4).

During the development Arduino is most frequently powered with an USB cable, a cable that serves for loading of the program and for debugging at the same time. However, powering through USB connector is inapplicable for the final product. For this reason a power supply source working in voltage range from 7 V up to 24 V was designed. Output voltage is stabilized 5 V. Thanks to this, it is possible to power the wattmeter with almost arbitrary power grid adapter. Almost all the electronics is at the same time directly separated from the grid voltage 230 V. The drawback of this approach is the necessity to use external adapter, which itself has certain consumption. Nevertheless, modern boost converters have high efficiency, over 80 %. Moreover, the wattage of the device itself does not exceed hundreds of miliamperes and thus it is an insignificant figure.

The source code had been created in Integrated Development Environment Arduino IDE 1.0.6. The language used for the programming is a simplified version of the programming language C. Every program for Arduino has to contain two methods. The first one is setup, in which the initial values are preset, and the other one is loop, which is an infinite loop of programming logic. The method setup is composed of several other method calls. First of all the method lcdInit is called for configuration of the LCD. Then the serial port UART is set up for the communication with the keypad. Afterwards the type of Ethernet configuration is loaded from the EEPROM. It can be set on automatic configuration from a DHCP server or statically. In the case of static configuration, the IP adress, subnet mask, default gateway and a DNS server adress is loaded from the EEPROM. Furthermore, the threshold value of consumption is read from the EEPROM. The additional step is the inicialization of the Ethernet and of the memory cards by the methods called initialize SDCard and initialize Ethernet. If the Ethernet can not be initialized according to the preset configuration, LCD display shows an error message and configuration dialogue for new settings. After the successful configuration of the Ethernet a server socket, which responds on requests from the mobile application, is established on the chosen IP adress and TCP port 3300. Then the connection is established with a NTP server, from which the exact time is being acquired. Then the program moves into the loop method. In the loop method in its infinite loop we increment variable count and when it reaches value 8000, the method sendNTPpacket is called to send a packet towards the NTP server. When the value reaches 10,000 (approximately the equivalent of 12 s of time) the variable count is zeroed and the methods updateTime, saveData, threshold and redraw are called in sequence. The updateTime method refreshes the time according to NTP server response. The saveData method reads the value of the electric current from the Hall effect sensor, converts it to electric wattage and saves it on SD card in "17:39:47 04-01-2015 337" format. The first figure is the time, the second is the date and the last figure is the wattage value of the measured device or appliance. A new file is created at the beginning of each hour of measurement in order to avoid having just one big file on the memory card. The filename contains the hour and the date. This approach simplifies consecutive reading and locating of the data. The threshold method compares measured consumption with the threshold that was set up by the user. If the threshold is surpassed, the piezo speaker makes a short beep. The next last method called redraw erases the content of the LCD display and overwrites the information of the threshold value, current consumption and the time. Pressing of a key is being controlled by the loop as well, and according to the sign, the LCD display shows the configuration. It is possible to change the Ethernet configuration or the consumption threshold directly on the wattmeter. The last method is check Ethernet, in which the arrival of the packet from the Android application is being checked. The packet contains information about the day and the hour to which the measured data from the SD card needs to be sent. The second part of the wattmeter is the application for the mobile platform Android. It was developed with the objective API version 14. Therefore it supports Android from version 4.0. It uses only one activity composed of fragments. The side bar from the material design of Navigation Drawer is used for the navigation in the application. This side bar is hidden on mobile phones and it is triggered by a gesture. Tablets have the side bar always visible. The application contains in its navigation menu items for displaying of charts, textual visual display, price calculations and settings. These fragments are represented by classes GraphFragment, TextFragment, PriceFraments and SettingsFrament. In order to be able to receive the data from the wattmeter, it is necessary to set up the IP address of the wattmeter in settings. Then the thread in class TCPclient is initiated, which periodically sends queries to the wattmeter demanding the measured values which are then stored in an integrated SQLite database.

4 Testing of the Developed Application

The hardware and the program of Arduino were tested from the point of view of energy consumption of the wattmeter itself, accuracy of the measurements and stability of the program. The consumption of the wattmeter was measured by a laboratory multimeter DIGITEK DT3800 on the input of the power supply 7–24 V, which stabilises the voltage on 5 V and supplies the hardware. The power source was plugged to 9 V grid adapter with 80 % effectivity. With LCD backlight off, the measured current is stabilized on 220 mA and with full LCD backlight on 250 mA. With regard to the power source effectivity the wattmeter wattage is 1375 W, or 15,625 W. Average price for kilowatt-hour in the Czech Republic is 4.80 CZK. Year-long running of the wattmeter costs at average 58, or 66 CZK. Manufacturers of commercially produced wattmeters do not indicate their consumptions, but we can presume that their consumption would be similar. Thus running the wattmeter is not financially demanding. Because of safety and difficulty the wattmeter accuracy was not tested on the grid voltages, but using safe direct voltage 12 V from hard laboratory power source MERA EZP-04-05. The constant for the multiplication of measured electric current was for this purpose changed from 230 to 12 in the code of the program. The consumption that was being measured was of classical Edison light bulbs with different denoted electric energy consumption. The values are compared with above mentioned multimeter. First of all, the current flowing through the light bulb was measured and then the voltage on the light bulb. The values were multiplied with one another. The results are shown in Table 1.

The measured values imply that the highest measured variance is 23 %. With increasing electric energy consumption the variance has tendency to diminish and stabilize itself. On condition that the grid voltage does not show any distinctive fluctuations, it is possible to claim that the measured valued correspond with results from commercial wattmeters. The last tested aspect of the wattmeter was its dependability. A short ungrounded conductor was attached to Arduino instead of Hall effect sensor. On the conductor is induced so called signal noise. Those are random values that were stored on a memory card instead of the consumption. Likewise the RAM occupancy was monitored through the serial port. The wattmeter was tested continuously for seven days. It turned out that the free memory does not change and stays on 5950 bytes. Thus it can be stated that the program does not contain any methods that could deplete the memory and cause the crash.

Table 1 Measured consumption

Indicated consumption (W)	DT 3800 (W)	The proposed wattmeter (W)	Variance (%)
6	5.76	7.11	23
12	12.48	11.88	5
50	52.4	55.2	5
100	98.5	105.72	7

Furthermore, the files on the memory card were checked. None of them was corrupted nor did it contain any wrongly stored data. The average size of the data file was 8 KB and those files were created every hour. Every day the card was filled with approximately 192 KB of data. Therefore it is possible to store the data for up to several years on memory card with 1 GB capacity.

5 Conclusions

The paper has presented a proposal, realization and testing of an innovative device designated for consumption monitoring and measuring in households. Firstly, there was introduced a proposal of a new solution, then the results of its testing. The solution was tested from the viewpoint of stability, operating expenses and accuracy of measurements. The results show that all the parameters match available commercial solutions.

Along with the proposal and realization of the new wattmeter, a controlling application for Android platform has been created. In the future, it is possible to expand the new solution in many ways; currently new supports for different mobile platforms are being developed and the layout of printed circuit is being optimalized for the final product.

The suggested solution is in accordance with current approaches of smart metering, which is based on measurements and control of the consumption in real time, to which the solution suggested by us corresponds. Our solution also enables the use of different electric energy consumption price tariffs and at the same time it implements immediate awareness of the user themselves, who in this way acquires other comprehensive information for their decisions regarding the optimalization of the electric energy consumption in the household.

Acknowledgement Support of the Specific Research Project of FIM UHK “SCM, Control of Markets and Production in Agent-based Computational Economics” is gratefully acknowledged.

References

1. Mahmood, A., Aamir, M., Anis, M.I.: Design and implementation of AMR smart grid system. In: IEEE Canada Electric Power Conference, pp. 1–6 (2008). doi:10.1109/EPC.2008.4763340
2. Rajasekharan, J., Koivunen, V.:Optimal energy consumption model for smart grid households with energy storage. IEEE J. Sel. Top. Sign. Proces., 1932–4553 (2014)
3. Ju, S.H., Muhammad, A., Lim,Y.H., Irfan, M., Choi, M.S., Baek, J.M., Lee, S.Y.: An efficient home energy management system based on automatic meter reading. In: IEEE International Symposium on Power Line Communications and Its Applications. IEEE, 2011, pp. 479–484 (2011). doi:10.1109/ISPLC.2011.5764445
4. Rodrigues, E.M.G., Caramelo, T., Mendes, T.D.P., Godina, R., Catalão, J.P.S.: Experimental wireless wattmeter for home energy management systems. In: IEEE International Symposium on Power Line Communications and its Applications. IEEE, p. 327 (2011). doi:10.1007/978-3-319-16766-4_35

5. Ramírez Muñoz, D., Moro Pérez, D., Sánchez Moreno, J., Casans Berga, S., Castro Montero, E.: Design and experimental verification of a smart sensor to measure the energy and power consumption in a one-phase AC line, Measurement, vol. 42, Issue 3, April 2009, pp. 412–419 (2009). ISSN 0263-2241
6. Djokic, B., So, E.: A high-resolution hybrid digital wattmeter for measurements at high voltage. In: IEEE Instrumentation and Measurement Technology Conference, vol. 1–5,978-1-4244-1540-3, (2008). doi:10.1109/IMTC.2008.4547353
7. Adamo, F, Attivissimo, F., Cavone, G., Di Nisio, A., Spadavecchia, M.: Channel characterization of an open source energy meter. IEEE Trans. Instrum. Meas. **63**(5), 1106–1115. doi:10.1109/TIM.2013.2286959
8. Kar, S., Hug, G., Mohammadi, J.: Distributed state estimation and energy management in smart grids: A consensus plus innovations approach. IEEE J. Sel. Top. Sign. Proces. 1932–4553 (2014)
9. Cataliotti, A., Cosentino, V., Di Cara, D.: A PC-based Wattmeter for high accuracy power measurements. In: Proceedings of IEEE International Instrumentation and Measurement Technology Conference I2MTC 2010, 978–1-4244-2833-5 (2010)
10. Arduino.: Arduino Mega 2560. [cit. 2014-12-17] (2013). Available from: http://arduino.cc/en/Main/arduinoBoardMega2560
11. Arduino.: Arduino Uno. [cit. 2014-12-17] (2013). Available from: http://arduino.cc/en/Main/arduinoBoardUno
12. Allegro MicroSystems: ACS712.: Fully integrated, hall-effect-based linear current sensor IC with 2.1 kVRMS voltage isolation and a low-resistance current conductor. Allegro Microsystems. [cit. 2014-12-19] (2014). Available from: http://www.allegromicro.com/en/Products/Current-sensor-ICs/Zero-to-fifty-amp-integrated-conductor-sensor-ICs/ACS712.aspx
13. Fransiska, R.W., Septia, E., M.P.: Electrical power measurement using Arduino uno microcontroller and LabVIEW. In: Proceedings of 2013 3rd International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering, 978-1-4799-1649-8 (2013)
14. Li, H.R., Pang, L.J., Zhang, X.F., Deng, G.: The application of FPGA and DSP techniques in intelligent wattmeter. In: Advanced Manufacturing Systems, PTS 1-3,978-3-03785-039-8 (2011)
15. Vanus, J., Novak, T., Koziorek, J., Konecny, J., Hrbac, R.: The proposal model of energy savings of lighting systems in the smart home care. In: IFAC Proceedings, vol. 12 (PART 1), pp. 411–415 (2013)
16. Vanus, J., Koziorek, J., Hercik, R.: Design of a smart building control with view to the senior citizens' needs. In: IFAC Proceedings Volumes, 12 (PART 1), pp. 422–427 (2013)
17. Machacek, Z., Slaby, R., Hercik, R., Koziorek, J.: Advanced system for consumption meters with recognition of video camera signal. Elektron. Ir Elektrotechnika. **18**(10), 57–60 (2012). ISSN: 1392-1215
18. Jančíková, Z., Koštial, P., Bakošová, D., Ružiak, I., Frydrýšek, K., Valíček, J., Farkašová, M., Puchký, R.: The study of electrical transport in rubber blends filled by single wall carbon nanotubes. J. Nano Res. **21**(16), 1–6 (2013). ISSN: 16625250
19. Krejcar, O., Spicka, I., Frischer, R.: Implementation of full-featured pid regulator in microcontrollers. Electron. Electr. Eng. **7**(113), 77–82 (2011)
20. David, J., Svec, P., Frischer, R., Garzinova, R.: The computer support of diagnostics of circle crystallizers. Metalurgija. **53**(2), 193–196 (2014). ISSN 0543-5846
21. Krejcar, O., Frischer, R.: Real time voltage and current phase shift analyzer for power saving applications. Sensors. **12**(8), 11391–11405 (2012). Published 21 August 2012. ISSN: 1424-8220
22. Pies, M., Hajovsky, R., Ozana, S.: Autonomous monitoring system for measurement of parameters of heat collection technology at thermal active mining dumps. Elektron Ir Elektrotechnika **19**(10), 62–65 (2013)

23. Hajovsky, R., Pies, M.: Complex measuring system for longtime monitoring and visualization of temperature and toxic gases concentration. Elektron Ir Elektrotechnika **122**(6), 129–132 (2012)
24. Machacek, Z., Slaby, R., Vanus, J., Hercik, R., Koziorek, J.: Non-contact measurement system analysis for metallurgical slabs proportion parameters. Elektron. Ir Elektrotechnika. **19**(10) (2013). ISSN: 1392-1215

A Methodology for the Enterprise Networks Architecture and Modelling

Josef Horalek, Vladimir Sobeslav, Ondrej Hornig and Ladislav Balik

Abstract The design of extensive computer networks is specific by its need of the deeper analysis of data flows and selected technologies, security strategy, network management, test and other phases of which goal is to suggest the best possible solution. Several frameworks or methodics implemented in many tools and systems in different areas of information technology are dealing with the design and administration of extensive and effective IT infrastructure. Based on analysis and comparison of frameworks used for creating of enterprise architecture, a T-NET methodics was created, which interconnects frameworks of enterprise architecture TOGAF with specifics from the analyses and designs of computer networks.

Keywords Enterprise IT architecture · TOGAF · ITIL · Enterprise architecture frameworks · Enterprise networks

1 Introduction

Communication is a basic means of understanding between people and at the same time it is an important prerequisite for functioning of the whole society. The massive advancement of information technologies in previous years was accompanied by the advancement of communication technologies. The exchange of data, information and knowledge is one of the fundamental corner stones of cognizant management and it is one of the prerequisites of effective functioning of any organization.

J. Horalek · V. Sobeslav (✉) · O. Hornig · L. Balik
Department of Information Technologies, Faculty of Informatics and Management,
University of Hradec Králové, Hradec Králové, Czech Republic
e-mail: vladimir.sobeslav@uhk.cz

J. Horalek
e-mail: josef.horalek@uhk.cz

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_70

The analysis and design of internet networks has to take into account many key factors, without which there is no way of effective network design. One of the factors is heterogeneous elements of the network connected into one logical complex. Companies and individuals use different communication technologies such as Ethernet, ATM, Frame Relay, DSL, on different media (optical fiber, mobile networks, metalic cabling, etc.), organized in different physical and logical topologies. With the increasing number of elements in the network the demands on transfer speeds, which currently climbed up to the gigabit transfer rates, are rising. The increasing of transfer rates on network interface layer poses higher demands on active elements, which ensure many functions related to the running of a network. With this kind of transfer rates the mere demand on packet delivery in the internet from a start to its destination presses upon a router to process millions of packets per second and to decide with the help of routing algorithms by which way the received packet should be sent. Some routing protocols such as for example RIP are not very suitable for extensive networks, where the choice of more sofisticated protocols such as OSPF, IGRP, BGP or ISIS [1] is essential. Security is a chapter of its own, which every network design should take into account emphasizing the protection of data from the user, of communication channels and of network services. One of the possible approaches to security is the concept of deep protection that analyses particular layers of the network according to their protocols and services [2]. Another factor is the reliability of the network, which forces the designers to create redundant topologies, to use intelligent active elements, to virtualize the network resources and so on. From the time perspective it is necessary to count with its future development, which means the possibility to extend the network with further elements and services, increase the transfer rates and implement new technologies.

From the above mentioned it is obvious that the analysis and design of computer networks is a difficult and dynamic problem that evokes the question whether it is possible to simplify or at least to support the process of the design itself. Specialized literature offers different concepts and approaches, which however often vary in recommended working procedures for the design itself, or they vary in utilized tools. Network technologies manufacturers publish different case studies, recommended working methods, methodologies, architecture designs, etc., which are logically wedded to products of particular manufacturer. The leader in the area of network technologies—CISCO company—developed these solutions:

- CISCO validated design—the sum of case studies and documentations of validated solutions,
- hierarchic approach to network modelling—division of parts of the network to core, distribution and access layer,
- SONA—architecture oriented on network services,
- PIDOO—life cycle of the network,
- certification system CISCO in the area of design.

A part of the issue is also the optimization of network topology, routing algorithms, measuring of network reliability and efficiency, quality of services, etc. The present paper does not primarily deal with the aforementioned areas; however, it is possible indicate several relevant sources:

- optimalization of network topology and routing [1, 3, 4–7]
- quality of services [8, 9–11],
- metrics of network reliability and efficiency [3, 4, 12–15]
- network security [2, 16, 17].

An effectively designed IT infrastructure enables to support enterprise processes, minimize expenses and maximalize effect of utilized technologies. In case of a big organization it is complex process, in which it is usually necessary to interconnect a larger number of heterogeneous systems, applications and devices. In such cases it is possible to use tools of enterprise (business) architecture.

2 Enterprise IT Architecture Frameworks

This area includes different tools, methodics, standards and approaches that lead up to the effective implementation of extensive technological infrastructure with a view to fulfill the business goals of the organization in the best possible way [18]. Presented solutions occur on different levels of implementation from abstract concepts up to methodics that define in detail the whole process of implementation [10]. A term Enterprise Architecture Framework was established for this kind of solutions, a term that can be loosely described as a general concept of business architecture. One of the famous concepts is The Open Group Architecture Framework (TOGAF), used in different areas of information technologies. Among certified tools belongs for example IBM Rational Architect or Sparx Enterprise Architect, which is used for UML modelling. There is a whole range of such frameworks; according to J. Schekkerman's classification [81], most important ones are:

- Extended Enterprise Architecture Framework (E2AF),
- Enterprise Architecture Planning (EAP),
- Federal Enterprise Architecture Framework (FEAF),
- The Open Group Architecture Framework (TOGAF),
- Treasury Enterprise Architecture Framework (TEAF),
- Zachman Framework,
- Integrated Architecture Framework (IAF),
- Joint Technical Architecture (JTA),
- C4ISR a DoDAF,
- Department of Defence Technical Reference Model (DoD TRM),
- Technical Architecture Framework for Information Management (TAFIM),
- Computer Integrated Manufacturing Open System Architecture (CIMOSA),
- Purdue Enterprise Reference Architecture (PERA),
- Standards and Architectures for Government Applications (SAGA).

Individual frameworks involve architecture creating methods or procedures that are formalised in the architecture metamodels. There are a large number of framework concepts in all possible forms from abstract descriptions of enterprise architecture principles up to detailed methodics with defined procedures in individual areas. The whole area could be divided into basic groups according to their origin and designation of individual frameworks. Detailed analysis and comparison of individual frameworks of course exceed the possibilities of this paper and thus in order to proceed it is necessary to select particular representatives appropriate for the analyses and design of networks, or more precisely limit their amount. Again, there are more ways of approaching this task, thus the selection of a suitable framework will be limited by the demand of generality and usability:

- **Generality**—enterprise architecture framework has to be designed in general terms in order to be usable in different companies. Some frameworks are designed for implementation of enterprise architecture in particular industry area or in public sector and individual parts and processes are thereby fundamentally influenced.
- **Usability**—enterprise architecture framework should be currently sufficiently put to use. The extent to which it is utilized in organizations reflects the amount of its acknowledgment by the community, consulting firms in the area of IT and it creates a groundwork for the following implementation of the methodics.

The demand of sufficiently high utilization of selected frameworks in practice leads to the analysis of currently used frameworks. According to the analysis of Microsoft company [19] the most important frameworks are currently Zachman framework, Federal Architecture Framework (FEAF), Treasury Enterprise Architecture Framework (TEAF), The Open Group Architecture Framework (TOGAF) and Department of Defense Architecture Framework (DODAF). Surveys of utilizations are published with various periodicity by government institutions in the USA but primarily by independent consulting organizations. Important sources of information about extension of enterprise architecture are for example researches from companies Capgemini, Gartner and Infosys. These companies deal with consulting services, implementations of enterprise architectures or tools for support of business processes, or they design their own framework (Gartner). According to the study from Infosys company [20] the most utilized enterprise architecture framework in companies is TOGAF (32 %) followed by Zachman framework (25 %). In public administration it is again TOGAF (44 %), followed by FEAF (12 %) (Fig. 1).

2.1 EA Frameworks Suitable for Network Analysis and Design

Even more frameworks and approaches exist in the area of enterprise architecture. Among the often utilized ones in business practice belong the following: COBIT, CORBA, ITIL or SOA. The aforementioned solutions, even though they are

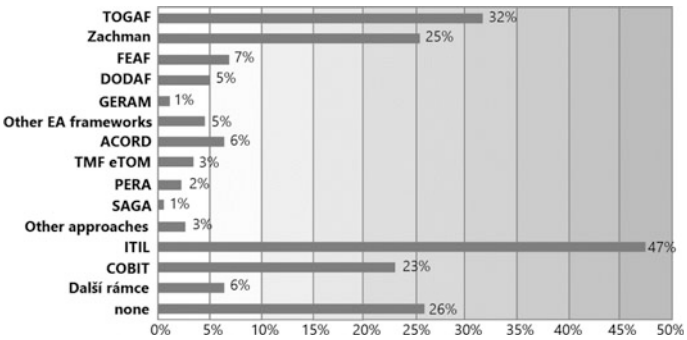


Fig. 1 Enterprise architecture framework utilization [6]

respected and spread in business practice, do not serve primarily for building of enterprise architecture but for more specific purposes. It is hardly possible to compare for example TOGAF and ITIL, although it is possible to encounter such comparisons, especially in less-founded sources on the internet. Accordingly, it is not possible to compare COBIT and ITIL with a goal to decide which framework should be used for the building of enterprise architecture. For consecutive decision on possible utilization of frameworks for creation of network architecture methodics it is necessary to somehow compare selected frameworks. There is a wide range of frameworks comparisons with diverse quality and assessment. The question is what parameters are feasible to be evaluated and mutually compared. Such characteristics could generally be divided into the undermentioned areas:

- way of architecture development planning,
- modelling and punctuality of architectonic viewpoints,
- life cycle and support of project management,
- utilization of information technologies,
- support of tools and standards.

Comparisons of framework characteristics are possible to be decomposed on the analysis of goals and means of their fullfilment, method of working with inputs and outputs which selected enterprise architecture provides. Report Comparative Analysis of Architecture Frameworks [21] analyses and compares most utilized enterprise architecture. The picture presented below shows modified output of this report with a view to today’s most utilized frameworks.

3 T-NET Methodology

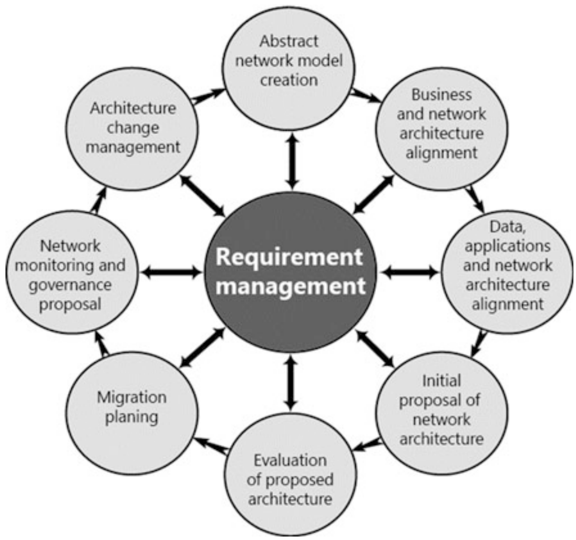
In the area of computer networks analysis and design there are a relatively small number of methodics or approaches used for the systematic support of building of network infrastructure. T-NET methodics is focused on systematic support in the course of analysis and design of computer networks. As it has been already implicated in the introduction, the term computer networks includes a wide range of heterogeneous technologies and solutions. The suggested methodics does not try to be helpful for the implementation of any network or communication technology in any organization, project or area. Such an approach is highly difficult and shifts the methodics to higher level of abstraction, in some cases to a verge of a concept that forms general framework for computer network building. Internet networks are thus utilized in such areas that fundamentally form the topological model, security, transfer technologies and many others. The most utilized enterprise architecture framework is currently TOGAF. The suggested methodics thus connects both areas in question, enterprise architecture and design of internet networks.

The utilization of this framework limits characteristics of T-NET methodics to a great extent. The essence of the whole methodics is a method called ADM (Architecture Development Method), which represents iterative approach. Partial phases of the methodics affect each other and process phases thus do not have a waterfall shape. Hence the phases are organized to a circle which enables them their mutual interaction through change management or continual access to a design of the network (Fig. 2).

T-NET methodics differentiates eight basic phases. However, the renaming of partial phases does not mean a fundamental declension from the essence of the framework, more likely it is an customization and concretization of particular area of information technologies. The concentration on the area of the design of computer networks allows to increase concreteness of partial phases and to simplify their passing through the phases. Individual phases are presented by those activities:

The creation of an abstract model of the network—the goal of this phase is a creation of an initial model of network infrastructure. To accomplish inicializative

Fig. 2 Process phases of T-NET methodics



distribution of network elements, transfer technologies and locial structure of the network. The goal of this phase is not detail configuration of network elements, protocols setups, permeability, the implementation of the system for network monitoring, etc. In this phase it is important to identify on abstract level the demands that iniciate a creation of a new network or its change. These demands have to be in compliance with the business vision, goals, information strategy or organization model of the company. This concordance is projected into the abstract model of the network and into partial steps of the phase.

The analysis of interconnection with enterprise architecture—this phase compares and analyses the concordance with the enterprise architecture. The goal of this phase is not to create enterprise architecture and related documentation. Outputs of TOGAF business architecture phase are in T-NET methodics input that serves for comparison and differential analysis. The main output of this phase is the finalization of the concordance with business architecture, evaluation of services and formulation of metrics on the business level.

The analysis of interconnection with application and data infrastructure—the application and data architecture is analysed in this phase. The goal is to create architectonic viewpoint that defines linking of the systems, applications and data traffic in the computer network. Network services are determined by demands of applications, information systems and data flows that they generate. In this are it is necessary to suitably analyse the data flows with regard to their security, application demands and other requests. The output of this phase is the analysis of the data flows, demands of applications on network infrastructure, formulation of priorities and quality of services.

The design of technological solution of the network—is divided into two main parts. The design of physical and logical network infrastructure. The physical infrastructure includes transfer technologies, the selection of passive and active network elements. On the contrary, the logical level contains mainly the design of the topological model, address and name structure, transfer protocols, logical security, etc. Inputs into this model are previous phases of T-NET methodics that create documents influencing the designed network solution. Among main outputs of this phase belong for example the design of technological solution in individual layers, utilization of outsourcing and cloud services or design of the computer network security [19, 22].

The evaluation of the designed solution—the main goal of this phase is to realize validation of the designed network model. Based on the differential analysis of the current state and the requested state of business, application, data and technological architecture, it is possible to conclude the design of the network or to proceed iteratively and execute the return to previous phases. If necessary, the redefinition of current design is realized. The main output of this phase is finalization of the design of the network in the form of documents, authorization of this design by the company management, preparation of migration plan and the design of network administration on abstract level.

The creation of migration plan—the migration plan presents a transformation from the initial state of the architecture to the objective state. It is suitable to use the principles of project control in this phase, or more precisely realize the mapping of the methodics on methodics and frameworks in this area. Furthermore, it is necessary to define and execute simulation tests before the initiation of the migration itself. The main output of this phase is the final implementation plan of the designed network.

The design of solution for control and monitoring of the network—This phase defines demands posed on the running of the computer network and the means of the running. Important steps are identification and utilization of current solution which serves for running of information technologies in the organization, definition of service and incidence approach to the operated network. The output of this phase is the definition of the computer network operation, the design of a monitoring system, of incidence policy or service management.

The revision of requests and changes in the network—this phase presents the continual approach to a network and enterprise architecture. The main goal of this phase is the identification of the maturity of the designed model of the computer network and identification of requests for changes. In this phase is it necessary to evaluate the current model of the network, execute its revision on strategic, business and technological level. The output of this phase is the determination of the concordance of the operated network with the enterprise architecture, evaluation of risks and creation of requests on changes of the operated network.

3.1 *Togaf and T-NET Comparison*

The comparison of TOGAF with T-NET methodics points out that the names of individual phases were changed as well as their content and partial functions. This fact is a logical outcome of the concentration on the area of computer networks. Main differences can be found in the approach inside individual phases, when outputs of original phase create inputs of the designed methodics. An example can be the information and system architecture where TOGAF defines the rules, requests and architectonic approaches to data and applications. On the contrary, T-NET methodics analysis standardized documents and other outputs for network design optimization. The original framework does not provide this point of view in individual phases, it merely points out the referential models and recommended rules for architecture building architecture. The suggested methodics results from the terminological basis of the original framework. This approach makes possible to maintain the compatibility with the original framework and to provide mutual interoperability among architectonic viewpoints, systems and TOGAF certified technologies.

Each phase is composed of several partial functions and parts that formalize the solution and operations necessary to be executed. The decomposition of phases enables to declare unambiguously the content, requests, partials steps and processes that are necessary to be absolved for successful accomplishment of the phase. Basic elements of T-NET methodics are goals, inputs, steps and outputs:

Goals—define the essence and elemental approaches in individual phases. They influence the content, focus and particular procedure in the given phase.
Inputs—fulfillment of the methodics goals is not an autonomic operation. Each of the partial phases is influenced both by external and internal factors that form the modelled architecture. These factors represent inputs in the form of documents, principles, rules, models or outputs from the previous phase.
Steps—define the sequence that has to be maintained in order to achieve the goals and conclude the phase. The steps include analyses, comparisons, utilizations of methods or external methodics.
Outputs—are the results of operation of the phase. All the outputs of the methodics have documented form and they are stored in architectonic repository. The outputs are thus documents comprising requests, basic principles, rules, strategic plans, analyses, architectonic viewpoints or further created tools, norms and standards.

It is obvious at the first glance that the elements of the methodics do not include criteria for the phase conclusion. However, this is a typical characteristic of iterative methodics that presuppose multiple passes through the phases or imperfection of initiatory passes. The iterative methodics often does not include such criteria or they are included only in the general level.

4 Conclusions

The analysis and design of computer networks have to take into account a wide range of aspects that have to be projected when network architecture is designed. The design itself includes for example a choice of transfer technologies, systems for ensurance of high accessability, security technologies, address and name plan, implementation of routing protocols, ensurance of network monitoring solutions and many others. Individual technologies and systems need to be integrated into the functional unit. Current solutions in the area of analysis and design of computer networks represent a technological point of view on network architecture building or interconnection of technologies from a specific manufacturer with selected frameworks. Not even the most superior technological solutions can guarantee the optimization of the design of a computer network, concordance with IT strategies and furthermore, it can not even testify anything about the efficiency of its interconnection with business, application and data architecture of the company.

The origin and utilization of the new T-NET methodics brings certain advantages and benefits. The basic benefit emerges from its very existence—similarly to other methodics, the main reason of its origin is systematic support of realization, design or implementation of certain solution. In case of T-NET methodics it is the systematic support in the course of the design of extensive computer network and its interconnection with enterprise architecture framework. Measurements of particular benefits of the methodics generally used to be the weakest link in the design of the whole methodics. The methodics itself is focuses, interconnects, evaluates and issues the recommendations in the specific area. Although this does not declare anything about its value and structure, methodologies originate and they are helpful in various areas of enterprise informatics. The greater the given project, the more the technological solutions have to yield under the pressure of other problems that needs to be solved. The verification of the T-NET methodics is currently in progress in multiple projects and case studies, for example in project interconnection of data centre and substation automation system") from the company ČEPS, a.s. (Czech Transmission System Operator).

Acknowledgement Support of the Specific Research Project of FIM UHK "SCM, Control of Markets and Production in Agent-based Computational Economics" is gratefully acknowledged.

References

1. Sportack, M.A.: Směrování v sítích IP. Vyd. 1. Computer Press, Brno, 351s (2004). ISBN 80-251-0127-4
2. Wenstrom, M., Velte, A.T.: Zabezpečení sítí Cisco: autorizovaný samostudijní výukový kurz. Vyd. 1. Computer Press, Brno, 753s (2003). ISBN 80-722-6952-6
3. Carroll, B.J., Ramasamy, K.: Network routing: Alogorithms, protocols, and architectures. 4th edn, Překlad Petr Matějů. Cisco Press, Amsterdam, 473 s. Exam certification guide series, (2008). ISBN 978-1-58720-211-7
4. Mccabe, J.D.: Network analysis, architecture, and design: pomocí OpenSource utilit a nástrojů. 3rd edn, Elsevier/Morgan Kaufmann Publishers, Boston, 473s, (2007). ISBN 978-012-3704-801
5. Penhaker, M., Darebnikova, M., Cerny, M.: Sensor network for measurement and analysis on medical devices quality control. Commun. Comput. Inf. Sci. **171**, 182–196 (2011)
6. Benikovsky, J., Brida, P., Machaj, J.: Proposal of user adaptive modular localization system for ubiquitous positioning. Lecture Notes Artif. Intel. **7197**, 391–400 (2012)
7. Mlynka, M., Brida, P., Machaj, J.: Modular positioning system for intelligent transport. In: 5th International Conference on Computational Collective Intelligence Technologies and Applications, Craiova, Romania, Studies in Computational Intelligence 513, vol. 513, pp. 115–124 (2014)
8. Balakrishnan, R., Hussain, F.: Advanced QoS for multi-service IP/MPLS networks, 1st edn. Wiley Publishing, Indianapolis (2008). 432s. ISBN 978-0-470-29369-0
9. Odom, W., Cavanaugh, M.J.: Cisco QOS exam certification guide: CCVP self-study, 2nd edn. Cisco Press, Indianapolis (2005). 730s. ISBN 15-872-0124-0
10. Schekkerman, J.: Enterprise architecture good practices guide: How to manage the enterprise architecture practice, 2nd edn. Trafford Pub, Victoria (2008). ISBN 978-142-5156-879
11. Schmitt, J.B.: Heterogeneous network quality of service systems. Kluwer Academic Publishers, Boston (2001). 234s. ISBN 07-923-7410-X
12. Nucci, A.: Design, measurement and management of large-scale IP networks: Bridging the gap between theory and practice, 1st edn. Cambridge University Press, Cambridge (2009). 394s. ISBN 978-0-521-88069-5
13. Vanus, J., Koziorek, J., Hercik, R.: Design of a smart building control with view to the senior citizens' needs. In: IFAC Proceedings Volumes (IFAC-PapersOnline), 12 (PART 1), pp. 422–427 (2013). ISSN: 14746670, ISBN: 9783902823533
14. Machacek, Z., Slaby, R., Hercik, R., Koziorek, J.: Advanced system for consumption meters with recognition of video camera signal. Elektronika Ir Elektrotechnika **18**(10), 57–60 (2012). ISSN: 1392-1215

15. Behan, M., Krejcar, O.: Modern smart device-based concept of sensoric networks. EURASIP J. Wireless Commun. Networ., Issue 1, No. 155, doi:10.1186/1687-1499-2013-155, Received 9 October 2012; Accepted 15 May 2013; Published 6 June 2013. ISSN 1687-1499
16. Dostálek, Libor. Velký průvodce protokoly TCP/IP: bezpečnost. 3. aktualiz. a rozš. vyd. Praha: Computer Press, 542s, (2002). ISBN 80-7226-675-6
17. Yusuf, F.H.: Network security technologies and solutions. 1st edn, Cisco Press, Indianapolis, 402s (2006). ISBN 15-870-5246-6
18. Müller, M.: Nástroje pro správu a řízení životního cyklu projektů [online], 2004-04, (2004). [cit. 2011-09-24]. http://www.lbms.cz/Reseni/_pdf/0704-ITS-MKT-IT-podpora-rizeni-projektu.pdf
19. Maresova, P., Klimova, B.: Investment evaluation of cloud computing in the European business sector. Appl. Econ. **47**(36), 3907–3920 (2015)
20. van den Berg, M., van Steenbergen, M.: Building an enterprise architecture practice: Tools, tips, best practices, ready-to-use insights. Springer, Dordrecht (2006). 203s (Enterprise series). ISBN 14-020-5606-0
21. Zuzák, F.: Podniková architektura – nástroj kvalitního řízení, [online], (2008). [cit. 2012-1-20]. http://si.vse.cz/archive/presentations/2008/podnikova-architektura-nastroj-kvalitniho-rizeni.pdf
22. Maresova, P., Halek, V.: Deployment of cloud computing in small and medium sized enterprises in the Czech Republic. E&M Ekonomie a Manage. **17**(4), 159–174 (2014)

Building a Measure to Integrate into a Hybrid Data Mining Method to Analyze the Risk of Customer

Huan Doan, Dinh Thuan Nguyen and Bao Quoc Ho

Abstract Analysis of customer risk is important task in business operation of the enterprise. This paper, first proposes the building of strength measure to rank clusters obtained from clustering algorithm. The next step builds an improved clustering algorithm by integrating of computing this measure into a clustering algorithm. This improved clustering algorithm is named Fuzzy C-Means-Rank (FCM-R). The final step proposes a hybrid data mining method to analyze the risk of the customer. The hybrid method includes two stages. In the first stage, we apply a classifying algorithm to classify customers into two groups: risk and no risk. In the second stage, we apply the improved clustering algorithm FCM-R to cluster risk customers into clusters and to rank clusters by measuring the risk level of clusters. The hybrid method has been applied to a real data set to generate clusters ranked according to the risk level from high to low. With such results, our proposed method will support enterprises to analyze customer risk and to propose appropriate risk management regulation for customers in each cluster.

Keywords Rank clusters · Risk analysis · Classifying algorithm · Clustering algorithm · Data mining

---

H. Doan (✉) · D.T. Nguyen (✉)
University of Information Technology, VNU-HCM, Ho Chi Minh City, Vietnam
e-mail: doan.huan@entersoft.com.vn

D.T. Nguyen
e-mail: thuannd@uit.edu.vn

B.Q. Ho (✉)
HCMC University of Science, VNU-HCM, Ho Chi Minh City, Vietnam
e-mail: hbquoc@fit.hcmus.edu.vn

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_71

1 Introduction

Customer segmentation is not the only important role in market prediction, but also in the risk analysis problem. In recent years, application of data mining methods for solving a risk analysis problem is interested by researchers and practitioners.

In this paper, we first propose the building of strength level measure of clusters to rank clusters obtained from clustering algorithm. Here, the strength level measure of clusters is risk level measure. A risk level measure of clusters is computed based on attributes had concern with risk of objects. The determining of attributes concerned object risk is proposed by domain experts. To facilitate for presenting, in this paper, we only consider attributes concerned risk, which had numerical data type.

Next, we build an improved clustering algorithm based on FCM++ [1] by integrating of computing of strength level measure of clusters into this clustering algorithm. This improved clustering algorithm is named FCM-R. FCM++ is an improved algorithm from FCM+ [2] and FCM+ is an improved algorithm from FCM [3, 4] by adding an indicator determined the appropriate number of clusters. The goal of building FCM-R is proposing a new algorithm not only making clusters, finding the appropriate number of clusters but also ranking clusters.

Finally, the paper proposes a hybrid data mining method to analyze the risk of the customer. The hybrid method includes two stages. In the first stage, we apply a classifying algorithm See5 [5] to classify customers into two groups: risk and no risk. The group of risk customer is input data of clustering algorithm FCM-R in the second stage. The classifying is to reduce data space because we only consider a group of risk customer in the risk analysis problem.

In the second stage, we apply the improved clustering algorithm FCM-R to cluster the risk customers into clusters and to rank clusters by measuring the risk level of clusters. To avoid affection of different measures of attributes to clustering result and strength level measure of cluster, the value of attributes of objects is normalized.

The proposed method experiment on a real data set of a company in Ho Chi Minh City, Vietnam, it obtains clusters ranked according to the risk level measure from high to low. The results obtained from the hybrid method will support to analyze customer risk and make a regulation of risk management suitably for avoiding loss.

The remainder of the paper is organized as follows. The next section reviews some existing methods about ranking cluster and segmenting customer by data mining technique. Section 3 presents proposed method, that comprises: building of strength level measure of clusters to rank clusters obtained from clustering algorithm, building the improved clustering algorithm FCM-R and proposing the hybrid method to analyze customer risk. Section 4 presents the experimental results on a real data set. The last section is the conclusion of the paper.

2 Related Works

The clustering algorithm generates clusters so that the objects in each cluster are similar based on a measure. However, the traditional clustering algorithm doesn't give a measure to rank clusters. In [6] introduced a strength measure for clusters-the integrated cohesion-which applies to optional weighted networks. The strength measure is applied to mining mailbox networks.

Analysis of customer risk regards as a special case of customer segmentation. In [7], agents are divided into two groups representing different confidence of adopting opinions in social networks. In [8] presented a two-phase model named 'Auto Insurance Customers Segmentation Intelligent Tool' to segment customers in insurance companies on the basis of risk by utilizing the concept of self-organizing map. In [9] proposed a procedure, joining quantitative value of RFM attributes and K-means algorithm into rough set theory (the LEM2 algorithm), to extract meaning rules for classifying the segmentation of customer value. In [10], the paper systematically integrates numerous data mining technologies to analyze customer value and thus promote customer value. First, the K-means and SOM methods are adopted to perform customer value analysis and segment customers based on customer value. Secondly, the decision tree is used to mine the characteristics of each customer segment. Third, different strategies are developed for differently valued customer segments and customer value is thus promoted. In [11] proposed a two-stage framework of customer behavior analysis using K-means clustering algorithm and Apriori association rule inducer. Now there are many works studied about customer segmentation, but there are very few works studied about analysis of customer risk by data mining techniques. This is the motivation for our proposing of a hybrid method to analyze customer risk by data mining techniques.

3 Method

3.1 Building of Strength Level Measure of Clusters to Rank Clusters

Here we discuss some problem concerning the ranking of clusters. For building of strength level measure of clusters to rank clusters obtained from clustering algorithm, first we have to determine ranking criteria concerned to objects in clusters. For example, with customer it can choose ranking criteria as the sales, profit or risk level. After a ranking criteria determined, we determine attributes of objects that concerns to this ranking criteria. The next will assess these attributes interrelated positively or negatively on ranking criteria. For example, with ranking criteria as profit, the sales will interrelate positively, but the cost will interrelate negatively. The final builds a formula of strength level measure of clusters according to the ranking criteria chosen.

In this paper, we choose ranking criteria of clusters is risk level. According to expert opinion, attributes of customer interrelate positively with risk level are: Total debt, Overdue total debt… Suppose that results of clustering algorithm are c clusters and G is set of attributes of customer that interrelate positively to risk level criteria. We build the risk ranking measure of clusters is followed as:

$$R_i = \frac{\sum_{i=1}^{ni} \sum_{j \in G} A_j}{n_i}, \quad \mathrm{i} = 1..\mathrm{c} \tag{1}$$

Here R_i is the value of the risk level measure of cluster i, A_j is attribute that belong attribute group G, n_i is the number of objects of cluster i.

3.2 Building of the Improved Algorithm FCM-R

To rank clusters in clustering algorithm, we integrate ranking measures of clusters additionally. Here we use algorithm FCM++ to integrate ranking measures of clusters. The improved clustering algorithm is named FCM-R and is presented as follows:

Algorithm FCM-R

Input: Data Set n objects x_i.

Output: c_{best} of data clusters ranked with c_{best} is the appropriate number of clusters

1) *Input n objects x_i, fuzzy parameter m > 1, epsilon small enough.*
2) *Input weighted vector w.*
3) *Input c_{min} and c_{max} (c_{min} >= 1, $c_{min} < c_{max} < n$).*
4) *Input the risk attribute group G*
5) *For c = c_{min} to c_{max}*
   a) Initialization matrix of members U_{cxn}
   b) Calculation of centre of cluster j C_j (j = 1,..,c)
   c) Update the distance matrix D (c x n)
   d) Update matrix of members U
   e) If the change of the matrix U is small enough compared to the previous step, then go to step f) otherwise go to step b)
   f) Based on the matrix U, x_i is arranged into clusters according to rules as follows: x_i will belong to any cluster that it has the greatest degree
   g) Save the c of data clusters at step c to disk
   h) Get all extremely marginal objects of all clusters
   i) Compute $\overline{\gamma}$ indicated the appropriate number of clusters
6) *Select c_{best} at neighbourhood when $\overline{\gamma}$ get a local minimum value after $\overline{\gamma}$ fluctuates fast and may start to have stable trend.*
7) *Take the c_{best} of data clusters at step c_{best} from disk.*
8) *For i = 1 to c_{best}*
      For j = 1 to n_i
         For k = 1 to |G|
            Compute R_i based on (1)
9) *Arrange clusters based on R_i, (i=1.. c_{best})*

3.3 Proposal a Hybrid Data Mining Method to Analyze the Risk of the Customer

The hybrid data mining method proposed includes two stages.

Stage 1 Selecting of attributes of customer is consistent with goal of analysis of risk. The next performs data preprocessing steps for building a training data set and a test data set. Applying of classifying algorithm See5 builds a decision tree. Based on the decision tree classifies customer data. Customers classified to risk group are input data of stage 2.

Stage 2 First, input data set is normalized. Next, Applying of clustering algorithm FCM-R on the input data set makes appropriate clusters and ranks clusters according to risk level measure from high to low. These clusters ranked are supplied to the user.

Model of proposed hybrid data mining method is depicted in Fig. 1.

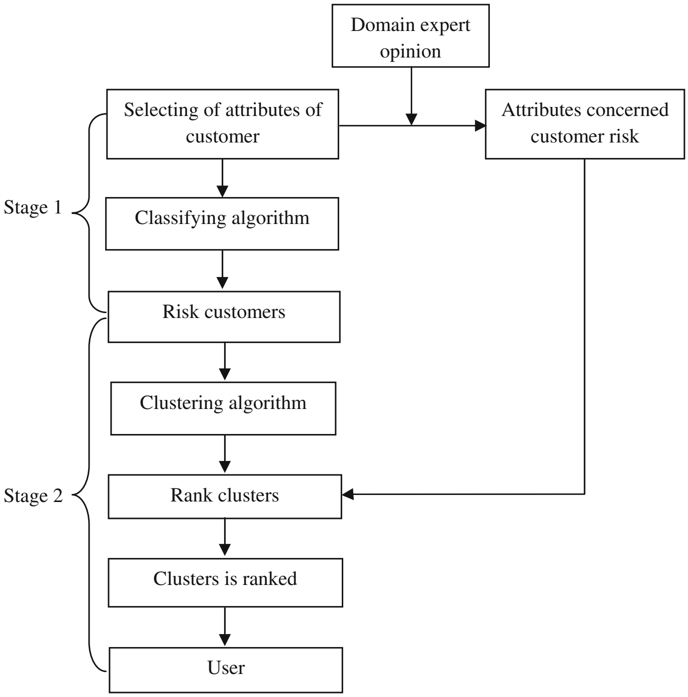


Fig. 1 Hybrid data mining model to analyze the risk of the customer

4 Experiment

We experiment our method on a real data set extracted in September, 2014 from the ERP database of a company in Ho Chi Minh City, Vietnam.

Stage 1: The original data set has 2848 debt invoices by customers. The data set after preprocessing steps has 443 customers with attributes: ID, Total debt, Average of the overdue number of days of debt, Overdue total debt, Over 30 days overdue total debt and a classification attribute: Risk. The training data set has 250 records and the test data set has 60 records. In both these data sets, based on expert opinion, the Risk attribute has two values: True or False.

Running the algorithm See5 with input data is above training data set and test data set, we obtained results as follows:

```
See5 [Release 2.10]    Sun Apr 19 11:59:00 2015

Read 250 cases (5 attributes) from DEBT 131 HK.data

Decision tree:
Overdue total debt <= 75000: 0 (67)
Overdue total debt > 75000:
:...Over 30 days overdue total debt > 1056000: 1 (90)
    Over 30 days overdue total debt <= 1056000:
    :...Overdue total debt > 6865000: 1 (51/10)
        Overdue total debt <= 6865000:
        :...Total Debt <= 2312780: 1 (16/3)
            Total Debt > 2312780: 0 (26/3)
Evaluation on training data (250 cases):
            Decision Tree
          ----------------
          Size      Errors
            5   16( 6.4%)   <<

           (a)   (b)    <-classified as
          ----  ----
            90    13    (a): class 0
             3   144    (b): class 1
        Attribute usage:
            100%  Overdue total debt
             73%  Over 30 days overdue total debt
             17%  Total Debt
Evaluation on test data (60 cases):
            Decision Tree
          ----------------
          Size      Errors
            5    7(11.7%)   <<

           (a)   (b)    <-classified as
          ----  ----
            21     2    (a): class 0
             5    32    (b): class 1
Time: 0.1 secs
```

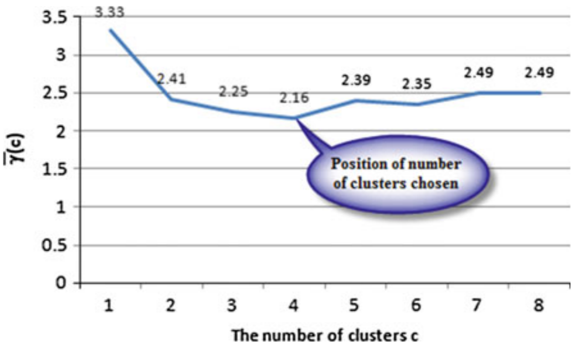


Fig. 2 The diagram shows the appropriate number of clusters

Table 1 Clusters ranked according to risk level measure from high to low

Risk level measure	Cluster	Customer ID	Total debt	Average of the overdue number of days of debt	Overdue total debt	Over 30 days overdue total debt
43.03 (very high risk)	4	C00375	871,842,518	53	749,311,518	605,225,418
9.39 (high risk)	3	C00386	437,437,000	3	437,437,000	0
		C00406	168,867,007	42	119,322,007	88,902,007
		C00322	267,697,720	57	195,787,720	137,816,840
		C00356	367,735,748	45	270,853,748	190,078,748
		C00395	227,663,183	39	111,872,183	111,872,183
		C00420	128,112,416	72	128,112,416	106,017,416
		C00431	395,218,413	28	207,261,353	98,518,953
		C00440	454,610,465	7	149,540,506	0
		C00443	384,591,982	17	131,148,182	49,229,182
		C00448	413,814,591	23	242,944,391	52,635,091
		C00454	1,067,174,224	14	527,971,524	0
5.42 (medium risk)	2	C00444	2,500,000	638	2,500,000	2,500,000
		C00428	248,053,729	638	248,053,729	248,053,729
		C00429	291,433,992	638	291,433,992	291,433,992
		C00430	975,000	1583	975,000	975,000
		C00433	92,229,600	496	92,229,600	92,229,600
		C00402	100,630,198	757	100,630,198	63,505,000
−0.80 (low risk)	1	C00403	62,880,320	6	46,135,640	0
		C00405	230,000	59	230,000	230,000
		C00398	20,760,000	127	20,760,000	20,760,000
		C00401	29,017,820	40	14,085,460	11,160,780
		C00423	116,433,000	37	64,093,500	64,093,500
		C00424	134,683,907	6	86,632,907	0
		C00425	108,987,254	51	84,822,254	60,239,254
		–	–	–	–	–

Based on the decision tree obtained, process of classifying the new customer data set included 133 customers, it obtains 72 customers belonged risk group.

Stage 2: Run algorithm FCM-R with the number of cluster c from 1 to 8 ($8 \approx \sqrt{72}$) with 72 customers data set belong to risk group (Risk = True), $\overline{\gamma}$ indicates the appropriate number of clusters that being 4 as in Fig. 2.

With the risk attribute group G included Total debt, Overdue total debt, Over 30 days overdue total debt and parameters m = 2, epsilon = 0.0001 and weights of each condition attribute is 0.25, we obtain 4 clusters ranked according to risk level measure from high to low as in Table 1.

In Table 1, we realize that cluster 4 (risk level measure = 43.03) is very high risk. The customer in the cluster 4 has values of attributes (Total debt, Overdue total debt, Over 30 days overdue total debt) that being much larger than the other. For cluster 4, we should have strong dealing to avoid loss. The cluster 3 (risk level measure = 9.39) also is high risk, although risk level is lower than cluster 4. For customers in this cluster, we also should have suitable dealing. Base on results obtained from our method, the user will have appropriate risk management regulation for customers in each cluster.

5 Conclusion

Applications of customer risk analysis have an important role in real life especially in enterprises. In the paper, we propose a hybrid data mining method to analyze the customer risk. To make a base for building the hybrid method, first we propose building of risk level measure to rank clusters obtained from clustering algorithm. Next, we propose an improved clustering algorithm, named FCM-R, by integrating a computing of strength level measure of clusters additionally into algorithm FCM++. The hybrid data mining method proposed includes two stages.

In the first stage, applying of classifying algorithm See5 classifies customers into two groups: risk and no risk and choosing of risk customer group is input data of stage two. In the second stage, applying of improved clustering algorithm FCM-R clusters the risk customers into clusters and ranks the clusters according to risk level measure.

We experimented with the proposed method on the real data set for illustrating for our method. The method has generated some useful results for the user.

References

1. Doan, H., Nguyen, D.T.: An adaptive method to determine the number of clusters in clustering process. In: Proceedings of The International Conference on Computer and Information Sciences, ICCOINS 2014, June 3–5, 2014, IEEE, Malaysia, pp. 1–6, ISBN 978-1-4799-4391-3 (2014)
2. Nguyen, D.T., Doan, H.: An approach to determine the number of clusters for clustering algorithms. ICCCI 2012, Part I, LNAI 7653. Springer, Heidelberg, pp. 485–494 (2012)
3. Bezdek, J.C., Ehrlich, R., Full, W.: FCM: The fuzzy c-means clustering algorithm. Comput. Geosci. **10**(2–3), 191–203 (1984)
4. Hathaway, R.J., Bezdek, J.C.: Recent convergence results for the fuzzy c-means clustering algorithms. J. Classif. **5**, 237–247 (1988)
5. http://www.rulequest.com/see5-info.html. Accessed May 2014
6. Bar-Yossef, Z., Guy, I., Lempel, R., Maarek, Y.S., Soroka, V.: Cluster ranking with an application to mining mailbox networks. Knowl. Inf. Syst. **14**(1), 101–139 (2008)
7. Shang, Y.: Consensus formation of two-level opinion dynamics. Acta Math. Sci. **34**(4), 1029–1040 (2014)
8. Hanafizadeh, P., Paydar, N.R.: A data mining model for risk assessment and customer segmentation in the insurance industry. Int. J. Strateg. Decis. Sci. **4**(1), 52–78 (2013)
9. Cheng, C.-H., Chen, Y.-S.: Classifying the segmentation of customer value via RFM model and RS theory. Expert Syst. Appl. **36**(2009), 4176–4184 (2009)
10. Liang, Y.-H.: Integration of data mining technologies to analyze customer value for the automotive maintenance industry. Expert Syst. Appl. **37**, 7489–7496 (2010)
11. Farajian, M.A., Mohammadi, S.: Mining the banking customer behavior using clustering and association rules methods. Int. J. Ind. Eng. Prod. Res. **21**(4), 239–245 (2010)

Design of Blind Adaptive Equalizer Using Frequency Domain Rectangular Contour Algorithm

Muhammad Haseeb Akram, Shahzad Amin Sheikh and Sana Abid Khan

Abstract Rectangular Contour Algorithm (RectCA) is a novel blind equalization technique which generally deals with the oblong QAM constellations. It uses the geometry of Rectangular QAM to correct the phase error within 180° hence improving the convergence rate. In this paper, we have improved the performance of existing RectCA in terms of computational complexity by implementing the frequency domain block based RectCA (FD-RectCA). Furthermore, convergence rate of FD-RectCA is improved by normalizing the update equation. Simulation results are carried out using 16-Rectangular QAM constellation which signifies superior performance of the proposed technique.

1 Introduction

In modern communication system, Inter Symbol Interference (ISI) deteriorates the system performance when the high speed data is transmitted over a band limited channels. To minimize ISI, adaptive equalization is usually employed. Conventionally, adaptive equalizers with training sequence are used but these equalizers have a disadvantage of consuming a large bandwidth due to the requirement of an initial training sequence [1]. To compensate the constraint of bandwidth, blind equalizers are used in which no training sequence is required.

Blind equalization algorithms mainly depend upon the cost function which consists of the transmitted data constellation statistics and equalizer output. This cost function is minimized by updating the equalizer weights using stochastic

M.H. Akram (✉) · S.A. Sheikh · S.A. Khan
National University of Sciences and Technology, Islamabad 44000, Pakistan
e-mail: haseebakram67@ee.ceme.edu.pk

S.A. Sheikh
e-mail: sasheikh@ceme.nust.edu.pk

S.A. Khan
e-mail: sakhan307@gmail.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_72

gradient descent approach [2]. There are several blind equalization algorithms available which include Constant Modulus Algorithm (CMA) [3], Reduced Constellation Algorithm (RCA) [4], Multimodulus Algorithm (MMA) [5], Square Contour Algorithm (SCA) [6], Improved Square Contour Algorithm (ISCA) [7], Modified Multimodulus Algorithm (MMMA) [8], and Generalized Cross Contour Algorithm (GCrCA) [9]. These algorithms give their best performance using square QAM [3–7] and cross QAM [8, 9]. Some existing algorithms are modified to be used with Rectangular QAM (RQAM). Rectangular Contour Algorithm (RectCA) has been proposed recently [10] which mainly deal with constellations with odd number of bits per symbol such as cross QAM and rectangular QAM. In this paper, by keeping in view the hardware implementation of RectCA, the block based implementation in frequency domain for RectCA has been proposed. This approach gives two-fold advantages in the form of lesser computational complexity and an improved convergence rate.

2 Blind Equalization Model

Consider the basic communication system as shown in Fig. 1, in which s_n shows the i.i.d (independently identically distributed) transmitted samples, v_n shows the AWG (additive white Gaussian) noise, x_n represent the input to the equalizer and a_n is the decision device output. The output of the equalizer y_n is defined as:

$$y_n = w_n^T x_n \tag{1}$$

where $w_n = [w_{0,n}, w_{1,n}, \ldots, w_{N-1,n}]^T$ and $x_n = [x_{0,n}, x_{1,n}, \ldots, x_{N-1,n}]^T$ are the weight vector and input vector of the equalizer of length N respectively.

To minimize the cost function with respect to weight vector is the main purpose of an equalizer. This minimization is carried out by updating the weights of the equalizer. The general form of weight update equation of an equalizer by using a stochastic gradient algorithm is given as:

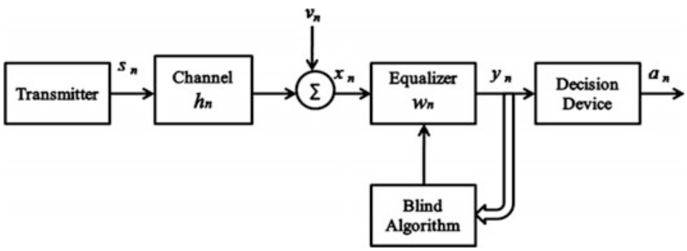


Fig. 1 Basic communication system

$$w_{n+1} = w_n - \mu e_n x_n^* \tag{2}$$

where e_n is the error function; which is different for every blind equalization algorithm. μ is the step size parameter which controls the convergence speed.

3 Use of Rectangular QAM

The use of rectangular QAM as a transmitted constellation in communication systems is very useful because of three main reasons. First is its capability of detecting a phase error within 180° [11]. Secondly, it gives a huge advantage over cross QAM for computation of average symbol error probability (SEP). Last and the most important reason is its ability to implement gray coding, which is required in performance enhancement of Personal Area Network (PAN) systems [12], where odd bit QAM is incorporated by gray coding. The disadvantage of using rectangular QAM is that the transmitted symbols require more energy which can be reduced by minimizing the adjacent points, distance in rectangular constellation [10]. Due to above mentioned reasons; Rectangular QAM is the main motivation behind the development of Rectangular Contour Algorithm [10].

4 Rectangular Contour Algorithm

Rectangular Contour Algorithm is a modified form of Square Contour Algorithm (SCA) which reduces the disparity between transmitted constellation and cost function to achieve the improved convergence rate for odd bit Rectangular QAM [10]. The Rectangular Contour Algorithm cost function is expressed by:

$$J_{RECT} = E\left[\left(\left(\left|by_{R,n} + ay_{I,n}\right| + \left|by_{R,n} - ay_{I,n}\right|\right)^p - \left(|ab|R_{RECT}\right)^p\right)^2\right] \tag{3}$$

Here y_R and y_I represents the real and imaginary part of the output symbol y (n) and p is a positive integer and its value is limited to 1 or 2. The parameters *a* and *b* are dependent parameters for Rectangular QAM constellation. Dispersion constant R_{RECT}^p can be formulated by considering perfect equalization i.e. y(n) = s(n) and given as:

$$R_{RECT}^p = \frac{E\left[\left(\left|bs_{R,n} + as_{I,n}\right| + \left|bs_{R,n} - as_{I,n}\right|\right)^p \cdot R''\right]}{|ab|^p \; E[R'']} \tag{4}$$

and,

$$R'' = \left(\left|bs_{R,n} + as_{I,n}\right| + \left|bs_{R,n} - as_{I,n}\right|\right)^{p-1} \cdot \left\{ \begin{aligned} & bsgn\left[bs_{R,n} + as_{I,n}\right] + bsgn\left[bs_{R,n} - as_{I,n}\right] \\ & + j\left(asgn\left[bs_{R,n} + as_{I,n}\right] - asgn\left[bs_{R,n} - as_{I,n}\right]\right) \end{aligned} \right\} \cdot s_n^* \tag{5}$$

In (5), s_R and s_I represents the real and imaginary parts of the transmitted symbol s_n and sgn represents the signum function. The error function of RectCA is given as follows:

$$\begin{aligned} e_{RECT} &= 4by_R\left(4b^2y_R^2 - (|ab|2R_{RECT})^2\right)A + j4ay_I\left(4a^2y_I^2 - (|ab|2R_{RECT})^2\right)B \\ A &= \frac{sgn(by_R)}{2}\left(sgn(by_R + ay_I) + sgn(by_R - ay_I)\right) \\ B &= \frac{sgn(ay_I)}{2}\left(sgn(by_R + ay_I) - sgn(by_R - ay_I)\right) \end{aligned} \tag{6}$$

5 Frequency Domain Rectangular Contour Algorithm

The frequency domain version of RectCA is presented in this section. The motivation behind this idea is to reduce the computational complexity as well as improvement of convergence rate of RectCA by converting the time domain RectCA into its frequency domain counterpart with block based weight update and error computation schemes [13]. In our case, FD-RectCA can be initiated by making the block size ‘β’ equal to the equalizer length ‘N’ i.e. N = β. From (1), it is clear that equalizer output is the resultant of convolution between its weight vector and input vector, so this equation can be transformed to multiplication if we implement it in frequency domain by using Discrete Fourier Transform (DFT). This transformation is carried out by implementing overlap save method of convolution in which input is arranged as overlapping blocks of symbols [13]. The block of input symbol, at each run, is given as:

$$x_{2\beta} = \left[x_\beta[n]^T x_\beta[n-1]^T\right]^T = \left\{ [x(n\beta); x(n\beta+1); \ldots; x(n\beta+\beta-1)]^T \times \begin{bmatrix} x((n-1)\beta); x((n-1)\beta+1); \ldots \\ \ldots; x((n-1)\beta+\beta-1) \end{bmatrix}^T \right\}^T \tag{7}$$

The above equation explains the structure of overlapping blocks of input vector $x_{2\beta}$ of length $2\beta \times 1$. The first β symbols are the new symbols and the last β symbols are the old ones arranged as [new old] block of data. This input vector $x_{2\beta}$ is changed to transformed domain through Fourier operation (FFT in simulation) to produce $x_{2\beta}$. Similarly, the block based version of equalizer output is given in frequency domain as:

$$Y_{2\beta} = X_{2\beta}[k]W_{2\beta}[k], \quad k = 0,1,\ldots,2\beta - 1 \tag{8}$$

$W_{2\beta}$ is frequency domain version of weight vector. As the overlap and save method is used, only last β symbols are correct. Thus, by taking inverse DFT of (2) and saving last β symbols the equivalent time domain equalizer output can be obtained as:

$$y_{2\beta} = F^{-1}(Y_{2\beta}) \tag{9}$$

$$y_{\beta} = [\,I_{\beta} \quad 0_{\beta}\,]y_{2\beta} \tag{10}$$

where I_β and 0_β are the identity matrix and the matrix containing all zeros of size β x β. The Error is calculated by using the following equation:

$$e_{\beta-RECT} = 4by_{\beta-R}\left(4b^2y^2_{\beta-R} - (|ab|2R_{RECT})^2\right)A + j4ay_{\beta-I}\left(4a^2y^2_{\beta-I} - (|ab|2R_{RECT})^2\right)B \tag{11}$$

where,

$$A = \frac{\mathrm{sgn}\left(by_{\beta-R}\right)}{2}\left(\mathrm{sgn}\left(by_{\beta-R} + ay_{\beta-I}\right) + \mathrm{sgn}\left(by_{\beta-R} - ay_{\beta-I}\right)\right)$$

$$B = \frac{\mathrm{sgn}\left(by_{\beta-R}\right)}{2}\left(\mathrm{sgn}\left(by_{\beta-R} + ay_{\beta-I}\right) - \mathrm{sgn}\left(by_{\beta-R} - ay_{\beta-I}\right)\right)$$

In error computation, we do not employ the direct frequency domain computation. Rather, we first calculate time domain error vector. This is because of the non-linearity of the error function in blind equalization in contrast to the non-blind equalizers which have a linear error function due to which direct implementation in frequency domain is possible [13]. Hence, in FD-RectCA, error is first computed in time domain and then converted to frequency domain error vector $E_{2\beta}$ by first appending zeros in $e_{\beta\text{-RECT}}$ and then performing the DFT operation given by:

$$E_{2\beta} = F\left[\, e^T_{\beta-RECT} \quad 0^T_{\beta} \,\right]^T \tag{12}$$

The weight vector in frequency domain is calculated as

$$W_{2\beta} = F\begin{bmatrix} w_{\beta}^{T} & 0_{\beta}^{T} \end{bmatrix}^{T} \tag{13}$$

Finally, frequency domain weight update equation can be given as:

$$W_{2\beta}[k + 1] = W_{2\beta}[k] - \mu Y_{2\beta}^{*}[k]^{*}E_{2\beta}[k] \tag{14}$$

5.1 Normalized Frequency Domain RectCA

FD-RectCA shows significant results as far as computational complexity is concerned but the convergence speed is almost the same as of its time domain counterpart. The improved convergence rate can be achieved by using normalized FD-RectCA, in which adaptive step size is used. The new update equation after normalization is given as [13]:

$$W_{2\beta}[k] = W_{2\beta}[k] + \frac{\mu}{\gamma_{2\beta}[k]} Y_{2\beta}[k]^{*}E_{2\beta}[k] \tag{15}$$

where γ is the input signal's power and given by:

$$\gamma_{2\beta}[k] = \lambda\gamma_{2\beta}[k] + (1 - \lambda)\left|Y_{2\beta}[k]\right|^{2}, \quad k = 0,1,\ldots,\beta - 1 \tag{16}$$

where λ is the forgetting factor.

5.2 Gradient Constraint

According to (13), $W_{2\beta}$ is obtained by adding β zeros in its time domain counterpart. To implement this constraint for every iteration, a modified weight update equation is formed by an additional operation known as "gradient constraint" [13]. In this operation, the resulting vector of the product of $E_{2\beta}$ and $X_{2\beta}$ is translated into time domain, by setting the last β samples equal to zero as to enforce (13) and the time domain product is again converted back to frequency domain. The new weight update equation after using gradient constraint is expressed as:

$$W_{2\beta}[k] = W_{2\beta}[k] - \mu F\begin{bmatrix} I_{\beta\times\beta} & 0_{\beta\times\beta} \\ 0_{\beta\times\beta} & 0_{\beta\times\beta} \end{bmatrix} F^{-1}\left(X_{2\beta}^{*}{}^{*}E_{2\beta}\right) \tag{17}$$

Gradient constraint involves two additional FFT/IFFT operations which slightly increase the computational complexity, but this arrangement gives the advantage of reduced steady state error as compare to unconstrained FD-RectCA in (15).

6 Computational Complexity

The major advantage of adopting frequency domain implementation is the reduction of computational complexity especially for large equalizers as compared to the time domain implementations. In each case, the computation of number of real multiplication per symbol is required to carry out the comparative analysis of computational complexity of FD-RectCA with RectCA. Starting from the time domain RectCA, in (1), the equalizer output requires N complex which corresponds to 4N real multiplications. Similarly, the error term in (6) consists of 8 real multiplications. Finally, the weight update equation requires N and 2N real multiplication which combines to make 6N real multiplications. So, the total number of real multiplication required in time domain RectCA is:

$$\mathrm{total}_{\mathrm{TD-RectCA}} = 10\mathrm{N} + 8 \tag{18}$$

In frequency domain, we deal with the DFT operations which are implemented as FFT operation in simulation. For the block size ‘β’, an FFT of size 2 β is required to compute $Y_{2\beta}$. Similarly, IFFT and FFT of size 2 β are required to compute equalizer output and $E_{2\beta}$. Finally, gradient constraint deals with a pair of FFT/IFFT of size 2β each, and each 2β point FFT operation, on a block of β samples, requires a total of $\beta\log_2(2\beta)$ complex multiplications. In our case, it requires $5\beta\log_2(2\beta)$ or $20\beta\log_2(2\beta)$ real multiplications. Similarly, equalizer output involves 8β real multiplications. Time domain error term in (11) needs 8β real multiplications. Lastly, weight update equation without normalization in (14) requires total of 12β real multiplications which includes product of $X_{2\beta}$ with $E_{2\beta}$ and multiplication with μ of the resultant term. So, the total numbers of real multiplications required for FD-RectCA are:

$$\mathrm{total}_{\mathrm{FD-RectCA}} = 20\beta\log_2(2\beta) + 28\beta \tag{19}$$

Hence, for each input it requires:

$$\mathrm{total}_{\mathrm{FD-RectCA}} = 20\log_2(2\beta) + 28 \tag{20}$$

For normalized weight update equation represented in (15), 4β complex or 16β real multiplications required. Similarly, (16) requires 8β real multiplications which constitutes the total of 28β real multiplications for normalized weight update equation. So, the total number of real multiplications required by NFD-RectCA is:

$$\mathrm{total}_{\mathrm{NFD-RectCA}} = 20\beta\log_2(2\beta) + 44\beta \tag{21}$$

Table 1 Computational complexity comparison for different lengths of equalizer

β	RectCA	FD-RectCA	NFD-RectCA
2	30	80	92
4	50	100	112
8	90	120	132
16	170	140	152
32	330	160	172
64	650	180	192
128	1290	200	212
256	2570	220	232
512	5130	240	252
1024	10,250	260	272

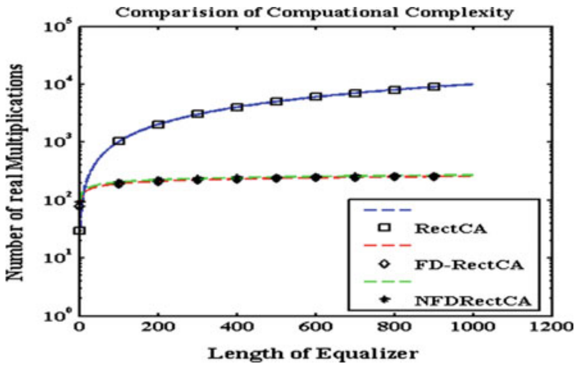


Fig. 2 Comparison of computational complexity for different algorithms

For a single input symbol:

$$\text{total}_{\text{NFD-RectCA}} = 20\log_2(2\beta) + 44 \tag{22}$$

In Table 1 and Fig. 2, comparison of computational complexity of RectCA, FD-RectCA and NFD-RectCA is illustrated. It is clear from the comparative analysis that FD-RectCA and NFD-RectCA is showing significant reduction in computational complexity as compare to time domain RectCA. Moreover, NFD-RectCA is exhibiting slight increase in computational complexity as compare to FD-RectCA but this can be indemnified by getting a significant improvement in convergence performance as explained in next section. Therefore, FD-RectCA and NFD-RectCA can be widely used in the real time scenarios where channels with large impulse responses are employed.

7 Simulation Results

To justify the performance of aforementioned algorithms, simulations have been carried out using 32-Rectangular QAM signal constellation with SNR of 30 dB and step size μ equal to 1e-9. A 9-tap equalizer is used with the value of 1 for the centre tap, other taps being set to zero [7]. Typical voice band communication channel has been employed with impulse response given in [7]. The transmitted signal using 32-Rectangular QAM, received signal after passing through the channel and the equalized output is shown in Fig. 3.

The performance of RectCA, FD-RectCA and NFD-RectCA in terms of MSE is shown in Fig. 4. MSE simulations are generated as an average over 20 Monte Carlo runs with dispersion constant and forgetting factor λ is taken as 1.9122 and 0.988 respectively. It can be seen that NFD-RectCA is giving fast convergence as compared to FD-RectCA and RectCA which illustrates the need of normalization of FD-RectCA. So, NFD-RectCA can give a better convergence performance by a slight increase in the computational complexity.

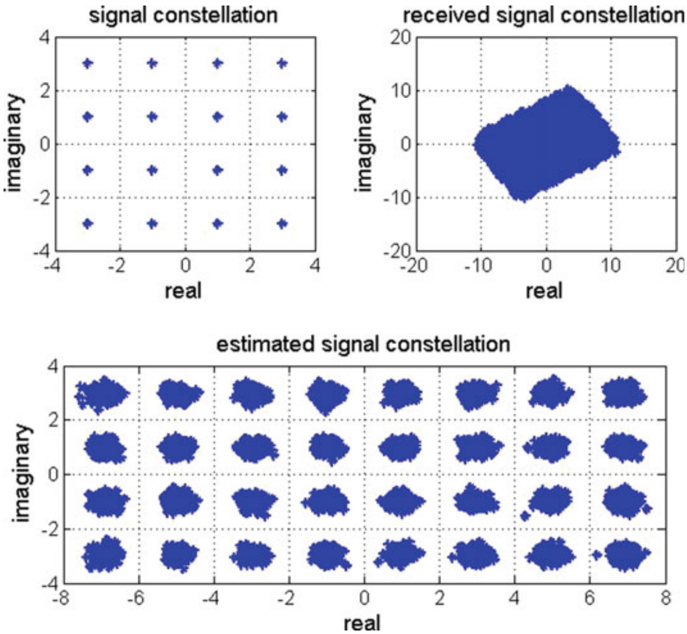


Fig. 3 Simulation results of FD-RectCA for 32 R-QAM

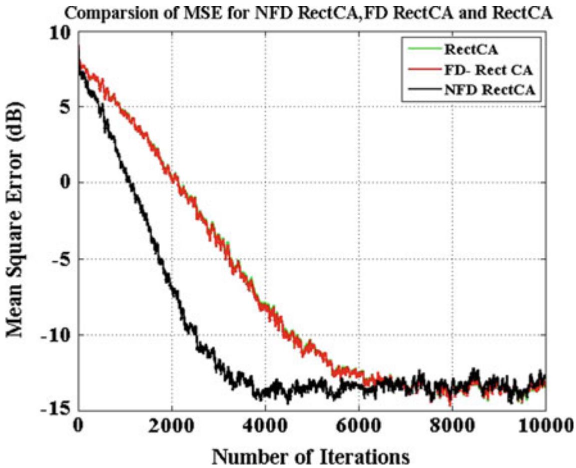


Fig. 4 MSE traces for RectCA, FD-RectCA and NFD-RectCA for 32 RQAM constellations

8 Conclusion

In this paper, the concept of Frequency Domain Equalization (FDE) has been applied on Time Domain Rectangular Contour Algorithm to design the Frequency Domain Rectangular Contour Algorithm. Simulation results confirmed that the significant reduction in computational complexity has been achieved by using the FD-RectCA as compare to time domain RectCA. Moreover, convergence rate is also improved by using the concept of normalization in the proposed technique.

References

1. Sharma, V., Naveen Raj, V.: Convergence and performance analysis of godard family and multimodulus algorithms for blind equalization. IEEE Trans. Signal Process. **53**(4), 1520–1533 (April 2005)
2. Syed, A.H.: Fundamentals of Adaptive Filtering. Wiley-IEEE, NewYork (2003)
3. Godard, D.: Self recovering equalization and carrier tracking in two-dimensional data communication systems. IEEE Trans. Commun. COM-28, 1867–1875 (1980)
4. Sato, Y.: A method of self-recovering equalization for multilevel amplitude-modulation systems. IEEE Trans. Commun. COM-23, 679–682 (1975)
5. Yang, J., Werner, J.J., Dumont, G.A.: The multimodulus blind equalization and its generalized algorithms. IEEE J. Sel. Areas Commun. **20**(5), 997–1015 (2002)
6. Thaiupatump, T., Kassam, S.A.: Square contour algorithm: a new algorithm for blind equalization and carrier phase recovery. In: Proceedings of 37th Asilomar Conference on Signals, Systems and Computers, pp. 647–651 (2003)
7. Sheikh, S.A., Fan, P.: New blind equalization techniques based on improved square contour algorithm. Digit. Signal Process. J. **18**(5), 680–693 (September 2008) (Academic Press, Inc., ISSN: 1051-2004)
8. Yuan, J.T., Tsai, K.D.: Analysis of the multi modulus blind equalization algorithm for non-square rectangular QAM signal constellations. In: ICSF"O4 Proceedings (2004)

9. Abrar, S.: A new cost function for the blind equalization of cross-QAM signals. In: 17th International Conference on Microelectronics, pp. 290–295 (December 2005)
10. Khan, Q.U.: Analysis of M-ary R-QAM and a new blind equalization algorithm RectCA using R-QAM, MS EE. Dissertation, NUST, EME (2011)
11. Yuan, J.T., Tsai, K.D.: Analysis of the multi modulus blind equalization algorithm for non-square rectangular QAM signal constellations. In: 7th International Conference on Signal Processing, pp. 302–306 (September 2004)
12. Victor Demjanenko Ph.D., Paul Marzec MS, and Alberto Torres Ph.D. IEEE P802.15 Working Group for Wireless Personal Area Networks (WPANs), Reasons to use non-squared QAM constellations with independent I&Q in pan systems, IEEE P802.15-15-03-0311-00-003a
13. Usman Gul, M.M., Sheikh, S.A.: New blind equalization techniques based on improved square contour algorithm, Elsevier Digit. Signal Process. J. 20(6), 1697–1710 (Dec 2010) (Academic Press, Inc., ISSN: 1051–2004)

Using Static Code Analysis for Improvement of Job Data Availability in Bobox Task Scheduling

Lukas Krizik, Zbynek Falt, Miroslav Cermak and Filip Zavoral

Abstract This paper proposes an optimization method for a task-based parallel framework called Bobox. The method is based on a static code analysis; it reduces a number of short running tasks. Since the Bobox scheduler has only limited runtime information, it can schedule a task even though the task does not have all input data. Such short-running tasks cause an unnecessary scheduling overhead. To achieve more effective scheduling, the proposed optimizer analyzes usage of inputs to remove short running tasks.

Keywords Parallelism · Optimization

1 Introduction

The Bobox project is a task-based framework for parallel computing. In such a framework, short-running tasks cause bigger CPU consumption by the framework. A static code analysis can be used to detect and eliminate such execution paths in the user code. Since the core and interface language of the Bobox framework is C++, static code analysis becomes more difficult in proportion to the lack of tools. This has been the biggest pitfall for any static analysis of C++ code. However, it has become less marginal with a growing support for tooling in the Clang compiler front-end [5]. Clang exposes C++ code as a user-friendly *Abstract Syntax Tree (AST)* structure.

L. Krizik · Z. Falt · M. Cermak · F. Zavoral (✉)
Charles University, Prague, Czech Republic
e-mail: zavoral@ksi.mff.cuni.cz

L. Krizik
e-mail: krizik@ksi.mff.cuni.cz

Z. Falt
e-mail: falt@ksi.mff.cuni.cz

M. Cermak
e-mail: cermak@ksi.mff.cuni.cz

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_73

In this paper, we propose a method for optimizing code using the Bobox framework. By analyzing AST, the optimizer is able to diagnose and potentially transform user code in order to eliminate short execution paths.

1.1 Bobox

Bobox is a framework [1–3] for task-based parallel programming. The framework uses a fixed number of worker threads, each one with its own scheduler. Communication between tasks uses a column-based data model, the most significant implementation detail that favors data processing problems. A task is scheduled to be executed when it has an unprocessed input [4]. The runtime environment handles the implementation details of the task-based parallel environment such as scheduling and the parallel execution of tasks, data transport and control flow. A programmer uses a declarative way to setup the environment with a *model* which defines the way individual tasks are interconnected. The model is used to create a *model instance* which is used for execution of the tasks.

The basic element of a model instance is a task called a box. Data is sent using an *envelope*, a column-based data structure. An empty envelope is a special type of envelope called a *poisoned pill*. When a box receives a poisoned pill in its input, there will be no more data sent to this input. All the paths of the model instance are required to end in another special type of box called When the *termination box* receives a poisoned pill, the execution is finished and the pipeline is deallocated.

Boxes are executed in three steps: (1) in *prologue* the box provides access to its inputs for the user code; (2) *action* is the main place for user code; (3) *epilogue* handles the scheduling of the next task based on two criteria - a task is scheduled again if it has got an unprocessed input and it has processed some input in the action step or if it requested to be scheduled again.

Boxes are main objects of the interest for optimization because they are the main location for the user code. Based on the static analysis of the action step, additional code can be injected in order to inform Bobox about the properties of the task.

2 Prefetch Method

Tasks in the Bobox framework are represented in form of boxes, which can have zero or more inputs. The problem is that the scheduler lacks information about a box execution, specifically about processing of its inputs. There are three cases of input data requirements for a meaningful task execution: (1) a task does not need data from any input at all; (2) a task needs data only from some inputs out of multiple inputs; (3) a task needs data from all inputs.

An execution of a task from the second and the third case without necessary input data adds significant overhead to the execution. If a task is executed before it has all necessary input data, it finishes its execution immediately. However, a developer can inform the scheduler about the number of envelopes on a specific input necessary for a meaningful box execution. The goal of the optimizer is to search for a usage of box inputs and inject prefetch member function calls accordingly.

2.1 Restrictions to Optimization

The optimization cannot be applied to the source code when some restricting conditions are satisfied. The algorithm for the prefetch optimization does not produce any runtime checks, but the static analysis checks whether it is safe to apply changes to the source code. First, the analyzer tests a box for various conditions whether it can be optimized at all, then it tests all box inputs for another set of conditions. If a box and its input pass all tests, the box input is prefetched.

The optimization of a box is discarded if at least one of these restrictions is satisfied:

(global.1) There are no functions with the user code for the action step, i.e., a class does not override any of functions representing the action step.
Rationale: If there is no user code in a class, no input is used. Improbable case, but it has to be taken into an account.

(global.2) There are no inputs.
Rationale: Nothing to optimize.

(global.3) There is no mapping of names to inputs created by using the Bobox helper macro.
Rationale: Currently, the optimizer identifies inputs by names associated by the Bobox helper macro. If there is no such mapping, the optimizer does not detect any input.

(global.4) A definition of the overridden `init_impl` function is inaccessible.
Rationale: This member function represents the initialization step of the box execution and it is the location for prefetch calls. If the analyzer cannot access its definition, there is no place to put function calls.

(global.5) The corresponding `init_impl` in the base class is private.
Rationale: The analyzer is able to override the initialization member function, but the programmer may assume that the corresponding initialization function from the base class is called. Therefore, there has to be a call to the base class function. However, if the function is inaccessible due to the protection level, the function call would break a compilation.

Single input optimization restrictions:

(single.1) There is already the prefetch call for the input.
Rationale: The programmer already handles the optimization.

(single.2) The optimizer cannot detect whether data from the input is likely to be necessary.
Rationale: The decision to prefetch the input is as good as the opposite decision. It can happen when data from an input is necessary only in a single branch of the code or not at all.

Overriding Initialization Step
Prefetch calls are placed in the box initialization step. If there is an accessible implementation of the initialization function, prefetch calls are injected into this definition. If the initialization function is not overridden, the optimizer is able to inject the overridden implementation by itself. The problem with an injection of the completely new overridden initialization function is that the previously overridden initialization function can prefetch inputs itself. Fortunately, if the prefetch call on the same input is called multiple times, only the last call has effect as it overrides the previous call. Therefore, the injected function calls prefetch functions on the beginning of the definition and the call to the previous overridden initialization function as the last statement, see Listing 1.1.

Listing 1.1. The generated box initialization function definition.

```
virtual void init_impl ( ) {
// prefetch_envelope for desired inputs
some base : : init_impl ( ) ; }
```

Calling the previous corresponding `init_impl` function as the last statement ensures that if there is a prefetch call, it is the one that counts.

2.2 Searching for Values in Code

To check the restriction *(single.1)*, the analyzer must search for prefetch calls on inputs in a code likely to be executed in the box initialization step. Furthermore, the restriction *(single.2)* describes searching for a usage of a box input in the box action step. Basically, the analyzer must search for values[1] that are present on all paths or paths likely to be executed in *Control Flow Graph (CFG)* of a specific function definition. Clang tooling libraries provide a developer with AST, but it is also possible to construct CFG using Clang static analyzer code.

Fortunately, the construction of CFG from AST is not necessary since a slightly modified default AST traversal can achieve the same result. In more details, the `RecursiveASTVisitor` template provides member functions with names

[1]A value is a too abstract notion. For example, such a value can be the name of a callee in a call expression represented by the `CallExpr` AST node.

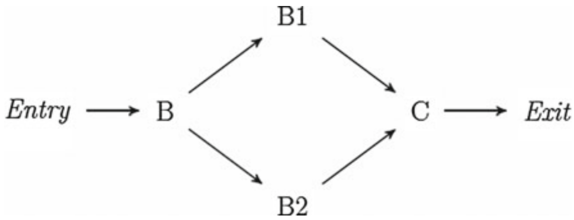
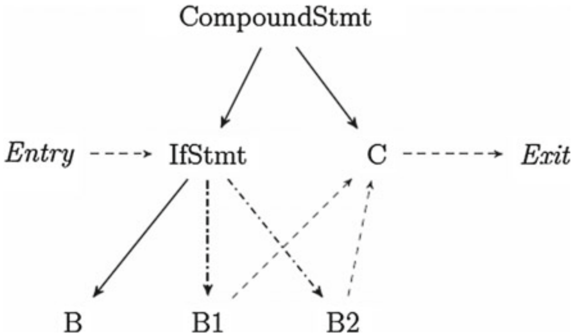


Fig. 1 The CFG representation of a code with a single if statement

Fig. 2 An example of AST tree traversal



starting with `Traverse*` ,[2] which are responsible for a traversal of the internal graph structure. Actually, these functions are responsible for traversing the structure kept internally in Clang as if it was AST. Those member functions can be *overridden* using CRTP.

Figures 1 and 2 show an example of a traversal of the same code in CFG and AST structures. Figure 1 shows CFG of the code with a single if statement with non-empty then and else branches followed by a non-empty block. B represents the condition expression block, B1 and B2 represent then and else branches of the if statement, and C is the last non-empty block on both paths from *Entry* to *Exit* blocks. Figure 2 shows the AST representation of the same code combined with nodes and edges from Fig. 1 with the simplification that IfStmt is followed by the block C in the CompoundStmt node. *Entry* and *Exit* nodes and dashed edges do not exist in AST. The only shared edges between CFG and AST are dashed-dotted edges to B1 and B2. For example, if the analyzer searches for a value on the path passing through block B1, assuming it starts in CompoundStmt, it visits node by node in the graph depth-first search algorithm:

1. `CompoundStmt`
2. `IfStmt`
3. *`Block B`*

---

[2] *represents a type of an AST node such as `TraverseStmt` for a statement or `TraverseCallExpr` for a call expression.

4. `Returns to IfStmt`
5. *`Block B1`*
6. `Returns to IfStmt`
7. `Returns to CompoundStmt`
8. *`Block C`*
9. `Returns to CompoundStmt and finishes`

The exactly same sequence of code blocks that would be searched in CFG is: block B, block B1 and block C. *Entry* and *Exit* blocks are empty thus not interesting for the optimization process.

Divide and Conquer
When searching a value in CFG, it would be necessary to either traverse the same path multiple times or remember which nodes and paths were already processed. On the other hand, divide and conquer algorithm design paradigm fits perfectly to the described custom AST traversal.

The implementation of the search algorithm in the optimizer tool enhances `RecursiveASTVisitor` functionality as it has already the well-established interface using the widely known pattern. The problematic part is to identify which AST nodes can affect a control flow of a program. There are relatively many classes for AST nodes. However, all language constructions that can affect the control flow are directly mapped to Clang AST nodes in the Stmt class hierarchy and its Expr sub-hierarchy.

Searching for a value in a linear program flow is straightforward. The algorithm visits node by node testing whether it contains a searched value. If the search algorithm encounters a selection statement, it runs itself on every branch. If a searched value is found in all branches, it is found for a current selection statement. If it encounters an iteration statement, it can continue searching in a loop body based on a tool configuration. Jump statements stop searching. A value is searched only in the left-hand side expression of logical expressions because of a short-circuit evaluation.

2.3 Loop with Fixed Number of Iterations

It was already mentioned that loop bodies are searched for values by default since they will likely be executed. But this option is configurable in the optimizer tool. A user can choose to disable search in loop bodies that cannot be proven to be executed at least once. The simple case of a loop where it can be proven that its body is executed at least once is a `for` loop with a fixed number of iterations which was widely used in an old C code.

If the analyzer can prove that i is not modified in the initialization and condition expressions, it can evaluate the condition as a constant expression. The tool is implemented on top of the compiler, which already has facilities necessary for operations such as the constant expression evaluation or the constant unfolding optimization. Clang exposes functions related to the constant expression evaluation in the `Expr` class. For example, it can evaluate an expression as a boolean condition, but it succeeds only if an expression is really constant for the compiler, which is not in this case. The tool can trick the compiler by setting temporarily the variable initialization declaration to be a constant expression. The same trick can be used to analyze even more complex loops.

2.4 Searched Values

This section describes what values are searched in the CFG; it is useful especially for the implementation of heuristics noted in Sect. 2.1.

Available Inputs

Inputs in box member functions are referred using `input_ index_type` which is constructed with an index of an input, or `inarc_index_type` which can be gathered from `input_index_type` using the specific `basic_box` member function. The Bobox framework also provides a helper macro for the assignment of names to inputs.

Currently, the optimizer works only with names of static member functions generated by the helper macro and identifies inputs by these names. In a future implementation, it can identify inputs by indexes, but it requires a more complex implementation with an extensive usage of the constant expression evaluation.

Prefetched Inputs

For already prefetched inputs, the overridden `init_impl` function is searched. The optimizer searches for `prefetch_envelope` member function calls. It checks function calls whether a callee is the one from the `basic_box` [3] class and collects input names that could be resolved from the first parameter. The first parameter is expected to be a call to the related static member function generated by the helper macro. Actually, a prefetch call is expected to look exactly as the injected prefetch call by the optimizer.

Used Inputs

Two member functions on a box are searched for a usage of inputs, `sync_body` and `sync_mach_etwas`. These member functions represent the action step. When a member function with such name is found, it is tested whether it overrides the `basic_box` member function.

[3] A function with the same name but a different signature can be implemented. Such function hides the base implementation.

There are two cases when data from input is considered to be necessary:

(1) if there is a call to the `pop_envelope` function on the `basic_box` class, or,
(2) if there is a helper variable of the type input_stream<> for working with a box input and there is a call to any member function on this variable.

3 Results

The goal of this optimization method is to reduce a number of short-running tasks. Specifically, tasks that are scheduled even though they do not have all input data for a meaningful execution. The gain in a speedup with this optimization method is the scheduling overhead. It only matters how often this situation occurs in user code.

In order to measure the scheduling overhead, it is necessary to maximize the number of wrongly scheduled tasks. A wrongly scheduled task has to have at least two inputs and there has to be some delay between getting data to its inputs. A bigger delay increases the possibility of the task being scheduled without all necessary data.

3.1 Model

There has to be a box with more than one input for the wrong scheduling and there can be more such boxes for a bigger effect. Now, it is necessary to create some mechanism to maximize the number of situations when a box is scheduled wrongly. This mechanism consist from a chain of boxes where each one creates input data for one specific input on all boxes that are supposed to be scheduled wrongly. All boxes from this chain, apart from the last box, then trigger scheduling of the next box in the chain that will create data for a different input on wrongly scheduled boxes.

Figure 3 shows the model used in measurements. However, the number of distribution and collection boxes as well as the number of their inputs and outputs may vary. The figure shows only a *pattern* used in the model construction. If all collection boxes prefetch only the first input or no input at all, the expected scheduling in the model instance constructed from this model is as:

1. control_box
2. distribution_box
3. 3x collection_box - pop data from the first input and wait
4. distribution_box
5. 3x collection_box - pop data from the second input and wait
6. distribution_box
7. 3x collection_box - pop data from the last input and produce an output
8. sink_box

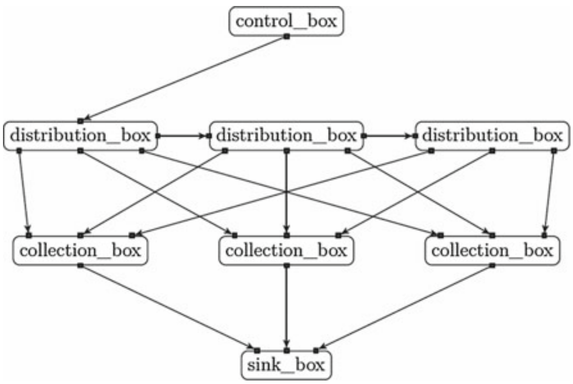


Fig. 3 The model used in measurements of the prefetch method optimization

If all collection boxes prefetch their all input data, step 3 and step 5 are skipped. These prefetch calls can save up to six schedules when boxes do nothing meaningful.

3.2 Benchmarks

There are two boxes in the model without any description. The control box repeatedly produces data for distribution boxes. It yields each time after data is produced. The sink box is there for a debugging purpose. All boxes do minimal work apart from a communication with the Bobox framework to maximize the impact of the scheduling.

The results for the model with 10 distribution boxes and 10 collection boxes shows that optimized code is very slightly faster in most measurements. However, the gain in a speedup is negligible even for a bigger number of iterations or a bigger number of boxes. Furthermore, the gain in a speedup is approximately the same for a bigger number of iterations as for a lower number of iterations. The speedup appears to be constant for measured results. These results did not correspond with the expected behavior.

3.3 Optimizer Tweak

It is necessary to look more into details of the model and the Bobox framework implementation to understand the low speedup. Both distribution and collection boxes are stateless. The Bobox framework allocates some objects of stateless boxes, initializes them by calling the init_impl member function and reuses those allocated objects all over again. But, init_impl is not called again when object is reused. Thus, prefetch calls added by this optimization method are useless when the framework reuses the boxes.

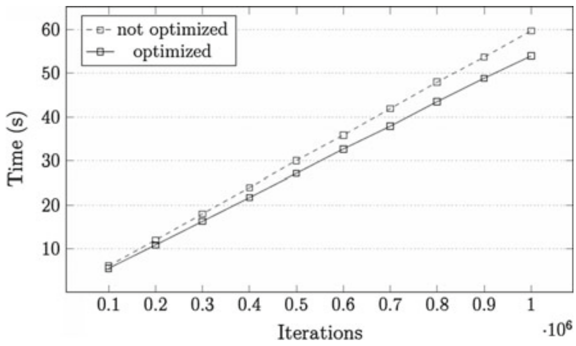


Fig. 4 Benchmarks for the prefetch optimization method with ten distribution and ten collection boxes and the *attach to execution body* tweak turned on

The prefetch optimization method is enhanced to add all prefetch calls for all used inputs to the end of the box execution member function body. This option is configurable and turned off by default. Figure 4 shows results with this optimization tweak turned on compared to not optimized code. The improvement in a speedup is clearly visible.

3.4 Discussion

The Bobox scheduler is greatly optimized. Even in the presented scenario with approximately 10 boxes scheduled 9 times out of 10 wrongly, the speedup with 1 million iterations is only 10 %. Furthermore, tests were measured on a single logical thread. It is much harder to achieve this scenario with multiple logical threads. The scenario itself is artificial to showcase the optimization potential. On the other hand, there are software products where every second matters and there is no reason to turn off this optimization.

However, there is still one factor that is not tested well in the presented scenario, cache usage. Since all boxes do not work with data excessively, their cache usage is minimal. Most data can be probably held in cache the whole execution time. On the other hand, a scheduled box without all input data probably immediately reschedules. Thus, its impact on a cache is negligible anyway.

The only concern about achieved results is the possible big number of false positives when assuming data from an input is necessary as it is described in the second case in the Sect. 2.4. A typical example is the situation when the input is *probably* necessary only in the single branch of the code, which is not a strong assumption. It was necessary to make such a soft assumption in order to make the optimization get the expected result using typical scenarios. An analysis of this issue can be significantly enhanced in the future.

4 Conclusion

The Bobox project is a framework for task-based parallel computing. The environment handles programming issues related to parallel computing and it does bring some overhead. This overhead concentrates in times when a task starts and finishes its execution. Thus, tasks should run for enough time to make this overhead negligible. The prefetch optimization method tries to reduce special cases of short-running tasks. Running a task without necessary input data causes this task to finish almost immediately. This method detects such tasks and informs the Bobox framework about their input requirements. The analysis makes some assumptions, which can result in false positives. However, if a task accesses input data using some of expected ways, this method detects it. The scheduling mechanism in the Bobox framework is highly optimized. The gain in a speed is not overwhelming even in cases where this optimization method excessively reduces an amount of unnecessary scheduling. On the other hand, there is no reason to refuse any gain in an application speed.

Acknowledgment This work was supported by the Czech Science Foundation (GACR), project P103-13-08195S, by Charles University Grant Agency (GAUK) project no. 472313, and by Specific Research project SVV-2014-260100.

References

1. Bednarek, D., Dokulil, J., Yaghob, J., Zavoral, F.: Parallelization framework for data processing. Adv. Inf. Technol. Appl. Comput. (2012)
2. Bednarek, D., Dokulil, J., Yaghob, J., Zavoral, F.: Data-flow awareness in parallel data processing. In: Intelligent Distributed Computing IV, Springer, Berlin (2012)
3. Falt, Z., Krulis, M., Bednarek, D., Yaghob, J., Zavoral, F.: Locality aware task scheduling in parallel data stream processing. In: Proceedings of the 8th International Symposium on Intelligent Distributed Computing—IDC'2014, Springer,Berlin, pp. 331–342 (2014)
4. Falt, Z., Bednarek, D., Krulis, M., Yaghob, J., Zavoral, F.: Bobolang—A language for parallel streaming applications. In: Proceedings of the 23rd International ACM Symposium on High-Performance Parallel and Distributed Computing, ACM, Vancouver, pp. 311–314 (2014)
5. "Clang": a C language family frontend for LLVM [online]. Clang homepage

The Design Trends of Keystream Generator for Stream Cipher for High Immunity Attacks

Saif Uldun Mostfa Kamal, Hayder Saad, Mustafa Musa Jaber, Mohammed Hasan Ali and Karam Dhafer

Abstract Due to the latest changes in observing the external network related threats within the stream cipher, it become necessary to address these threats in order to identify the generator suitable for avoiding such threats. In this paper, the researcher addresses the current threats of immunity attacks in the stream cipher. Such attacks are resulted from the correlation within the key stream's multiplexer. The key stream generators are also introduced in order to clarify its working process in avoiding attacks. After all, a comparison of key stream generators are resulted where it can be used as guidelines for other researches in designing key stream generator for high immunity attacks in the cipher.

Keywords Stream cipher · Immunity attacks · Generator based multiplexer · Key stream

The original version of this chapter was revised. The erratum to this chapter is available at 10.1007/978-3-319-24584-3_114

S.U. Mostfa Kamal (✉) · H. Saad · K. Dhafer
Faculty of Computer Science, UKM, Selangor, Malaysia
e-mail: saifaldeenalsamir@gmail.com

H. Saad
e-mail: haydersaad@gmail.com

K. Dhafer
e-mail: karamdhafer@gmail.com

M.M. Jaber
Biomedical Computing and Engineering Technologies (BIOCORE)
Applied Research Group, Universiti Teknikal Malaysia Melaka (UTeM),
Melaka, Malaysia
e-mail: mustafamusajaber@yahoo.com

M.H. Ali
Faculty of Computer Science and Software Engineering, Universiti Malaysia Pahang,
Pahang, Malaysia
e-mail: eng.mohammed12@yahoo.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_74

1 Introduction

Cryptography in general consists of both block and stream ciphers where the main settings of block cipher is explained by symmetric key cipher that run on a group of bits with a fixed-length in order to facilitate the block calling. However, the stream cipher is differently designed where the symmetric key cipher is integrated bits of plaintext with a pseudorandom bit stream which known as a key stream. Such process is usually not corresponded to the clear-cut functioned by some modes based block cipher primitive which acts efficiently as a stream cipher [1].

Generating a secure key stream has been always considered to be the most essential factor in securing the cipher in which it comes under the cryptography. In this regards, the security of cipher keys was reported in the processing and transferring data where key stream works randomly so that it is unfeasible to multiply or guess them, which carries the main standard for generating the number randomly [2, 3].

Researchers such as [4] reported on the fact that attackers are usually depends on the length of these keys in order to predicate the level of security that ciphers designed for. Therefore, those researchers encouraged the development of stream cipher security and determine the main threats in contrast to period, speed, and other elements.

Alhamdan and Wong [5] stated that ensuring a secure cipher is depending on the design of its key stream generator where immunity threats are usually take place. Hence, key stream generator is used to offer stream cipher reliable related sectary settings against different cryptographic attacks. This led us to conclude that the period required for processing data in key stream is a key factor in which it make it impossible for attackers to recover the cipher's key or internal state from the key stream.

This is a part of linear feedback shift register (LFSR) which it considered to be the main unit of the cipher where it depends on the linear function of random bits as an input. Current research showed that there different configurations for setting LFSR for the aim of sampling the cipher execution. Along with the effectiveness of LFSR, there are several issues regarding the ability of LFSR against the high immunity attacks [6–8]. Therefore, LFSRs are currently used in nonlinear linked by Boolean function for avoiding the inherent of LFSR linearity in which it works on categorizing it into nonlinear of LFSRs. This result output of several LFSR to generate a secure key stream. In addition, there is the nonlinear filter generator that performs based on a different approach and one single LFSR, the case is not the same for other LFSR input-based filter. In same cipher, both types are begin used in order to withstand certain particular cryptographic attacks [9].

Even though, bit based LFSRs are disreputably slow in processing some cipher requests. Therefore, iterations are used to foster data processing using one bit of key stream. This paper focuses on the key stream design for stream cipher in order to avoid high immunity attacks with good statistical properties.

2 Issues in Designing Keystream

The stream ciphers are widely utilized nowadays for securing shared information that gives a consequential advantage in enhancing the security in the coming years [10].

In order to provide an effective key stream generator, it is important to include several linear feedback shift registers to get a key stream with required statistical properties. Basically, the attack occurs when the correlation is detected over number of known sequences with a definite output with the LFSR used as a divider and conquers attacks [11–13].

In most cases, the correlation attack is carried on generator of the multiplexer in line with the correlation attacks on the data shift register. The high immunity resulted in correlation attack make the most of the bitwise correlation between data of output sequence and other input sequence [14].

However, [14] reported on the main aspects of correlation attack on the data LFSR in the MUX generator, they utilized bitwise correlation between the output of MUX and data input of K sequences along with the MUX. This in turn result two different attacks where the first one depends on the state of the MUX which predicated to minimize the hamming distance between the output segment and the corresponding data input segment. While the second attack was found to be based on the posteriori probabilities between the output segments of key stream and data input segments.

Therefore, few researchers have recommended applying a fast correlation attacks in stream cipher within the multiplexer by depending on the data resulted from the feedback polynomial of LFSR. This was also addressed to be as a threat where any increment in the number of similarity check increases the possibility of successful retrieving of the LFSR data [15].

Meanwhile, there are other potential threats associated with the key stream bit in which attackers may guess the multiplexer address by using a proper calculation method based registry. In such case, it may lead to conflict for the registry where attackers may only guess half bits of the registry. In addition, this may cause the resulting of recursion linear of these bits via feedback polynomial especially if polynomial data is small number of guessed bits which imply a much larger of sequence bits.

Based on the aforementioned possible threats, determining the limit of key stream design in avoiding these threats is highly recommended. Meanwhile, since the stream ciphers are limited to encrypt plaintext without generating errors propagation that could be a significant in some cases where transmission errors are immensely expected [16]. Thus, this paper consists on reviewing the characteristic of current key stram for stream ciphers that are less complex hardware circuitry and more suitable than block ciphers in hardware.

2.1 Pseudorandom Number Generator (PRNG)

Pseudorandom generator is known as a facilitator for increasing the short random sequence into a longer sequence that still "random" to opponents with limited elements [17]. The short random sequence is known as the key of the cipher [18].

PRNG provides consistently distributed and independent sequence of keys for the cipher to allocate the random numbers generated by the Truly Random Number Generator (TRNG) [19]. The working process of this generator consists on the process that produces the independent elements in an random way [20]. As such, it relays on the bit oriented keys that provide a different forms of sequences described in the following section.

3 Key-Stream Generators

There are huge quantity of key stream generator based bit proposed the Following discuses the most popular which are Product generator, geffe generator, stop and go generator, alternating step generator, A5/1 generator, shrinking generator, Knapsack Generator and Multiplexing generator and review most characteristics of them that related with linear complexity and period.

3.1 Linear Feedback Shift Register

The LFSR consists of the length in the n stage along with a linear feedback function. The conceptual process of this function depends on changing the storage unit to the next sequence where the result will be shifted to the storage unit as shown in Fig. 1. In registry, the output sequence presented by the LFSR has the maximum period of $2^L - 1$ if the feedback function of the LFSR is primitive and its initial state is a non-zero state.

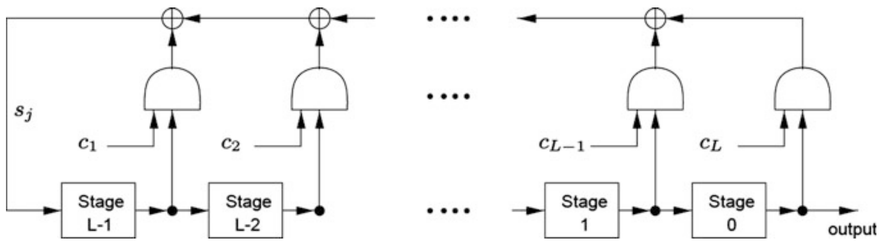


Fig. 1 LFSR of length L

Several considerations need to be stated during each unit of time [21] as shown in Fig. 1. The stages of LFSR were initialed in Fig. 1 as (111)2, which result the dimensions of s(n) as [11100101110010111001O……]. The number of states is $2^3 - 1$. The condition of (000)2 cannot be determined as the initial condition. However, the output of the s(L) stages LFSR is determined by the following Eq. (1):

$$\mathrm{S(n)} = \sum_{i-1}^{L} c(i) \cdot s(n-i) = c(1) \cdot s(n-1) \oplus c(2) \cdot s(n-2) \oplus \cdots \oplus c(L) \cdot s(n-L) \tag{1}$$

The properties of the LFSR can be represented by f(x) polynomial over GF(2). For the Lc LFSR explained as follow:

$$f(x) = 1 + \sum_{i=1}^{L} c(i) \cdot x^i = 1 + c(1) \cdot x^1 + c(2) \cdot x^2 + \cdots\cdots + c(L) \cdot x^L \tag{2}$$

Researchers stated that the primitive polynomial of order n help the process of the resulted sequence where it repeat itself at every $2^L - 1$ with considering the initial condition except for zero values.

A primitive polynomial of degree n is an irreducible polynomial that divide $x^P + 1$, where $p = 2^L - 1$, but does not divides $x^d + 1$ for any $d < 2^L - 1$ [22].

The polynomial cannot be adjusted if it is divisible by total of 1 or itself. This led to suggest that If f(x) and g(x) are two polynomials over GF(2), they are irreducible if GCD (f(x), g(x)) = 1 [23]. Examples of reducible and irreducible polynomials are:

$$f(x) = 1 + x^1 + x^2 + x^3 = (1 + x)\,(1 + x^2)\ \text{(reducible)}$$
$$f(x) = 1 + x^2 + x^3 \text{(irreducible)}$$

This generator represents main threats where m-sequences process a low linear complexity in nonlinear sequences.

3.2 Product Generator

The product generator is represented by the maximum length of m-LFSRs with length $L_1, L_2 \ldots L_{m,}$ where $m \in Z^+$ is pair wise relatively prime, with AND combining function:

$$\mathrm{F}(y_1,\ y_2, \ldots y_m) = y_1 \cdot y_2 \cdot y_3 \cdots y_m = \prod_{j=1}^{m} y_j \tag{3}$$

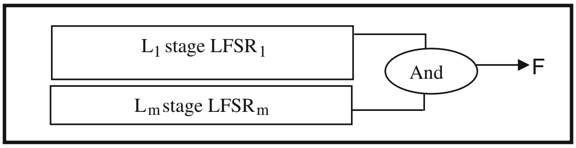


Fig. 2 Product generator

The key-stream generated has period $\left(2^{l_1}-1\right)\cdot\left(2^{l_2}-1\right)\ldots\ldots\left(2^{l_m}-1\right)$ and linear complexity $L = L_1L_2 \ldots L_m$. This generator is reported to be a weak provider for a randomness, despite of its good linear complexity [24] as shown in Fig. 2.

3.3 Geffe Generator

The Geffe generator is functioned based on the grouping of a nonlinear with a three LFSRs within the lengths of the standard N_1, N_2, N_3 that are pairwise comparatively with nonlinear functions as explained by Eq. (4) [25].

$$F\ (Y_1,\ Y_2,\ Y_3) = Y_1Y_2 \oplus (1 +\ Y_2)\ Y3 = Y_1Y_2 \oplus Y_2Y_3 \oplus Y_3 \tag{4}$$

Where the period represented by $(2^{N_1}-1)\cdot(2^{N_2}-1)\cdot(2^{N_3}-1)$ with linear complexity of

$$L = N_1N_2 +\ N_2N_3 + N_3 \tag{5}$$

as shown in Fig. 3.

Schneier [25] reported that Geffe generator is not suitable for high immunity attacks with complicated statistical channels, he linked such lack to the limitation of data begin generated from $LFSR_1$ and $LFSR_3$ into the output sequence where it simplify correlation attacks.

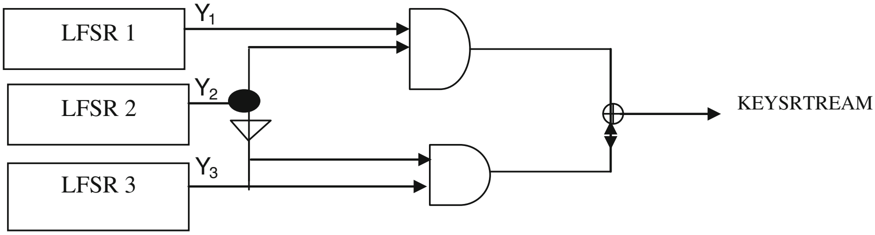


Fig. 3 Geffe generator

3.4 Stop and Go Generator

Figure 4 shows the working process of stop and go generator where the main aim is to reduce the threat level in the regular clock of LFSR based on nonlinearity by clocking it irregularly. It relay on the output of LFSR to adjust the clock of another LFSR in the same cipher where the input of the clock in the second LFSR is controlled by the output of LFSR1 as shown in Fig. 4.

The period is measured in this generator by $(2^k - 1)^2$ where k refers to the first $LFSR_1$ and second $LFSR_2$ with considering the linear complexity k $(2^k - 1)$.

Researchers found that using such generator for high immunity attacks is not recommended, this is because that $LFSR_2$ is not clocked, the bit from LFSR2 keep acting as the output of the generator where it was detected by the alternating step generator [11, 26].

3.5 Alternating Step Generator

This generator is focusing on applying one bit from LFSR1 to control the clocking of both $LFSR_2$ and $LFSR_3$ to be clocked at each step as shown in Fig. 5 with the dimensions of the register X, Y, Z:

- The clock is initialed for the Register X.
- X is resulting to 1 then Y is clocked and Z is repeated.
- If X_{output} is 0 then Z is clocked and Y is repeated.
- Finally, the output bits of Y and Z are XORed.

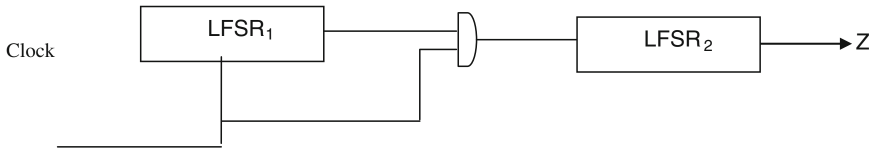


Fig. 4 Stop and go generator

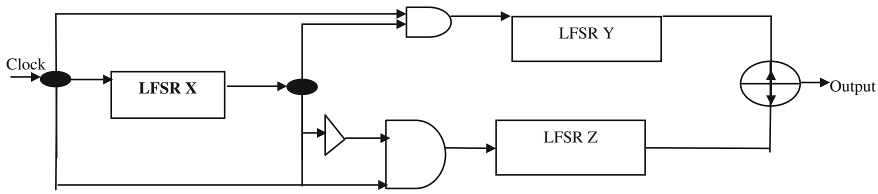


Fig. 5 Alternating step generator

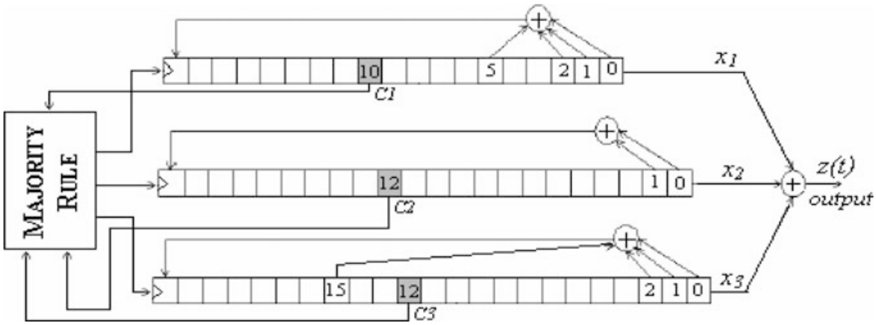


Fig. 6 A5/1 generator

The sequence from the final step (x) consists of period of $2^{n_1} \cdot (2^{n_2} - 1) \cdot (2^{n_3} - 1)$ that lead to linear complexity of L(x) generated by the following (6):

$$(n_2 + n_3) \cdot 2^{n_1 - 1} < L(x) \leq (n_2 + n_3) 2^{n_1} \tag{6}$$

In addition, this generator is open to the divide and conquers attacks that consist on guessing the initial state of $LFSR_1$ on the control register X [4, 27].

3.6 A5/1 Generator

A5/1 contains three LFSRs, each one process a certain range of control bit. This followed by the clocked of LFSR when the control bit consent with major of the three LFSRs, which leads to execute at least two clocks of LFSRs for the entire step. Also LFSR enables one bit to be XORed that generates the keystream bit. This customization indicates that LFSRs of the A5/1 are connected with each other; therefore, breaking LFSRs is not applicable that most attacks get benefits from the small 64-bit size [4, 28]. Figure 6 shows the A5/1 generator flow.

3.7 Shrinking Generator

The shrinking generator, a control LFSR X is used to select a portion of the output sequence of a second LFSR Y. The keystream produced is, therefore, a shrunken version (also known as an irregularly decimated subsequence) of the output sequence of Y, as specified in following steps and depicted in Fig. 7.

- Registers X and Y are clocked.
- If the output of X is 1, the output bit of Y forms part of the keystream.
- However, if the output of X is 0, the output bit of Y is discarded.

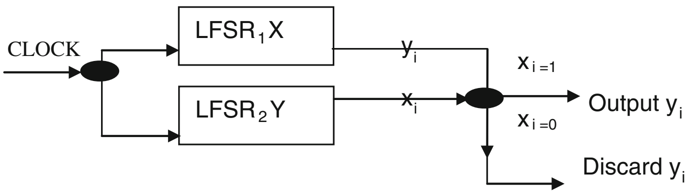


Fig. 7 Shrinking generator (Mollin, 2007)

It is inconvenient to use the shrinking generators in practice since they do not generate key-stream at a constant rate [4, 29].

3.8 *Knapsack Generator*

This is a real implementation of the first kind of the nonlinear filter generator which presents a set of weights that were adapted as part of the key. This leads to utilize additional integer arithmetic which used as a part of the keystream that generated from the additional integer. The generated cycles obtain better period and linear complexity properties that initialized based on the following procedure (Fig. 8):

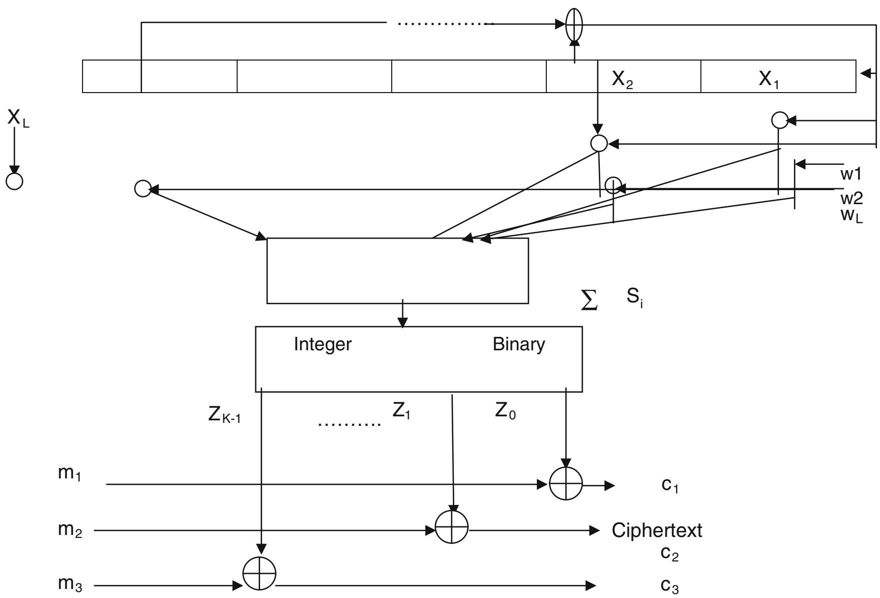


Fig. 8 Knapsack stream cipher

Input: LFSR < $G;H$ > with the initial state $H_0 = (h_1^0;h_2^0 ; : : : h\ x_L^0) \bmod Q$;
L weights $w_1;w_2; : : : ;w_L$.
For i = 1; 2; : : :, do:

- Compute the i-th state of the LFSR X_i.
- Determine the sum where represents $S_i = \sum_{k=1}^{L} h_k^i w_k \bmod Q$.
- Extract some bits of S_i to be Z_i.

Output: Z_i, for i = 1; 2;:::.

The security of the nonlinear filter, the knapsack sum, lies in the hardness of the knapsack problem [30]. In addition, the period and linear complexity of the generator is K $(2^L - 1)$.

3.9 Multiplexer Generator

The classification of multiplexer generator is belongs to the second nonlinear sort, while the filters are belongs to the basic ideas design of keystream generators: that mainly calls technology-driven protocols. In this generator, it is possible to use h = m if $2^h - 1 \leq n$ and if $h \neq m$ then (h) must also suit $2^h \leq n$. After choosing (h), it's easier to indicate the (h) stages $a_{x1}, a_{x2}, \ldots, a_{xh}$ of SR1 at any time (t) when the binary h-tuple $(a_{x1}(t), a_{x2}(t), \ldots, a_{xh}(t))$ is interpreted as a symbol of a natural number that symbolized earlier by N_t, as $1 \leq N_t \leq 2^m - 1$ if h < m in a reversible mapping as follow:

$$\theta : \{0, 1, \ldots, 2^h - 1\} \rightarrow \{0, 1, \ldots, n - 1\}$$

if h = m

$$\theta : \{1, \ldots, 2^h - 1\} \rightarrow \{0, 1, \ldots, n - 1\}$$

Based on the indications of h x1, x2…xn and θ, it define a new sequence u(t) called multiplexed sequence, by $u_t = b\theta_{(N_t)}(t)$. The multiplexer then choose one stage of the SR2 stages at each time (t), which means $u_t = b_{t+}\theta(N_t)$ for every t.

This generator provides a correlation between sequence key-stream generator and $LFSR_2$ output that are cryptographically weak and falls to correlation and guess and determinations. A Sequence that results from this algorithm has the following feature Period length is $((2^N - 1) \cdot (2^M - 1))$ (Fig. 9):

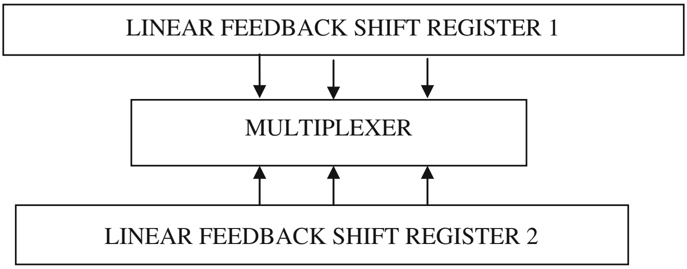


Fig. 9 Multiplexer generator

– Linear complexity is $M\sum_{i=0}^{n}\binom{N}{i}$
– It passes all randomness tests [30, 31].

4 Comparison of Key-Stream Generators

A wide range of keystream generators have not been mentioned by the researcher in this paper, but it has summarized the most influenced generators found in the literature. Table 1 compared between the mentioned generators in terms of period and linear complexity for immunity attacks.

Table 1 Key-stream generators based period and linear complexity

No	Key-stream generator	Periods	Linear complexity
1	Linear	$(2^{k_1}-1)(2^{K_2}-1)$	$K_{1+}k_2$
2	Product	$(2^{k_1}-1)\cdot(2^{K_2}-1)\ldots\ldots(2^{k_n}-1)$	$K_1k_2\ldots k_n$
3	Gaffe	$(2^{k_1}-1)(2^{K_2}-1)(2^{k_3}-1)$	$L = k_1k_2 + k_2k_3 + k_3$.
4	Multiplexer	$(2^{k_1}-1)\cdot(2^{k_2}-1)$	$K_2\sum_{i=0}^{h}\binom{k_1}{i}$
5	Knapsack	$K(2^n-1)$	$K(2^n-1)$
6	Stop and go	$(2^k-1)^2$	$K(k2^k-1)$
7	Multiplier	$(2^{k_1}-1)\cdot(2^{k_2}-1)$	$(2^{k_1}-1)k_2$
8	Alternating step generator	$2^{L_1}\cdot(2^{L_2}-1)\cdot(2^{L_3}-1)$.	$(L_2+L_3)\cdot 2^{L_1-1} < L(x) \le (L_2+L_3)\,2^{L_1}$.
9	A5	$(4/3)(2^{23}-1)$	$L_1+L_2+L_3$
10	Shrinking	$(2^{L_2}-1)\cdot 2^{L_1-1}$	$L_2\cdot 2^{L_1-2} < L(x) \le L_2\,2^{L_1-1}$.
11	Threshold(3 LFSR)	$(2^{k_1}-1)(2^{K_2}-1)(2^{k_3}-1)$	$K_1k_2 + k_1k_3 + k_2k_3$
12	Integer addition sequence	$(2^{k_1}-1)\cdot(2^{k_2}-1)$	$(2^{k_1}-1)\cdot(2^{k_2}-1)$

5 Conclusion

After all, we found that applying a big size keystream is the most effective way for preventing higher community attacks by considering the use of techniques that determine the feedback shift register to get imply to others bits. This in order result a new challenge for predicating the other bits in different period. In addition, we found that small non-linearity of the output function of multiplexer is the main reason for resulting the correlation between data LFSR and keystream generator. In such case, the attackers has the change to guesses a certain part of the address input in the multiplexer, which in turn my led them to guess the output data as well. We have found that A5/1 is the best generator in which the LFSRs is not applicable that most attacks get benefits from the 64-bit size. Future studies can be carried on the design of immunity multiplexer for enhanced keystream in stream ciphers.

References

1. Cusick, T.W., Ding, C., Renvall, A.R.: Stream Ciphers and Number Theory (vol. 66). Elsevier, Amsterdam (2004)
2. Ghani, M.K.A., Jaber, M.M., Suryana, N.: Barriers faces telemedicine implementation in the developing countries : toward building Iraqi telemedicine framework. ARPN J. Eng. Appl. Sci. **10**(4), 1562–1567 (2015)
3. Ghani, M.K.A., Jaber, M.M., Suryana, N.: Telemedicine supported by data warehouse architecture. ARPN J. Eng. Appl. Sci. **10**(2), 415–417 (2015)
4. Wu, H., Preneel, B.: Cryptanalysis of the Stream Cipher ABC v2, pp. 56–66 (2007)
5. Alhamdan, A., Wong, K.K.: Slid Pairs in the Initialisation of the A5/1 Stream Cipher (2013)
6. Fujiwara, H., Fujiwara, K., Tamamoto, H.: Secure scan design using shift register equivalents against differential behavior attack. In: Proceedings of Asia South Pacific Design Automation Conference ASP-DAC, pp. 818–823 (2011)
7. Liu, Y., Wu, K., Karri, R.: Scan-based attacks on linear feedback shift register based stream ciphers. ACM Trans. Des. Autom. Electron. Syst. **16**(2), 1–15 (2011)
8. Tsaban, B., Vishne, U.: Efficient linear feedback shift registers with maximal period, vol. 267, pp. 256–267 (2003)
9. Deepthi, .P.P, Sathidevi, P.S.: Hardware stream cipher based on LFSR and modular division circuit. Int. J. **2**(10), 632–640 (2008)
10. Courtois, N., Meier, W.: Algebraic Attacks on Stream Ciphers with Linear Feedback, pp. 9–11 (2003)
11. Golić, J., Bagini, V., Morgari, G.: Linear cryptanalysis of Bluetooth stream cipher. In: Advances in Cryptology—EUROCRYPT …, pp. 238–255 (2002)
12. Handschuh, H., Preneel, B.: Key-recovery attacks on universal hash function based MAC algorithms. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 5157 LNCS, pp. 144–161 (2008)
13. Hermelin, M., Nyberg, K.: Correlation Properties of the {Bluetooth} Combiner Generator, pp. 17–29 (1999)
14. Golić, J.D., Morgari, G.: Optimal correlation attack on the multiplexer generator. Inf. Process. Lett. **109**(15), 838–841 (2009)

15. Dawson, E., Clark, A., Goli, J., Simpson, L., Millan, W.: The LILI-128 keystream generator. In: 7th Annual Workshop on Selected Areas in Cryptology, pp. 248–261 (2000)
16. Ali, M.H., Othman, M.A.: Towards a exceptional distributed database model for multi DBMS. Lect. Notes Electr. Eng. **315**, 553–560 (2015)
17. Sudan, M., Trevisan, L., Vadhan, S.: Pseudorandom generators without the XOR lemma. In: Proceedingsof the Fourteenth Annual IEEE Conference Computing Complex. (Formerly Structure in Complexity Theory Conference (1999)
18. Coppersmith, D., Halevi, S., Jutla, C.: Cryptanalysis of stream ciphers with linear masking. In: Advances in Cryptology—CRYPTO …, pp. 1–23 (2002)
19. Sunar, B., Martin, W.J., Stinson, D.R.: A provably secure true random number generator with built-in tolerance to active attacks. IEEE Trans. Comput. **56**(1) (2007)
20. Paar, C.: Applied cryptography and data security. Obstet. Gynecol. **123**(2 Pt 1), 369 (2005)
21. Katz, J., Lindell, Y.: Introduction to Modern Cryptography (2007)
22. Maximov, A.: Some words on cryptanalysis of stream ciphers. Lund University (2006)
23. Hariprakash, R., Ananthi, S., Padmanabhan, K.: Secured image transmission using wavelets for spread spectrum communication in a remote surveillance system. In: Proceedings of 3rd International Conference on Advances in Human-Oriented and Personalized Mechanisms, Technologies and Services, CENTRIC 2010, pp. 61–66 (2010)
24. Schneier, B.: Applied Cryptography: Protocols, Algorithms, and Source Code in C. Wiley, London (2007)
25. Schneier, B.: Applied Cryptography: Protocols, Algorithms, and Source Code in C (1996)
26. Mollin, R.A.: An Introduction to Cryptography. CRC Press, Boca Raton (2006)
27. Golić, J.D., Menicocci, R.: Edit probability correlation attacks on stop/go clocked keystream generators. J. Cryptol. **16**(1), 41–68 (2003)
28. Azri, Z., Muhamad, B., Algaet, M.A., Teknikal, U.: Comparative Study of Performance in Cryptography Algorithms (Blowfish and Skipjack) Ali Ahmad Milad, 2 Hjh Zaiton Muda, Department of Computer System and Communication, Department of Computer Science," vol. 8, no. 7, pp. 1191–1197 (2012)
29. Schneier, B.: Schneier on Security. *wwwschneiercom*, p. 336 (2008)
30. Qi, H.: Stream ciphers and linear complexity (2008)
31. Zaidan, A.A., Othman, F., Zaidan, B.B., Raji, R.Z., Hasan, A.K., Naji, A.W. : Securing cover-file without limitation of hidden data size using computation between cryptography and steganography. Engineering **I** (2009)

Validation of Questionaire for Website Usability (QWU) Instrument Based on Experts Review

Nur Sukinah Aziz and Adzhar Kamaludin

Abstract Nowadays, website is used widely all over a world for medium of communication for information or services and usability principles was implemented in web environment and not for software only. Organization uses website to market their product and services. There are many methods to evaluate the website such as heuristic evaluation, testing, survey and many more. This study is aimed to develop the questionnaire for measuring website in context of usability. The instrument is consist of 9 constructs and 60 item of questionnaire. Content validity by experts is used as one of methodology to validate the instrument. The result shows that from 60 item of instrument, 9 item need to be removed based on experts review. Therefore, it shows some justification for the content validity of items when some items are not focusing to the research target.

1 Introduction

There are many factors or characteristic to determine the quality of the website [1, 2]. Usability is one of the most important factors in website or software quality. There are many definitions or terms about usability. In Human Computer Interaction (HCI) term, usability is more to usable user interface or in other word to make the system easy to learn and easy to use [3]. Based on ISO 9241-11 in the HCI field, usability is defined as the "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" [1, 4]. Refer to the definition of ISO

N.S. Aziz (✉)
Faculty of Computer, Media and Technology Management, TATI University College, Kemaman, Terengganu, Malaysia
e-mail: nursukinah@tatiuc.edu.my

N.S. Aziz · A. Kamaludin
Faculty of Computer Systems and Software Engineering, Universiti Malaysia Pahang, Kuantan, Pahang, Malaysia
e-mail: adzhar@ump.edu.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_75

9241-11, the criteria of usability are effectiveness, efficiency and satisfaction. This definition has 3 components that can divide such as "specified users", "achieve specified goals" and "specified context to use". This definition is more clear what usability is mean and many researchers use this definition [3]. Questionnaires in the survey are the most frequently used tools for usability evaluation. It's easier and cheaper than other approaches for gathering user feedback [5]. The researcher can choose either made it online survey or paper-and-pencil instruments. Before distributing the questionnaire to respondents in the actual study, the questionnaire needs to evaluate on validity and reliability. One of the approaches is based on content validity based on experts review. There are a few previous research that addressed the content validity and included the evaluation of the content validity based on experts review when developing the questionnaire instrument [6] especially in Information System research [7]. The important of experts review is to determine the items of the questionnaire instrument are appropriate to the constructs and in the context of the research. The feedbacks from the experts are used to delete ambiguous, redundant, or unrelated items in the questionnaire instrument. It also necessary since the questionnaire instrument is based on literature review, therefore the questionnaire instrument need to evaluate by experts and to have confidence in any inferences using the questionnaire instrument.

2 Website Usability Evaluation Approach

There are many approaches for usability evaluation such as testing, inspection, inquiry, analytical modeling and simulation [8]. All the evaluation methods have one common characteristic that is dependent on user judgement. Usability lab testing is focused on the experience and comment from users that used the web sites or in scenario-based environment [9]. Usability lab testing only involves a small group of user [10]. It records the user behavior and cognitive processes to see what users actually think and how users use the website. There are a few approaches in usability testing such as: Thinking aloud protocol, Co-discovery learning, Performance measurement, Coaching method, Remote testing, Eye-tracking and many more. Usability inspection is generic name for a set of methods based on having evaluators inspect or examine usability-related aspects of a user interface [11]. Requires usability specialists or software developers, users and other professional to examine or judge either the prototype or each element of interface follows established usability principles. Cognitive Walkthrough, Heuristic Evaluation, Feature Inspection, Pluralistic Walkthrough and Guideline Checklist are a few example of usability inspection. Usability inquiry involves experts to get information about the user requirement for the system by communicating with them or observed them while users are using the system. A few approach in usability inquiry such as Field observation, Interviews, Focus groups, Proactive field study, Logging actual use and Surveys. Table 1 shows the advantage and disadvantage between usability evaluations.

Table 1 Advantage and disadvantage between usability evaluations

Categories	Method	Advantages	Disadvantage
Inquiry	Individual interview	i. Enable to learn about things that cannot be directly observed ii. Allow for probing	i. Time consuming ii. Expensive
Inquiry	Focus group	i. Can provide speedy results ii. Structured data can be collected iii. Planning can enable in-depth discussion	i. Recruitment can be expensive, time consuming ii. Controlled settings may affect behaviours iii. Data can be difficult to analyse
Analytical modelling	Task analysis	i. Help attain understanding of processes and resources to complete each task ii. Helps make recommendation regarding changes to the system	i. Require time and resources ii. Skills required for efficient analysis of the task
Analytical modelling	Card sorting	i. Easy to conduct ii. Identifies items that are likely to be difficult to categorize iii. Understanding of how real users categorize	i. Takes time to complete ii. Can be expensive to hire participants
Inspection	Prototype	i. Issues in design can be identified ii. Complete functionality can the tested	i. Time consuming to create ii. More expensive to develop
Inquiry	Survey	i. Quick and cost effective ii. Gather a lot of data iii. Can be administered on large population	i. Time consuming to collect data
Inspection	Heuristic evaluation	i. Easy to perform; cheap ii. No planning required iii. Able to find many problems (both major and minor problems) iv. Not time-consuming (no users involved)	i. Focus on problems
Testing	Thinking aloud	i. Able to find why the problems occur ii. Small number of test users iii. Low time in relation to other evaluations methods iv. Direct interaction of the users with the transactional web application	i. Time consuming ii. Can be expensive to hire participants
Testing	Formal evaluation	i. Objective method ii. Provide substantive depth in quantitative data	i. Time consuming
Testing	Query technique	i. Provide qualitative and quantitative data ii. Simple and cheap	i. Time consuming

There are many types of standard post-test questionnaires based on the specific field and criteria such as SUS (System Usability Scale), QUIS (Questionnaire for User Interface Satisfaction) [12], CSUQ (Computer System Usability Questionnaire), Words (adapted from Microsoft's Product Reaction Cards) [13] and other more. All this questionnaires are widely used and readily available. Basically questionnaires are evaluations that access the perception from the user's point of view of the websites. There are many approaches to evaluate the questionnaire instrument before distribute in actual study such as content validity by experts, pre-test and pilot test. Content validity means the degree that the instrument covers the content that it is supposed to measure [14, 15]. Content validity is an important method in developing the questionnaire instrument. Content validity by experts are important to see the view of experts on the important of construct and item in the questionnaire [16]. It also provides the evidence about the items extent to which the constructs in the questionnaire instrument are related to and representative. The approach can ensure the construct validity and give confidence to the researcher about the instrument before distributing it for pilot study or actual study. The feedback and comments from the experts give the researcher to see the instruments in the width direction and focusing to the specific element to evaluate the website usability.

3 Methodology

This study implementing a post-test questionnaire. The survey instrument used in this study is a 60 item questionnaires. Content validity by experts is conducted to see the validity based on experts view before conduct the pilot study or actual study. About 20 experts were invited thru email to participate in this study. The experts were choose based on their experience and participated in website or usability. Only 5 experts give a feedback. Five experts for content validity are sufficient because the recommendation is to select at least 3 experts for evaluation [15–17]. The instrument has undergone review process by 5 experts in website and usability studies. The experts are selected based on their experience in website and usability field. This process required about a month to gather all the experts present the result of content validity.

3.1 Initial Instrument

Questionnaire from the Computer System Usability Questionnaire (CSUQ) and WAMMI were adopted and also include a few questions from previous research. The questionnaire instrument is known as Questionnaire for Website Usability (QWU). Table 2 shows the list of 60 item in the initial questionnaire. The instruments divide into 9 parts from 60 item that reflect to 9 constructs that are used in

Table 2 The initial list of questionnaire instrument

Construct	Code	Item	Reference
Efficiency	EY1	When I use the web site there is very little waiting time between my actions and the web site's response	[18]
	EY2	It is easy to find the information that I need	CSUQ
	EY3	I am able to efficiently complete my work using this website	CSUQ
	EY4	I can effectively complete my work using this website	CSUQ
	EY5	I believe I became productive quickly using this website	CSUQ
Effectiveness	EV1	On this website, it is simple to accomplish the task I want to accomplish	[8]
	EV2	I find it easy to get this web site to do what I want it to do	[19]
	EV3	I am able to complete my work quickly using this website	CSUQ
	EV4	It was simple to use this website	CSUQ
	EV5	The information is effective in helping me complete the tasks and scenarios	CSUQ
Satisfaction	S1	I feel comfortable using this website	CSUQ
	S2	This website has all the functions and capabilities I expect it to have	CSUQ
	S3	I am satisfied with how easy it is to use this web site	CSUQ
	S4	I am satisfied with this web site	CSUQ
Learnability	L1	Learning to operate the Web site is easy for me	[18]
	L2	I find the Web site easy to use	[8]
	L3	All the material is written in a way that is easy to understand	WAMMI
	L4	Using this website for the first time is easy	WAMMI
	L5	The contents provided by the website are easily understood	[20]
	L6	The website is designed for easy understanding	[20]
	L7	The information provided by the website is easy to understand	CSUQ
	L8	It was easy to learn to use this website	CSUQ
Accessibility	AC1	The website offers customization	[8]
	AC2	It was easy to move from one page to another	[21]
	AC3	The text on the website is easy to read	[18]
	AC4	It takes time to open the web page or download the web page	

(continued)

Table 2 (continued)

Construct	Code	Item	Reference
	AC5	It has a accessibility function on the web site (can resize text, change the background colour etc.)	
	AC6	The website's wording is clear and easy to understand	[20]
	AC7	The website uses colors and structures that are easy on the eyes	[20]
	AC8	The pages download quickly on this website	[8]
Navigation	N1	I can easily navigate this site	[22]
	N2	This site provides good navigation facilities to information content	[22]
	N3	I like the way hyperlinks are embedded in this site's design	[23]
	N4	I feel in control when I'm using this web site	WAMMI
	N5	I get what I expect when I click on things on this website	WAMMI
	N6	The navigation and labels on this Web site were clear	[24]
	N7	Links are consistent and easy to identify	[25]
	N8	The website provides multiple search features (e.g.: search engine, menu bar, go-back-and-forward button, etc.) to obtain the target information	[20]
	N9	It was easy to move from one page to another	[21]
Content	C1	I trust the Web site to keep my personal information safe	[18]
	C2	I can trust this website	[22]
	C3	I trust the information presented on this website	[22]
	C4	The information provided at this site is sufficient	[22]
	C5	The website adequately meets my information needs	[22]
	C6	I find the information on this site to be well organized	[22]
	C7	I feel this Web site clearly stated its purpose for using the site	[24]
	C8	The website provides up-to-date information	[20]
	C9	The information (such as online help, online messages, and other documentation) provided with this website is clear	CSUQ

(continued)

Table 2 (continued)

Construct	Code	Item	Reference
	C10	The organization of information on the website pages is clear	CSUQ
Interface/design	ID1	The website repeats the same structure, components and overall look across pages	[20]
	ID2	Web pages in the website are consistently designed	[20]
	ID3	This web site is presented in an attractive way (i.e. colors, images, layout etc.)	WAMMI; [22]
	ID4	The pages on this website are very attractive	WAMMI; [22]
	ID5	The layout of pages made tasks easier	[24]
	ID6	The interface of this web site is pleasant	CSUQ
	ID7	I like using the interface of this web site	CSUQ
Intention to reuse	ITU1	I intend to use this website again	[21]
	ITU2	I would be willing to visit this website again	[24]
	ITU3	I feel this website reflects most current trend(s) and provides nice design for the site visit	[24]
	ITU4	I will reuse this website again	[21]

this study which is *Effectiveness, Efficiency, Satisfaction, Learnability, Accessibility, Navigation, Content and Interface/design* for Independent Variable (IV) factors and *Intention to Use* as a Dependent Variable (DV).

The first part of the questionnaire contains a demographic profile of expert, including gender, age, current position, education level, experience in teaching or in industry and involment in website or usability. A five-point Likert-type scale ranging from (1) "Extremely unimportant"; (2) "Unimportant"; (3) "Less Important"; (4) "Important" and (5) "Extremely Important" was used to evaluate the 60 item of the questionnaires. This instruments also include suggestion or comment in each of constructs.

4 Findings and Discussion

Based on the results shown in Table 3, 9 items from 60 items in the questionnaire instrument have a low result of the feedback from experts review. The 9 items need to remove from the instruments are EY4, EY5, AC4, N4, C1, C2, C3, C7 and ITU4. The results are based on the mean value that the item need to have 3.5 and above to consider important to the constructs in the questionnaire. Besides that, each item selected and consider important if 4 experts give more than 3 [6]. This called

Table 3 The result from experts review

Code	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Mean	Mode	Median	SD
EY1	5	5	4	5	4	4.6	5	5	0.548
EY2	5	5	5	2	4	4.2	5	5	1.304
EY3	5	4	2	5	4	4	5	4	1.225
EY4	4	5	2	1	**4**	3.2	4	4	1.643
EY5	4	4	2	1	4	3.0	4	4	1.414
EV1	4	4	3	4	4	3.8	4	4	0.447
EV2	5	5	3	1	4	3.6	5	4	1.673
EV3	4	4	3	1	5	3.4	4	4	1.517
EV4	5	5	4	1	4	3.8	5	4	1.643
EV5	5	5	3	5	4	4.4	5	5	0.894
S1	3	5	3	5	4	4.0	3	4	1.000
S2	4	4	4	5	4	4.2	4	4	0.447
S3	4	5	5	5	4	4.6	5	5	0.548
S4	4	4	5	5	4	4.4	4	4	0.548
L1	5	4	4	4	4	4.2	4	4	0.447
L2	4	5	5	1	4	3.8	4	4	1.643
L3	4	5	4	4	4	4.2	4	4	0.447
L4	5	5	4	1	4	3.8	5	4	1.643
L5	5	5	4	5	4	4.6	5	5	0.548
L6	5	5	3	5	5	4.6	5	5	0.894
L7	5	5	3	3	4	4.0	5	4	1.000
L8	5	5	5	1	4	4.0	5	5	1.732
AC1	3	4	5	1	5	3.6	5	4	1.673
AC2	5	5	4	1	4	3.8	5	4	1.643
AC3	5	5	4	4	4	4.4	4	4	0.548
AC4	1	5	4	4	2	3.2	4	4	1.643
AC5	3	4	4	4	5	4.0	4	4	0.707
AC6	5	4	2	3	4	3.6	4	4	1.140
AC7	5	5	4	1	4	3.8	5	4	1.643
AC8	5	5	4	4	4	4.4	4	4	0.548
N1	5	4	5	4	4	4.4	4	4	0.548
N2	5	5	3	4	4	4.2	5	4	0.837
N3	4	4	2	4	4	3.6	4	4	0.894
N4	2	5	2	3	4	3.2	2	3	1.304
N5	5	5	4	4	4	4.4	4	4	0.548
N6	5	5	4	4	5	4.6	5	5	0.548
N7	5	5	3	4	5	4.4	5	5	0.894
N8	5	4	2	4	5	4.0	5	4	1.225
N9	5	5	4	4	4	4.4	4	4	0.548

(continued)

Table 3 (continued)

Code	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Mean	Mode	Median	SD
C1	5	2	2	1	2	2.4	2	2	1.517
C2	5	2	5	1	2	3.0	5	2	1.871
C3	5	3	4	1	2	3.0	N/A	3	1.581
C4	5	4	3	4	4	4.0	4	4	0.707
C5	5	4	4	5	4	4.4	4	4	0.548
C6	5	5	3	5	5	4.6	5	5	0.894
C7	5	4	2	1	4	3.2	4	4	1.643
C8	5	5	5	4	4	4.6	5	5	0.548
C9	4	5	5	4	4	4.4	4	4	0.548
C10	4	5	2	4	4	3.8	4	4	1.095
ID1	4	4	5	4	4	4.2	4	4	0.447
ID2	5	5	5	5	4	4.8	5	5	0.447
ID3	5	5	4	5	4	4.6	5	5	0.548
ID4	5	5	3	4	4	4.2	5	4	0.837
ID5	5	5	4	5	4	4.6	5	5	0.548
ID6	5	5	4	4	4	4.4	4	4	0.548
ID7	5	4	3	4	4	4.0	4	4	0.707
ITU1	5	4	5	5	4	4.6	5	5	0.548
ITU2	5	4	5	5	4	4.6	5	5	0.548
ITU3	5	4	3	4	4	4.0	4	4	0.707
ITU4	5	4	2	4	2	3.4	4	4	1.342

universal agreement [15]. So most of the experts need to give mark 3 or above to consider the item is important [6]. The construct that involve in elimination of item are efficiency, accessibility, navigation, content and intention to use. All 9 items do not achieve the two scenarios.

For efficiency, there are 2 items did not achieve the 2 scenarios that need to be considered as important for website usability. EY4—*I can effectively complete my work using this website* that has mean value 3.2 and EY5—*I believe I became productive quickly using this website* has mean value 3. AC4 is in accessibility construct. AC4—*It takes time to open the web page or download the web page* has mean value 3.2. The statement for this item is in negative statement. Some of the experts did not agree with the negative statement because the entire items in this instrument are in a positive statement. The item also has same meaning with AC8. AC8—*The pages download quickly on this website.* Item N4 in navigation has mean value 3.2. N4—*I feel in control when I'm using this web site.* This item is about control. The experts did not agree that the item is important in this instrument. The C1, C2 and C3 in the content construct are more on trust. C1—*I trust the Web site to keep my personal information safe*, C2—*I can trust this website* and C3—*I trust the information presented on this website*. The mean value for C1, C2 and C3 are 2.4, 3.0 and 3. Most of the experts do not agree that trust is an important element when evaluate the website university. It is because website for university is trusted website. So did not need to evaluate on trust element for the content in the website. For C7—*I feel this Web site clearly stated its purpose for using the site* has mean value 3.2. The experts did not agree that the item is important in the instrument. The higher education institution or university website are clearly stated the purpose and function the website. ITU4 is in Intention to Use construct has mean value 3.4. ITU4—*I will reuse this website again.* Some of the experts did not agree the wording of reuse. Only 3 experts that gave scale 3.0 and above. After analyzing the data using SPSS software in this phase, 51 items are remaining in the questionnaire instrument.

Based on the results content validity from the experts review, there are a few issues that are identified. Most of the experts are disagree and give lowest scores the items in the context of trust in the questionnaire instrument to measure the university website. Some of the reasons are the university website is trusted website and all the information in the website is valid. The experts also give lowest scores on the item in the context of control and purpose. They think that the items are irrelevant in the questionnaire instrument. Some of the experts did not agree with the negative statement because the entire items in this instrument are in a positive statement and only one item is negative statement. Based on the research [26], the researcher should avoid the negative statement. Other items that need to be removed are more about the wording of the sentences.

5 Conclusion and Future Recommendation

Content validity is an important method in developing the questionnaire instrument. The approach can ensure the construct validity and give confidence to the researcher about the instrument before distributing it for pilot study or actual study. There are a few previous research that addressed the content validity and included the evaluation of the content validity based on experts review when developing the questionnaire instrument [6] especially in Information System research [7]. The feedback and comments from the experts give the researcher to see the instruments in the width direction and focusing to the specific element to evaluate the website usability. Besides that, the finding could contribute to support the construct validity of the instrument. Based on the result from the content validity, there are justifications on the related item to the constructs in the context of research because most of the constructs and items are based on previous literature. Content validity measures the comprehensiveness and representation of the content on a scale and to obtain some justification for the content validity of items when some items are not the focus point of the research.

To ensure the instrument is trusted and valid, the next phase is the questionnaire instrument will evaluate by focus group. In this pre-test, researchers will involve direct with the focus group to gather the feedback from them. After doing pre-test with the focus group, then pilot study will conduct before the final questionnaire instrument distribute to the actual study. The instruments will have a few stages of evaluation to ensure the item and construct covers the content and suppose to measure. The analysis based on the experts review contributes to the body of knowledge in term of evaluation of the website usability focusing on higher education institution or university website.

Acknowledgments The authors would like to extend sincere appreciation to *Faculty of Computer, Media and Technology Management (FKMPT)*, TATI University College and *Faculty of Computer Systems & Software Engineering (FSKKP)*, Universiti Malaysia Pahang for the support and encouragement in making this study a success.

References

1. Fernandez, A., Insfran, E., Abrahão, S.: Usability evaluation methods for the web: a systematic mapping study. Inf. Softw. Technol. **53**(8), 789–817 (2011)
2. Dubey, S.K., Rana, A., Mridu A.K.: Analytical Comparison of Usability Measurement Methods, vol. 39, pp. 11–18 (2012)
3. Leventhal, L., Barnes, J.: Usability Engineering Process, Products, and Examples. Pearson Prentice Hall (2008)
4. De Marsico, M., Levialdi, S.: Evaluating web sites: exploiting user's expectations. Int. J. Hum. Comput. Stud. **60**(3), 381–416 (2004)
5. Elling, S., Lentz, L., De Jong, M.: Website Evaluation Questionnaire : Development of a Research-Based Tool Evaluation, pp. 293–304 (2007)
6. Hardesty, D.M., Bearden, W.O.: The use of expert judges in scale development. Implications for improving face validity of measures of unobservable constructs. J. Bus. Res. **57**(2), 98–107 (2004)
7. Straub, D., Boudreau, M.-C., Gefen, D.: Validation guidelines for IS positivist research. Commun. Assoc. Inf. Syst. **13**(1) (2004)
8. Wang, J., Senecal, S.: Measuring perceived website usability. J. Internet Commer. **6**(4), 97–112 (2008)
9. Batra, S., Bishu, R.R.: Web usability and evaluation : issues and concerns. Hum. Factors, 243–249 (2007)
10. Ai-wabil, A., Ai-khalifa, R.: A framework for integrating usability evaluations methods : the Mawhiba web portal case study methodology (2009)
11. Nielsen, J., Mack, R.L.: Usability Inspection Methods. Wiley (1994)
12. Madan, A., Dubey, S.K.: Usability evaluation methods: a literature review. Int. J. Eng. Sci. Technol. **4**(2), 590–599 (2012)
13. Tullis, T.S., Stetson, J.N.: A comparison of questionnaires for assessing website usability, pp. 1–12 (2004)
14. Yaghmale, F.: Content validity and its estimation. J. Med. Educ. **3** (2003)
15. Haynes, S.N., Richard, D.C.S., Kubany, E.S.: Content validity in psychological assessment: a functional approach to concepts and methods introduction to content validity. Psychol. Assess. **7**(3), 238–247 (1995)
16. Delgado-rico, E., Carretero-dios, H., Ruch, W.: Content validity evidences in test development. Int. J. Clin. Heal. Psychol. **12**, 449–459 (2012)
17. Crocker, L., Maria Llabre, A., Miller, M.D.: The generalizability of content validity ratings. J. Educ. Meas. **25**(4), 287–299 (1988)

18. Loiacono, E., Watson, R., Goodhue, D.: WebQual: an instrument for consumer evaluation of web sites. Int. J. Electron. Commer. **11**(3), 51–87 (2007)
19. Palmer, J.W.: Web site usability, design, and performance metrics (2002)
20. Lee, Y., Kozar, K.A.: Understanding of website usability: specifying and measuring constructs and their relationships. Decis. Support Syst. **52**(2), 450–463 (2012)
21. Poelmans, S., Wessa, P., Milis, K., Bloemen, E., Doom, C.: Usability and acceptance of e-learning in statistics education, based on the compendium platform. Inf. Syst. J., 1–10 (2008)
22. Cyr, D.: Website design, trust and culture: an eight country investigation. Electron. Commer. Res. Appl. **12**(6), 373–385 (2013)
23. Zhang, X., Keeling, K., Pavur, R.: Information quality of commericial web site home pages: an explorative analysis. In: ICIS'00: Proceedings of the 21st International Conference on Information Systems, pp. 164–175 (2000)
24. Downing, C.E., Liu, C.: Assessing web site usability in retail electronic commerce. In: 2011 IEEE 35th Annual Computer Software Applications Conference, pp. 144–151 (2011)
25. Meyers, P.J.: 25 point website usability checklist (2009) [Online]. Available: http://drpete.co/blog/25-point-website-usability-checklist. Accessed 20 Apr 2015
26. Sauro, J., Lewis, J.R.: When designing usability questionnaires, does it hurt to be positive?. In: Proceedings of 2011 Annual Conference Human Factors Computing Systems—CHI'11 (2011)

Performance Evaluation of Preloading Resources for Web Pages

Lukáš Čegan

Abstract In the first part, this paper describes the modern web application and the key parameters affecting the speed of web page loading. The following part focuses on the description of the proposed solution for preloading of web page resources. The paper presents the static and dynamic approach. The impact of the static preloading model on page load time is evaluated by a benchmark test. In the last section of the paper the results of measurement are discussed, it gives a conclusion and future research is outlined.

1 Introduction

The web has dramatically evolved over the last 25 years. Demands of the users on the quality of the website have increased significantly. Users expect rich content that is delivered to them in a very short time. Unfortunately, sometimes the network environment does not allow delivery of web content within a reasonable time. The reasons are due to the physical characteristics of networks such as a high latency or low bandwidth. Despite these limitations, however, we can ensure fast page loading using preloading resources of a website.

After introducing the context, the web page performance optimization factors are presented in more detail in Sect. 2. Next, in Sect. 3, a proposed solution for decreasing the page load time by intelligent preloading web page resources is described. In Sect. 4 practical experiments, benchmark settings and test results are described. Finally, the last section gives conclusions and future research opportunities.

L. Čegan (✉)
Department of Information Technology, Faculty of Electrical Engineering and Informatics, University of Pardubice, Pardubice, Czech Republic
e-mail: lukas.cegan@upce.cz

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_76

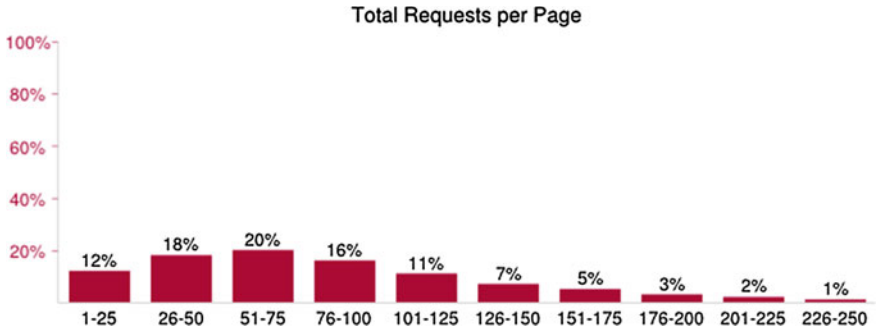


Fig. 1 Total requests per page [1]

2 Performance Optimization

Modern web application consists of many resources. The current average web page size is 1958 kB and it is composed of 94 resources. The chart on Fig. 1 shows the average number of requests for a single web page [1]. The average number of TCP connections that were opened during page loading is about 38, and only 14 % of all requests are done over https.

2.1 *HTTP Waterfall Sequence*

The waterfall chart provided detailed information about web page resources and shows them in time ordering, with the duration of each load. This information can be used for identifying performance bottlenecks (see Fig. 2). The waterfall sequence shows us the inner workings of how a browser loads a web page and its resources.

HTTP practically allows only one outstanding request per TCP connection, which is considerably limiting in terms of page load time. The HTTP standard recommended a limit of 2 parallel connections per host, but a modern browser does not comply with this recommendation. The current average number of parallel TCP connections, which web browsers create, is from 6 to 8 per one domain [2]. Another technique for faster loading page is domain sharing. We can spread web page resources over multiple domains, which allow an increased number of parallel TCP connection.

2.2 *Critical Rendering Path*

One of the most important concepts to make a page load faster is recognition of the critical rendering path. A critical rendering path is defined as codes and resources,

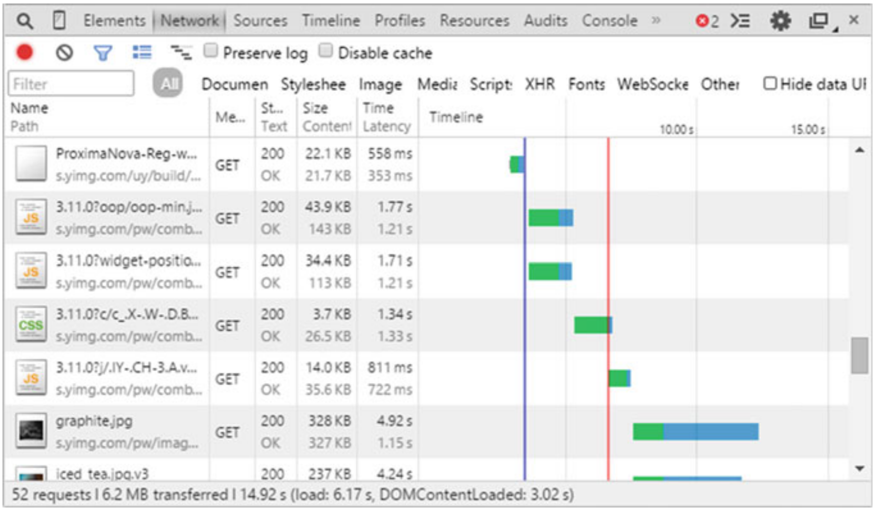


Fig. 2 HTTP waterfall chart in Chrome DevTools

required to render the initial view of a web page. The rendering path consists of events that occur to make web page appear on a browser. First, it's necessary to load the source code of page. Then, the web browser creates Document Object Model (DOM). This process consists of several steps. The raw data translates to a HTML character, then the browser converts strings of characters into distinct tokens. Tokens are then converted into objects and finally, the created objects are linked in a tree data structure (see Fig. 3).

When DOM is ready to use, the page can be rendered to the screen. But there are certain type of resources that actually block the render of a web page. One of these types is an external CSS stylesheet, because CSS carries information about how the element should look when it is rendered. The next type of resource which blocks rendering is JavaScript. JavaScript can modify the content and style of the web page and for this reason DOM construction is blocking when a script tag is encountered and until the script has finished executing [3]. Hence, if a browser discovers necessary resources to render the page, it immediately dispatches requests for these resources. All major web browsers use, "preload scanner/speculative parser" for this discovering tool which is designed to look ahead in the requested document and discover critical resources [4, 5]. These types of resources can be loaded with

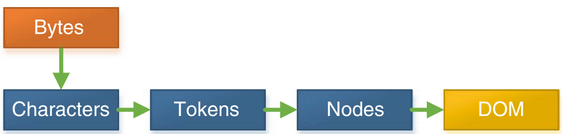


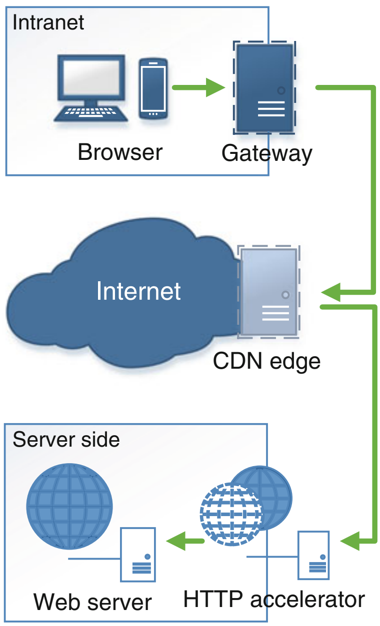
Fig. 3 DOM construction

higher priority which has a positive impact on total rendering page time. Measuring of performance characteristics of each step of the critical rendering path can be done using "Navigation Timing" [6] which provides a browser's API for obtaining performance data of every request in the browser using JavaScript. Based on this data it is possible to optimize a critical rendering path.

2.3 Cache

Loading web page resources over the network is, in most cases, slow. Therefore we have to optimize content efficiency. Web pages contain some resources that don't change very often, like images, CSS files, etc. Storing previous seen data in cache, speeds up page load time. On the roundtrip between the client and server (see Fig. 4) are used in the WWW scope various kinds of caches. A browser cache eliminates network time and it is shared only by one user. The gateway provides resources from the faster company intranet and it is shared by all company users. Next, CDN provides edge caching which reduces roadtrip time, because it is closer to the user. This type of cache is shared by all users. The last type of cache is HTTP accelerator which reduces server load, provides faster response turnaround and it is shared by all users. It is important to remember that all types of caches have some limitations. The default recommended cache size of a desktop web browser is about 250 MB, which is equivalent to a 125 full pages. Mobile web browsers have a

Fig. 4 Cache types



smaller cache. It is about 85 MB, which equates to approximately 50 pages (average size of mobile site is 1.7 MB). A browser caches have another big limitation, as we still haven't script interfaces for smart caching. One of the ways to solve this drawback is to build a scriptable browser cache by using JavaScript and HTML5 localStorage. LocalStorage is supported by all major web browsers (except Opera mini [7]), but the size limit in the browser is generally 5 MB.

A cache is controlled by HTTP headers. In response, we can use Expires, Pragma (HTTP/1.0) [8], Cache-control, ETag and Last-Modified (HTTP/1.1) [9]. The problem occurs when the data in a cache are outdated. We haven't a server-side tool to refresh thesis data. To avoid this situation, we can use a versioning of selected resources. It is commonly CSS and JavaScript files.

3 Proposed Solution

Ordinary users of the website are impatient. For this reason, we must show them the website as soon as possible. Today's web page consists of many resources. In an effort to send these sources to the client, we encounter the limitations of HTTP protocol, the size of bandwidth and latency. The proposed solution how to cope with these limitations is preloading the resources that will be necessary to render the following page. The proposed solution can be divided into two different approaches: static model and dynamic model.

3.1 *Static Model*

Each website consists of many pages that refer to themselves variously. This structure can be expressed as a navigation graph in which we can define relationships predecessors and successors (see Fig. 5). Each node (page) in this graph binds sources such as images, stylesheets and scripts (see Fig. 5).

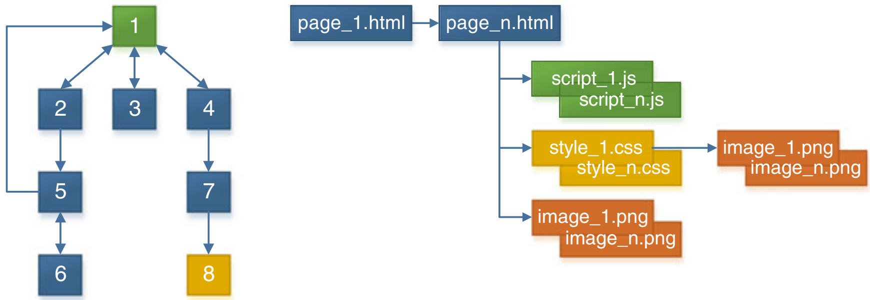


Fig. 5 Navigation graph and page resources

A graph of a website and their resources can be created manually or by automatic builder. The manual approach allows us to have full control over graph creation, but with a larger range of pages, this approach becomes unfeasible. Also, the manual approach is disadvantageous for frequently updated websites.

To eliminate sending already cached data it is necessary to keep a history of visited pages and loaded resources. This history is stored in cookies, and its value is updated if the node of graph is passing.

3.2 Dynamic Model

Preloading all possible sources can be extremely inefficient because the number of followers in the graph can be considerable. For this reason, we must select only those resources that the user will most likely use. To be able to dynamically determine which resources are most appropriate for the user, we need to know its behavior. To understand user behavior we must collect all performance data (transfer size, load time, etc.), web browser date (type of user agent, screen size, etc.) and data about visited ancestors. Based on the statistical data processing, we can create a predictive model, which shall recommend appropriate resources for preloading.

A website navigation model can be modeled as a Markov chain. The web pages represent the state of the Markov chain, and the relations (hyperlinks) between web pages represent transitions. The Markov model predicts the next user's step by only looking at the last user's action. Unfortunately, lower-order Markov models are not very accurate in predicting future user's steps, because the Markov chain is a stateless mathematical model. To achieve greater accuracy of prediction higher order Markov models can be used. However, these higher order Markov models are extremely complicated due to their large number of states, reduced coverage, and sometimes even poor prediction accuracy [10–12].

Our proposed solution reduced the complexity by clustering of selected nodes into the subgraphs. Clustering is performed based on user segmentation. Distribution of users into segments is performed according to how they work with the website content. For this purpose, we measure and collect data expressing:

- type of platform/device,
- screen resolution,
- type of network connection,
- demographics,
- day and time,
- navigation flow,
- page interaction time,
- applied resource,
- interest level/user target.

Based on statistical analysis of data we create website navigation patterns.

The newcomer website users are properly classified into groups according to static information such as browser type, screen size, language versions, network latency, bandwidth, etc. Each group has their own set of navigation patterns. Assigning users to a specific navigation pattern depends on their own session history. Some patterns have identical subgraphs (e.g. the first sequence of navigation) and therefore the user can be on a certain vertical level included in several navigation patterns.

The prediction algorithm is dynamic in time. It must respond to changing conditions such as the growth of websites (e.g. new blog posts), change of users' interest (e.g. Olympic games website—first, users are focused on sports facilities and then, during the games, users are focused sports results), etc. Recalculation of navigation patterns run in a predetermined period.

4 Benchmarking

We evaluated the performance of the first generation of a dynamic model solution on dummy test website. The website consisted of 50 pages, 300 different resources (CSS, JavaScript and images) and the website graph contained 150 relationships. For the experimental part of this paper, we created a data set of random users's access sequence. The first generation of the dynamic model is based on Markov chains, and to determine the subsequent step of the user, utility function that combines transition probability and the total size of the subsequent pages is used.

4.1 Testbed Platform

The TestBed platform consists of the physical machine Dell Latitude E6410, Intel Core2 i5 M460@2.54 GHz, 4 GB RAM, Windows 8 32 bit and virtualization platform VMware Workstation. The virtual machine host server provides computing resources, such as processing power, memory, disk and network I/O, and so on. The guest is a completely separate and independent instance of the operating system. The virtual machine host represents a desktop client with web browser Chrome 41 and it is connected with Android 4.0.4 smartphone via USB cable for remote debugging. The guest represent the server side with operation system Debian 7.8 and the web server Apache 2.4 with "mod_spdy" module.

4.2 Experiment Methodology

For performing benchmark tests, it is necessary to create an environment in which it is possible to simulate network characteristics. In our experiment we used the Linux tool Netem (Network Emulator) which provides functionality for variable delay, loss, duplication and re-ordering. The second tool that we use is TBF (Token Bucket Filter), which is network and processor friendly traffic shaper. This tool is a good solution to slow down transmitted traffic to the specified rate. To configure network traffic in the Linux kernel we use command tc (Traffic Control) and add rules for Netem and TBF. In our experiment, we simulate various network speeds:

- Cable—100 Mbit/s bandwidth and 50 ms latency,
- Fiber—unlimited Mbit/s bandwidth and 25 ms latency,
- 3G—2 Mbit/s bandwidth and 300 ms latency,
- and LTE—15 Mbit/s and 100 ms latency.

The performance metric in our experiment is the page load time. Measurement is performed using "Navigation Timing" which provides a browser's JavaScript API for obtaining performance data of every request. The results of measurements are stored for statistical evaluation in DB on the server side.

We evaluated the performance of proposed solutions on a test platform which consists of Apache server, Nginx server, browser Google Chrome. Processing of unsecured requests is performed by accelerator Varnish. The secure requests are terminated by a proxy server Nginx and forwarded to Varnish (see Fig. 6).

4.3 Results

Testing was performed in two different scenarios. Scenario (A) contained only a dummy test website, scenario (B) contained a proposed solution form static model. For each test scenario, we made five measurements (each measurement contained 10 random sequences) with a Selenium tool and we simulated a variety of bandwidths and delays. The resulting measured values are shown in Tables 1 and 2.

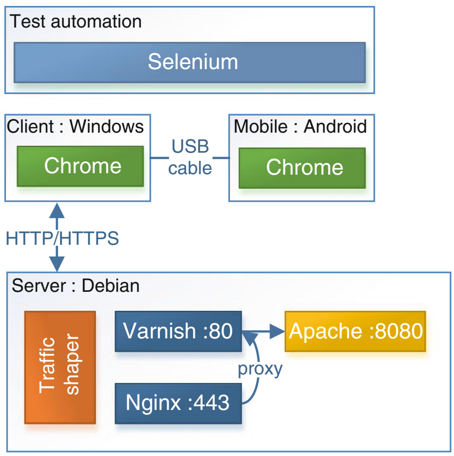


Fig. 6 Communication schema

Table 1 Scenario A—median of measured page load time [ms]

Network	1	2	3	4	5
Cable	381	389	348	315	390
Fiber	288	366	321	177	258
3G	4031	3721	4295	2324	3176
LTE	890	901	911	577	771

Table 2 Scenario B—median of measured page load time [ms]

Network	1	2	3	4	5
Cable	942	1024	813	790	980
Fiber	562	621	601	527	629
3G	7840	7914	8214	7421	7625
LTE	1782	1793	1942	1617	1646

Table 3 Number of unnecessary files and their total size [Mb]

Network	1	2	3	4	5
Files	30	25	26	20	25
Size	2.22	1.83	1.9	1.51	1.85

By applying dynamic model, the server can sending files to the client that the client will never use. For larger sites it can be numerous files that unnecessarily burden the web server. Table 3 shows the total number of files and their sizes that were not used by the client.

5 Conclusion and Future Work

In the beginning, in this paper we described the issue of web page loading and we described the basic optimization techniques that are used in this area. We identified a problem with website performance in the environment with high latency and low bandwidth. Therefore we proposed and described a solution based on preloading of web page resources.

Our solution is divided into two approaches. The first approach is only aimed at evaluating the static structure of the websites. However, this solution has a number of shortcomings and can be used only in a small percentage of cases. The second approach is based on user segmentation and user navigation patterns, which are dynamically made up of statistical processing of historical data on real user sessions. For this purpose a tool was created that can measure and store values of key parameters that characterize the behavior of users on the websites. To assess the impact of different variants of the proposed solutions, we have created a testbed platform and we have established methodology of experiments. In the first project stage we tested the functionality of the whole platform on pre-generated dummy data. Currently we measure user behavior on several production servers. This data will be used for the next phase of the project in which we will examine in more detail algorithms for accurate segmentation of users.

As future work, we are also planning to handle the proposed solution as a module for various web servers, which would, in an automated manner, sufficiently accurately predict the next user step, and would be able to preload all the necessary web page resources.

Acknowledgment This paper is published thanks to the financial support Faculty of Electrical Engineering and Informatics, University of Pardubice specific research project.

References

1. Total requests per page. In: HTTP Archive: Tracks How the Web is Built [online]. San Francisco, CA 94118: Internet Archive, 2001—[cit. 2015-04-07]. Received from http://httparchive.org/interesting.php#reqTotal
2. Browserscope [online]. [cit. 2015-01-12]. Received from http://www.browserscope.org/?category=network&v=top
3. Critical rendering path. In: Grigorik, I. (ed.) Optimizing Performance [online]. [cit. 2015-02-10]. Received from https://developers.google.com/web/fundamentals/performance/critical-rendering-path/?hl=en
4. Optimizing your pages for speculative parsing. In: Mozilla Developer Network [online]. [cit. 2015-04-27]. Received from https://developer.mozilla.org/en-US/docs/Web/HTML/Optimizing_your_pages_for_speculative_parsing
5. Speculative optimization with Chrome's predictor. In: Grigorik, I. (ed.) High Performance Networking in Google Chrome [online]. [cit. 2015-02-12]. Received from https://www.igvita.com/posa/high-performance-networking-in-google-chrome/#predictor
6. Wang, Z., Jain, A.: Navigation Timing. Technical Report https://dvcs.w3.org/hg/webperf/raw-file/tip/specs/NavigationTiming/Overview.html, W3C, Jan 2013
7. Storage. In: HTML5 Features [online]. [cit 2015-03-15] http://www.html5rocks.com/en/features/storage
8. Berners-Lee, T., Fielding, R., Frystyk, H.: Hypertext Transfer Protocol—HTTP/1.0 Technical Report RFC1945, IETF, May 1996
9. Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., Berners-Lee, T.: Hypertext Transfer Protocol—HTTP/1.1. Technical Report RFC2616, IETF, June 1999
10. Dongshan, X., Juni, S.: A new Markov model for web access prediction. IEEE Comput. Sci. Eng. **4**, 34–39 (2002)
11. Deshpande, M., Karypis, G.: Selective Markov models for predicting web page accesses. ACM Trans. Internet Technol. **4**, 163–184 (2004)
12. Borges, J., Levene, M.: Evaluating variable-length Markov chain models for analysis of user web navigation sessions. IEEE Trans. knowl. Data Eng. **19**(4) (2007)

Cost Benefit Analysis Approach for Cloud Computing

Petra Marešová

Abstract The use of cloud computing services appears to offer significant cost advantages. Cost calculations are complicated due to the number of variables comprising inputs to the utility billing model of cloud computing. The aim of this paper is to conduct and describe the theoretical evaluation model of cloud computing with regard to the application in business practice in evaluating the effectiveness of investments. We applied a multi-method approach (systematic literature review, analysis of real cloud computing services, expert interview, case study) for the development and evaluation of the formal model.

1 Introduction

One of the more frequently used terms to describe international business in the twenty-first century is globalisation [1]. In connection with the economic crisis, the emphasis is more than ever laid on the improve the work performance, efficient functioning of processes in enterprises, well-incurred costs and the ROI (Return on Investment) [2]. It has been obvious for quite a long time that it also applies to investments in information technologies (IT). Several surveys indicate that the issue of measuring the benefits of IT investments is a concern in many organisations. Measuring IT benefits and value is frequently reported as one of the most important issues for senior IT management [3–6]. Although this issue has been addressed for many years, it is still relevant in the context of new technologies in enterprises. One of the major trends in recent years is cloud computing. As with previous technologies, it is also associated to the key problem with the evaluation of returns on cloud computing investment [7]. Another problem is the quantification of numerous highly qualitative variables. Cloud computing is a new model of providing sources

P. Marešová (✉)
Faculty of Informatics and Management, University of Hradec Kralove, Rokitanskeho 62, 50003 Hradec Kralove, Czech Republic
e-mail: petra.maresova@uhk.cz

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_77

and an independent user access to the service platform [8]. The definition that has gained wide industry recognition, was created by The National Institute of Standards and Technology (NIST) [9]. According to the NIST Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

The use of the cloud computing approach in organisations is a complex process that affects various organisational units, internal processes, enterprise data security and many others at strategic, tactical and operational level. The complexity of the problem causes low deployment of cloud computing. Company managers do not understand the exact concept of cloud computing and cannot directly evaluate the efficiency of investment in this technology [10]. Although there are a number of metrics that evaluate the efficiency of technologies, they're not used in business practice. This fact may be due to many reasons:

- lack of metrics knowledge for evaluating IT by business executives,
- lack of motivation to deal with these metrics,
- too technically oriented indicators and little connected with normal business practice of evaluating the effectiveness of investments,
- do not provide the comprehensive information needed for managers and their decisions,
- lack of evaluation of the security risks of the migration of existing technologies to cloud,
- solves only a partial area related to IT.

The aim of this paper is to conduct and describe the theoretical evaluation model of cloud computing with regard to the application in business practice in evaluating the effectiveness of investments. Within this model the Cost benefit Analysis method is used. The model aims to eliminate the above mentioned weaknesses.

2 Methods

To fulfil the objective and propose the evaluation model of cloud computing, the literature review and expert interviews were used. The literature review is conducted on methods in the area of measuring the effectiveness of IT, then directly in relation to cloud computing. Interviews with experts were carried out in order to properly design the structure of the evaluation model of cloud computing in relation to the characteristics of this technology and to verify the suitability of the selected indicators and metrics. All of this with regard to the possibilities of use in business practice. The experts included IT specialists and experts in the field of business.

3 Cost Benefit Analysis for Cloud Computing

Cost benefit analysis (CBA) method was selected as a basis for a comprehensive evaluation of cloud computing. Multicriteria evaluation methods of alternatives was used, where experts evaluated suitability of this method. This evaluation is described in study [10].

Cost benefit analysis (CBA) is a systematic approach for estimating the strengths and weaknesses of alternatives that satisfy transactions, activities or functional requirements for a business. It is a technique that is used to determine options that provide the best approach for the adoption and practice in terms of benefits in labour, time and cost savings etc.

Its application in the deployment of advanced technologies, information systems, or new software (e.g. [11, 12]) is relatively frequent. The negatives ("costs") of the project include for example software price, cost of consultants, installation and user training. The positives ("benefits") of the project include improved business process leading to savings in production costs, improvement of the decision-making process and increased employee morale due to better feeling of working with something new. Basic concepts that can occur when using this method are: the effect arising from investment, costs, benefits and utility.

Effects arising from investments—all the effects on the surveyed entities, which are brought by the realisation of the investment projects. They can occur in financial as well as non-financial (or intangible) form. From the perspective of an entity, they may have positive (benefits), negative (costs) or neutral (the entity is not affected) nature.

Costs—all negative impacts to the research subject(s), or their group. These are the negative effects resulting from the investment [13].

Benefits—any positive impacts on the research subject(s), or their group. These are the positive effects resulting from the investment [13].

Utility—the emergence of new values (e.g. increasing the value of a property) as a result of the project implementation, reducing the costs associated with the relevant processes as a result of the project implementation.

Based on interviews with IT experts, the general steps of the CBA, method mentioned in [14, 15] were modified for cloud computing purposes. These include:

- specification of deployment of cloud computing,
- deciding which subjects are related to the impacts of the project,
- description of the differences between current IT solutions and cloud computing,
- learning and possible quantification of all relevant costs and benefits,
- expression of all the consequences of the introduction of cloud computing in monetary units,
- calculation of criteria indicators,
- sensitivity analysis and assessment of the project.

Given the scope of this paper, in the description of the individual CBA steps, the attention is focused especially on the specification of the costs and benefits of cloud computing, indicators and methods for their economic assessment.

3.1 Specifications for Cloud Computing Deployment

The basic solution to deploy the correct form of cloud computing is the identification of possible current shortcomings in the IT workplace and the required functionality. There are several indicators that can highlight the current inefficiencies in the IT workplace. These include a deviation from past performance, a deviation from the plan or outside criticism.

At this stage they should identify the deployment model: public, private, hybrid or community. Furthermore, a distribution model which will be used, or that will be offered under the cloud computing (IaaS, SaaS, PaaS model). Specification of the desired functionality should be the most accurate. For this purpose a system of criteria was created and designed, which is divided into three levels: economic, operational and technical criteria. The model is based on the literature review and interviews in companies (15) and also on the opinions of five experts in the field of implementing this technology in the Czech Republic.

Economic-financial criteria are the primary indicators of economic efficiency of the target state in relation to the current solution. The most quantifiable criteria include investments in infrastructure, which especially includes hardware (servers, network devices, security technology, data storage, connectivity and peering into the Internet or virtual private networks) and software solutions (solutions for virtualisation of computer nodes, platform for authentication, adducting and accounting of the operation, user front-end for access to infrastructure, platform and applications, monitoring system). Economic criteria are fundamentally affected by the selected model of cloud computing services, at this stage it is necessary to analyse the different levels of integration of cloud services from migration of the entire infrastructure, or platforms up to the operation of the selected applications.

Operational criteria are very important indicators for the effectiveness of operation of the current/target solution. Different methodologies and approaches deal with the area of efficient design and operation of information technologies in organisations. The connecting link of these approaches is the emphasis on connection of technological and economic perspective of the implementation of information technology in business processes [16]. According to [17] the operational criteria are often overlooked and may result in problematic operation of the entire solution. The main qualitative factors include user-friendliness and user support system, risk and incident management of events, platform flexibility, speed/difficulty developing new applications and requirements. Service Level Agreements (SLAs) are part of service contracts and are agreements between usually two parties (service provider and customer), which formally define the services. Service contracts use the percentage of service availability as a unit [18].

A service-level agreement is an agreement between two or more parties, where one is the customer and the others are service providers. This can be a legally binding formal or an informal "contract" (for example, internal department relationships). SLAs commonly include segments to address: a definition of services, performance measurement, problem management, customer duties, warranties, disaster recovery, and termination of agreement [19].

Technical criteria represent a wide range of parameters and characteristics of information technologies and infrastructure, which are usually possible to precisely quantify or specify its nature. The fundamental problem, however, is their connection with the economic and operational needs of the organisation. Performance and capacity of a particular element of the IT architecture of the solution is not a guarantee of success or effectiveness of the solution. Success of cloud services depends on required functionality and other characteristics like availability, respond time, latency, performance, timeliness, scalability and high availability. All of these characteristics can be covered by term Quality of Cloud Service (QoCS), which comes from general QoS [20].

3.2 The Decision on Which Subjects Are Related to the Impacts of the Project

The result of this step should be a structured list of subjects to investigate the effects of the introduction of cloud computing. For this purpose, it is possible to use the so-called Stakeholder model. The approach sees a company as a place of conflict of interests of various stakeholders, who give a certain deposit of their scarce resources and goods into the relation with the organisation. The model identifies six groups of stakeholders, which include: owners, employees, creditors, suppliers, customers and the state, which represents all non-market organisations around the company [21] (Fig. 1).

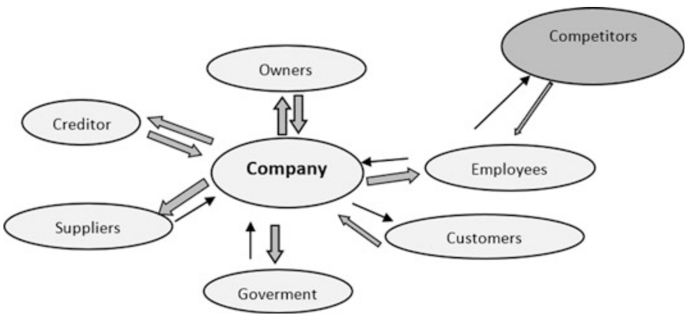


Fig. 1 Company as a point of impact of different entities. *Source* Author according to [22]

3.3 Description of the Differences Between Current IT and Cloud Computing

All the effects and impacts arising from the investment plan will be better identified when two variants stand next to each other. The first variant is the one in which the investment plan is described, i.e. investment variant, and the second one will be describing the original state, i.e. zero variant. The difference between these two states includes all consequences resulting from the investment, which can be used for their valuation. The subsequently determined effects of the introduction of cloud computing will include the impacts, which individual entities wouldn't record in the case of zero variant.

3.4 Identification and Quantification of All Relevant Costs and Benefits

Within this step, a structured list of the costs and benefits should be developed. This step can be considered as crucial and should be facilitated by former precise specification of the zero and investment variants, as well as the list of entities covered by the impacts of introducing the new technology. For illustration, some impacts are structured in Table 1.

3.5 Expression of All the Consequences of the Introduction of Cloud Computing in Monetary Units

The next step within the CBA is converting all impacts of the project to the form of cash flows, on which the criteria indicators of investment evaluation will be applied. All the implications of the project, which in its primary form gain the forms of income or expenditure (realised revenues, expenses paid, etc.), are already in financial terms. However, the area of cloud computing brings many of (see chapter "Real-Valued Negative Selection Algorithms: Ensuring Data Integrity Through Anomaly Detection") those that are expressed in non-monetary form. In the event that there is a market, which would set the price for such a product, it is possible to use direct valuation based on the market price. The essence of the method is finding an identical or very similar asset, which enters the market and is traded at a market price. This procedure cannot be used for a full range of effects resulting from the introduction of knowledge management. In this case it is necessary to use other methods. Methods used in the method of Cost Benefit Analysis is shown in Table 2.

A more detailed procedure for using these methods is provided in the above-mentioned sources.

Table 1 Cost types and related cost factors, non-quantifiable and quantifiable benefits

Cost type	*Cost factors*
Strategic decision, selection of cloud computing services and cloud types	Expenditure of time, consulting services, information for decision-making
Evaluation and selection of service provider	Expenditure of time, consulting services, information for decision-making
Service charge *IaaS*	Computing power, storage capacity, inbound data transfer, outbound data transfer, provider internal data transfer, number of queries, domain, SSL certificate, licence, basic service charge
Service charge *PaaS*	User-dependent basic charges, storage capacity (for the developer team), inbound data transfer, outbound data transfer, provider internal data transfer, extra user data storage capacity, extra user document storage capacity, queries to the Application Programming Interface, sent emails, database, secured logins, connections with other providers' applications
Service charge *SaaS*	Access to the service system, user
Implementation, configuration, migration	Expenditure of time, porting process
Support	Expenditure of time, support costs, problem solving
Benefits	*Non-quantifiable and quantifiable*
Quantifiable	Reduction in operating costs of IT department, energy saving, consolidation of number of physical machines and their replacement for virtual (servers, network devices, security systems), saving of staff and appreciation of the consumption of information sources
Non-quantifiable	Simplification of management of the information system, concentration of it staff on key activities, increased flexibility of it architecture and response to business needs, improving the resilience of systems and increase the availability of services through internet technology, online technical support available, increased collaboration with customers

Source According to [23]

3.6 Calculation of Criteria Indicators

The evaluation project of the economic effectiveness provides key information for assessing the profitability of the project and deciding on its implementation [25]. The fundamental division of these methods is based on whether the method takes into account the time factor:

Table 2 Overview of evaluation methods used in the CBA

Type of a method according to the source data		Method
Preferential methods		Contingent valuation method
Market methods	Comparative methods	Method of shadow prices
		Method of asset value
		Method of hedonic prices
		Study of labour markets (a method of wage risk)
	Cost methods	Travel cost method
		Method of cost prevention
		Consumer behaviour study
		Method of cost recovery
		Method of alternative costs
		Method of opportunity costs
	Benefit methods	Method of blanks
		Method of past income
Experts and other methods		

Source Own processing by Shahid [24]

(a) Static methods—these methods ignore the time factor,
(b) Dynamic methods—respecting the time factor in the investment decision-making.

Most common used methods within cloud computing are: profitability indicators (ROI—Rentability of investment, ROA, and ROE), NPV, TCO (Total cost of ownership) and productivity per employee. In this step, companies can decide for themselves what economic indicators will be used. At this stage all the necessary input values are available.

3.7 Sensitivity Analysis and Assessment of the Project

In case it is not possible to determine the exact input parameters for evaluating the effectiveness of investment projects, you can use a sensitivity analysis to help determine how much the tested project is sensitive to changes in various factors that may affect it. The purpose of sensitivity analysis is to determine the sensitivity of certain economic criteria of the project (e.g. its net present values) depending on the factors that affect this criterion. It is a determination of changes in certain parameters (volume of production, price of products, basic raw materials and energy, investment costs, interest and tax rates, foreign exchange rates, the life of the project, discount rates, etc.) in response to changes in the factors that affect these variables.

Table 3 Relationship between traditional methods of investment evaluation and CBA

	The traditional approach to the assessment of investments	CBA
Type of method	Static (profitability indicators payback period, indicators based on sales)	This analysis compares the benefits and costs of the project with respect to qualitative variables. These variables are converted to quantitative expression. In conclusion, the method includes qualitative variables to standard methods of investment evaluation
	Dynamic (NPV, IRR, EVA)	
	Indicators for IT (TCO, ITIL, IT scorecard)	
Benefits	Fast processing	The possibility of evaluation of qualitative factors
	Familiarity of business environment in the European business sector	A comprehensive method for decision making
	By using multiple methods—high explanatory value	Method is not completely new for companies
Disadvantages	The need for interpretation of the results in the context of other processes in the company	Relatively long duration of methods processing
	The required combination of multiple indicators	Companies are not used to process this methods
	High demands on the exactness of inputs	

Source Author

4 Discussion—Differences of CBA to Other Investment Evaluation Methods

Comparison of classical methods of investment decision in business economics and CBA method is described in Table 3.

Fundamental benefits of the CBA method in relation to the common methods of investment evaluation in ICT area are: complexity, the possibility of using both static and dynamic methods, relation to the characteristics of the company and its activities, available applications of this method in other areas and the conversion of qualitative variables into quantitative terms.

5 Conclusion

The use of cloud computing services appears to offer significant cost advantages [26, 27]. The aim of this paper was to propose a theoretical model evaluating the effectiveness of investments in cloud computing in business practice. The Cost Benefit Analysis method was selected, as a suitable framework for the evaluation of this technology. The steps of this method were modified and characterised in relation to characteristics of cloud computing. The main benefit of the method compared with the commonly used indicators (only ROI and TCO are currently used in practice) is the inclusion of non-financial benefits and their conversion expressed in terms finances. Thus, these factors can be involved into a decision about cloud computing. In many cases it happens that the amount of financial investment is the same as the updating of the current state of IT in companies. Soft benefits can be often crucial for decision-making.

At the present time the model is tested in selected enterprises that have implemented this technology within the Czech Republic. Based on this verification, specification of each method steps are expected. As part of further research we also expect the creation of a specific freely available web application that would allow companies and other entities (after other possible adjustments within the test operation) to carry out an independent initial evaluation of the potential of cloud computing. The question of web architecture is currently being solved.

Acknowledgements This paper is created with support of the specific research project "Economic and Managerial Aspects of Processes in Biomedicine" and project Excellence 2015 funded by the University of Hradec Kralove.

References

1. Brune Thornton, R., Bureš, V.: The cross-cultural management: establishing a Czech benchmark. E + M Ekonomie Manag. **15**(3), 46–62 (2013)
2. Křenek, J., Kuča, K., Krejcar, O., Marešová, P., Bartůšková, A.: Artificial neural networks in biomedicine applications. In: Proceedings of the 4th International Conference on Computer Engineering and Networks. Springer, Shanghai (2015)
3. Brancheau, J.C., Janz, B.D., Wetherbe, J.C.: Key IS issues in information systems management: 1994–95 SIM Delphi results. MIS Q. **20**(2), 225–242 (1987)
4. Brancheau, J.C., Schusterand, L., March, S.T.: Building and implementing an information architecture. Database **20**(2), 9–17 (1989)
5. Watson, R.T., Kelly, G.G., Galliers, R.D., Brancheau, J.C.: Key IS issues in information systems management: an international perspective. J. Manag. Inf. Syst. **13**(4), 91–115 (1997)
6. Laudon, K.C., Laudon, J.P.: Management Information Systems: Managing the Digital Firm. New Jersey (2004)
7. Dedrick, J., Gurbaxani, V., Kraemer, K.L.: Information technology and economic performance: a critical review of the empirical evidence. Comput. Surv. **35**(1), 1–28 (2003)
8. Sosinsky, B.A.: Cloud Computing Bible, p. 497. Wiley, Chichester (2011)
9. Mell, P., Grance, T.: The NIST Definition of Cloud Computing (Draft): Recommendations of the National Institute of Standards and Technology [Online]. Available from http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf
10. Marešová, P., Hálek, V.: Deployment of cloud computing in small and medium sized enterprises in the Czech Republic. E + M Econ. Manag. **17**(4), 159–174 (2014)
11. Harrigan, P.O., Boyd, M., Ramsey, E., Ibbotsonand, P., Bright, M.: The development of e-procurement within the ICT manufacturing industry in Ireland. Manag. Decis. **46**(3), 481–500 (2008)

12. Gabler, E.: Economics of an Investments: Analysis Helps Agencies Target Limited Transportation Resources to Their Best Uses, Innovations and Technology, p. 66. Roads and Bridges (2004)
13. Sieber, P.: Analýza nákladů a přínosů metodická příručka [Cost-benefit analysis methodological guide], Ministerstvo pro místní rozvoj. 2004, [on line]. Available from http://ww.strukturalni-fondy.cz/uploads/old/1083945131cba_1.4.pdf
14. Boardman, E., Greenberg, D.H., Vining, A.R., Weimer, D.L.: Cost-Benefit Analysis: Concepts and Practice, 3rd edn, p. 493. Prentice Hall, Upper Saddle River, New Jersey (2006)
15. Nas, T.F.: Cost-Benefit Analysis. Theory and Application, p. 220. SAGE Publications, Thousand Oaks, California (1996)
16. Steinberg, R.A.: Measuring ITIL: Measuring, Reporting and Modeling the IT Service Management Metrics that Matter Most to IT Senior Executives, 1st ed, p. 154. Trafford, Victoria (2006)
17. Mcgovern, J.: A Practical Guide to Enterprise Architecture, p. 306. Prentice Hall, Upper Saddle River, NJ, Professional Technical Reference (2004)
18. Strassmann, P.A.: The Squandered Computer: Evaluating the Business Alignment of Information Technologies, p. 426. The Information Economics Press (1997)
19. Service Level Agreement Zone (DT): An Outline of the Core Elements of an SLA. 2014, [online]. Available from http://www.sla-zone.co.uk/
20. CISCO: Cisco IOS IP SLAs Configuration Guide 2008, [online]. Available from <http://www.cisco.com/en/US/docs/ios/12_4/ip_sla/configuration/guide/hsla_c.html>
21. Marešová, P., Kuča, K.: Assessing the effectiveness of cloud computing in European countries. In: Proceedings of the 4th International Conference on Computer Engineering and Networks. Springer, Shanghai (2014)
22. Šiška, L.: Možnosti měření a řízení efektivnosti a úspěšnosti podniku [Options measurement and management effectiveness and success of an enterprise], Working Paper č. 14/2005, Centrum výzkumu konkurenční schopnosti české ekonomiky
23. Walterbusch, M., Teuteberg, F.: Costing of cloud computing services: a total cost of ownership approach. In: Proceedings of the 45th Hawaii International Conference on System Sciences, pp. 1563–1572. Maui, Hawaii, USA (2012)
24. Shahid, S.: A Business-Driven Evaluation of Distributed-Computing Models, 2008, [online]. Available at http://msdn.microsoft.com/en-us/library/cc984967.aspx
25. Bartuskova, A., Krejcar, O., Selamat, A., et al.: Framework for managing of learning resources for specific knowledge areas. In: 13th International Conference on Intelligent Software Methodologies, Tools, and Techniques (SoMeT), Frontiers in Artificial Intelligence and Applications, vol. 265, pp. 565–576 (2014)
26. Horalek, J., Sobeslav, V., Pavlik, J.: Statistic and analysis of service availability in cloud computing. In: IDEAS'14. Byte Press, Porto, pp. 310–313 (2014)
27. Horalek, J., Sobeslav, V., Matyska, J., Suba, P.: Energy efficiency measurements of data center systems. In: Elektro 2014. Faculty of Electrical Engineering, University of Zilina, Žilina, pp. 41–45 (2014)

Electrical Technologies in Healthcare—Cost of Development of Linear Octapolar Tool in the Czech Republic

Petra Marešová and Marek Penhaker

Abstract Whereas in countries like the USA, Japan or China the development of medical devices has been on the rise, in Europe it has stagnated. Nevertheless, these products' significance has been growing in developed countries because of their ageing population. This contribution describes significant aspects of medical electrical equipment development in relation to investment in research and development. It also contains a case study looking into the costs of developing a linear octapolar tool for rapid ablation in time and mass of tissue, which was designed in the Czech Republic in 2013.

1 Introduction

The development of medical devices has been rising in the USA, Japan or China, whereas in Europe it has stagnated. The development of medical devices is expected to affect the future consumers' demand and behaviour and the expectations of the future medical device market are high. Several available forecasts indicate growth in the market for medical devices 20 % for 2022 [1].

Concerning medical technology patents, USA has enjoyed an unchallenged clear long-term leading position. Europe is the only region reporting a decline, whereas China experiences a sharp growth (Fig. 1). The development of these patents can be seen from different angles [2, 3]. On the one hand, the growth of this indicator means at the same time the growth of the market in the country in question. On the other hand, there will be much stronger competitive pressure on the mentioned market.

P. Marešová (✉) · M. Penhaker
Faculty of Informatics and Management, University of Hradec Kralove, Rokitanskeho 62, 50003 Hradec Kralove, Czech Republic
e-mail: petra.maresova@uhk.cz

M. Penhaker
e-mail: marek.penhaker@uhk.cz

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_78

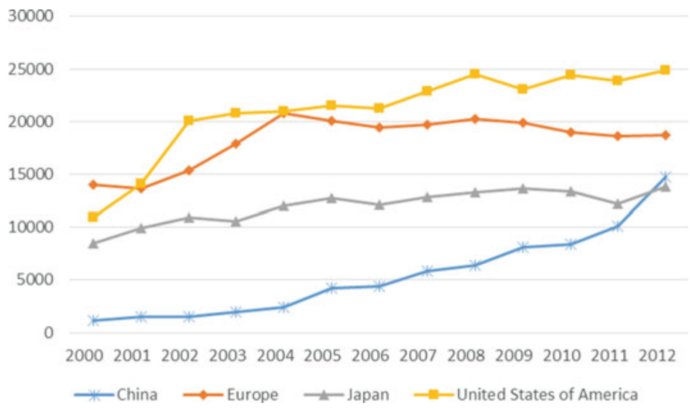


Fig. 1 The number of patents in China, Europe, Japan and United States. *Source* According to [26]

There are reasons why the existing competition is very strong, namely [4]: low product differentiation in certain markets, hospital's low switching costs, various changes (reimbursement, regulatory), biological challenges, strong competition in the market may also result in declining venture capital investment.

In developed countries with their ageing population medical devices become more and more significant. This contribution aims to describe certain aspects of developing medical electrical equipment in relation to science and research investments as well as to present a case study on costs of developing a linear octapolar tool for rapid ablation in time and mass of tissue, which was designed in the Czech Republic in 2013.

2 Theoretical Background

2.1 Definition and Regulations of Medical Device Equipment

Medical devices are diverse products. They are divided into specific classes, whose regulatory requirements differ. A more precise specification is as follows: a medical device is any instrument, appliance, apparatus, material, software or other article to be used for human beings with following purposes: diagnosis, monitoring, prevention, alleviation or treatment of disease [5].

Various regulatory bodies classify medical devices in different ways. They may take into account, for instance, the potential hazards of using a device or its breakdown or failure; the duration of its contact with human body; its degree of invasiveness; its local versus systematic effects. Despite a lack of any universally approved classification, medical devices in general are put into categories according to the level risk involved in using them. The higher are the risks associated with using a device, the higher is their class and the greater level of assessment is required by regulatory bodies. Presently, medical device regulations vary considerably, which makes compliance to any set of common rules difficult to achieve. Nevertheless, there are two important regulations to be considered by developers; namely, the EC Medical Device Directive 93/42/EEC and the US Food and Drug Administration regulations. Compliance with them is required by the European Community (through the CE mark) and the USA, respectively [6]. Medical device regulation in the EU is based on and follows the three below-listed EU directives and their amendments:

- Council Directive 90/385/EEC on Active Implantable Medical Devices (AIMDD) (1990).
- Council Directive 93/42/EEC on Medical Devices (MDD) (1992).
- Council Directive 98/79/EC on In Vitro Diagnostic Medical Devices (IVDMD) (1998).

These directives are supported by guidelines promoting, according to the European Commission (EC), "common approach by manufacturers and notified bodies involved in the conformity assessment procedures according to the relevant annexes of the directives, and by the competent authorities charged with safeguarding public health".

All medical devices launched into the EU market must comply with the corresponding directive. The MDD covers a wide range of medical devices, including walking frames, first-aid bandages, implantable devices or CT scanners. The level of medical device assessment varies based on the perceived risks involved in using the device. Medical devices are classified in the following way:

- Class I—generally regarded as low risk
- Class IIa and IIb—medium risk
- Class III—high risk.

Companies medium and high risk devices are required to have their quality systems and technical documentation reviewed by a notified body before their products have been launched onto the market. The notified body is responsible for making sure that all requirements are met.

2.2 *Development of Medical Devices—A Review*

Research and development is essential in all sectors [7]. Large medical device companies usually develop variations of existing devices. It is venture-backed start-ups that come up with most new device categories. Only few ideas are conceived in academic medical centres financed from federal funds or other grants. Few academic centres are capable of developing a device beyond its early prototype stage. Intellectual property is often out-licensed to another company or start-up to

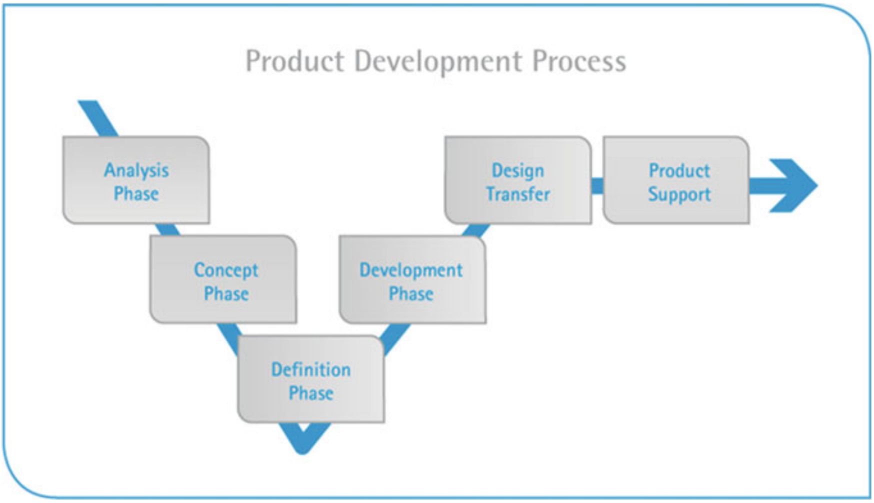


Fig. 2 Scheme of product development process. *Source* [27]

be further development [8]. Medical device development follows the algorithm depicted in Fig. 2.

Case studies in the development of medical electrical equipment concentrate on technical aspects and the system itself. For instance, Marani and Perri [9] describe a medical electronic-computerized platform for diagnostic use, allowing doctors to carry out a complete cardio-respiratory examination on remote patients in real time. The system has been developed to enable real-time rescue in case of emergency without data being constantly monitored by a medical centre, which leaves patients free to move. Therefore, the system has been equipped with advanced firmware ensuring automated functioning and making complex decision-making possible. In case of emergency detected through the real-time diagnosing system, the system informs by means of a warning message persons capable of arranging the patients rescue, providing information about the patient's position. All this happens automatically without the patient's intervention. The system might be used by sportspeople, too. Moreover, in this study is also described a microcontroller-based digital electronic system, which monitors the patient's respiratory cycle and relevant ventilator setting. The system allows accurate processing, effective auscultation, and detailed visualization (temporal and frequency graphs) of any sound produced by lungs. Last but not least, it can be used for continuous real-time monitoring of breathing functions, which may prove useful in order to diagnose respiratory pathologies. Finally, the study presents the digital Cardio-Holter with multiple leads and a system for ECG transmission by Bluetooth.

Another study [10] outlines some existing as well as emerging biomedical applications and provides a list their particular performance requirements. It highlights the power constraints and performance of biomedical devices. It also presents circuit techniques to achieve complete systems operating down to power levels of microwatts. Approaches that exploit advanced technology trends are emphasized. Devices like pacemakers must be highly energy-constrained as it is required to operate on a single battery charge for years in order to avoid repeated surgeries. However, there are also emerging applications like neural recording systems or retinal stimulators. Researchers concentrate on processing platforms as well as challenges in the interface and acquisition electronics. To sum up, as biomedical devices become common, the demand and need for these devices grows as well as they should support the formation of body-area networks, which allow for individual devices to communicate and cooperate with one another.

Other research papers treat the economic impact of financial support to medical device development. This report looks into potential waste in healthcare that could be eliminated if medical devices better interoperated and if commonly accepted standards for interoperability were adopted. Waste reduction facilitated by greater medical device interoperability would lead to higher efficiency, better quality and more affordable care. Commonly accepted standards can potentially accelerate medical device interoperability and reduce the cost of achieving interoperability [11].

Another study [12] focuses on functionality and performance of medical devices. Clinical evaluation reports, or CE reports, verifying safety and performance of medical devices are currently a required part of the CE marking process in the developed countries, including Europe.

The text above suggests that most case studies deal with technical aspects of medical device product development. Other studies concentrate on economic aspects.

On the other hand, the case study below describes functionality, purpose and costs linked with developing linear octopolar tool for radiofrequency ablation. This tool was developed at Technical University of Ostrava in Czech Republic. The costs were therefore covered by the Ministry of Education, Youth and Sports of the Czech Republic.

3 Case Study in Czech Republic—Development of Electric Medical Device—Linear Octapolar Tool for Radiofrequency Ablation

Radiofrequency ablation is a minimally invasive procedure. It can be used to treat nonresecabile primary and secondary tumours. Radiofrequency ablation (RFA) technique is a kind of intervention during which the electrode or electrode system is laparoscopically introduced by a physician directly to the tumour. Is done surgically or percutaneously through the skin [13]. Principle of operation of electrodes varies depending on types of tools. In the last decade, various devices have been developed and designed in order to minimise the blood loss during laparoscopic liver parenchymal transection [14]. In spite of these efforts, bleeding remains the main cause of conversion to a laparotomy in order to limit the haemorrhage during the liver resection [15–17]. One of the used tools is the laparoscopic Habib™ 4X.

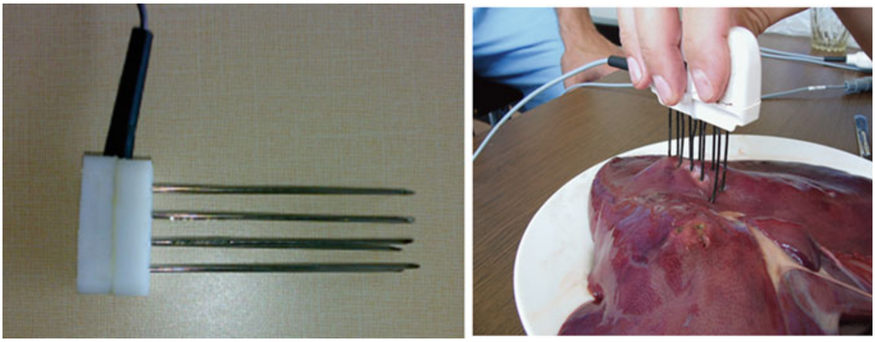


Fig. 3 Radiofrequency linear octopolar tool. *Source* According to [19]

This tool measures 45 cm in length with a protected insulated area to allow 5 cm in length for RF. The device can be introduced via a 10–12 mm laparoscopic port and is connected to a 500-kHz generator. It allows to measure the time, temperature, and generator output and tissue impedance. There is also a pneumatic foot pedal used to switch the RF energy on and off. The generator can be operated in automatic mode or manual mode. On connecting the device and turning the generator on, the RF power setting defaults to 125 W. This can be modified based on the user's experience and thermal requirements of the individual tissue types [18].

Octapolar tool for rapid ablation in time and mass of tissue was designed using the expert knowledge of surgeons. Moreover, the existing tool HABIB 4X served as another source of reference. The distance between lateral needles is 10 mm and axial needle distance is 8 mm. All needles are 80 mm long. Chrom molybden coated steel staple is used (see Fig. 3). There were also used RITA Generator 1500× and beef liver in 520 g [19].

The following text informs about the costs linked to the two-year-long development of this tool in the Czech Republic.

3.1 Cost to Develop Linear Octapolar Tool

Costs are not known for all phases of development. The attention will be focused on the development phase within the product development process (Fig. 4).

The costs in the first year of solving the project included personal costs linked to researchers' salaries, cost of materials for histopathologic examination, purchase of animals, housing the animals during experiment period, fodder, treatment of these animals, also two laparoscopic operations, haematological examination and an autopsy. In the second year, the structure of costs was similar. Beside the above mentioned items, remunerations for surgeons doing animal operations were included. The total cost was 58,444 EUR. These costs were paid for the University of Ostrava, Czech Republic (Table 1).

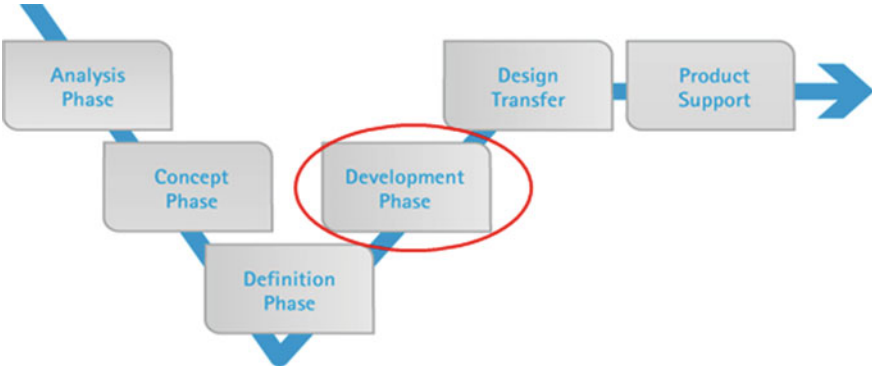


Fig. 4 The development phase within the product development process. *Source* Own processing according to [27]

Table 1 Cost to develope linear octopolar tool on the development phase (EUR)

	1st year	2nd year
Personal costs	12,963	11,111
Cost of material	1296	1296
Literature	185	185
Services	15,704	15,704
Total	30,148	28,296

4 Summary and Future Research

Currently, there are very few studies specifying medical devices development costs with respect to a particular application area. In the Czech Republic there are almost no available public data about the development of various medical devices. The area of the methodology for evaluating the effectiveness is crucial for deeper cooperation between scientific institutions and companies. The entities which are affected by the development of medical devices are described by Fig. 5. He costs

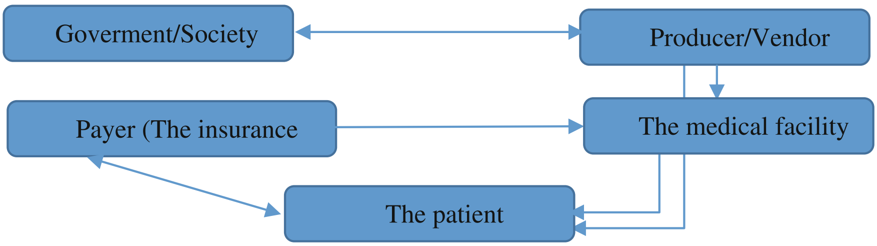


Fig. 5 Financing among individual entities. *Source* Own

from the payer (direct medical expenses), patient (lost wages, co-pays), hospital (operating room capacity, length of stay in hospital), and societal (lost productivity) perspectives should be examined.

Reasons for creating the system of assessing the investment in medical device development in the future in the Czech Republic are as follows:

- Limited financial resources of governments in developed countries and an effort to spend effectively on research and development.
- Investments in the development and production of medical devices in relation to the growth of this market seems to be prospective in the private sector.
- Annual accrual of healthcare spending caused by using new technologies.
- A growing number of patients who live longer thanks to the high quality healthcare.

5 Conclusions

Producing medical devices has become significant due to the demographic trends leading to the ageing of population in developed countries. The objective of this contribution was to describe aspects linked to developing medical electrical equipment in relation to investments into science and research. Another goal was to present a case study on the costs of developing the linear octapolar tool for rapid ablation in time, which was designed in the Czech Republic (University of Ostrava) in 2013. Currently accessible case studies concentrate either on technical aspect of developing medical devices [20–23, 24] or on macroeconomic views of the medical device industry, or on managerial procedures describing individual stages of electrical device development [25].

In comparison with other case studies in this area, the presented case study uses other data, for instance, the costs. There was also illustrated the cost of one stage of developing a medical device at an EU university.

The follow-up research will compare cost items for similar types of medical devices in various EU countries. Furthermore, the researchers plan to describe the costs in other stages of developing other medical devices in the Czech Republic.

Acknowledgements This paper is created with support of the specific research project "Economic and Managerial Aspects of Processes in Biomedicine" funded by the University of Hradec Kralove.

References

1. Happich, J.: Medical devices push microfluidics and electrodes markets to reach $9.5 bn in 2022, says Lux. In: EETimes Europe Analog. Available from http://www.analog-eetimes.com/en/medical-devices-push-microfluidics-and-electrodes-markets-to-reach-9.5bn-in-2022-says-lux.html?cmp_id=7&news_id=222905697 (2014). Accessed Jan 19 2015
2. Marešová P., Kuča K.: Porter's five forces on medical device industry in Europe. Military Medical Science Letters (2014)
3. Marešová, P., Mohelská, H., Kuča, K.: Cooperation policy of rare diseases in the European Union. In: 5th International Conference on Education and Educational Psychology, pp.134–144 (2014)
4. Wipperfurth, A., Savary, K., Gilchrist, A.: Minnesota Medical Device Cluster. Available from http://www.hhh.umn.edu/centers/slp/economic_development/documents/MNMedicalDeviceCluster.pdf (2010). Accessed 19 Jan 2015
5. Medtech: The European Medical Technology Industry Study of Medtech Europe. Available from http://www.eucomed.org/uploads/Modules/Publications/20140219-mte-data-brochure-the_eu_medtech_industry_in_figures.pdf (2014). Accessed 19 Jan 2015
6. Martin, J.L., Norris, B.J., Murphy, E., Crowe, J.A.: Medical device development: the challenge for ergonomics. Appl. Ergonomics **39**(3), 271–283 (2008)
7. Marešová, P.: Analýza volby na trhu pojistných produktů. Hradecké ekonomické dny 2011: Ekonomický rozvoj a management regionů, pp. 189–195. Gaudeamus, Hradec Králové (2011)
8. Kaplan, A.V., Baim, D.S., Smith, J.J., Feigal, D.A., Simons, M., Jefferys, D., Fogarty, T.J., Kuntz, R.E., Leon, M.B.: Medical device development, from prototype to regulatory approva. In: Dartmouth Drug and Device Development Symposium (3DI) Held at the Woodstock Inn, pp. 3068–3072 (2003)
9. Marani, R., Perri, G.: Design of advanced electronic biomedical systems. Int. J. Adv. Eng. Technol. **4**(1), 15–25 (2012)
10. Chandrakasan, A.P., Verma, N., Daly, D.C.: Ultralow-power electronics for biomedical applications. Annu. Rev. Biomed. Eng. **10**, 247–274 (2008)
11. West Health Institute: The Value of Medical Device Interoperability. Available from https://S3.Amazonaws.Com/Wwhi.Org/Interop/Whi-The_Value_Of_Medical_Device_Interoperability.Pdf (2013). Accessed 19 Jan 2015
12. McLain, D.E.: EMBASE—White Paper—Conducting Medical Device Safety and Performance Assessments. Available from http://www.elsevier.com/__data/assets/pdf_file/0003/180840/MedDevice-Embase-White-Paper.pdf (2012). Accessed 19 Jan 2015
13. Klener, P.: Klinická onkologie. Galén (2002)
14. RITA Medical Systems: Medical Systems. Available from http://www.cattus-ritamedical.com (2014). Accessed 19 Jan 2015
15. Fong, Y., Jarnagin, W., Conlon, K.C., DeMatteo, R., Dougherty, E., Blumgart, L.H.: Hand-assisted laparoscopic liver resection: lessons from an initial experience. Arch. Surg. **135**, 854–859 (2000)
16. Kaneko, H.: Laparoscopic hepatectomy: indications and outcomes. J. Hepatobiliary Pancreat. Surg. **12**, 438–443 (2005)
17. Dagher, I., Lainas, P., Carloni, A., Caillard, C., Champault, A., Smadja, C., et al.: Laparoscopic liver resection for hepatocellular carcinoma. Surg. Endosc. **22**, 372–378 (2008)
18. Pai, M., Navarra, G., Ayav, A., Sommerville, C., Khorsandi, S.K., Damrah, O., Jiao, L.R., Habib, N.A.: Laparoscopic Habib 4X: a bipolar radiofrequency device for bloodless laparoscopic liver resection. HPB (Oxford) **10**, 261–264 (2008)
19. Penhaker, M., Hlavackova, M., Horyl, P., Vavra, P., Prokop, L., Sikora, T.: Linear octapolar radiofrequency tool for liver ablation. In: 15th IEEE International Conference on Intelligent Engineering Systems (INES), 2011, pp. 405–409 (2011)

20. Pajic, M., Jiang, Z., Lee, I., Sokolsky, O., Mangharam, R.: Safety-critical medical device development using the UPP2SF model translation tool. ACM Trans. Embed. Comput. Syst. **5**, 1–25 (2013)
21. Behan, M., Krejcar, O.: Modern smart device-based concept of sensoric networks. EURASIP J. Wirel. Commun. Networking **2013**(1) (2013)
22. Krejcar, O., Jirka, J., Janckulik, D.: Use of mobile phone as intelligent sensor for sound input analysis and sleep state detection. Sensors **11**(6), 6037–6055 (2011)
23. Krejcar, O., Frischer, R.: Smart intelligent control of current source for high power LED diodes. Microelectron. J. **44**(4), 307–314 (2013)
24. Horalek, J., Sobeslav, V., Matyska, J., Suba, P.: Energy efficiency measurements of data center systems. In: Elektro 2014, University Of Zilina, Faculty Of Electrical Engineering, Žilina, pp. 41–45 (2014)
25. Horalek, J., Sobeslav, V.: Data networking aspects of power substation automation. In: Communication and Management in Technological Innovation and Academic Globalization, pp. 147–153 World scientific and engineering academy and society, Athens, (2010), ISBN 978-960-474-254-7
26. WIPO: Tge website of The World Intellectual Property Organization. Available from http://ipstats.wipo.int/ipstatv2/IpsStatsResultvalue (2014). Accessed 19 Jan 2015
27. SensileMedical: Product Development. Available from http://sensile-medical.com/product-development-63.html (2014). Accessed 19 Jan 2015

The Design of Railway Network Infrastructure Model for Localization of Rolling Stock with Utilization Technology Oracle Spatial and Dynamic Database Views

Jan Fikejz and Emil Řezanina

Abstract In its first part, this article discusses various options for the localization of rolling stock, with an emphasis on regional lines. The following part focuses on the description of the three-layer rail network model design reflecting an undirected graph. The proposed model is used for simulation of rolling stock traffic and induction of various non-standard situations, detection of which could be used as a supplementary aid to dispatching traffic control. Further attention is aimed on optimizing search operations ORACLE database and designing of the optimization using dynamic views.

1 Introduction

Localization of rolling stock has been a largely discussed issue involving a wide range of subjects. The issue of rolling stock localization could be generally divided into two main areas of interest. Localization for the needs of (i) signalling technology and localization for the needs of (ii) information and telematics systems. While the former underlines the safety and reliability, these systems usually call for higher implementation costs as they often require the complementing of railway infrastructure with additional communication or identification elements/equipment.

J. Fikejz (✉) · E. Řezanina
Department of Software Technologie, FEI, University of Pardubice,
532 10 Pardubice, Czech Republic
e-mail: Jan.Fikejz@upce.cz

E. Řezanina
e-mail: Emil.Rezanina@upce.cz

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_79

2 Possible Types of Localization

Rolling stock localization can be divided into three main parts:

- localization without the use of GNSS,
- GNSS using localization,
- GNSS-based, involving further support systems.

2.1 Rolling Stock Localization Without the Use of GNSS

This type of rolling stock localization often requires complementing the rail network infrastructure with additional construction elements, which entails higher costs of the actual implementation. On the other hand, this type of localization shows a high accuracy and reliability and is often used in the railway signalling technology. Essentially, it relates to the system of:

- ETCS [1, 2]
- Automatic train control [3, 4]
- Track circuits [5]
- RFID

2.2 Rolling Stock Localization Using GNSS

If we use GNSS for various application levels, we need to take into account an indicated position error, which is generally based on the nature of the satellite navigation. If we use systems that operate with the position information only on an informative level, we can tolerate a certain error; however, such inaccuracy is unacceptable in the railway signalling technology. However, various additional systems can be implemented to eliminate the error (completely or at least partially), thus making the position of the tracked object more accurate. The following systems can be listed in this group:

- EGNOS [6]
- Differential GPS [7]

2.3 GNSS Based Localization Involving Additional Support Systems

As mentioned above, precise localization of rolling stock using GNSS, especially for the needs of signalling technology, is a priori impossible. Nevertheless, the position of a rail vehicle can be put significantly more precisely with the use of additional systems. This concerns especially the solutions using inertial systems [8], but also less known systems such as those based on GNSS and contactless eddy current measurement [9].

3 Localization Within Regional Tracks

There are many different solutions around the world that deal with localization of rail vehicles on regional tracks; most of these systems use GNSS, balises or odometry. Wireless transmission of information is then realized mainly through mobile GSM-R networks (Global System for Mobile Communications—Railway) or GSM. Namely, these are [10]

- ERTMS Regional—Sweden,
- LOCOPROL—France,
- 3InSat—Italy,
- SATLOC—Romania.

In the Czech Republic, Radioblok System, developed by AZD, is currently being tested on the track Číčenice—Volary. This system is based on sending messages (via the cellular network) on location information into the radio block centre and vice versa, the driver of the rail vehicle receives information for permission to go from the radio block centre, while satellite navigation is used here to control the movement of the vehicle.

Rolling stock localization based solely on GNSS can be used, for example, to identify the position within single-track regional lines in the Czech Republic that, unlike the main corridors, do not have such a high degree of technical equipment such as advanced security systems ETCS. With regional lines we can often only register that a train has departed/arrived from/to a station, but with an open line the localization of a rail vehicle without supplementing the infrastructure with identification elements is considerably more complex. As previously mentioned, localization using GNSS is burdened always with an error that stems from the nature of satellite navigation. Nevertheless, this type of localization can be advantageously used, for example, as an additional aid to the dispatching control or in various information systems operating with positions of rolling stock.

Within the EGNOS system, SoL service (Safety of Life Service) has a positive impact on the localization of rolling stock, allowing more efficient use of satellite navigation thanks to greater precision and integrity service. At the moment, however, SoL service is applied rather in the air transport. Certain improvements in this area could be brought by the planned project Interfejs EGNOS-R for ground-based systems with high safety integrity. In principle, EGNOS-R is supposed to be an additional ground-based diagnostic unit to monitor the integrity of GNSS signals.

One of the basic methods of locating a rail vehicle is using a communication terminal with GNSS, with which selected rail vehicles are equipped [11]. These terminals periodically send defined message that include information on the position of the rail vehicle. Using UDP protocol, data messages are subsequently transmitted to the control centre. The data is primarily transmitted by the GSM-R transmission network, and if it is unavailable, the data is transmitted using the conventional GSM network.

4 Designing a Railway Infrastructure Network Model

Undirected graph, as defined graph theory, is a natural candidate for a railway network model. Based on an analysis of data provided by the company SŽDC-TUDC (consisting of service regulations, passports and codebooks), sets of algorithms were subsequently created, with which it was possible to generate a three-layer model of the rail network [12, 13]. Roughly speaking, the track can be divided into individual so called supertracks, which consist of definition supra-sections (TDNU), where each supra-section contains track definition sections (TUDU) with mileposts (in hectometres). Basic aspects of the description of the rail network are collectively shown in Fig. 1.

Mileposts (in hectometres) are shown in figure with the distance in kilometres and are graphically represented using gray points. TUDU is recorded using a six-digit code (163105, 163106, 16307, 173202) and are graphically represented using solid lines (red, black, orange, brown). Individual supra-sections (CLS 007, CLS008, REG023) are shown in light blue and supertracks (B421021 1 and B421021 1A) are shown in dashed lines. A place significant in terms of transportation (branch line) is symbolized by a green square.

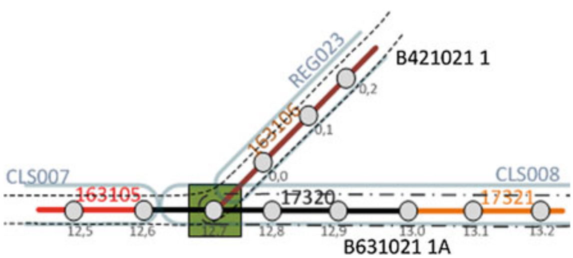


Fig. 1 Basic aspects of the description of the rail network

4.1 Creating Data-Micro Data Layer

Since all tracks in railway stations belong to a single definition section, the so called TUDU, the reduction of the basic set of mileposts (in hectometres) was performed so that only one continuous track was considered through a railway station under one TUDU.

The second modification of the basic table mileposts (in hectometres) then arose from the rail embranchment in the rail network. All junctions are always signalled by a flag value V (railroad switch) in the table of mileposts (in hectometres). The value K (track) then signals the point with no rail junction.

The data obtained through an algorithm (saved in the table NODES) represent information about vertices for the future data structure reflecting an undirected graph (i.e. railway network model). The edges of the graph considered need to be based on tables of super-tracks and NODES on the following algorithm, the result of which is saved in the table EDGES (Fig. 2).

The next phase of the algorithm was to create edges within each definition section and then define the edges between vertices that belong to the related definitional sections, which could be determined from the properties and description of super-tracks. The result of the second and third phase of the algorithm is shown in Fig. 3.

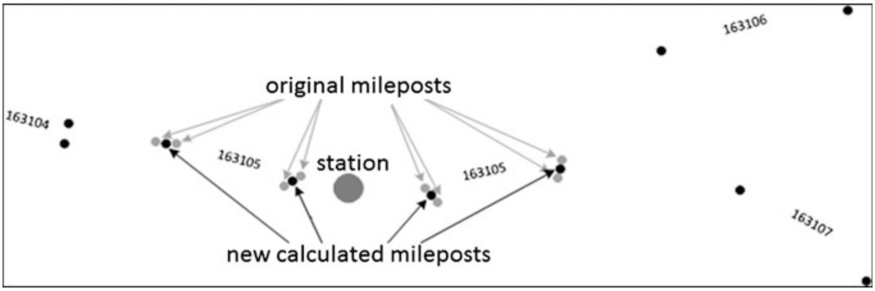


Fig. 2 Calculating new vertices

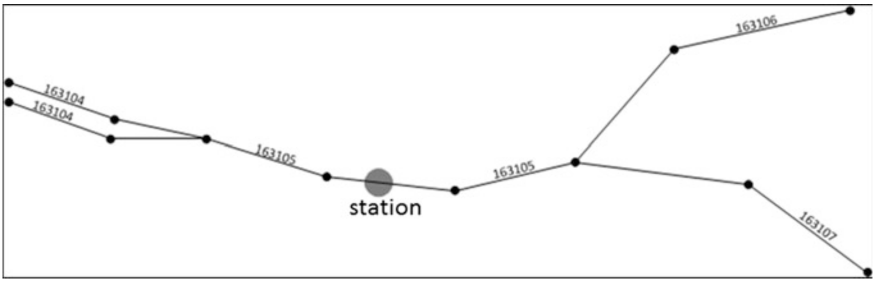


Fig. 3 Graph representation of the railway network model—Data-micro

4.2 Creating a Data Layer Data-Macro

To monitor the rolling stock within higher units of the railway network, it was necessary to create another railway network model with a higher level of abstraction. At this level, we consider the so-called super-edges to be higher units, i.e. a section of the track where there are no rail junctions. This data layer is obtained through a simple algorithm that selects those elements from the table of vertices of the data layer Micro-data that are incident with more than two edges. This gave us a table of vertices and edges reflecting the railway network with a higher level of abstraction. The result of this operation is shown in Fig. 4.

4.3 Creating a Data Layer Data-Mezo

In the first data layer (Data-micro), the length of the edge is approximately 100 m, whereas in the second data layer with a higher level of abstraction (Data-macro, the length of the edge (i.e. the so called super-edge) is several tens of kilometres, regardless whether there is a railway station at the edge. The third data interlayer Data-mezo tries to remove this lack by decomposing the super-edge.

The two original layers used an identical data base for vertices, table of mileposts (in hectometres) while other types of vertices representing individual stations entered the new layer Data-mezo. Creating a new layer Data-mezo consists of three parts:

(a) Preparing input data and creating the layer
(b) Filling the layer with data
(c) Generating additional data.

The proposed algorithm [13] was implemented directly on the database level using PL/SQL language. However, the algorithm had to be adjusted and generalized several times since there are various nonstandard conditions in the data, such as jumps in the mileposts (nonlinear growth of the kilometre succession between the

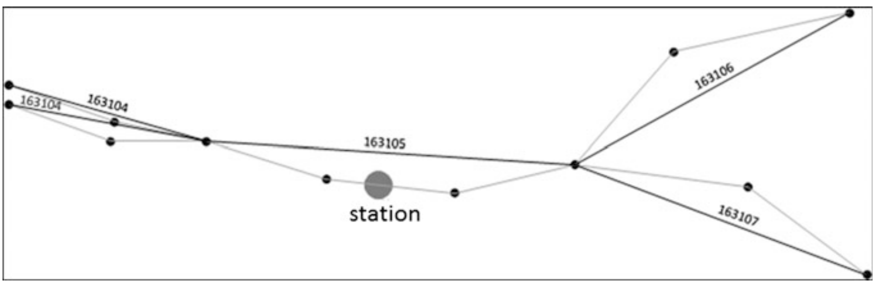


Fig. 4 Graph representation of the railway network model—Data-Macro

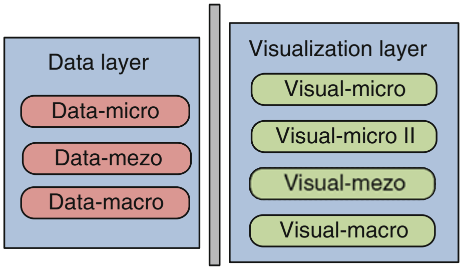


Fig. 5 Overall concept of the railway infrastructure network model

mileposts) or change of an increasing kilometre sequence into a decreasing one and vice versa. Figure 4 shows the overall concept of the data and visualization layers of the railway network model (Fig. 5).

5 Visualization

For the purposes of visualization of proposed railway infrastructure model visualization tool MapViewer developed in the JAVA language [14] was employed. MapViewer represents J2EE service for displaying maps using the spatial data (e.g. using object data type SDO_GEOMETRY) managed by ORACLE Spatial. This technology enables to compose extensive map layers with various levels of details of information displayed [11].

It is possible to use various operators and functions above spatial objects of ORACLE database with option component Spatial [15]. One of them is SDO_NN (Near Neighbour) that enables to specify the nearest geometry (so called neighbour), in our case the nearest vertex or rather edge of non-oriented graph. Provided that we have GPS information about actual position of a rolling stock it is possible to use this operator for specifying the nearest vertex/edge and subsequently for example to visualize the location on a map composed by means of MapViewer technology.

6 Simulation of Trafic

The selected train vehicles are equipped with communication terminals, which broadcast data including current GPS coordinates of the rolling stock. When the vehicle is in motion, this communication terminal sends information about its position every 30 s.

Designed simulation model contains the core of discrete simulation utilizing standard calendar of process messages, which were, during the simulation, executed based on their time stamp. This model was consequently implemented into the software demonstrator InfraRail which is intended for the additional support of dispatching control.

Simulation of traffic of rolling stock can be divided into two parts. Simulation based on:

- real historical data (emulation of operation),
- generated data.

7 SQL Optimization Queries

If we use the database and SQL language there should be focused on the optimization. In this case we use SDO_NN operator so that we focus on optimization of SQL queries which are called directly from JAVA application. We can use the following:

- Statement Object,
- Prepared Statement Object,
- Calling a functions that are stored in the database.

The Statement Object is base of universal query but its main disadvantage is periodical assembling of query. Use of Prepared Statement Object and PL/SQL function brings the use of so called bind variables. In this case the database gets still the same SQL statement (with still same hashcode) but with different values. This means that in the database machine is used always the same execution plan.

7.1 Design of the Optimization Using the Dynamic Views

Despite the advanced optimization techniques of database ORACLE, the question arises as to whether is necessary to perform queries to the whole bases of vertices respectively edges. Whereas we are inquiring on the position of rolling stock (RS) within railway network which from the logical view, are not able to change the position by more than tens of meters, then we can reduce the base of vertices/edges by using the dynamic view. The main ideas are based on the following assumptions:

- for each new RS is created an initial dynamic view
- queries are periodically performed into the current view
- if the RS is not already in the current view then the dynamic view is recalculated regarding to the azimuth of a moving train.

This means SQL query is performed into the relevantly reduced base of vertices/edges. This leads to optimization and the overall time saving of query.

8 Testing

Testing of proposed optimization of selected search operations should show whether the using of dynamic view are correct. For the purpose of testing there has always been generated thousand points with GPS coordinates in the selected area, wherein the base of data consists of about one hundred thousand entries. For the time measuring of each query there were used information from database system tables ORACLE *v$sql* and values from column *ELAPSED_TIME*. For secondary comparison there was used time measured directly in application that reflect the time overhead of communication between application and database machine. For each test the time was measured as following:

- average time,
- max/min time,
- median/modus/deviation,
- total time in database/application.

Firstly there were tested all query techniques without dynamic views. As was expected the Statement Object showed approximately 3.5× worse time of processing then Prepared Statement Object and query in the function.

The second area tests used the dynamic views for Prepared Statement Object and query in the function for different sizes variants of dynamic views. The test results for area 40 × 20 km are shown in Table 1. Time is measured in seconds.

Although the total times in database are quite similar the different situation is in the total time in JAVA application. The ORACLE function against Prepared Statement has significantly worse time. The reason is that the ORACLE function requires additional time to the calling and takeover result. The average times for different approaches is shown in Fig. 6.

The optimization of SQL queries based on dynamic views was subsequently integrated into a demonstration application InfraRAIL for localization of rolling stock. The running application using dynamic views within a railway network model is illustrated in Fig. 7.

Table 1 Times of two different approaches for view 40 × 20 km

Method	Oracle function 40 × 20	Prepared statement 40 × 20
Average	0.007475	0.007537
Min	0.003242	0.002815
Max	1.354342	1.157581
Median	0.005647	0.005347
Modus	0.005491	0.003494
Diversion	0.0427506	0.036910
Total time DB	7.4746270	7.537129
Total time app	54.835112	39.70152

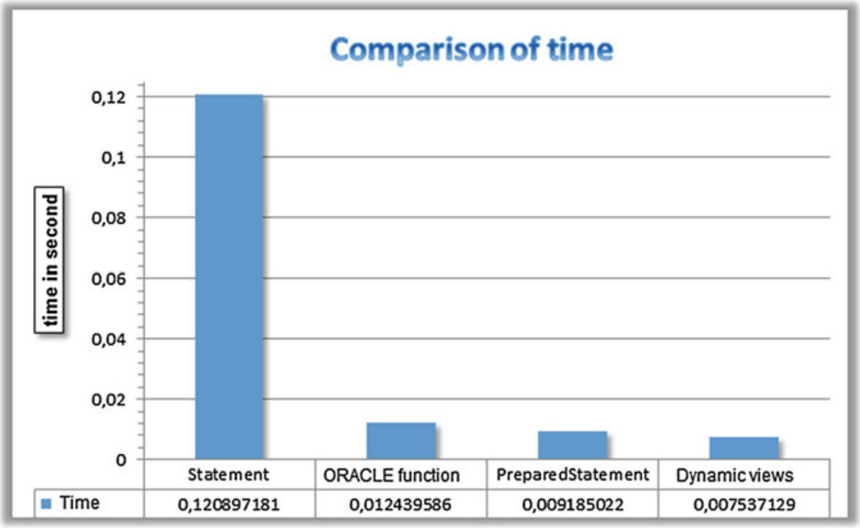


Fig. 6 All tested approaches

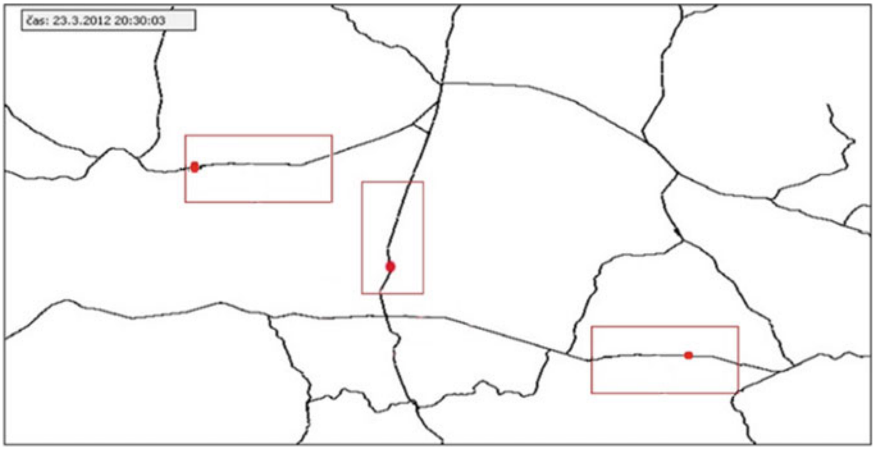


Fig. 7 Running application

Within the final evaluation of results was found that for hundreds active trains on the railway network when the each train sends information about position each ten seconds is total time savings up to a few seconds.

Acknowledgments This work has been supported by the project "SGSFEI_2015008 Models of infrastructure and operation of land transport systems" (financed by the University of Pardubice).

References

1. GHAZEL, M.: Formalizing a subset of ERTMS/ETCS specifications for verification purposes. In: Transportation Research Part C: Emerging verification purposes. In: Transportation Research Part C: Emerging Technologies. Elsevier Limited, pp. 60–75 (2014). ISSN: 0968-090X
2. Chudaček, V., Lochman, L.: Vlakový zabezpečovací systém ERTMS/ETCS. In: Vědeckotechnicky sborník ČD, č. 5/1998
3. Lieskovský, A., Myslivec, I.: ETCS a AVV—poprvé společně. In: EuroŽel, Žilina (2010)
4. Lieskovský, A. Automatické vedení vlaků Českých drah. In: Automatizace. Praha: Automatizace, roč. 10 (2004). ISSN: 0005-125x
5. Dorazil, P.: Základní vlastnosti kolejových obvodů bez izolovaných styků. Pardubice. Bakalářská práce. Univerzita Pardubice. Vedoucí práce Milan Kunhart (2008)
6. Senesi, F.: Satellite application for train control systems, In: The Test Site in Sardinia, Journal of Rail Transport Planning and Managemt. Elsevier BV, pp. 73–78, (2012). ISSN: 2210-9706
7. O'Connor, M.L.: Carrier-phase differential GPS for automatic control of land vehicles. In: Dissertation Abstracts International, Vol. 59–06, Section: B, p. 2876; 158 pp. Stanford University, USA (1997). ISBN: 9780591909272
8. Stadlmann, B.: Automation of operational train control on regional branch lines by a basic train control. In: Proceedings of IEEE Intelligent Transportation Systems Conference, Toronto, Canada, (September 2006), pp. 17–20
9. Becker, U., Poliak, J.: DemoOrt repositions trains with satellite. In: EURAILmag Business and Technology (18. vyd), pp. 216–219. Blue Line & bro, France (2008)
10. Kolář, Petr.: Řízení provozu na vedlejších železničních tratích. Seminář ZČU Plzeň—Fakulta elektrotechnická (2014). Accessed 12 Dec 2014
11. Fikejz, J., Kavička, A.: Modelling and simulation of train positioning within the railway network. In: KLUMPP, Matthias. In: The European Simulation and Modelling Conference, pp. 366–376. EUROSIS—ETI, Ostende (2012). ISBN: 978-9077381-73-1
12. Fikejz, J., Kavička, A.: Utilisation of computer simulation for testing additional support for dispatching rail traffic. In: European Simulation and Modelling Conference, pp. 225–231. EUROSIS—ETI, Ostende (2011). ISBN: 978-90-77381-66-3
13. Fikejz, J., Řezanina, E.: Utilization of computer simulation for detection non-standard situations within the new data layer of railway network model. In: The 26th European Modeling and Simulation Symposium, Bordeaux, pp. 371–377 (2014) (ISBN: 978-88-97999-32-4)
14. Murray et al.: ORACLE® Fusion Middleware: User's Guide for ORACLE MapViewer 11g Release 1 (11.1.1) (2010). Accessed 07 Sept 2012
15. Kothuri, R., Godfrind, A., Beinat, E.: Pro Oracle Spatial for Oracle Database 11g. Distributed to the book trade worldwide by Springer, New York, Handbook 2007, vol. xxxiv, 787 pp. Mill Valley, University Science (1989). ISBN: 15-905-9899-7

Computer Support in Building-up a Consistent Preference Matrix

Martin Gavalec, Hana Tomášková and Richard Cimler

Abstract A method of a computer support for an expert who is creating a matrix of pairwise preferences in a decision making problem is described. The final preference matrix should be antisymmetric and consistent. While it is easy to control the antisymmetricity of the entries, the consistency of the inserted values, on the other hand, is far from obvious. The suggested computer support is based on the idea that the consistent hull of the previously inserted entries is maintained in the computer, and in every step the human expert gets the information whether the intended preference value can be chosen independently of the previous inputs. If the opposite case, then computer recommends the unique consistent value. Still, the expert can decide differently according to his/her own opinion. Then the optimal consistent approximation of all previous entries including the last input is computed and maintained for further steps. The computer support uses the optimal approximation algorithm due to the authors. The new method is illustrated by examples.

Keywords Consistency · Preference matrix · AHP · Optimization · Algorithms · Pairwise comparison · Computer support

---

M. Gavalec · H. Tomášková · R. Cimler (✉)
Faculty of Informatics and Management, University of Hradec Králové,
Rokitanského 62, 50003 Hradec Kralove, Czech Republic
e-mail: richard.cimler@uhk.cz

M. Gavalec
e-mail: martin.gavalec@uhk.cz

H. Tomášková
e-mail: hana.tomaskova@uhk.cz

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_80

1 Introduction

The basic method used in solving multi-criteria decision problems is the Analytic Hierarchy Process (AHP) method suggested by [12]. The AHP approach has been studied and modified by many authors, see e.g. [3, 6, 8, 10, 13, 14]. AHP uses a specific model of the problem, which helps to find the best decision from the set of possible alternatives. The method starts with defining the structure of the problem, then continues with quantifying the relative preferences, computing the priorities and with computing the final evaluation of all considered alternatives. Subjective pairwise comparison values are assigned to individual components of the model, showing their relative importance.

One of the most important questions in AHP decision making is how to find the appropriate preference matrix for a set of alternatives. The subjective preferences given by human experts are often inconsistent and do not reflect all deeper relations between the alternatives, see [1, 2, 5, 7, 9, 10]. The standard approach to the problem of finding the relative importance vector out of an incosistent preference matrix uses the so-called *inconsistency index*, and the preference matrix is applied if the index value does not exceed 0.1, see [11]. In fact, this approach empirically decides that the expert's preferences are 'not too inconsistent' for the practical purposes.

A more exact method of treating the inconsistency of a given relative preference matrix has been suggested in [4]. The additive form of expressing the relative importance for every pair of alternatives is used and the consistency of a matrix is characterized by the conditions that the diagonal elements are equal to zero, and the columns of the matrix are parallel, in the sense that the difference of every pair of columns is a constant vector. Analogous characterization of the consistency is also possible in the multiplicative form, taking ones instead of zeros and the quotients instead of differences. However, both approaches are equivalent and can standardly be transformed to each other. The advantage of the additive form for optimisation purposes, is that linear programming methods can directly be applied.

The problem of finding the optimal consistent approximation, i.e. a matrix with the minimal distance from a given preference matrix has been formulated in [4] as a linear optimisation problem. The algorithm described there can be characterized as 'postoptimisation'. That is, first the preference matrix is created by an expert, and only afterwards the best consistent approximation to the expert's preferences are computed.

During the expert's work, the computer support is provided at every step, in the dialogue form. Not only the antisymmetricity of the entries is automatically controlled, but, which is substantial, the consistent hull of the previously inserted entries is continuously maintained in the computer. In every step the expert gets the information whether the intended preference value can be chosen independently of the previous inputs. If it is not the case, then computer recommends the consistent value, which is, in fact, uniquely determined. The decision is left on the expert who can choose the weight of the recommendation, and put in a different preference value according to his/her own opinion. The computer then emerges the recommended and inserted value with respect to the chosen weight, and computes the optimal consistent approximation of all previous entries including the emerged input. The result is then maintained for further steps.

The layout of the paper is as follows. The basic facts on the consistency of preference matrices are presented in Sect. 2. The consistency notion for partial preference matrices is studied in Sect. 3. Further section contains description of the algorithm OCA for computing the optimal consistent approximation of a given preference matrix.

2 Consistency of Preference Matrices

A given set of alternatives A_1, A_2,…, A_n is typically considered in multicriterial decision making. Till the end of the paper N will denote the set $\{1, 2, \ldots, n\}$ and $\mathcal{R}$ the set of all real numbers. The quantified judgments on pairs A_i, A_j are represented by an $n \times n$ matrix $A = (a_{ij})$, $i, j \in N$. Every entry a_{ij} is real number, which is interpreted as an evaluation of the relative preference of A_i with respect to A_j, in the additive sense. That is, A_i is considered to be by a_{ij} better than A_j. Consequently, A_j is by $-a_{ij}$ better than A_j and $a_{ii} = 0$ for every $i \in N$. Then A is called *additive preference matrix* (for short: preference matrix) of the alternatives A_1, A_2,…, A_n. The basic properties of preference matrices are defined as follows

- A is *antisymmetric* if $a_{ij} = -a_{ji}$ for every $i, j \in N$,
- A is *consistent* if $a_{ij} + a_{jk} = a_{ik}$ for every $i, j, k \in N$.

More symmetric equivalent definitions are

- A is *antisymmetric* if $a_{ij} + a_{ji} = 0$ for every $i, j \in N$,
- A is *consistent* if $a_i + a_{jk} + a_{ki} = 0$ for every $i, j, k \in N$.

In particular, for $i = j$,

- if A is *antisymmetric*, then $a_{ii} + a_{ii} = 0$, i.e. $a_{ii} = 0$ for every $i \in N$,
- if A is *consistent*, then $a_{ii} + a_{ik} + a_{ki} = 0$, i.e. $a_{ik} + a_{ki} = 0$ for every $i, k \in N$.

Hence, if A is consistent, then A is antisymmetric, but the converse implication does not hold. E.g., $A = \begin{pmatrix} 0 & 1 & 2 \\ -1 & 0 & 3 \\ -2 & -3 & 0 \end{pmatrix}$ is antisymmetric, but it is not consistent, because $a_{12} + a_{23} = 1 + 3 = 4 \neq a_{13}$.

The entries in a preference matrix are usually submitted by an expert in the given field. The antisymmetricity can easily be verified by formulas $a_{ii} = 0$, and $a_{ij} = -a_{ji}$. On the other hand, the consistency is difficult to be seen directly from the data.

Consistent matrices are characterized in Theorem 1, see [4]. Vectors $x, y \in \mathcal{R}(n)$ are called *parallel*, if there is $c \in \mathcal{R}$ such that $y_i = c + x_i$ for every $i \in N$.

Theorem 1 *If all diagonal entries of a matrix A are zero, then the following assertions are equivalent* [4]

(i) *A is consistent,*
(ii) *every two columns of A are parallel.*

3 Consistency in Partial Preference Matrices

Denote $\overline{\mathcal{R}} = \mathcal{R} \cup \{-\infty\}$. By *partial preference matrix* we understand an $n \times n$ matrix A with entries $a_{ij} \in \overline{\mathcal{R}}$, $i, j \in N$. The value $-\infty$ is used to indicate the missing or unknown entries. $D(A) = \{(i \cdot j) \in N \times N | a_{ij} > -\infty\}$ denotes the domain where the entries of A are well defined. The antisymmetricity and consistency conditions are only applied to entries over $D(A)$.

By definition, every preference matrix is partial preference matrix with domain $N \times N$. Such a matrix is also called *full preference matrix*.

Furthermore, we say that partial preference matrix $A \in \overline{\mathcal{R}}(n, n)$ is *consistently closed*, if for every finite sequence of indices $i_1, i_2, \ldots, i_k \in N$ of length $k > 2$

$$(i_1, i_2), (i_2, i_3), \ldots, \left(i_{k-1,k}, i_k\right) \in D(A) \Rightarrow (i_1, i_k) \in D(A). \tag{1}$$

The consistent hull of $A \in \overline{\mathcal{R}}(n, n)$ is a partial matrix $A^\star \in \overline{\mathcal{R}}(n, n)$ fulfilling the following conditions

(i) $D(A) \subseteq D(A^\star)$,
(ii) $A^\star$ is consistently closed,
(iii) $A^\star$ has the smallest domain of all matrices fulfilling (i) and (ii).

It is easy to see that the following theorem holds true.

Theorem 2 *A consistent matrix* $A \in \overline{\mathcal{R}}(n, n)$ *is consistently closed if and only if there exist disjoint subsets* $C_1, C_2, \ldots, C_k \subseteq N$, $k \geq 1$ *such that*

(i) $N = \bigcup\{C_s | s = 1, 2, \ldots, k\}$,
(ii) $D(A) = \bigcup\{C_s \times C_s | s = 1, 2, \ldots, k\}$,
Moreover, in the positive case
(iii) *the diagonal submatrix of A restricted to* $C_s \times C_s$ *is a consistent full preference matrix, for every* $s \in \{1, 2, \ldots, k\}$,
(iv) *if for fixed* $s \in \{1, 2, \ldots, k\}$, the *diagonal submatrix of A on* $C_s \times C_s$ *is substituted by any consistent full preference matrix, then the whole matrix on* $N \times N$ *remains consistent.*

Remark 1 Theorem 2 presents a detailed characterization of the consitent hull for any partial preference matrix. An important special case is described below.

Theorem 3 *Suppose that A is consistently closed and* $C_1, C_2, \ldots, C_k \subseteq N$, $k \geq 1$ *fulfill the conditions (i) and (ii) in* Theorem 2. *If* $B \in \overline{\mathcal{R}}(n,n)$ *is an antisymmetric extension of A with* $(i,j) \in N \times N \backslash D(A)$, $D(B) = D(A) \cup \{(i,j),(j,i)\}$ and $b_{ij} = -b_{ji}$, *then*

- (i) *there are* C_s, C_t *such that* $s \neq t$, $i \in C_s$, $j \in C_t$,
- (ii) *B is consistent*
- (iii) *the consistent hull* $B^\star$ *is characterized by disjoint subsets* $C_1^\star, C_2^\star, \ldots, C_{k-1}^\star \subseteq N$ *similarly as in* Theorem 2, *where* $C_1^\star = C_s \cup C_t$ *and* $\{C_2^\star, \ldots, C_{k-1}^\star\} = \{C_1, C_2, \ldots, C_k\} \backslash \{C_s, C_t\}$,
- (iv) *the entries of the consistent hull* $B^\star$ *are* $b_{kl}^\star = b_{ki} + b_{ij} + b_{jl}$ *and* $b_{lk}^\star = b_{lj} + b_{ji} + b_{ik}$ *for* $k \in C_s$, $l \in C_t$.

Proof The assertions follow from the definitions and from Theorem 2.

4 Optimisation Algorithm OCA

Consider the optimisation problem: given a (possibly inconsistent) matrix A, find a consistent matrix $\tilde{A}$ which will be as close to A as possible. Such a matrix $\tilde{A}$ is called *the optimal consistent approximation* of A.

A consistent preference matrices is closely related with the vector showing the weights (importances) of the alternatives. In the additive notation, vector $w \in \mathcal{R}(n)$ is called the *balanced vector* if $\sum_{i \in N} w_i = 0$. When alternatives $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n$ are considered, then w_i is interpreted as the weight of $\mathcal{A}_i$ for every $i \in N$. The differences of weights are the entries of the corresponding matrix of relative preferences $A(w)$ with $a_{ij}(w) = w_i - w_j$ for $i, j \in N$. We say that $A(w)$ *is induced by* w.

Theorem 4

- (i) *For any* $w \in \mathcal{R}(n)$, *the induced matrix* $A(w)$ *is consistent.*
- (ii) *If A is a consistent matrix, then there is unique balanced vector w such that* $A = A(w)$.

Proof

- (i) If $w \in \mathcal{R}(n)$, then $a_{ij}(w) + a_{jk}(w) + a_{ki}(w) = (w_i - w_j) + (w_j - w_k) + (w_k - w_i) = 0$ for every $i, j, k \in N$. That is, $A(w)$ is consistent.
- (ii) Assume that $A \in \mathcal{R}(n,n)$ is consistent. Define $w \in \mathcal{R}(n)$ by putting

$$w_i = \frac{1}{n}\sum_{k \in N} a_{ik} \quad \text{for every } i \in N. \tag{2}$$

for every $i \in N$. Using the consistency assumption in computation

$$\sum_{i \in N} w_i = \frac{1}{n}\sum_{i \in N}\sum_{k \in N} a_{ik} = \frac{1}{n}\left(\sum_{i \in N} a_{ii} + \sum_{\{i,k\} \subseteq N, i \neq k} (a_{ik} + a_{ki})\right)$$
$$= \frac{1}{n}(0+0) = 0,$$

we verify that w is a balanced vector. By further computation we get

$$a_{ij}(w) = w_i - w_j = \frac{1}{n}\sum_{k \in N} a_{ik} - \frac{1}{n}\sum_{k \in N} a_{jk} = \frac{1}{n}\sum_{k \in N}\left(a_{ik} + a_{kj}\right)$$
$$= \frac{1}{n}\sum_{k \in N} a_{ij} = a_{ij} \quad \text{for every } i, j \in N,$$

that is, $A(w) = A$. The uniqueness of w then follows from 2.

It has been shown in [4] that the closest consistent matrix to a given preference matrix A can be found by looking for a suitable balanced vector w such that the distance between A and $A(w)$ will be as small as possible. The distance d is measured by the well-known metric formula

$$d(A, A(w)) = \max_{i,j \in N} |a_{ij} - a_{ij}(w)|. \tag{3}$$

Theorem 5 *If A is antisymmetric, then* [4]

$$d(A, A(w)) = \max_{i,j \in N}\left(a_{ij} - w_i + w_j\right). \tag{4}$$

As a consequence of Theorem 5, the best consistent approximation of a given preference matrix given by an expert can be formulated as the following minimisation problem.

OCA (optimal consistent approximation)
Input: antisymmetric matrix $A \in \mathcal{R}(n, n)$
Variables: $m \in \mathcal{R}$, $w = (w_1, w_2, \ldots, w_n) \in \mathcal{R}(n)$

$$\begin{aligned} &\text{Minimise} \quad z = m \to \min \\ &\text{subject to} \\ &\qquad \sum_{i \in N} w_i = 0, \\ &\qquad m + w_i - w_j \geq a_{ij} \quad \text{for } i, j \in N. \end{aligned} \tag{5}$$

Theorem 6 *If A is antisymmetric and* (m, w) *is an optimal solution to OCA, then* [4]

(i) *A is consistent if and only if* $m = 0$,
(ii) *m is the minimal possible distance of a consistent matrix to A*,
(iii) *w is a balanced weight vector and the induced matrix* $A(w)$ *is the best consistent approximation of A with* $d(A, A(w)) = m$.

5 Computer Support Algorithm

The aim of the computer program is to support the person creating the matrix of relative preferences. The support is provided in the dialogue form and the user can choose the level of the support. The basic functionality only fills in the opposite values to the cells on symmetric positions automatically. When a cell of the matrix is filled in (by the user or by the program) the symmetric cell is filled in with the opposite value immediately. The advanced support level helps to compute all values in the clique based on the last filled-in cell. The full support provides the consistency supervision of the matrix in the case when the user changes one of the previously filled-in cells.

If the matrix is consistent and the value in one of the non-empty cells has been changed, then the matrix becomes inconsistent. The newly input value of the cell will be merged with the original value recommended by the computer, and the user is asked to choose the (percentage) weight of his input. From the original value and the value filled-in by the user, the weighted arithmetic mean is then computed. The mean value is put into the cell and its opposite appears in the cell on the symmetric position. The matrix is inconsistent, therefore the OCA—Optimal consistency approximation is used to make the matrix consistent (Fig. 1).

6 Example

The values automatically computed by the program are printed in bold.

1. Matrix 3×3 is generated, cliques (1); (2); (3)

$$\begin{pmatrix} 0 & & \\ & 0 & \\ & & 0 \end{pmatrix}$$

2. The user fills in $a_{(1,2)}$—program automatically fills $a_{(2,1)}$, cliques (1,2); (3)

$$\begin{pmatrix} 0 & 3 & \\ \mathbf{-3} & 0 & \\ & & 0 \end{pmatrix}$$

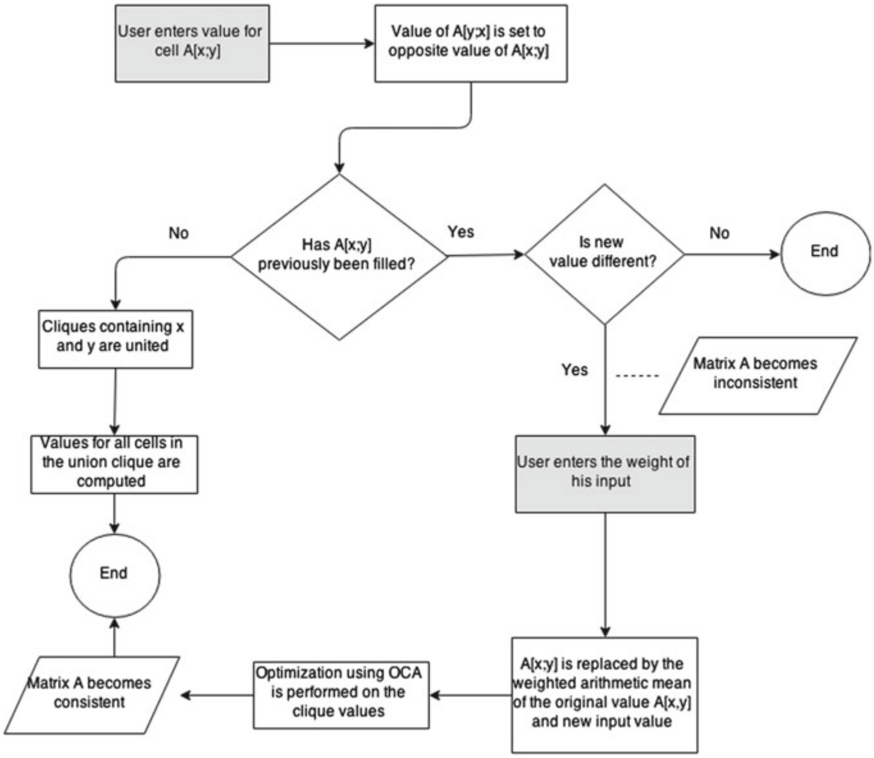


Fig. 1 Support algorithm flow chart

3. The user fills $a_{(1,3)}$—program automatically fills $a_{(3,1)}$ and the value for $a_{(2,3)}$ is computed based on the other values in clique then $a_{(3,2)}$ is filled automatically, cliques (1,2,3)

$$\begin{pmatrix} 0 & 3 & 10 \\ -3 & 0 & \mathbf{7} \\ -\mathbf{10} & -\mathbf{7} & 0 \end{pmatrix}$$

4. The user changes value $a_{(2,3)}$

$$\begin{pmatrix} 0 & 3 & 10 \\ -3 & 0 & \mathbf{8} \\ -10 & -\mathbf{8} & 0 \end{pmatrix}$$

5. The user is asked for the weight of his change

   (a) If the weight is 0 % then result is original matrix

$$\begin{pmatrix} 0 & 3 & 10 \\ -3 & 0 & 7 \\ -10 & -7 & 0 \end{pmatrix}$$

   (b) If the weight is 25 % then result is:

$$\begin{pmatrix} 0 & 2.92 & 10.08 \\ -2.92 & 0 & 7.17 \\ -10.08 & -7.17 & 0 \end{pmatrix}$$

   (c) If the weight is 50 % then result is:

$$\begin{pmatrix} 0 & 2.83 & 10.17 \\ -2.83 & 0 & 7.33 \\ -10.17 & -7.33 & 0 \end{pmatrix}$$

   (d) If weight is 75 % then result is:

$$\begin{pmatrix} 0 & 2.75 & 10.25 \\ -2.75 & 0 & 7.5 \\ -10.25 & -7.5 & 0 \end{pmatrix}$$

   (e) If weight is 100 % then result is:

$$\begin{pmatrix} 0 & 2.67 & 10.33 \\ -2.67 & 0 & 7.67 \\ -10.33 & -7.67 & 0 \end{pmatrix}$$

Even if the user chooses the weight 100 %, his input may be changed by the subsequent optimization.

7 Conclusions

The problem of finding a consistent preference matrix for a set of alternatives was previously solved empirically by so-called inconsistency index expressing the belief that the expert preferences are 'not too inconsistent', or by 'postoptimisation' approach, where the preference matrix is first created by an expert, and only then the best consistent approximation is computed.

Computer support in the dialogue form is offered to the expert during the creation process of the preference matrix. The antisymmetricity of the entries is automatically controlled, and the consistent hull of all the previously inserted entries is maintained in the computer. In every step the expert gets the information whether the intended preference value can be chosen independently of the previous inputs, or the previous entries will be influenced. In such case the computer recommends the unique existing consistent value. Still, the decision is left on the expert who can accept the recommendation, or put in a different preference value according to his/her own opinion. The computer then emerges the recommended and inserted value with a chosen weight, and computes the optimal consistent approximation of all previous entries including the emerged input. This actualization is the starting point for further steps.

The advantage of the presented approach is that an immediate feed-back is provided to the expert's inputs. The method allows numerous variations and can be adjusted to fit the expert's personal convenience, the type of the data, or the character of the decision problem. The work of the suggested algorithm is illustrated by numerical examples.

Acknowledgement The support of SPEV UHK FIM is gratefully acknowledged.

References

1. Aguaron, J., Escobar, M.T., Moreno-Jimenez, J.M.: Consistency stability intervals for a judgement in AHP decision support systems. Eur. J. Oper. Res. **145**(2), 382–393 (2003)
2. Carlsson, C., Fuller, R.: Fuzzy multiple criteria decision making: recent developments. Fuzzy Sets Syst. **78**(2), 139–153 (1996)
3. Dytczak, M., Szklennik, N.: Principles and applications of AHP/ANP-based multiple MCDA methods approach. In: Proceedings of ISAHP (2011)
4. Gavalec, M., Tomášková, H.: Optimal consistent approximation of a preference matrix in decision making. Int. J. Math. Oper. Res.
5. Gavalec, M., Ramk, J., Zimmermann, K.: Decision Making and Optimization: Special Matrices and Their Applications in Economics and Management, vol. 677. Springer, Berlin (2014)
6. Jalao, E.R., Wu, T., Shunk, D.: A stochastic AHP decision making methodology for imprecise preferences. Inf. Sci. **270**, 192–203 (2014)
7. Leung, L., Cao, D.: On consistency and ranking of alternatives in fuzzy AHP. Eur. J. Oper. Res. **124**(1), 102–113 (2000)
8. Mls, K., Gavalec, M.: Multi-criteria models in autonomous decision making systems. In: Proceedings of the 10th International Symposium on the Analytic Hierarchy/Network Process, pp. 1–8 (2009)
9. Ramik, J., Korviny, P.: Inconsistency of pair-wise comparison matrix with fuzzy elements based on geometric mean. Fuzzy Sets Syst. **161**(11), 1604–1613 (2010)
10. Ramik, J., Perzina, R.: A method for solving fuzzy multicriteria decision problems with dependent criteria. Fuzzy Optim. Decis. Making **9**(2), 123–141 (2010)
11. Saaty, T.L.: Decision-making with the AHP: why is the principal eigenvector necessary. Eur. J. Oper. Res. **145**(1), 85–91 (2003)
12. Satty, T.L., et al.: The analytic hierarchy process (1980)
13. Vargas, L.G.: Reciprocal matrices with random coefficients. Math. Model. **3**(1), 69–81 (1982)
14. Zahedi, F.: The analytic hierarchy process: a survey of the method and its applications. Interfaces **16**(4), 96–108 (1986)

Effects of Phase Offset Error for Asymmetric Communication Link in Heterogeneous Mobile Network

Mohammad Kamrul Hasan, Ahmad Fadzil Ismail, Wahidah Hashim, Shayla Islam and M.H. Khairolanuar

Abstract Due to augmentation of data traffic, imperfect coverage in indoor surroundings small base stations for instance femtocells (HeNodeBs) are deliberated and executed to extend coverage also capacity. The arbitrary placement of HeNodeBs are highlights the significance of interference mitigation to apprehend the ultimate. Timing and synchronization are identically imperative for control persistence for HeNodeBs in 4G LTE classifications. In order to conserve HeNodeBs for the control tenacity, IEEE 1588 be presented. Though, attributable to backhaul sufficiency for extensive time synchronization disputes to the neighbor HeNodeBs which leads to misalignment of the offsets. A tangible exploration has completed in order to substantiate the challenges of current systems. The model result indicates that the improved enhanced IEEE 1588 methodology is better in expressions of offsets error.

Keywords IEEE 1588 · Femtocell · LTE-Advanced

1 Introduction

At the present time, Heterogeneous Mobile Network (HMN) achieves the capacity as well as coverage in indoor though deploying HeNodeB small cells. These HeNodeBs are plug-and-play states Customer premises Equipment's (CPEs) which

M.K. Hasan (✉) · A.F. Ismail · S. Islam · M.H. Khairolanuar
Department of Electrical and Computer Engineering, International Islamic University, 50728 Gombak, Malaysia
e-mail: hasankamrul@ieee.org

A.F. Ismail
e-mail: af_ismail@iium.edu.my

W. Hashim
College of Information Technology, Universiti Tenaga Nasional, 47000 Kajang, Malaysia
e-mail: Wahidah@uniten.edu.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_81

are associated over IP backhaul to macro-eNodeB. Advanced Mobile communication practices packet switching technique, whereas transfer of time and frequency precise functioning is essential. For time synchronization of HeNodeBs in HMN, there has few modules to be function which are point wise as below.

(a) Transfer control
(b) Time synchronization
(c) Fluctuation measurement
(d) Clock.

The time synchronization module uses a received time synchronization packet to synchronize the clock. The fluctuation measurement module determines an accuracy of a time contained in the received time synchronization packet based on a result of comparison between the time contained in the received time synchronization packet and a time of the clock.

From main control server to HeNodeB, packets are carrying the control and necessary information. In HMN, HeNodeBs are depends on internet backhaul so the propagation delay differences raised unexpectedly due to varying traffic congestion. Timing and synchronization are significant mechanism of competent HMN. As a result the entire indoor network become asynchronous and communication link become asymmetric. These cause the dramatic effect of the efficiency and capacity of the network, and more strongly disrupt the calls or data sessions. Most importantly owing to these link asymmetry HeNodeBs are required to arrange the received signal to abate interference for the multiple accesses while ensuring an offset of the carrier within acceptable limits. As a result asynchronous HeNodeBs delay on its transmission time by a period effects the total degradation of service to the UEHeNodeBs. Moreover, as the improvement of technology from 3G HeNodeBs to LTE-Advanced (LTE-A), HeNodeBs are bounded to fulfill the timing and synchronization standards. This is also required to ensure unbroken handovers in synchronized networks, especially in Time-Division Duplex (TDD) and LTE-A systems. The standards and requirements for the frequency synchronization are also critical to sustain the correct order of frequency in macro-eNBs and HeNBs. Furthermore, time synchronization in multi-hop heterogeneous system likewise needs to stay away from impedance among the range, particularly for Frequency-Division Duplex (FDD) [1–2].

The heterogeneous system has more prohibitive timing requests. Despite the fact that LTE FDD frameworks require just the rate of repeat synchronization, LTE TDD systems give a further step necessity of 1.5 micro seconds (μs). Intended for more updated LTE-A systems, the requirement is more particular. In Table 1, the timing necessities of dissimilar access technologies [3, 4] are concise.

All through ordinary 3G systems for pico-eNodeB and macro-eNB, synchronization was conveyed utilizing TDM line timing or Global Navigation Satellite System (GNSS). Then again, for 4G systems, for example, HeNB was synchronized utilizing IP based third party internet backhaul including IEEE 1588 [5], and IEEE 1588 PTP [6], improved IEEE 1588 Block Burst Protocol (IEEE 1588 BBP) [7]. Nonetheless, by and large, where there is absence of bandwidth otherwise

Table 1 Classification of HMN timing prerequisite

Methods/systems	Phase (μs)
Frequency division duplexing	–
Time division duplexing	1.5
Enhanced inter-cell interference coordination	1.32
Long-term evolution-advanced	0.5

constrained connectivity, the 1588 master-slave method devises limitations since, it synchronizes the hubs utilizing pair wise informing framework through reliable medium, for example, switches, gateways and wireless networks. Subsequently, the delay builds up and system turns asymmetric. At the point when pulse edges are non-distinguishable, this recurrence blunder becomes a link asymmetry. The asymmetry causes offset problems, which are shown in Fig. 1a and b. In any case, a synchronized heterogeneous network the pulse edges ought to remain indistinguishably equivalent on that period.

On the other hand, administrators pick GNSS for conveying synchronization and location identification [1]. On the contrary, for indoor situations GNSS signal is feeble due to building wall thickness. Moreover, a few propositions are made which require additional hardware, such as local Grandmaster. By the by, additional equipment may expand the expenses.

The main aim of this paper is to analyses the offset errors over different link speed for the enhanced IEEE 1588, improved enhanced IEEE 1588 BBP, and IEEE 1588 Precision Timing Protocol (PTP) approaches.

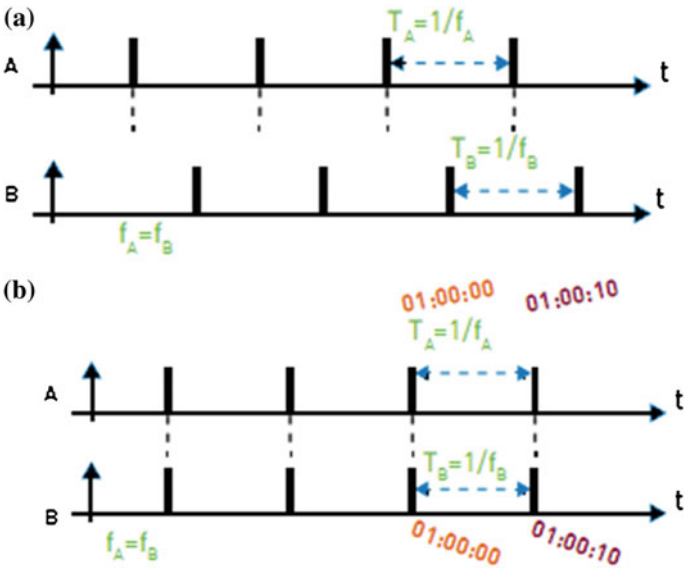


Fig. 1 **a** Represents the mismatching pulse edge indicates the frequency error. **b** Matched pulse edges with no error

The rest of the paper is organized as: the synchronization methods and problems to address the issue is presented in Sects. 2, 3 analysed the performances, and Sect. 4 determines the paper conclusion.

2 Synchronization Algorithms

To synchronize HeNodeBs, macro-eNodeB assistance has considered through broadcasting frame by using the Poisson clock [8]. The approach shows that each of the HeNodeBs needs to update their clocks with macro-eNodeB. The approach is able to synchronize the HeNodeBs. On the other hand, the method has accomplished a 0.5 s offset which cannot support LTE-A based systems.

Microsecond synchronization impreciseness remains obviously a long way as of perfect for LTE-A cellular systems as the time span of a solitary sub-frame in LTE-A is only 1 μs [9]. It is additionally observed that utilizing upgraded Inter-Cell Interference Cancellation (eICIC) even with a flawed synchronization between neighboring macro-eNBs. By means of asymmetry in HeNBs, the accomplished execution could be further debased even with diverse obstruction subdual techniques. Hence obstinate precision is essential for the Heterogeneous network.

2.1 IEEE 1588 PTP

Precision Time Protocol (PTP) is based on IEEE 1588 standard which enables time synchronization for packet switching HMN. As enhanced IEEE 1588 BBP, PTP is also master-slave synchronization protocol. The timestamp appliance is functional on PTP for the master to slave time synchronization. In PTP master clock sporadically transmits Sync messages to slave clock. The PTP transmitted messages from master clock to slave clock can be varied in payload size. The PTP well-thought-out to appraisal clock offsets by means of the subsequent equation [5, 6].

$$\delta_{T-offset} = \frac{(T_2 - T_1) - (T_4 - T_3)}{2} - \frac{(T_{M \to S} - T_{S \to M})}{2} \quad (1)$$

where, T_1 to T_4 is timestamps, $T_{M \to S}$ presents the association of delay from central server to destination slave clock, and otherwise represents through $T_{S \to M}$. (Fig. 2).

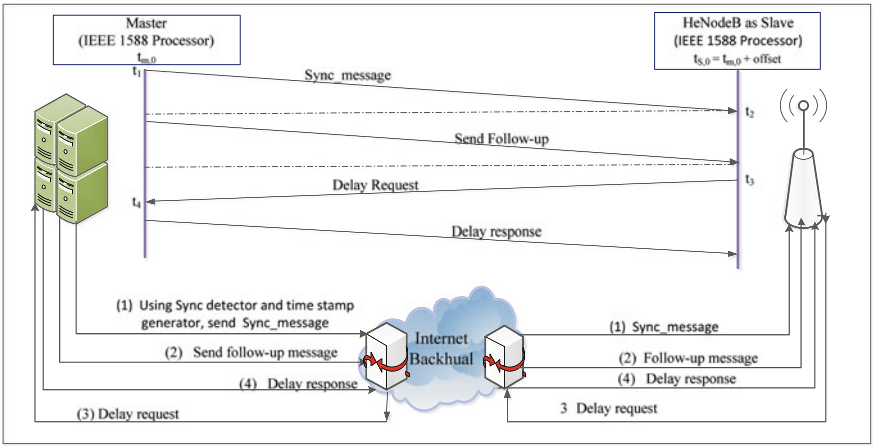


Fig. 2 Demonstrations of PTP estimation [5, 6]

2.2 Enhanced IEEE 1588 BBP

A superior time synchronization tactic devises anticipated for IP based (ADSL as well as VDSL) over IEEE 1588 in paper [7] to decrease the phase inaccuracy. It is observed that there has a destination clock which sends *Asymm_Check_Req* to main clock towards conducts pretend packets denotes equally BB. These BB appears as a *Loop_Req_M*essage (*L_R_M*) and it encloses a symptom of status. Additionally, this *L_R_M* clenches pretend stream of traffic statistics which concurrent to the central clock server. Once the first *L_R_M* is thrash out at destination, it acknowledge to the main server. The main clock will accumulate the time $T_{(S.4)}$ and stored. Whereas, for the preceding *L_R_M* stores as $T_{(M.4)}$ and communicates *Asymm_Check_Resp_*Message to the destination clock with précised $T_{(M.3)}$ and $T_{(M.4)}$. Figure 4 characterises the better BB irregular ratio assessment procedure (Fig. 3).

$$\delta_{T-offset} = \partial^{delay}_{M \to S} - \partial^{delay}_{downlink} \tag{2}$$

where, $\partial^{delay}_{downlink}$ is the downstream relations delay for one-way packet communication is expressed in Eq. (3) in where, R shapes for irregular ratio of main clock to destination clock communication (Eq. 4). At that juncture the total offset for the slave clock can be estimated smearing Eq. (5). Where, $\left(\partial^{delay}_{T_{M \to S}} + \partial^{delay}_{T_{S \to M}}\right) = \hat{T}_{rtt}$ states the round-trip-time of master clock to slave clock. In Eq. (6), the real offset error is displayed, where the predictable offsets are barred.

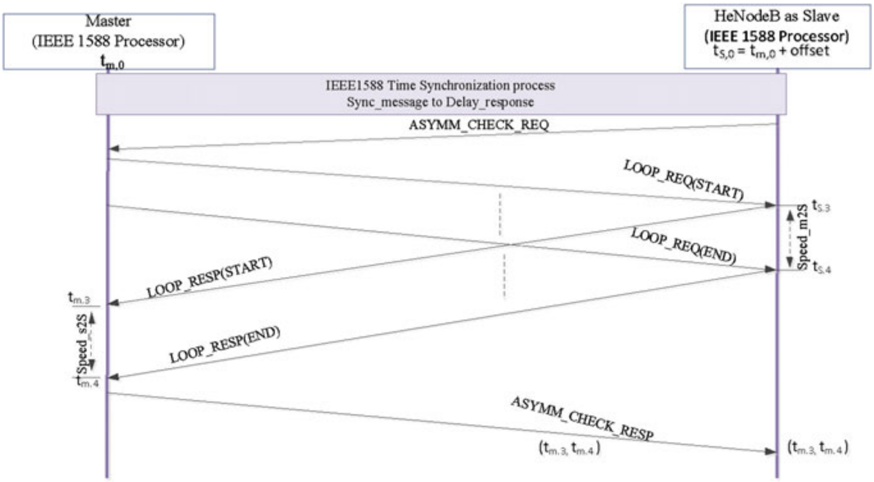


Fig. 3 Presents the enhanced IEEE 1588 BBP

$$\partial_{downlink}^{delay} = \frac{\left(\partial_{T_{M\to S}}^{delay} + \partial_{T_{S\to M}}^{delay}\right)}{1 + \mathrm{R}} \tag{3}$$

$$\mathrm{R}_{scale} = \frac{M_{(T_4 - T_3)}}{\mathrm{S}_{(T_4 - T_3)}} \tag{4}$$

$$\tilde{\delta}_{T-offset} = \partial_{M\to S}^{delay} - \frac{\hat{T}_{rtt}}{1 + \mathrm{R}} \tag{5}$$

$$\tilde{\delta}_{Offset}^{Error} = \tilde{\delta}_{\mathrm{Real}}^{Offset} - \tilde{\delta}_{T-Offset}^{estimated} \tag{6}$$

However, the procedure has protracted from [10] for wireless link in view of the distinct association transmission delays [7] of sync as well as delay_req. Aimed at this individual delay assessment the process has considered the fixed message payload. In view of that delays of the enhanced BB algorithm is enabled to address the exact clock offset and also can possible to decrease the bias errors. Conversely, the extended process which functions through, payload, insignificant data rate of the frame, $\hat{T}_{rtt}$ to minimize the bias error and likewise it can appraisal the transmission association of speed ratio.

3 Performance Analysis

To accomplish the aim of synchronization, the performance matric is well-known as clock offset which is also known as phase. The enactment assessment of the techniques is conceded consuming Monte Carlo simulation. The simulation parameters are in Table 2 [4–7, 11–13]. The outcome for the phase correction is attained for the HetNet is improved than the requisite timing is revealed.

In order to analyses the performance of enhanced IEEE 1588 BBP [7], improved enhanced IEEE 1588 [10], and PTP the software time stamps are used for analytical evaluation. Considering symmetric link speeds to address the offset error for slave nodes in HMN. The offset error for the slave HeNodeB is depicted in Figs. 4, 5 and 6 for enhanced IEEE 1588 BBP [7], improved enhanced IEEE 1588 BBP [10], and PTP [5, 6]. The clock offset error is estimated in terms of link speed, it is observed from the Figs. 4, 5 and 6 that if the link speed changes offset effects are imposed. It can be seen from the result that the PTP and enhanced IEEE 1588 BBP [7] created significant offset error, whereas the improved enhanced IEEE 1588 BBP [10] performs better in terms of less offset error. However, the improved enhanced IEEE 1588 BBP also marginally better, and still have offset error, which may create asymmetry.

For extreme analytical point of view, the enhanced IEEE 1588 BBP [7], and improved enhanced IEEE 1588 BBP [10] both of them considered transmission issued and egress time stamps are equivalently 0 and for the received and ingress timestamps considered as payload; which may greatly impact on $\hat{T}_{rtt}$ and link delays negatively.

Table 2 Simulation parameters

Parameters		Data
Network size		250 × 250 m^2
Distance (d)		20 m
Radius		30–40 m
Delay variation (σ) for delay		10, 3 µs
Number of progression		15/25
Range		30 m
Number of samples		1024
Carrier bandwidth		20 MHz
Time bandwidth factor		100 MHz
Payload size		118 Bytes
Initial offset		50 ms
Payload size		118 Bytes
Master clock transmit speed		1 Mbps
Slave clock transmit speed		11–20 Mbps
Communication link speed	Low speed	10 ms
	Medium speed	100 ms
	High speed	1 ms

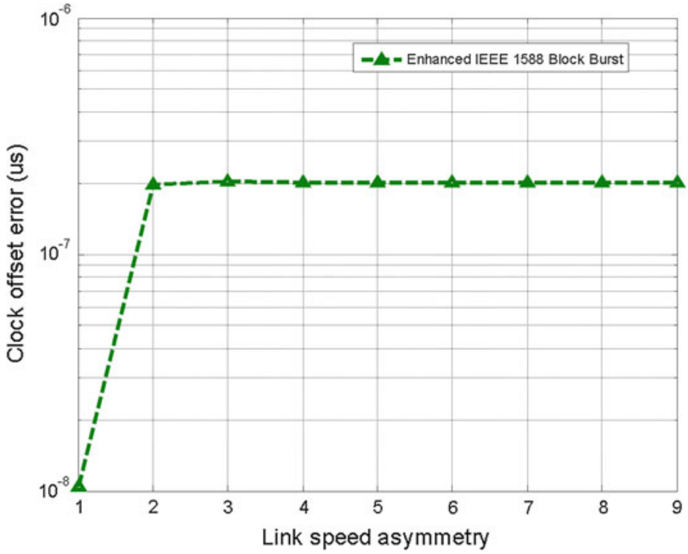


Fig. 4 Illustration of offset error for BBP on slave

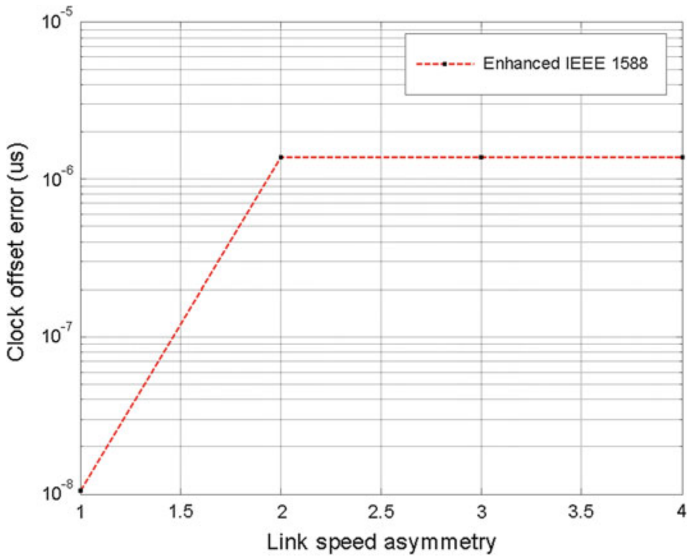


Fig. 5 Offset error for enhanced IEEE 1588

This is because of in HMN, the preamble and headers are used to send lowest data rate, while the payload are higher rate thus chances of raising $\hat{T}_{rtt}$. Since, HeNodeBs are fully depends on internet connectivity, thereby this type of link delay will impact the performance negatively.

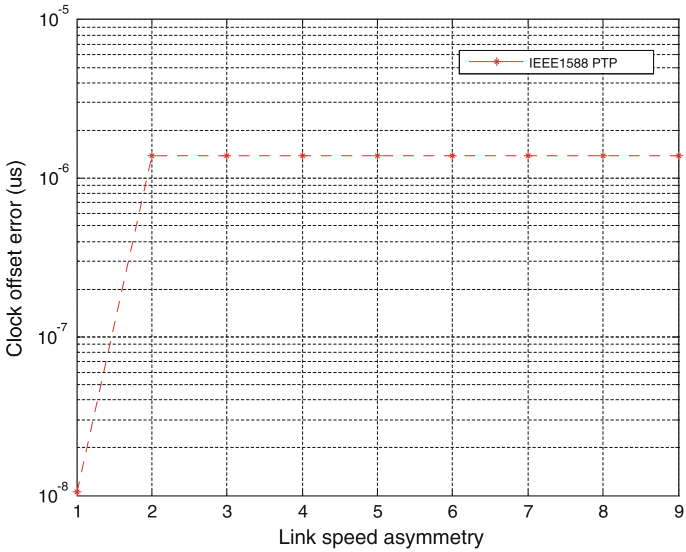


Fig. 6 Offset error for PTP

In OFDM based HeNodeBs are using multi-tone modulations with 4 microsecond (μs) symbols (IEEE 802.11g) and the frame length is an integer multiple of 4 μs. As a result, the assessed and average data rates will be different. In consequence, it can be point out that the physical layer frame duration is approached as non-linear function of the payload divided by the transmission rate.

4 Conclusion

This paper investigated time synchronization algorithms of enhanced IEEE 1588 BBP, improved IEEE 1588 BBP and PTP. The performance analysis suggests that if the synchronization signal exactness is higher than the time synchronization, can be précised.

A synchronized HeNBs are prerequisite for control of exchange, phase synchronization, fluctuation estimation module and exact clock to deploy it in heterogeneous network. Subsequently there is a major probability to force distinctive delays for transmit and receive path, in outcome asymmetry has been created, and this asymmetry cannot be assessed in a pairwise synchronization framework. However, due to plug and play of HeNodeB by the customer premise, may time connectivity is challenging. Thus there has a big possibility to impose different delays for transmit and receive path, in consequence asymmetry has created, and this asymmetry cannot be estimated in a pairwise synchronization system. Thereby the future recommendation is to develop time synchronization approach for plug-and play HeNodeB neighbor network.

Acknowledgement A distinct acknowledgements to Ministry of Education for the education funding, and appreciations to Research Management Center (RMC), International Islamic University Malaysia.

References

1. Pesyna, K.M., Wesson, K.D., Heath, R.W., Humphreys, T. E.: Extending the reach of GPS-assisted HeNodeB synchronization and localization through tightly-coupled opportunistic navigation. IEEE International Workshop on HeNodeB Networks, pp. 242–247. doi:10.1109/GLOCOMW.2011.6162445
2. Okada, M., Hara, T., Saitue, N., Wada, T.: Frame synchronization among base stations for TDD systems, IEEE. 4th International Symposium on Communications, Control and Signal Processing (ISCCSP), pp. 1–4. doi:10.1109/ISCCSP.2010.5463476
3. HetNet Synchronization: Online available www.http.com//qulsar.com/Applications/Telecommunications_and_Networks/HetNets.html
4. Hasan, M.K., Saeed, R.A., Hashim, A.H.A., Islam, S., Alsaqour, A.R., Alahdal, A.T.: Femtocell network time synchronization protocols and schemes. Res. J. Appl. Sci. Eng. Technol. **4**(23), 5136–5143 (2012)
5. Andre, V., Dominik, S.: Clock synchronization in telecommunications via PTP (IEEE 1588). IEEE International Frequency Control Symposium, pp. 334–341 (2007). doi:10.1109/FREQ.2007.4319093
6. Vallat, A., Schneuwly, D.: Clock synchronization in telecommunications via PTP (IEEE 1588). In Frequency Control Symposium, 2007 Joint with the 21st European Frequency and Time Forum. IEEE International, pp. 334–341. IEEE, 2007
7. Sungwon, L.: An enhanced IEEE 1588 time synchronization algorithm for asymmetric communication link using block burst transmission. IEEE Commun. Lett. **12**(9), 687–689 (2008). doi:10.1109/LCOMM.2008.080824
8. Lien, S.-Y., Lee, H.H., Shih, S.-Y., Chen, P., Kwang-Cheng, C.: Network synchronization among femtocells. IEEE GLOBECOM, pp. 248–253 (2011). doi:10.1109/GLOCOMW.2011.6162446
9. Wang, Y., Pedersen, K.I.: Performance analysis of enhanced inter-cell interference coordination in LTE-Advanced heterogeneous networks. IEEE VTC Spring, pp. 1–5, May 2012
10. Lee, S., Lee, S., Hong, C.: An accuracy enhanced IEEE 1588 synchronization protocol for dynamically changing and asymmetric wireless links. IEEE Commun. Lett. **16**(2), 190–192 (2012)
11. Kaushik, G., Prasad, V.: Network time synchronization in time division—LTE systems. M.S. thesis, School of Electrical Engineering Aalto University, Finland (2013)
12. Kerö, N., Müller, T., Kernen, T., Deniaud, M.: Analysis of precision time protocol (PTP) locking time on non-PTP networks for generator locking over IP. SMPTE Motion Imaging J. **123**(2), 37–47 (2014)
13. Exel, R.: Mitigation of asymmetric link delays in IEEE 1588 clock synchronization systems. IEEE Commun. Lett. **18**(3), 507–510 (2014)

A Framework of an Automatic Assessment System for Learning Programming

Nuraini Abdul Rashid, Lau Wei Lim, Ooi Sin Eng, Tan Huck Ping, Zurinahni Zainol and Omar Majid

Abstract Automatic assessment system is a system that will evaluate assignment submitted automatically by the students with the minimum input from human. Studies show that existing works handled only one or two programming languages. They are web-based platform, using marking scheme and strategy, are embedded in the learning management system and provides graphics user interface. This paper presents a proposed framework of an Automatic Assessment System for learning object oriented programming language. We initiate the research by first conducting survey, which investigates the needs of the system in the department. The survey was conducted on two main groups, which are the first year programming students and the lecturers who have experience in teaching programming. The results from the survey show that the Assessment system can help students in learning programming language as well as reduce the lecturers' workload. We proposed a framework for Automatic Assessment system, which consists of five modules. The modules are teaching module to help students learn and submit program code; real time detector module to show existing error and supply hints to solve problem;

N.A. Rashid (✉) · L.W. Lim · O.S. Eng · T.H. Ping · Z. Zainol
Pusat Pengajian Sains Komputer, Universiti Sains Malaysia, Penang, Malaysia
e-mail: nuraini@usm.my

L.W. Lim
e-mail: wlim.ucom12@student.usm.my

O.S. Eng
e-mail: oseng.ucom12@student.usm.my

T.H. Ping
e-mail: thping.ucom12@student.usm.my

Z. Zainol
e-mail: zuri@usm.my

O. Majid
Pusat Pengajian Pendidikan Jarak Jauh, Universiti Sains Malaysia, Penang, Malaysia
e-mail: momar@usm.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_82

marking module to compile, mark, execute, and pre-set evaluation criteria; analysis and reporting module to give instant feedback and social module for interactions between the users.

Keywords Programming · Automatic assessment · One-to-one tutor · Feedback · Web application

1 Introduction

Automatic assessment system assesses assignments submitted automatically with little input from human. Automatic assessment also allows instant feedback without the need to reduce programming exercises. It is an important tool that is needed by any tutor or lecturer in teaching large programming classes. Students need to implement many programming exercises to master the programming languages more effectively. Continuous assessment during a programming course ensures that the students get enough practices as well as feedbacks on the quality of their solutions [1] instantly. However, this practices somehow does not work very well in the past, especially in large classes. In classes with more than 100 students, tutor could not manage to evaluate all the exercises done by the students and it is an impossible mission to assess all the exercises in such a short time.

In order to give many exercises to large number of students, a big number of tutors and huge resources are required. This is not really convenient and practical way to teach in programming courses, since the resources are one of the important aspects that need to be considered. Therefore the existence of an automatic assessment system is very helpful and useful in teaching and learning process. With the primary aim of improving the teaching and learning of computer programming, we have proposed a Web-based framework, for the use in programming courses. This paper presents our experiences with the review of the current existing system, and discusses its impact on our practices in teaching and learning. We initiate the research by first understanding the needs of the first year programming students and the lecturers who are currently teaching programming. The rest of this paper is organized as follows. Section 2 discusses the related works. It outlines the review and analysis of the current existing automatic assessment system. Section 4 presents results and analysis from the survey given out to both students and lecturers, and shares our experiences from both the perspectives of students and instructors. Section 5 outlines the proposed automatics assessment system, which is the outcome of the literature study and finally we compare our framework to existing system.

2 Related Works

Assessment system has been revolutionized through many generations [2] since programming language was created. The main goal is to improve learning efficiency and reduce tutor's workloads. We conducted the literature study with five main questions in mind. They are (1) Does the system handles one or more programming languages, (2) On what platform are they implemented, (3) What are the strategies used in online marking (4) Can this system detect plagiarism and (5) Are the system embedded in any learning management system.

Most of the assessment system targeted only for Java and C or C++ [1] since these two languages are more popular in the past years and only a small number of assessment systems support other languages. C-Marker system, for example, has a very static assessment module because it can only mark one type of language and tested only the C Code structure and examine on the correctness of the output [1]. This might not be suitable for education institutions that teach different languages although those languages have the same paradigm, thus the cost to redevelop different systems for them are high. GAME system has successfully overcome this problem, by recognizing the common characteristics and behaviors of programming source code [1]. Sub classes of different languages can be evaluated by using inheritance concept from a common class, and also provide concrete implementation of their own language [1]. This will reduce cost since Java and C++ have a lot in common, especially in basic programming. General class concept can also be applied to others language paradigms like Matlab, HTML5, and etc.

In terms of platform, web-based system is important to enhance the effectiveness of tutors and students. Ludwig system allowed student to edit program in text editor, then submit it for grading [3]. Encapsulated text editor in Java Applet does not allow student to cut and paste codes from outside the editor. Stylistic analysis will also be provided, which includes programming style i.e. proper indentation, cyclometric complexity, global variables and etc. [3]. Tutor has to create solutions for the expected output based on selections of input. The students' program is then executed and the output will is compared. Students can also edit their mistakes and re-submit the code again. Questions are randomized [3] so that each of the students will receive different version of questions. An advantage of web-based system is compatibility where it can run on any operating system or devices without much trouble thus this will save development cost. However, performances will decreased compared to an installed application, since the page has to be loaded every time a user needs to access to the page. Security issues and network traffic issues are also a concern since hacking or unintentional mistake in coding will always happen.

Marking scheme and strategy to assess programming assignment can be divided into two main types; peer assessment and automatic assessment. Peer assessment method considers the correct program when marking [4]. Each student's assignment input is run against different tests to check correctness of the functions of code. Quality of the program is marked by student's peers which is not directly defined and variety of interpretation is offered. Average mark of each student will depend on tutors and students. Peer assessment helps in improving the programming's skill among students since they can learn from each other's code. However, this needs commitment from both the tutors and students, even if the tutor alone would be enough for grading purposes. By using automatic assessment, marking scheme contains assignment files, marks, and marking criteria [4]. Student's program contains file to execute, which are input, correct output, and instruction files. Correct output file contains expected correct output; output file contains the output from students. Instructions for program's operation are held inside instruction file. Different types of assessment require different marking criteria, to allow different marking strategy. "Keyword" strategy looks for a keyword in student's output. "Ordered-keyword" strategy looks out for ordered set of keywords [5]. Since there are different kinds of approaches to solve programming problem, the quality expect of the program cannot be marked against any template. Comparison of expected output (from tutors) and student's output is by far the most efficient and cost-saving strategy, especially for beginner's programming assignment, which contains only simple structure.

Plagiarism has always been a problem for automatic assessment [6]. Plagiarism detection is important when it comes to automatic assessment. To do this, a kind of abstract tree, the result of reduced version of individual assessment, is proposed [7]. Then, the names of variables are removed, sorting commutative primitives' definitions and arguments. Lastly, the resulting trees are used to compare among each other. Copying other's work is a serious issue, but mutual review of codes can increase efficiency of learning, therefore students are encouraged to write their own codes after reference from friends rather than direct copying.

Learning management system can help tutors and students to manage the submission in suitable order [5]. After the assessment submission, the results and reports of students will be displayed in a summary format. This format includes source code structure mark; correct mark based on output and related compiling errors and warnings. The assessment system calculates the final mark by summing the marks of three sub-parts: block comments number compared to functions number, variables declaration validation, amount of "magic numbers" used in the code [1]. At further stages, time for completing the assignment can also be part of criteria of marking. Various kinds of feedback will be given to students. A set of possible outcomes will be given by each test in configuration files. For each outcome, points and comments is specified by the configuration files [7]. In some systems, students are allowed to edit or make correction to their code, if mistakes exist. Then, they are allowed to re-submit the code again for next evaluation [8]. By allowing resubmission of program, it can actually increase the confident of student in coding. Programming Assignment aSsessment System (PASS) can help to mark certain specific problems, and return the result with feedback of performance, for student to do correction, to resubmit and to remark. However, students complain about less-attractive interface and textual matching problem. Simple, consistent and good visual appeal interface is important for student with or without information technology background to learn programming.

3 Methodology

A survey was conducted in order to collect valuable feedback from respondents before we proposed a improved framework. In order to get a different perspective from different category of participants, participants are divided into two categories, which are students and lecturers. Participants are from the School of Computer Sciences, Universiti Sains Malaysia. Students come from different years of exposure to programming languages; this could be helpful in data collection. The lecturers that were chosen to answer the survey question have vast experiences in teaching programming languages courses. This could provide a precise and quality feedback regarding the proposed framework. Other than giving out survey questions, literature review and study on existing automatic assessment system is being done at the earlier stage before proposing an enhanced framework for the assessment system.

4 Preliminary Study

This section outlines the analysis of the survey distributed to our participants. Target participants were given sets of questions in order to get their perspective on the automatic assessment system for learning programming. A total of 54 students and 12 lecturers have responded to the survey questions.

Figure 1 presents the result of the survey distributed to the student. The first question asked is if the students felt the necessity for the lecturers to give instant feedback of the given assignment. The second question is to gauge how long the students expected the lecturer to return their assignments.

Feedbacks from students reflect that they need the automatic assessment system in the learning process. The respondents feel that it is better to have feedbacks as soon as possible from the lecturers once they submitted their assignments. Without

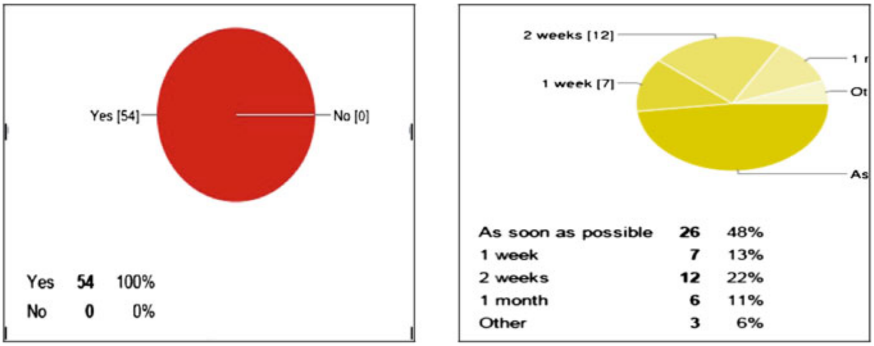


Fig. 1 The need of feedback and the expected response time from lecturer for a given assignment

the automatic assessment system, lecturers are not able to give the feedback to students promptly. Students preferred to have one-to-one tutor in their programming courses. This could be achieved only with large number of human resources. Hence, respondents are showing their willingness to use the automatic assessment system in the future. Based on several survey questions given out to respondents, it clearly shows the importance of the automatic assessment system for learning programming languages. Out of 54 respondents, 37 respondents agree that functionality is the most important feature than interface and performance in the automatic assessment system.

Lecturers were asked to suggest functions that they need. Below are the suggestions from the respondents.

- Syntax error detector—similar to compiler but in a simpler form
- Logic error detector—difficult to implement but good to have
- Code compare—check similarity of the code submitted (to detect cheating/Plagiarism)
- Code suggestion—point out the error and hint how to get correct answer
- A module for understanding the program
- A module for automatically explaining the error.

5 Proposed Framework

An overall framework as shown in Fig. 2 has been proposed to solve the needs of new automatic assessment system. Web-based platform is used since it requires no installation procedures. Tutors and students can log in into the system by using the interfaces and perform operations based on the privileges set by the system or owner. Teaching module is highly coherent with real time detector, marking module and analysis and report module. These three modules will focus on basic learning and supports for students as well as tutor's management for assessment. In addition, social module is an extension module, which provides features to students to increase the efficiency of learning using social network applications.

5.1 Teaching Module

In this module, the tutor will provide materials to be learnt, including sample programs and test questions.

Programming Language: We selects Java, C# and C++ because they have many common behaviors, characteristics and are in the same paradigm concept.

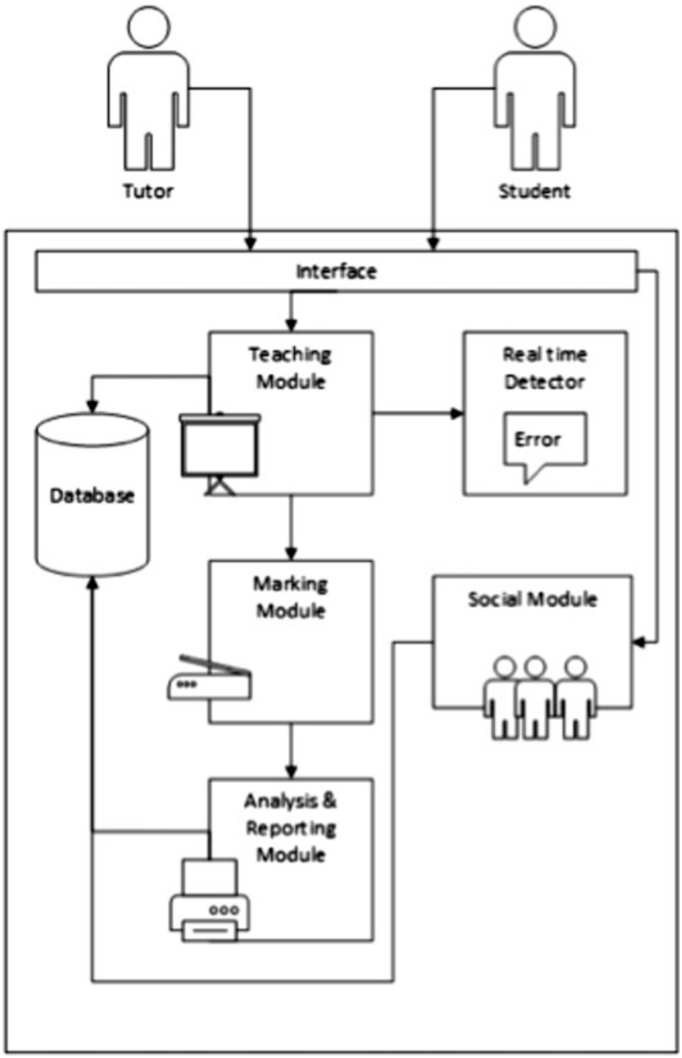


Fig. 2 The proposed framework of an automatic assessment system

Tutorial based on chapter: A tutorial based on programming language chosen by the students is added so user can learn some basic concepts before taking a test. Student will have to learn the tutorial chapter by chapter to enhance the foundation of programming. Videos will be added into this tutorial section if necessary.

Test Question: A set of questions can be added by tutor and is stored in database. Test question can be divided into several categories: basic input output, algorithm, structure, procedural and others. Tutors have to set the expected output of the question and it will be compared with student's output to determine the accuracy of code. The tutor can also determine running time limit or memory limit to prevent students from using brute-force on solving particular algorithm question. Test sequencing is random. The system will auto randomize the questions that are extracted from the database so that every student will get different set of question. The lecturers created the questions earlier. This can prevent the students from cheating during the test.

5.2 Real Time Code Suggestion

To have real time code detection and suggestion feature, Java Applet is to be implemented as extension to this system. Java Applet is able to detect syntax error in real time when compared to the normal editor in web platform. Furthermore, encapsulated text editor in Java Applet can prevent student from copy and paste directly from other sources. By using Java Applet extension for editor, system does not have to compile the code again when the student submits the code. The system just has to detect any form of error before the student submits the code. This will save a lot of system resources and also processing time.

Detect error and Suggest Hints to solve the problem: The system will point out the errors made by student. There are two types of errors detected which are syntax error and logic error. The system will give hint to the students on how to rectify the error.

5.3 Marking Module

After the student submits the code, a series of operations will be conducted in order to determine the quality of the code and a score is given according to marking scheme and criteria. Manual assessment will be used to give the student score for the visual interfaces and graphics design, but automatic assessment system will have many limitations to do so.

Enhanced marking scheme: Marking scheme contains assignment file, marks and marking criteria. Marking criteria focuses on output correctness, programming concept, creativity and performance; time to complete the project, and simplicity of the program. In terms of output correctness, student's output file will be compared with tutor's output file, which contains the correct or expected output. Percentage of correctness will then be calculated and reported. In terms of programming concept, codes will be required to be broke down into a kind of abstract tree, and eliminates unnecessary information. At the end, the resulting trees will be used to compare with the correct one. Percentage of match between two trees will be calculated. In terms of simplicity, the most direct and easiest way would be calculating the lines of codes submitted by students. A program with fewer lines of codes reflects that it is simpler and will be given more marks for its simplicity.

Marking strategy: "Keyword" strategy can be used to look for a keyword in student's output while, "Order-keyword" strategy is looking output for ordered set of keywords. These two strategies will ignore the irrelevant part of codes.

Plagiarism screening: A kind of abstract tree is used to represent the reduced version of individual assessment. Then, the names of variables are removed, sorting commutative primitives' definitions and arguments. Lastly, the resulting trees are compared. The 95 % similarity threshold is used to filter the plagiarized work.

5.4 Analyses and Reporting Module

This module allows the student to self-improved. It also allows tutor to manage their students depending on the score and the understanding level on certain subject in programming.

Reporting Information: Students are able to get their reports and results of programming assessment in a summary format. This format includes series of marks, which depends on the marking criteria set earlier by tutors and also (if exist) compiling error and warnings. If allowed, or the deadline has not been reached yet, the students can resubmit the code for evaluation again, if they can identify the mistakes made.

Feedback: Based on the submitted solution, automation and various kind of feedbacks will be given to students. A set of possible outcomes will be given by each test in configuration files. For example a test run with a set of possible answers. For each outcome, the configuration files will specify points and comments. The basis of individual tests will give comments to user. Error message, warnings and encouragement will also be part of the comments to students.

5.5 Social Module

A social platform is also introduced in our proposed system so that the user are able to discuss and contribute ideas or make correction on error on programming. This social module is an extra feature where the user can easily get help when they faced problem during programming.

Social platform: A social platform will also be introduced in our proposed system to enable users to post their problems to the platform. This platform is like a forum where user can ask questions in the forum using the web based system. Students around the world can give suggestions to help them through the forum.

Game: To increase the interest of student in learning programming language, a game concept will be inherited to the system. Some mini games, quiz, and questions will be added. Students can level up by finishing the quiz. Intra-competition with friendly environment will make process of learning programming to be fun.

One to one tutor: In this feature students can make an appointment with tutor, this will enable the tutor to arrange one to one tutoring slot to give one to one tutorial. Our proposed system will have an online compiler that support multi-user editing.

5.6 Database

Cloud database will be used and run on a cloud-computing platform. Users can run databases on the cloud independently, using a virtual machine image. SQL Data Model is suggested since relational database is more suitable for this system.

Basically our proposed system has all the functions and features of the three systems. We also enhance our proposed system with some extra feature (see Table 1). Social features and one to one tutoring have been added to help student learn programming language.

Table 1 Comparing the proposed system with existing systems

System functions	PASS	Ludwig	Game	Proposed system
Programming languages	C	C++	Java C++	Java C++ C#
Feedback	Instant feedback Informative	Instant feedback	Instant feedback Give return of procedure	Instant feedback Detect error Give code suggestion
Platform	Web-based	Web-based	Desktop	Web-based
Plagiarism detection	Not stated	Cannot copy paste from outside	Yes	Yes Questions are randomise
Social platform	No	No	No	Yes GAME platform Multi-user
One to one	No	No	No	Provide tutor through Web-Cam

6 Conclusion

In this paper, we found out that the student respondents in the School of Computer Science, USM would like to have an automatic assessment system. Only 58 % of lecturer respondents think that USM needs an automatic assessment system although 67 % of them think that this automatic assessment system can reduce their workloads. A web based automatic assessment system is convenient to students and lecturers. In order to ensure quality of programming project of student, the marking scheme of the automatic assessment system must focus on correctness, creativity, time to complete, programming concept and simplicity. With a good automatic assessment system to be developed, student can learn programming in a different way. Unfortunately, only half of the lecturers who participated the survey said they are willing to take part in the development project of this automatic assessment system.

Acknowledgement We would like to acknowledge USM RU grant Pkomp/1001/817065 and School of Computer Sciences, USM for the supporting the publication of this work.

References

1. Blumenstein, M., Green, S., Nguyen, A., Muthukkumarasamy, V.: GAME: a Generic automated marking environment for programming assessment. In: Proceedings of International Conference on Information Technology: Coding and Computing 2004, ITCC 2004, vol. 1, pp. 212, 216, 5–7 April 2004
2. Douce, C., Livingstone, D., Orwell, J.: Automatic test-based assessment of programming: a review. J. Educ. Res. Comput. **5**(3), 4 (2005)
3. Shaffer, S.C.: Ludwig: an online programming tutoring and assessment system. SIGCSE Bull. **37**(2) (2005)
4. Jirarat Sitthiworachart, J., Joy, M.: Effective peer assessment for learning computer programming. In: Proceedings of the 9th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education (ITiCSE'04). ACM, New York, pp. 122–126 (2004)
5. Pihantola, P., Ahoniemi, T., Karavirta, V., Otto Seppl, O.: Review of recent systems for automatic assessment of programming assignments. In: Proceedings of the 10th Koli Calling International Conference on Computing Education Research (Koli Calling'10). ACM, New York, pp. 86–93 (2010)
6. Ala-Mutka, K.M.: A survey of automated assessment approaches for programming assignments. Comput. Sci. Educ. **15**(2), 83–102 (2005)
7. Saikkonen, R., Malmi, L., Korhonen, A.: Fully Automatic Assessment of Programming Exercises. Department of Computer Science and Engineering, Helsinki University of Technology, Finland (2001)
8. Yut, Y.T., Poon, C.K., Choy, C.K.: Experiences with PASS: developing and using a Programming Assignment aSsessment System. Department of Computer Science City, University of Hong Kong (2006)

Parallel Image De-fencing: Technique, Analysis and Performance Evaluation

Madiha Khalid and Muhammad Murtaza Yousaf

Abstract Image de-fencing is a real-life problem in digital photography, where fences from images are removed in a seamless way such that it appears as if fence was never existed on the image. Often such fence objects are unwanted but unavoidable in real photography. Several attempts has been made to automatically de-fence an image, but the problem with most of the techniques is that the whole image needs to scan many times before Inpainting the unwanted block of pixels, and even it can cause more inefficiency and degrade in performance if the fence detection procedure is automated. Thus there should be a balance between efficiency and accuracy. One way to achieve significant performance gain is through parallelism, but it depends on how well the algorithm lends itself to parallelization. In this paper, we present different parallel versions of a feature based image de-fencing technique that has a lot of potential for parallelism. We implemented a parallel version that uses modern GPUs and NVIDIA CUDA framework to increase the overall efficiency of the algorithm with approximately no noticeable effect on visual quality of the results. We have also determined average speed up of parallel algorithm over serial algorithm through relative experiments.

1 Introduction

As parallel processing has become the most promising solution to the computing requirements of the real life applications. Image processing is one such area that can be highly promoted by enabling higher degrees of parallelism on multiple CPUs or by the use of GPUs. Although there are a number of existing image-defacing [1]

M. Khalid (✉) · M.M. Yousaf
Punjab University College of Information Technology,
University of the Punjab, Lahore, Pakistan
e-mail: madiha.khalid@pucit.edu.pk

M.M. Yousaf
e-mail: murtaza@pucit.edu.pk

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_83

techniques but none of them have parallel implementations. The reason may be their high complexity or the inherent sequential nature.

Existing image de-fencing techniques detect fences on the basis of colors [2], intensities [3], shapes [1] or some other feature sets [4]. Due to the better performance and greater potential for data parallelism of feature based de-fencing technique, we have chosen feature based image de-fencing [4] for parallel implementation. The crux of feature based image de-fencing is to compute a robust feature set that can distinguish fence pixels from non-fence pixels. The approach that we follow is proposed in [4] that uses pixel based classifier which is rooted on a robust feature set. The classifier runs on the image and returns a fence mask. The fence mask is then refined and removed from the image using a hierarchal approach that uses Gaussian pyramids [5]. After a careful analysis of the approach, we identified feature computation as the most time consuming and parallelizable part of the algorithm while post processing and restoration takes considerably less time. So we considered this step for parallelization. We discussed different parallel approaches for distributed memory and shared memory architectures. Furthermore, we implemented a parallel approach on GPU and compared the serial execution time with parallel execution time. We gained an average speed up of 15. Section 2 presents various parallel approaches towards the algorithm. Section 3 focuses on implementation details while Sect. 4 discusses experimental setup and test beds. Section 5 demonstrates the experimental results and Sect. 6 concludes our work.

2 Parallel Approaches

Typically, a serial algorithm is evaluated on the basis of execution time defined as a function of the length of the input. However, efficiency of a parallel algorithm is not only measured by the size of the input but also by the number of processors, their running times and the inter-process communication. Hence, the parallel algorithm development is a critical task of problem solving that cannot be achieved without considering the underlying architecture and additional dimension of parallelism extraction. There are several parallel programming models in common use including threads model, distributed memory model/message passing and shared memory model. In the following section, parallel approaches for image de-fencing with respect to distributed memory model and shared memory model will be presented and illustrated.

2.1 Distributed Memory/Message Passing

The algorithm in [4] uses a variant of HOG (Histogram of Oriented Gradients) [6, 7] as feature set. Several attempts have been made to efficient the performance of HOG [8–12]. The HOG itself can be parallelized in different ways and there are a

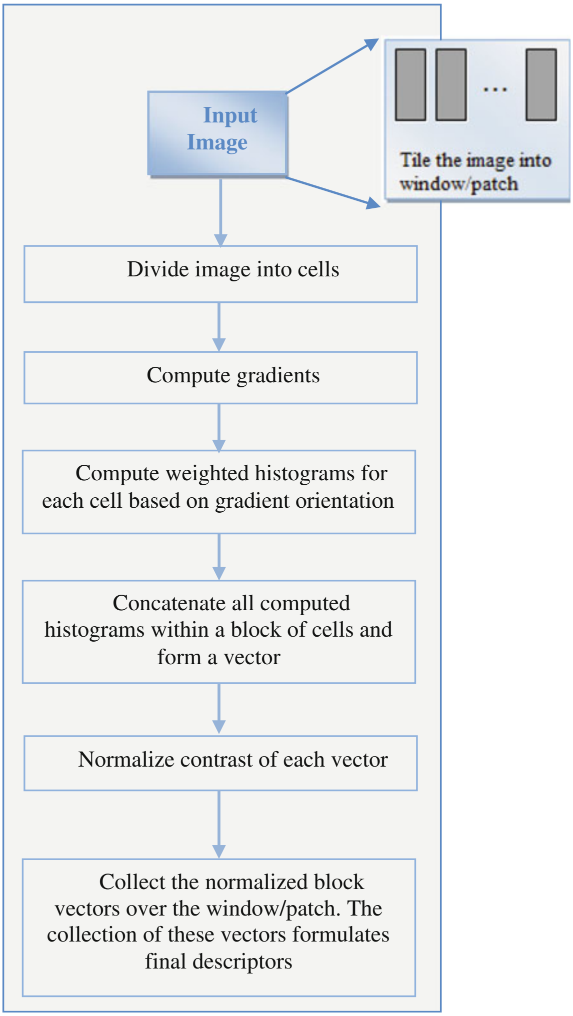


Fig. 1 A step-by-step description of HOG features computation

couple of parallel implementations already proposed in literature [8, 9] but for our purpose the way HOG is used is parallelized in a different way. The step by step description of feature computation is shown in Fig. 1. In feature computation phase, the HOG descriptor is computed for each pixel over a window of size x × y thus every descriptor computation is independent of other descriptor computation and can be executed in parallel. Each process needs to perform similar operation on the data assigned to it and thus, satisfying the applicability of data parallel model. Therefore, we used data parallel model for decomposition and static mapping.

First Approach—One way to parallelize the algorithm is to decompose data in row major order. In this firstly proposed version we will consider parallelizing all the steps in one fashion. In order to compute features for both training and detection purpose, we can decompose the image into set of rows and map each row set to a processing element. For an image of size $m \times n$, divide the total number of rows by the total number of processes. Let the total number of rows be m and total number of processes be p, this will give m/p data partitions to be mapped to p processes. For each pixel in each data partition, compute the feature descriptor. The feature descriptor is computed over a window expressed in terms of rows and columns in an image. All the p processes will compute feature descriptors for each pixel in their m/p data partition, although every process will require some values from neighboring processes to perform its computation. So neighboring processes will send those required values, after getting all the required values from neighbors the descriptor vector will be computed. To reduce the communication cost of providing required values we will ensure the partition size to be at least equal to the number of rows in the window, thus creating the m/xp partitions where x is the number of rows in a window as shown in Fig. 2a. For all the pixels in first row of each partition the HOG descriptors will be computed without requiring any additional values from other processes. Whereas it is worth noting that, all the pixels in other rows require some values from the neighboring processes i.e. the 2nd row of each process will require the top most row of data chunk assigned to next process, the 3rd row of each process will require 2 rows from the next process similarly the last row requires $x - 1$ rows from the neighboring process. Thus each process requires $x - 1$ rows to be sent to its neighboring process except the first process that will only receive the data.

The first step of HOG computation is to compute differentials along x-axis and y-axis. To reduce the communication cost this step can also be done in parallel. Each process can compute the gradients on its chunk of data. All the subsequent computations will also be done in parallel. Data sharing among processors can be done using send and receive calls from MPI libraries.

Second Approach—This approach is similar to first approach, except that it partitions data in column major fashion. Instead of mapping m/p rows to each process, now n/p columns will be mapped to each process. To complete its processing, now

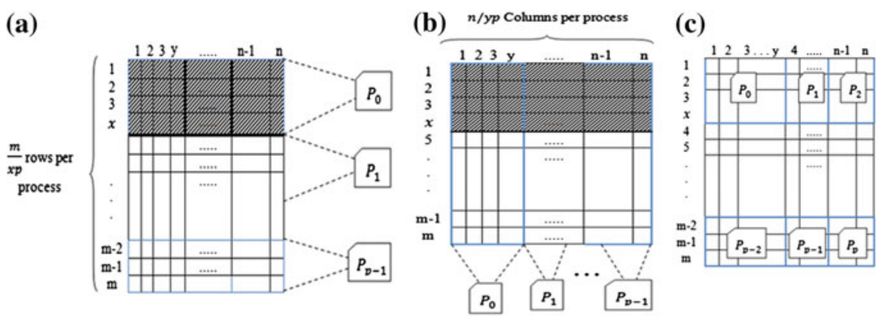


Fig. 2 **a** Data division in approach 1, where *m/xp* rows are mapped to each process. **b** Illustration of data mapping in approach 2. **c** Illustration of data mapping in approach 3. Each detection window is mapped to a process

every process requires pixel vales from neighboring columns. All other computations and communication will remain same as of first approach.

Third Approach—The third way of data decomposition is window by window. In this way, each process will receive small sub-matrices of size x × y for processing. A process may not need any more values from other processes to perform computation for one pixel, but for other x − 1 and y − 1 rows in the sub matrix the process again needs data from neighboring rows as well as from columns.

2.2 Shared Memory

It is notable that the communication costs taken by each of the above described approaches is considerable. Every process needs many values from the other processes to complete its computation. Thus their performance is greatly affected by the communication cost. One way to reduce cost is to use shared memory model where the communication cost is implicit. The approach used for shared memory model can be divided into two parts: the one that is to be done sequentially on host and the other that is to be done in parallel by the GPU. Feature computation is the GPU part while classification and restoration is the host part. Figure 3 shows the two parts executed by CPU and GPU. As this approach clearly reduces communication cost while keeping the efficiency gained by parallel computation so we implemented this algorithm.

3 Implementation

3.1 CUDA Architecture

Modern GPUs support thousands of threads that can run in parallel and uses NUMA (Non-Uniform Memory Access) memory access model. We used NVIDIA CUDA framework for our implementation. CUDA is NVIDIA's parallel computing architecture that allows writing C functions that are to be executed by multiple threads. Such functions are termed as "kernels" in CUDA. Many threads can execute a single kernel but a thread can execute only one kernel at a time. The CUDA architecture group threads in blocks, and blocks are grouped in grids. Threads in a block can cooperate through shared memory and can be synchronized as well. Grids are executed on devices while thread blocks are executed on multiprocessors. Each thread block is executed on a single multiprocessor but a multiprocessor can execute several thread blocks.

A GPU have six types of memories that a thread can access. Register, constant, shared memory, global memory, texture memory and local memory. Global memory, constant memory and texture memory are accessible for all threads while shared memory is accessible for threads within a block. Each thread has its own local memory and registers while each processor has shared memory plus its own registers. All the memories can be accessed for updating data except constant memory and texture memory, they have global but read only access. Also, constant and texture memories are cache optimized and thus gives faster access to data.

3.2 The Algorithm

For feature computation i.e. GPU part, our implementation idea is to place the image into global memory that every process can access and decompose image matrix onto detection window size such that each thread perform its execution on one window and can access the data of its own part.

Once the gradient matrices (magnitude and direction) are computed, the original image intensities are no longer required because all the subsequent computations are performed on the gradient magnitude and orientation values instead of original pixel intensities. Thus, instead of transferring the original image to GPU memory, the gradient matrices are copied in the GPU texture memory (texture memory is used because it gives global access to all threads, is faster due to cache optimization and is read-only) so that every thread in every block can access it. Then a kernel is executed with the number of threads equal to the number of windows in an image. Each thread is responsible to compute the HOG descriptor for its own window. After computing descriptors, each thread transfers its descriptor on a specified place in the global memory corresponding the pixel location in the image. In this way the encoded image is ready to feed the classifier. Figure 3 gives an overview of the proposed parallel implementation. The classifier performs fence/non-fence classification. This step is done sequentially.

4 Experimental Setup

We implemented and tested our algorithm on a core i5 laptop machine with Intel processor running at 2.00 GHz. The physical processor chip contains 48 cores with NVIDIA GeForce 410M graphics card. The sequential implementation of the algorithm was done in Matlab.

Furthermore, Matlab offers a facility to integrate C code into the Matlab code with the help of mex functions and CUDA provides a way of calling the device kernel from Matlab. Therefore, the parallel implementation was done in Matlab and C on GPU with NVIDIA CUDA framework and we used mex functions and feval functions for managing inter language calls.

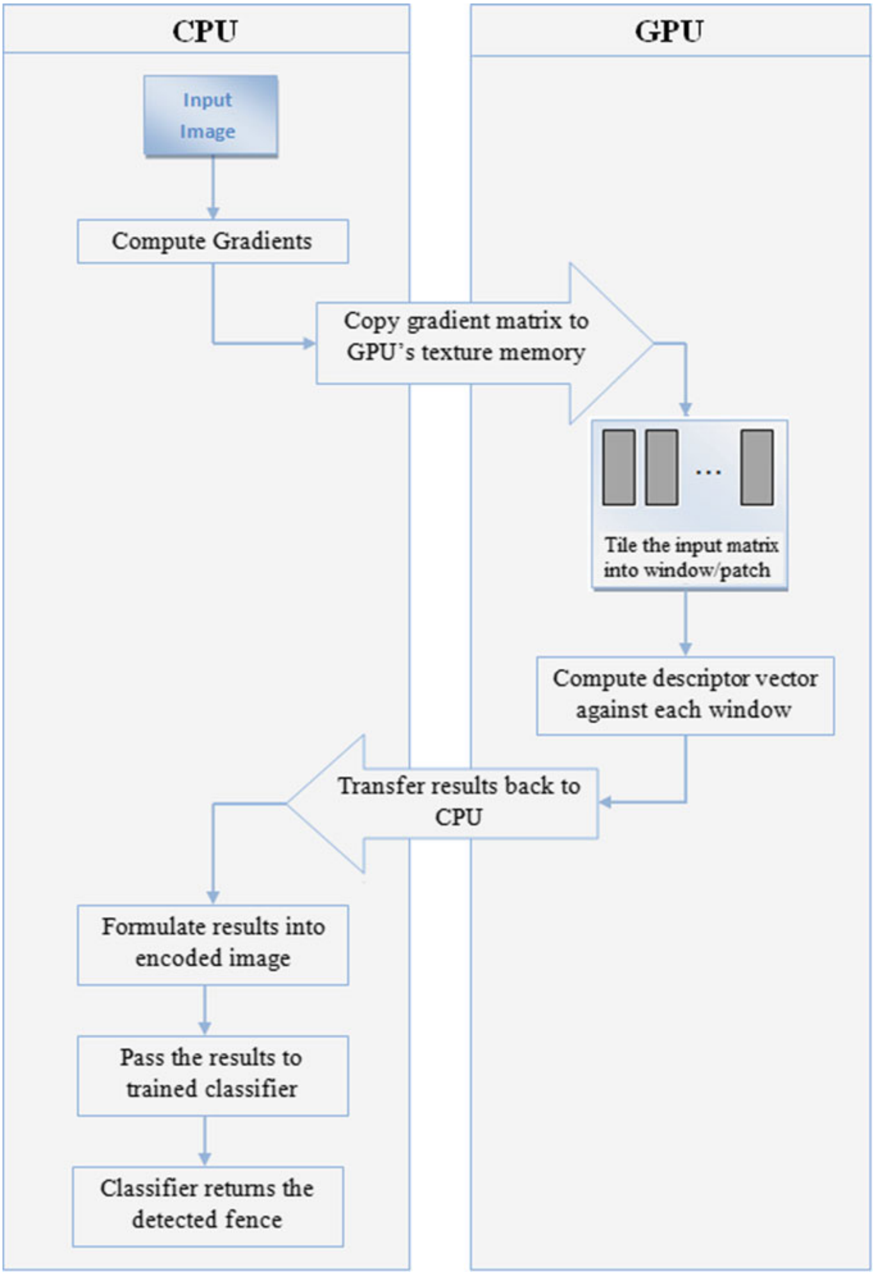


Fig. 3 An overview of GPU implementation

5 Results

The runtime analysis of the algorithm is made on both CPU and GPU. We considered 8 images for tests (as shown in Fig. 4). Table 1 shows the runtime comparison of both implementations. Comparison of CPU runtime measurements and GPU runtime measurements shows that the parallel implementation outperforms serial implementation in all cases with an average speedup of 15.

Figure 5 compares the execution times of the images against the image size. It has clearly shown that runtime is significantly decreased as the image size

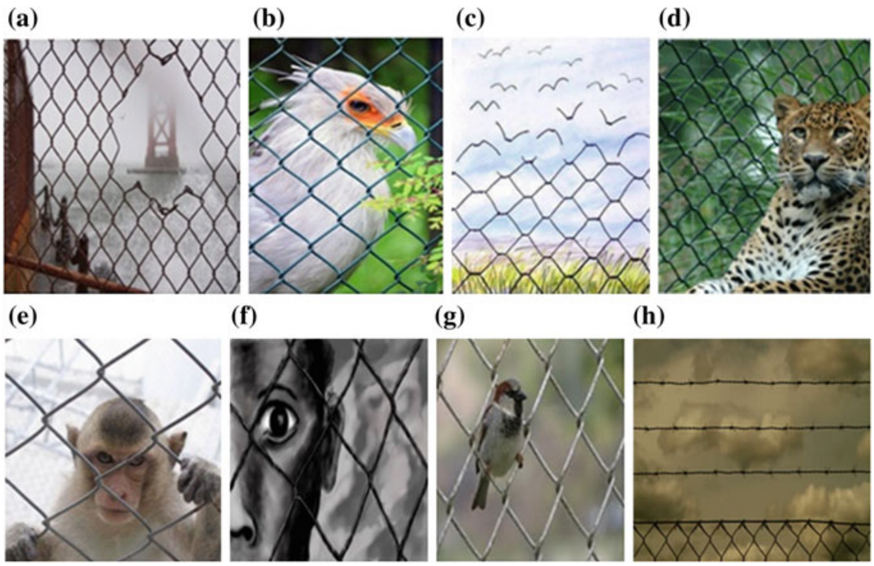


Fig. 4 Images used for runtime analysis. **a** Sea port, **b** White Pigeon, **c** Sky, **d** Lion, **e** Monkey, **f** Human Face, **g** Sparrow, **h** Clouds

Table 1 Runtime analysis of proposed technique

Exp.	Image name	Size (pixels)	CPU time (min)	GPU time (min)	Speedup
1	Sea port	425 × 640	46.3	3.28	14.1
2	White Pigeon	640 × 427	49.5	3.39	14.6
3	Sky	373 × 345	22.6	1.30	17.4
4	Lion	263 × 224	10.0	0.40	25.0
5	Monkey	446 × 452	39.5	2.42	16.3
6	Sparrow	194 × 259	8.28	0.46	18.0
7	Human Face	183 × 275	8.44	1.27	6.65
8	Clouds	224 × 400	14.3	1.42	10.1

Results of 8 images with their sizes in pixels, parallel and serial execution times and the speedup achieved against each image

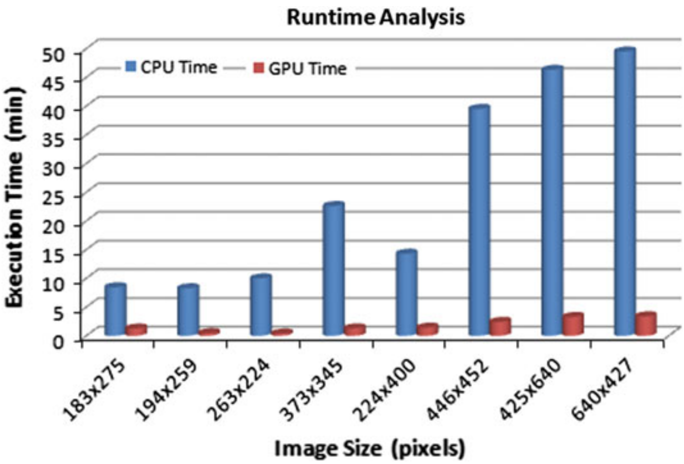


Fig. 5 A comparison of parallel versus serial execution time of the proposed technique (Execution time against image size)

is increased. Furthermore, the parallel implementation is given while keeping the scalability of the system in mind, therefore the implementation can further reduce the GPU execution time as the number of cores increases.

6 Conclusion

A parallel implementation of a feature based image de-fencing technique is proposed and the results have shown that the parallel technique is computationally more efficient. The parallel algorithm was implemented on GPUs using matlab and CUDA and an average speedup of 15 was observed over sequential implementation.

Additionally, we theoretically compared parallel approaches on different platforms. These approaches are not implemented and the real time results are not presented because this is beyond the scope of this paper. This would be a subject of entire worthwhile research.

For future work, a parallel implementation of the complete framework of the algorithm can be given, as the algorithm also involved inpainting method and training classifier that can be computationally intensive for large number of images.

References

1. Liu, Y., Belkina, T., Hays, J., Lublinerman, R.: Image de-fencing. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Anchorage, Alaska, pp. 1–8 (2008)
2. Hassan Khan, M.: Some New Approaches to Image Inpainting. M. Phil. thesis, Punjab University College of Information Technology, University of the Punjab, Lahore, Pakistan, pp. 9–24 (2011)
3. Park, M., Brocklehurst, K., Collins, R.T., Liu, Y.: Image de-fencing revisited. In: Proceedings of the 10th Asian Conference on Computer Vision, ACCV'10, vol. IV, Berlin, Heidelberg, pp. 422–434 (2010)
4. Khalid, M.: Parallel Image De-fencing: Technique, Analysis and Performance Evaluation. M. Phil. thesis, Punjab University College of Information Technology, University of the Punjab, Lahore, Pakistan, pp. 17–36 (2013)
5. Farid, M.S., Khan, H., Mahmood, A.: Image inpainting based on pyramids. In: IEEE 10th International Conference on Signal Processing 2010 (2010)
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of IEEE Conference Computer Vision and Pattern Recognition, San Diego, USA, pp. 886–893 (2005)
7. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Proceedings of the European Conference on Computer Vision, Graz, Austria (2006)
8. Prisacariu, V., Reid, I.: FastHOG—A Real-Time GPU Implementation of HOG. Technical Report 2310/09, Department of Engineering Science, Oxford University (2009)
9. Sugano, H., Miyamoto, R., Nakamura, Y.: Optimized parallel implementation of pedestrian tracking using HOG features on GPU. In: Conference on Ph.D. Research in Microelectronics and Electronics (PRIME) (2010)
10. Farhadi, M., Motamedi, S.A., Sharifian, S.: Efficient human detection based on parallel implementation of gradient and texture feature extraction methods. In: Machine Vision and Image Processing (MVIP) (2011)
11. Zhu, Q., Avidan, S., Yeh, M.C., Cheng, K.T.: Fast human detection using a cascade of histograms of oriented gradients. In: Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, pp. 1491–1498 (2006)
12. Luo, H., Xin-Rong, Y., Hong-Mei, L., Qing-Hai, D.: A method for real-time implementation of HOG feature extraction. In: Proceedings of SPIE, The International Society for Optical Engineering (2011)

Design of Folded Planar Inverted-F Antennas with Stair-Shaped Radiator for LTE700

Muhammad Feroze Akbar J. Khan, Ahmad Rashidy Razali, Aslina Abu Bakar, Mohamad Amir Imran b. Mohd Hasli and Mohd Amininudin Murad

Abstract A miniaturized Planar Inverted Antenna F-Antenna for LTE700M service band is presented. The main radiator of PIFA antenna is been folded into three radiating arm which provide optimized dimensions. Then, stair-shaped geometry have been used in order enhance impedance bandwidth for both lower band and higher band. Finally, the size of the finite ground is gradually reduced to achieve optimum size reductions. In the experimental results, the antenna achieves lower band impedance bandwidth between 725 MHz and 1.18 GHz at 6 dB return loss. At upper band, the antenna covers bandwidth between 2.33 and 3.82 GHz at 10 dB return loss. The final antenna #2 achieved reduction in overall dimension by 46.6 % in length, 33.7 % in width, and 5.19 % in height as compared to its reference antenna. Hence, this miniaturized size of proposed PIFA antenna is a promising candidate for several wireless services such as LTE700, GSM850/900, WLAN2.45G and WiMAX3.5G.

Keywords LTE · PIFA antenna · Miniaturizations · Folded and stairs shape radiator

1 Introduction

The challenges for modern portable devices are getting higher as most of devices are going for compact and slimmer design. Most of cellular communication system are using microstrip and printed antenna that is flat in nature of appearance. Planar Inverted-F antenna (PIFA) is widely used due to its flat nature and PIFA capable to

M.F. Akbar J. Khan (✉) · A.R. Razali · A.A. Bakar · M.A.I.b.M. Hasli · M.A. Murad
Antenna and Microwave Research Group, Universiti Teknologi MARA, 13500 Permatang Pauh, Penang, Malaysia
e-mail: feroze.aman@gmail.com

M.F. Akbar J. Khan
Motorola Solution Malaysia Sdn Bhd, 11900 Bayan Lepas, Penang, Malaysia

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_84

cover several wireless communications services such as WiMax, WLAN, Long Term Evolutions (LTE), Bluetooth and many more [1–5].

PIFA has been introduced in 1987 by Taga for GSM Band Applications [6, 7]. PIFA was known for its inverted F-shape antenna with a shorting arm to its ground plane [2, 3, 8]. PIFA is an improvement from monopole antenna where the top radiating arm is been folded to reduce its size which forms F-shape [8]. In late 2009, authors in [9] has proposed Coplanar Inverted-F Antenna (CIFA), where the radiator and the ground plane are on the same planar substrate [9, 10, 5]. Figure 1 shows fundamental antenna of PIFA design. In this antenna, the folded radiator creates capacitive loading with the ground edge. Thus, an inductive shorting pin to ground is introduced to compensate the imbalance reactance.

On the other hand, compared to monopole, rod and helix antenna, PIFA has an advantage of easy integration with most of handheld devices [2]. However, compact PIFA antenna that covers lower WWAN band such as LTE700M are difficult to design due to its large radiator and ground size. As the resonant frequency goes lower, the overall antenna dimension becomes larger [8], and it will be difficult to integrate it into portable devices such as smart phones.

The limitation in PIFA is due to its narrow bandwidth achievement at lower band [1–4, 6, 8] and its overall dimensions. In literatures, several techniques have been studied in order to enhance the impedance bandwidth; one of the technique that are used is ground slot near the main radiator in order to improve the operating bandwidth [1–4, 9–13]. Other bandwidth enhancement techniques such as addition of stairs near the main radiator have been explained in [14–16]. Addition of stair creates longer electrical current path, hence creates better impedance bandwidth. Furthermore, fractal and spiral radiator techniques have also been used to achieve compact and multiband features as reported in [9, 17].

In this paper, two antennas of new PIFA designs that mainly design for LTE700M with additional capabilities to support others wireless services such as GSM900, WLAN2.5G and WiMAX3.5G are proposed. The antenna introduces a stair-shaped PIFA radiator design to increase the bandwidth. The antenna provides dual band operations with nearly omnidirectional radiation characteristics and good impedance matching.

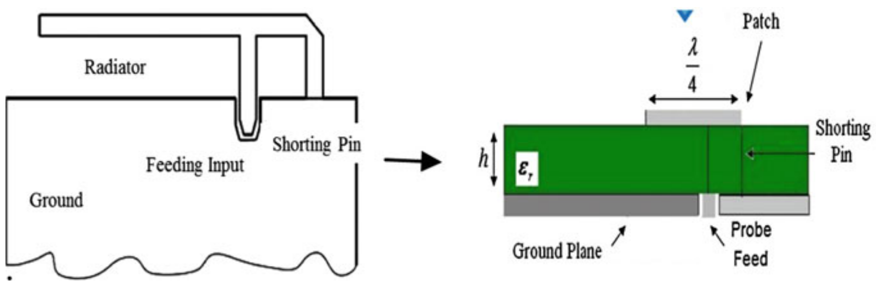


Fig. 1 PIFA configuration. Basic design (*left*) and antenna cross-section (*right*)

2 LTE700M PIFA Antenna Designs

PIFA is created at quarter wavelength of the lowest desired frequency, in this case, 690 MHz. The radiator is folded into three stages in order to reduce its size, which named as Arm1, Arm2 and Arm3. Figure 2 shows the initial design of PIFA. Folding main radiator into several arms creates capacitive loading which shift the resonant frequency. A shorting arm is introduced to compensate the capacitive loading that is created after folding the main radiator parallel to the ground at the bottom layer. The antenna is created on an FR4 with dielectric constant $\varepsilon = 4.3$ and thickness = 1.6 mm. The antenna design was modeled using CST Microwave Studio. Following subsection are the details of proposed PIFA designs.

2.1 Antenna #1: LTE700M PIFA Antenna

In this antenna, there are two approaches used that are the width arm increment and addition of stairs at radiator in order to enhance the bandwidth for LTE700M. Figure 3 shows the front and back view of PIFA antenna structure. This overall antenna dimension (*l x w x h)* is 125 × 60 × 1.634 mm, where the copper thickness is 0.017 mm and substrate thickness is 1.6 mm. At bottom layer, the finite ground size is 93 × 60 mm. Meanwhile, at the top layer, there are two stairs added at the radiator. The stair 1 dimension is 16 × 9 mm, and stair 2 dimension is 14 × 10 mm.

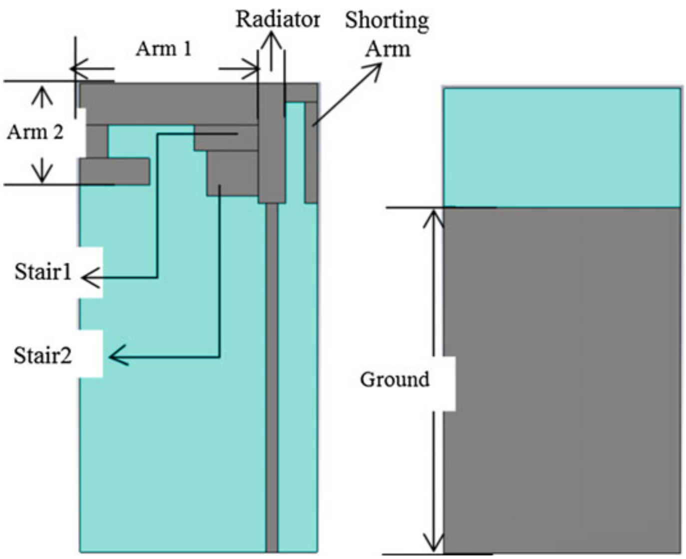


Fig. 2 Proposed LTE antenna design descriptions

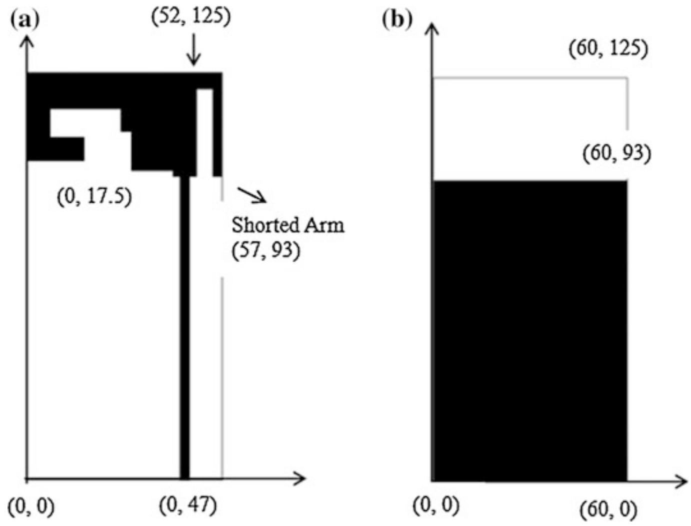


Fig. 3 Antenna #1 design geometry. **a** Front view. **b** Back view

The PIFA is fed by 50 Ω microstrip feedline that is positioned at (0, 47) and width of 3 mm. The shorting arm was located at (57, 93) and width of 3 mm.

2.2 Antenna #2: Miniaturized of LTE700M PIFA Antenna

Based from antennas #1, the size of the antenna finite ground plane is miniaturized in order to have compact design. In the miniaturization, the length of the ground is reduced from 125 to 115 mm, a 10 mm length reduction. Due to this reduction, the antenna stair-shaped geometry is re-optimized to achieve the original LTE700M band resonance. After miniaturization, the overall antenna (l × w × h) is 115 × 60 × 1.634 mm and stair 1 and stair 2 dimension is 22 × 6 mm and 15 × 13 mm respectively (Fig. 4).

3 Simulation and Fabrication Results

3.1 Results of Antenna #1

With the introduction of stair-shaped geometry at the radiator, there is a significant frequency shifting and bandwidth enhancement that can be observed in Fig. 5a

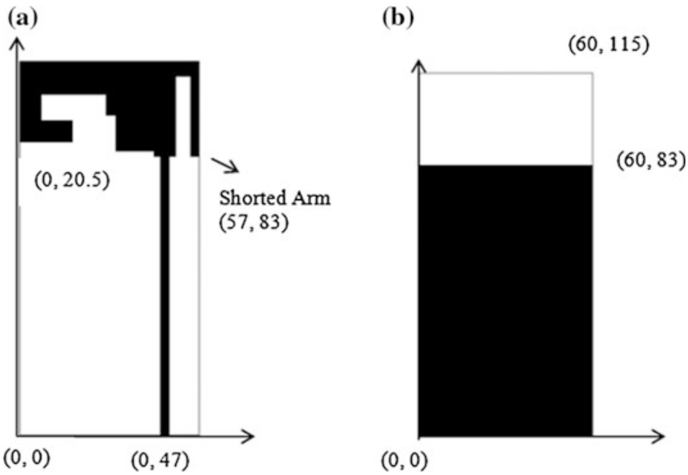


Fig. 4 Antenna #2 design geometry. **a** Front view. **b** Back view

compared to the antenna without the stair-shaped geometry. As simulated data shows the bandwidth achieves at −6 dB is from 0.68 to 0.96 GHz, which met the LTE700M requirements.

Figure 5b shows both the simulated and measured results for the input reflection coefficient of antenna #1. Good correlation has been observed between measured and simulated results.

The measured results show that minor shift in the lower band LTE700M which may probably due to fabrication error. The measured results of antenna #1 at −6 dB return loss is covers from 725 up to 980 MHz, and at −10 dB return loss it cover from 2.6 up to 4.2 GHz.

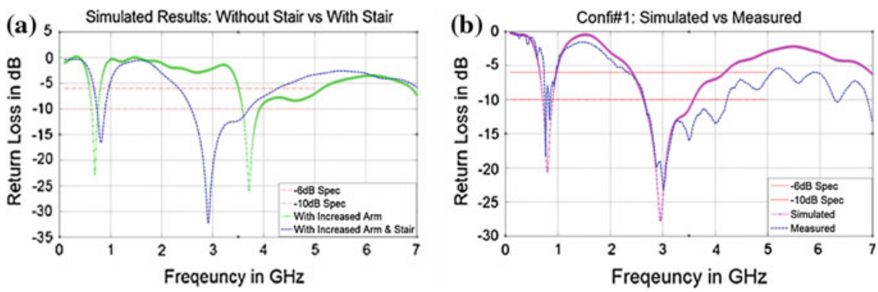


Fig. 5 Antenna #1 results. **a** Simulated results with and without stairs. **b** Simulated versus measured data

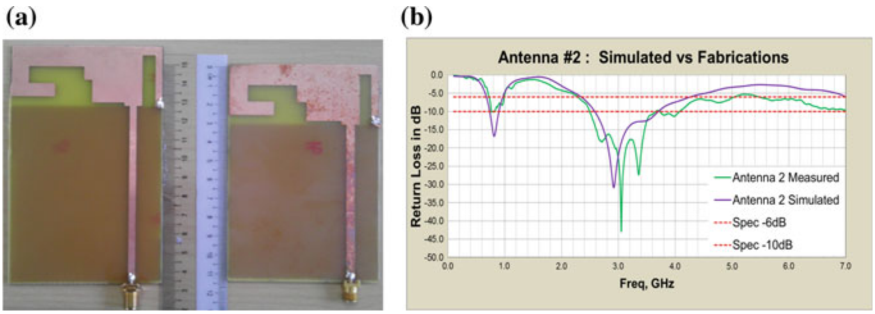


Fig. 6 Antenna #2 results and antenna design. **a** Antenna #1 and #2 design geometry. **b** Simulated versus measured data

3.2 *Results of Antenna #2*

This antenna is the improvement of antenna #1, whereby the total length was reduced by 10 mm. Figure 6a shows the geometry difference observed between antenna #1 and antenna #2.

Figure 6b shows both the simulated and measured results for the input reflection coefficient of antenna #2. The measured results show that minor shift in the lower band LTE700M. The measured results of antenna #2 at −6 dB return loss is covers from 720 up to 980 MHz, and at −10 dB return loss it cover from 2.38 up to 4.1 GHz.

4 Analysis and Discussion

The folded mechanism is basically for the purpose of miniaturization process where it help to reduce antenna overall dimension. The shorting pin plays an important role to achieve zero reactance at the desired resonance frequency. Figure 7a shows

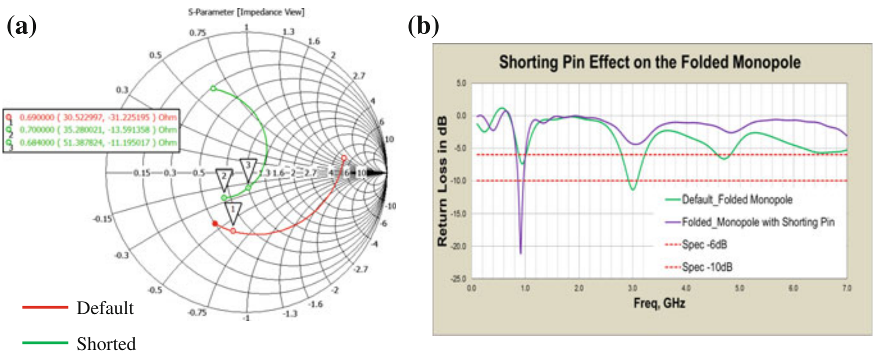


Fig. 7 Shorting pin effect analysis. **a** Smith chart characteristic. **b** Shorting pin effect in term of S11 response

the Smith Chart characteristic where default impedance point is located at −ve reactance (capacitive). As mentioned by author in [9, 10], by adding shorting pin to the ground, it help to improve reactive excitation thus the overall impedance point near zero. Figure 7b shows the S11 parameter response of shorting pin effect.

By having the shorting pin, PIFA achieve better impedance matching. The introduction of stairs act as series capacitance which helps broaden the bandwidth from 690 to more than 960 MHz. Figure 8 shows the effect of addition of stairs in term of Smith Charts.

The simulated surface current distribution also been studied for PIFA antenna #2 for three frequency band which are f_1 = 700 MHz, f_2 = 800 MHz and f_3 = 2400 MHz. As shown in Fig. 9a–c, the excited surface currents flowing near the stairs have maximum strength and begin to decreases slowly before turn to max peak again at Arm3. The longer path taken by this surface current creates better impedance matching at particular frequencies.

Figure 10 shows the simulated VSWR characteristic of both antennas. Typical antenna design required VSWR characteristic of less than 3:1, whereby lower VSWR indicates better reflection coefficient, hence better impedance matching.

In term of excitation position, antenna #1 center resonant frequency occurs at 764 MHz and 3.01 GHz, while antenna #2 center resonant frequencies occurs at 834 MHz and 3.03 GHz. Table 1 shows the design summary of both PIFA antennas.

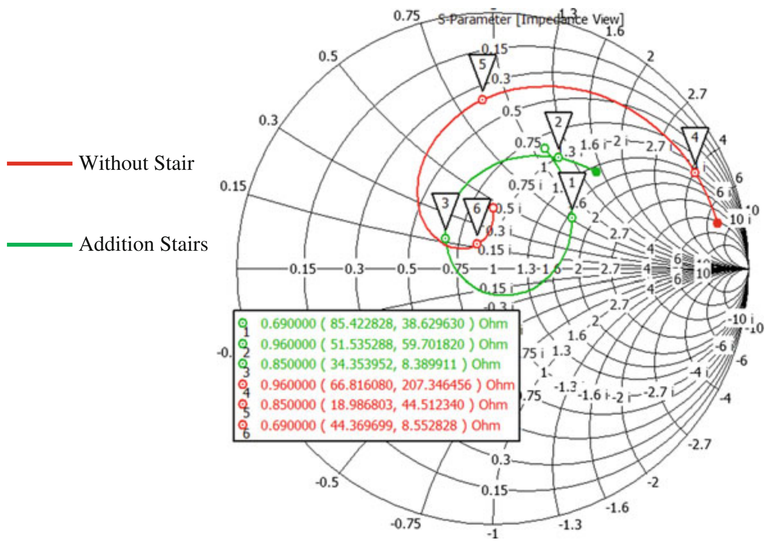


Fig. 8 Addition of stair effect based on Smith chart

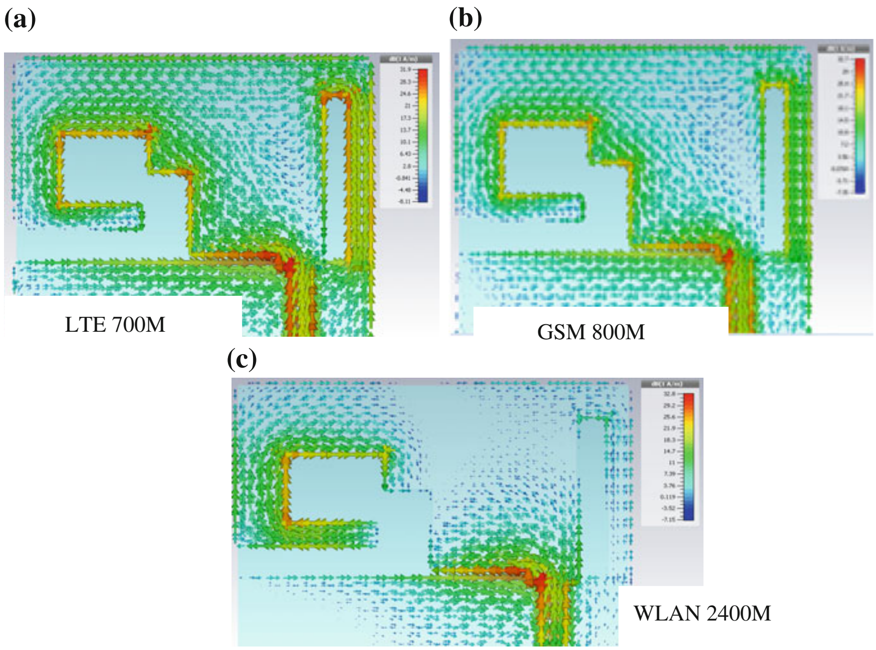


Fig. 9 Simulated surface current distribution on PIFA Antenna #2. **a** f = 700 MHz. **b** f = 800 MHz. **c** f = 2400 MHz

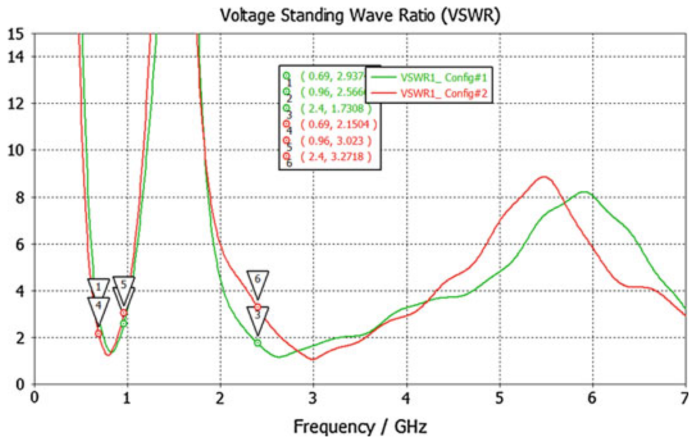


Fig. 10 Simulated VSWR characteristics for proposed antennas

Table 1 Summary of fabricated results

Type of antenna	Antenna #1		Antenna #2		Units
	Simulated	Measured	Simulated	Measured	
Bandwidth for LTE700M	0.65–0.96	0.715–0.97	0.689–0.96	0.725–1.18	GHz
Bandwidth for higher resonance	2.66–3.54	2.672–4.23	2.34–3.30	2.33–3.82	GHz
Dimension (*w x l x h*)	60 × 125 × 1.634		60 × 115 × 1.634		mm
Applications	LTE700, GSM8/900 and WLAN		LTE700, GSM8/900, LTE2600M, Bluetooth, WLAN, WiMAX		

4.1 Radiation Pattern: Antenna #1

The simulated radiation pattern of the antenna is shown in Figs. 11 and 12. The patterns are nearly omnidirectional which is suitable for wireless portable devices.

4.2 Radiation Pattern: Antenna #2

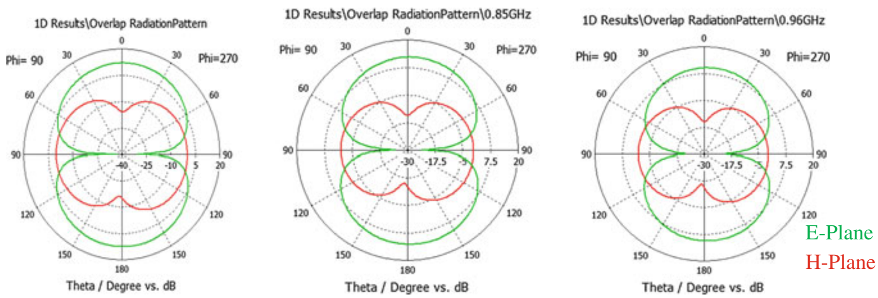


Fig. 11 Simulated 2D radiation pattern for 700, 850 and 960 MHz in antenna #1

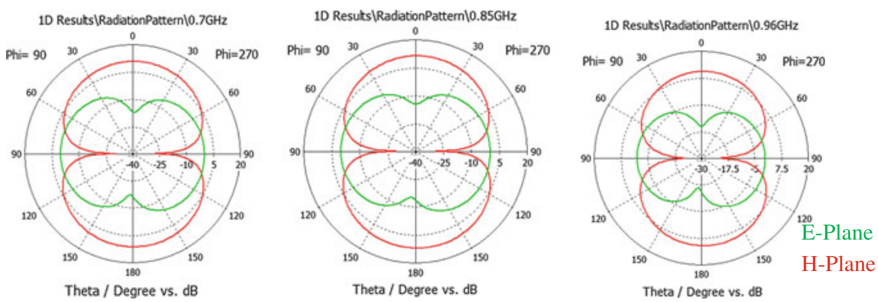


Fig. 12 Simulated 2D radiation pattern for 700, 850 and 960 MHz in antenna #2

5 Conclusions

In this paper, two new PIFA that operates at LTE700M have been presented. PIFA Antenna #1 operates at 0.715 up to 0.96 GHz at 6 dB return loss reference and at 2.672 up to 4.23 GHz at 10 dB return loss. Meanwhile, PIFA Antenna #2 operates at 0.72–1.18 GHz at 6 dB return loss and also operates 2.32–3.82 GHz at −10 dB return loss reference. The overall Antenna #2 size has achieved reduction in volume by 46.6 % in length, 33.7 % in width, and 5.45 % in height when compared to its reference antenna [18]. Therefore, with its compact size and easy integration with modern handheld devices, the antenna is a promising candidate for several wireless services such as LTE700, GSM850/900, WLAN2.45G and WiMAX3.5G.

Acknowledgement The authors would like to acknowledge the Ministry of Science and Technology (MOSTI), Malaysia through EScience Research Grant Project No. 04-01-01-SF0620 that supports this work.

References

1. Wang, H., Zheng, M.: A multi-band internal antenna. In: Loughborough Antennas and Propagation Conference, UK, (2011)
2. Kwak, W.I., Park, S.O., Kim, J.S.: A folded planar inverted-F antenna for GSM/DCS/Bluetooth triple-band application. IEEE Antennas Wirel. Propag. Lett. **5**(1), 18–21 (2006)
3. Lee, C.T., Wong, K.L.: Uniplanar printed coupled-fed PIFA with a band-notching slit for WLAN/WiMAX operation in the laptop computer. IEEE Trans. Antennas Propag. **57**(4), 1252–1258 (2009)
4. Gomez-Villanueva, R., Jardon-Aguilar, H., Linares-y-Miranda, R.: Broadband PIFA antenna for mobile communications terminals. In: IEEE, Campeche, pp. 1–6, 29 Sept 2014–3 Oct 2014
5. Bialkowski, M.E., Razali, A.R., Boldaji, A.: A folded quarter-elliptical wideband antenna for portable devices. In: IEEE, Toronto, ON, pp. 1–4, 11–17 July 2010
6. Taga, T., Tsunekawa, K.: Performance analysis of a built-in planar inverted F antenna. IEEE J. Selected Areas Commun. **5**(5), 921–929 (1987)
7. Taga, T., Tsunekawa, K.: A built-in diversity antenna for 800 MHz band portable radio units. In: Antennas and Propagation Society International Symposium, 1986, vol. 24, pp. 705–708, Philadelphia, PA (1986)
8. Ray, J.A., Chaudhuri, S.R.B.: A review of PIFA technology. In: Indian Conference on IEEE Antenna Week (IAW) 2011, Kolkata, pp. 1–4, Dec 2011
9. Razali, A.R., Bialkowski, M.E.: Coplanar inverted-F antenna with open-end ground slots for multiband operation. IEEE Antennas Wirel. Prop. Lett., **9**, 1029–1032 (2009)
10. Razali, A.R., Bialkowski, M.E.: Design of a dual-band Microstrip-Fed Meandered-tail PIFA for WLAN Applications. In: Antennas and Propagation Society International Symposium, 2009. APSURSI'09. IEEE, Charleston, pp. 1–4, 1–5 Jun 2009
11. Alja'afreh, S.S., Huang, Y., Xing, L.: A compact, wideband and low profile planar inverted-L antenna. In: 8th European Conference on Antennas and Propagation (EuCAP), 2014, The Hague, pp. 3283–3286, 6–11 Apr 2014
12. Chen, Z., Ban, Y.-L., Chen, J.-H., Li, J.L.-W., Wu, Y.-J.: Bandwidth enhancement of LTE/WWAN printed mobile phone antenna using slotted ground structure. Prog. Electromagnet. Res. **129**, 469–483 (2012)

13. Razali, A.R., Bialkowski, M.E., Boldaji, A., Tsai, F.-C.E.: Invention on a mulitband microstrip-fed inverted-F antenna with ground slot. Microw. Opt. Technol. Lett. **58**(8), 1713–1719 (2010)
14. Sobli, N.H.M., Abd-El-Raouf, H.E.: Design of a compact band-notched antenna for ultrawideband communication. In: Antennas and Propagation Society International Symposium, 2008. AP-S 2008. IEEE, San Diego, pp. 1–4, 5–11 July 2008
15. Islam, M.R., Alam, A.Z., Khan, M.F.A.J., Alkaraki, S..: Design of microstrip patch antenna using slotted partial ground and addition of stairs and stubs for UWB application. Cyber J. Multidisciplinary J. Sci. Technol., JSAT (2012)
16. Raj, R.K., Gurhar, S., Sharma, M.: Design of stair and slotted UWB antenna using stepped-feed with modified slotted ground plane. Int. J. Comput. Appl.® (IJCA) (0975–8887), 22–25 (2014)
17. Nie, Y., Song, L.: A compact triband fractal PIFA antenna for mobile handset applications. In: 6th International Congress on Image and Signal Processing (CISP) 2013, vol. 03, pp. 1468–1472, Hangzhou, 16–18 Dec 2013
18. Loutridis, A., John, M., Ammann, M.: Dual band LTE planar inverted-F antenna for M2M applications. Microw. Opt. Technol. Lett. **55**(12), 2925–2929 (2013)

Establishing the State of the Art Knowledge Domain of Cloud Computing

Madiha Khalid, Muhammad Murtaza Yousaf, Yousra Iftikhar and Noreen Fatima

Abstract Cloud is an inflection point for customers to transit the burden of infrastructure and platform management to a service provider. There is virtual centralization in cloud computing. This paper provides the state of the art knowledge domain of cloud to developers and researchers. It gives a brief overview of cloud computing and how it relates with grid, cluster, distributed, parallel and GPU computing. We have surveyed different cloud deployment models and essential characteristics of cloud that are required to build cloud architecture. A state of the art study is presented that explores several cloud architectures including discussion of applications, platforms and shared infrastructure of cloud that unveils the sharing mechanism of physical services, networking capabilities and storage. Furthermore, this study covers the primary business service models for instance Infrastructure as a Service (IaaS), Platform as a service (PaaS) and software as a service (SaaS) with vendors of each layer. On the other hand some issues of clouds are also highlighted such as availability of cloud and challenges which include issues regarding security, lack of standards, privacy and location concerns. We analyze several cloud architectures on multiple performance measuring parameters such as scalability, security, cost, performance, reliability, maintenance, and mobile accessibility.

M. Khalid (✉) · M.M. Yousaf
Punjab University College of Information Technology, University of the Punjab, Lahore, Pakistan
e-mail: madiha.khalid@pucit.edu.pk

M.M. Yousaf
e-mail: murtaza@pucit.edu.pk

Y. Iftikhar
Oxford University Press, Oxford, UK
e-mail: Iftikhar.yousra@yahoo.com

N. Fatima
i2c Inc., Lahore, Pakistan
e-mail: nfatima@i2cinc.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_85

1 Introduction

Over the past few years, cloud computing has become the source of attention for the people. It is a computing paradigm which is widely accepted. Cloud computing is widely adopted by the industries. The main focus of this paper is on the state of the art knowledge regarding cloud. There are so many issues regarding cloud computing which are still not addressed. So our agenda is to focus those issues as well. The rest of the paper is organized as follows: Sect. 2 gives a comparative analysis of different computing domains like Parallel and Distributed Computing, Cloud Computing, GPU Computing, Cluster Computing and Grid Computing. Section 3 gives a complete overview of cloud computing. Section 4 presents a suitable and most comprehensive architecture of cloud covering all the basic and important layers (like SaaS, PaaS, IaaS). Section 5 consternates on the main features of cloud computing. Sections 6, 7 and 8 covers architecture, vendors and essential characteristics respectively. Sections 9 and 10 focus on opportunities and risks associated with cloud. Finally Sect. 11 concludes the study.

2 Cloud and Related Domains

In this section we have provided a brief overview of common computing models and how they are related to each other. We focused on the similarities and some of the differences of these computing with each other. These computing models are:

- Distributed computing
- Grid computing
- Parallel computing
- Cluster computing
- GPU computing

Cloud computing [1] makes the services like software, information, architecture and other resources available to users over the network. Cloud computing have key characteristics which include quick access, user friendly application interface, security, reliability, cost effectiveness, performance, scalability and maintenance. Some cloud issues are related to privacy, legal compliance, open source, open standards, security and sustainability.

In distributed computing [2], multiple computing nodes communicate with each other by passing messages to achieve a common goal through a network. In this computing, it is not know in advance that how to organize the network links and these network links might change during the execution of a distributed program. The main focus in distributed computing is on communication operations than computational steps. The most simple and the basic model of distributed computing on which it works is synchronous system, the communication pattern it follows is, receive messages from its neighboring nodes, perform computation relevant to message and send the response message to neighbors. Challenges that distributed computing is facing include fault tolerance and understanding the asynchronous nature of distributed systems. The architectures used for distributed computing maybe hardware or software included 3-tier, n-tier, client server, distributed nodes. Work can be communicated in distributed systems among simultaneous processes through message passing protocols.

Grid computing [3, 4] is similar to distributed systems in which different computers interact with each other by connecting to a network. Middleware is used in grid computing to divide sub-tasks of a program to many different computers. Grid computing collects different computers from many places to fulfill the mutual goal while solving a job. The participating nodes must have the trust on central system.

Parallel computing [5] divides larger problems or applications into smaller sub-problems and then they are solved in parallel. Parallel computing gives direct access of shared memory to all processors. Parallel computing has multi-core processors (single computing component with more processors) within a single machine while grid computing uses multiple computers for a single task to be performed. These programs are difficult to write because everything is working in parallel so we need to synchronize the tasks. One problem is to break a task into several small sub-tasks so that each processing node can execute one sub-task. There is shared as well as distributed memory. Also, there is local memory as well as the non local memory while the access to the local memory is faster than the non local.

Cluster computing [6], many computers are connected through each other making a form of a single computer. Their components are mostly connected through the local area network. Clusters improve the availability as well as the performance of computers. They are cost effective. When the user's request comes then it is divided among all the computers which form a cluster resulting in balanced computational work for better performance.

GPU computing offers the use of graphical processor unit to accelerate the performance of general purpose applications. GPU generally works with CPU, CPU offload compute intensive part to GPU to achieve higher performance gains. GPU have massively parallel architecture that has thousands of simple homogenous cores designed to execute several thousand threads in parallel. There are memory hierarchies ranging from local memory to shared memory in GPUs for handling efficient data storage. Table 1 compares all the discussed cloud related computing models on different features like computation, cost, centralization, reliability, scalability, maintenance and speed.

Table 1 Comparison of cloud related domains

	Cloud computing	Grid computing	Parallel computing	Distributed computing	Cluster computing	GPU computing
Distributed application	✓	✓	✓	✓	✓	✓
Parallel Computation	It supports parallel computations	It supports parallel computations	It supports parallel computations	It supports parallel computations	It supports parallel computations	It supports parallel computations
Cost	Lower cost	Lower cost	Cost effective, expensive if shared	Cost effective on power consumption	Cost effective	Cost effective
Centralization	Centralization of infrastructure	Centralization of infrastructure	Centralized	Centralized	Centralized	Centralized
Reliability	Multiple redundant sites are used so it is reliable	Hardly reliable, due to wide area distributed characteristics and heterogeneous data	Hardly reliable, parallel working	Hardly reliable, due to distributed characteristics	Hardly reliable, due to distributed characteristics	More reliabile due to the use of GPU
Scalability	Easily scalable	Easily scalable	Easily scalable	Easily scalable	Easily scalable	Not scalable fixed number of GPU and CPU are used in it
Security	Partially secured. If data is distributed over wide area then security issues arise	Several security policies are implemented in grid computing	It is secure and privacy is maintained	Several security mechanisms work together to secure distributed systems	It is secured. As the redundant nodes are used to reduce the occurrence of failure	It is secure and can identify virus for thousands of data objects in parallel

(continued)

Table 1 (continued)

	Cloud computing	Grid computing	Parallel computing	Distributed computing	Cluster computing	GPU computing
Performance	Performance is monitored and consistent	It gives high performance	High performance computing	High performance	High performance	Very high performance
Middleware	There are data centers and users can access their data through web application	There is middleware between the hardware and the software	No middleware	Use of communication network that connects several computers	MPI and PVM are being used as middleware	Middleware used in software development tools for GPU
Speed	High speed	High speed	High speed	High speed	High flops but whole data can't be accessed at once	Together use of GPU and CPU makes it high speed
Models	SaaS, PaaS, IaaS	It uses SaaS only	Shared memory model, threads model, data parallel model, message passing model, hybrid model	Can be implemented using message passing model	–	CUDA parallel programming model

3 Cloud Computing

Cloud computing is a general term that refers to providing services to the users via web technology. These IT services are scalable, efficient, agile, fast and elastic. It makes the services such as resources, infrastructure, software, and information available to computers over the network not requiring the location of end user. Cloud represents a delivery model in IT services comprise on Internet protocols that provides the easy and reliable access to remote computing sites available on internet, that acts as a web based tool and we can make it available on local computers if the software is installed. Cloud computing have key characteristics which include agility, application program interface, cost, device, reliability, scalability, performance, security, maintenance. There are five layers in cloud computing that includes software as a Service (SaaS), Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and servers that consists of computer hardware and softwares typically designed for the delivery of cloud services. These are also considered as a service model for cloud.

4 Deployment Models

Clouds can be deployed in following forms [7]:

- Public cloud (available to arbitrary companies)
- Community cloud (operated by group of companies)
- Private cloud (operated by single organization)
- Hybrid cloud (mix two or more clouds)

Armbrust et al. [8] proposes definitions for public cloud as cloud assessable in a pay-as-you-go manner to the general public. He says about private cloud as "Internal data center of a business or other organization not made available to the general public." According to him a community cloud is "shared by several organizations and supports a specific community that has shared concerns (e.g., mission, security requirements, policy, and compliance considerations)." When a private cloud is complimented with computing capability from public clouds, it acts as hybrid cloud. It depends on the nature of the organization that which cloud model or models can be implemented in that organization.

5 Service Models

Layered architecture of cloud consists of three general layers of cloud architecture named as SaaS, PaaS and IaaS [9]. Along with these layers, some more layers are also identified in literature as Data as a Service (DaaS), Hardware as a Service (HaaS) and Communication as a Service (CaaS). Physically Cloud architecture is similar to the stack and pyramid [10], in which services are classified in different layers with respect to their role in cloud architecture. Here we have discussed the service level layers with the help of examples to understand the cloud architecture. The layers are managed in the way that, the higher layers contain the services of the lower layers. More than one layer can be on the same level, if they provide same services.

5.1 Software as a Service

The first layer of cloud architecture, in which central location is provided to users and they can access services through internet. Cloud applications act as a substitute for general applications. It is the layer in cloud architecture which is visible to the users and they can actually access these services on the concept of pay as you use.

A model in which, software's access are provided on the central location and users can access it through internet. Users are not allowed to make changes in these softwares. Software that have very less or no interaction with other systems and can perform simple tasks fall in the category of SaaS. There are some other advantages of SaaS which include its customization factor of software for unique customers, SaaS application developer can earn more as compared to the previous methods of creating software and providing it in market and SaaS includes software access through web which makes it more reliable. Problems related to SaaS are its flexibility that can create issue for application vendors for providing users the opportunity to customize application up to their needs. Examples for SaaS are virtual Microsoft windows, web email, Google Docs and Microsoft.

5.2 Platform as a Service

It is resource on demand providing platform that provides web based access of software. Its basic task is to provide platform for making a complete cycle of developing web applications and service on internet without installing any software on local systems. PaaS can be used for supporting the operating systems and the development of software. Certain problems which PaaS is facing due to its portability, software developer cannot easily switch its cloud provider. He may have to pay high cost for it. Examples of PaaS are Google App Engine (GAE), Microsoft's Azure etc.

5.3 Infrastructure as a Service

It is an out source, through which customers can have access to their storage, hardware, servers and networking components by providing web hosting services. Providers are responsible for its service and maintenance. Advantages of IaaS are simultaneous access of the same data to multiple users across the world, online software for data storage and maintenance purpose. It contains flexibility of computing resources. It's most important benefit is in business because user has to pay according to usage. It maintains all the factors of low cost, easily manageable and low risks. Cloud providers can make infrastructure according to the customer requirements by virtualization.

5.4 Communication as a Service

It is difficult for business owners to select the way of communication now-a-days, traditional methods of communication are costly and less reliable. CaaS model of cloud is best suited in such situations, with the benefit of pay as you use. It is scalable and flexible for small business as well. In case of CaaS, it is provider's responsibility to offer the latest version of hardware and software for the users to upgrade them. According to Gartner [11] CaaS provides annual market of 2.3 billion in year 2011.

6 Cloud Architecture

In this work several cloud architectures have been investigated. According to Rimal et al. [12] cloud architecture is in four layers and three modes. Four layers for cloud architecture are SaaS, PaaS, IaaS and HaaS [13]. Above four layers with three modes for cloud architecture are public cloud, private cloud and hybrid cloud.

According to Cisco cloud computing data center strategy [14] the cloud architecture is described in three layers and one base layer for this architecture. Cisco provides an eminent work in Infrastructure as a Service and particular work in Software as a service and Platform as a service layer. It represents the architectural layers of cloud computing connected through API's and repositories. Four main layers in this architecture are SaaS, PaaS, IaaS and IT foundation.

Youseff et al. [15] described cloud architecture in five layers. Bottom layer for cloud architecture is hardware layer which constitute on actual physical components for cloud system. Most top layer is of cloud application layer which acts as an interface layer for cloud users. Five layers in this architecture are Application (SaaS), Platform (PaaS), Infrastructure (Computations, Storage, and Communications), software kernel and Hardware layer (HaaS).

According to Lenk et al. [16] cloud computing technologies is distributed in two groups named as Object Management Group's CORBA and the Open Group's DCE. They provide programmable interfaces to facilitate complex remote procedure calls. The basic factor that differentiate cloud computing from those which were described earlier is that many web based successful applications launched from this cloud phenomenon like Facebook, Amazon and Google. According to the author in [16] the architectures which were proposed earlier are consists of very little details therefore they provide the first comprehensive architecture which entails the complete categorization of cloud layers similar to the complete stack. "Everything as a Service" is the name provided by the researchers for their reference stack. The complete cloud architecture provided by these researchers consists of HaaS, IaaS, SaaS, PaaS and Human as a service.

7 Vendors of Cloud

In this section we are providing a comparison of some of the vendors which are considered best in their respective layer. Google Apps is considered to be the best vendor among all that implements the SaaS layer. Salesforce is the vendor that implements the PaaS layer. Oracle [17] is the vendor that implements the IaaS layer and IBM is considered to be the best vendor for the storage as a service layer. These vendors are compared on the basis of some of the attributes such as availability, interoperability, maintainability, manageability, performance, reliability, storage, scalability, security, cost effectiveness and support. On the basis of these features a comparison in given Table 2.

8 Essential Characteristics

There are some essential characteristics of cloud computing that are discussed below [18].

On demand self-service: Without interacting with the human or the service provider you can easily get the services like email, application, network or the server services. These services can be Amazon Web Service, Google, Microsoft, Salesforce.
Broad network access: The capabilities of cloud can be accessed through the network and can be used using the standards and mechanisms. It promotes thin and thick platforms of client like mobile phones and laptops etc.
Rapid elasticity: The services of cloud can be elastically provisioned. One can quickly scale out and scale in, in some cases. The capabilities for the provisioning can be purchased by the consumer any time.
Resource pooling: By using multi-tenant model, many consumers are served by pooling together the computing resources. Physical and virtual resources are dynamically assigned and reassigned according to the demand of consumer. The resources are storage, processing, email services, virtual machines and network bandwidth.

Table 2 Comparison of different cloud venders

	Google Apps (SaaS)	SalesForce (PaaS)	Oracle (IaaS)	IBM (storage as a service)
Availability	✓	✓	✓	✓
Maintenance	Minimal maintenance is required	Automatic maintenance is supported	Automatic maintenance is supported	data is automatically maintained
Manageability	Management is easy	Effective management of your business	Self-managing database is used	Provides integrated web content management capabilities
Performance	Malware scanners, syncing large documents and slow network connection can impact the performance	Better performance	Better performance due to concurrency	High performance
Reliability	99.9 % uptime guarantees SLA. Top priority of Google	Provides reliable and secure services	Gives complete range of reliability features	Highly reliable
Storage	25 GB/account	Easy storage	Storage solutions are provided like SAN, NAS and Tape storage etc.	Storage with the databases and storage oriented application software
Scalability	Highly scalable web applications, easy to scale the traffic. Automatic scalability is there	Highly scalable	There is scalability, concurrency and portability	There are administrative and scalability features
Security	Secure HTTPS access, customizable policies	Provides better security measures	Provides secure services	There is security systems integration as well as security access managers
Cost effectiveness	It is cost efficient	Easy customizations with fewer costs	Reduced infrastructure costs and risks	Reduced cost of deployment
Supportability	Self-service online support	Supports complex database functions	Many essential support services are provided	Support tools and frameworks. Virtualization and multiple profiles are also supported

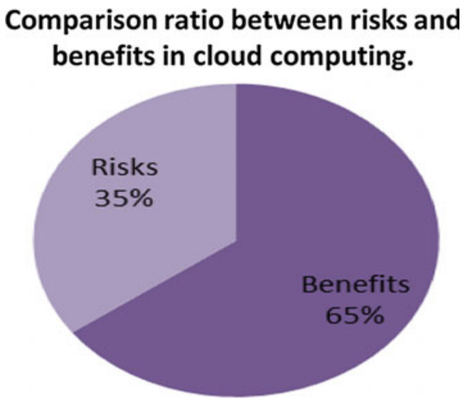
Measured service: The resource usage can be measured and controlled by providing the transparency for the consumer as well as for the provider. A metering capability is being used for the cloud computing services which are used for the controlling and optimizing the use of resources. There is the term used for the IT services and it is pay-per-use. Higher will be the bill, the more you utilize it.

9 Advantages of Cloud Computing

Cloud computing have various benefits, which make it favorable now a days. We have drawn a comparison ratio between cloud benefits and risks as shown Fig. 1. The most important benefits are as follows:

Scalability: Cloud provides a solution for the situations when number of users face increased computing needs. The solution is, User can buy extra CPU cycles and he has to pay for how much he will use, instead of buying a new CPU.
Simplicity and high speed: Cloud computing provides a simple use of software immediately as compare to those without cloud. Multiple Users can work on same software simultaneously because cloud provides a high speed network infrastructure.
Low cost: Services provided by cloud are offered at very low cost. Softwares can be downloaded for free or at a minimal cost as compared to commercial pricing. Infrastructure on cloud can be used on pay per use basis.
Reduce distance: Most favorable benefit of cloud is the shared services of cloud that reduces distance among people. They can share their data online, they can chat and share send data with each other.
Online storage: Through cloud computing, users have advantages to save their data on web that are secure too. Example is drop box and sky-drive etc.

Fig. 1 Comparison ratio between risks



Provide extra processing power: User has now advantage of buying extra CPU cycle to perform their tasks. For example, Oxford University Centre for Computational Drug Discoveries is using more than one million systems from around the world to cure cancer. Systems from all over the world donate few CPU cycles through screen-saver time.

10 Issues with Cloud Computing

Legal issues in cloud computing are wide ranging [19, 20]. One should go through the risks before moving their business into cloud. These risks include:

Security risks: Hacking and stolen user passwords or usernames is common these days. Managing secure passwords and usernames is a must for securing data at online sites such as cloud.
Data protection concerns: If any personal data of customer is being stolen then company has to face consequences. It is being prosecuted by national authorities and customers. One should have complete knowledge of data store so that in situation of any disaster owner of data should be able to have the backup of data [21].
Failure in provider security: There is a possibility that the service provider fails in securing data. Service provider is the one who is responsible for the data on cloud and sometimes there is failure in data security or availability. To ensure data security one should verify and monitor the service provider's security measures.
Legal and regulatory issues: One should evaluate the legal issues before shifting business to cloud. Sometimes there are geographical issues as the new location may have different rules and regulations.
Cloud outages: In outages the server crash happens. If the server crashes cause temporary outages then it is acceptable but the data should not be lost. Data backups must be provided by the administrators and should maintain the data. Outages from public clouds are frequently happen, Amazon, Google, Microsoft and salesforce are few names that are being facing the outages for many times.
Loss of data: In cloud structures there are redundant systems that keep data backups and can recover data from disasters. Another solution to prevent from data loss is to use the hybrid cloud deployment model in which non critical data is placed at the public cloud and other critical data is stored at private cloud.
Operational issues: There is also a risk of operational issues in which there is vendor lock-in and its failure. There is a question that is it possible for you to transfer your data to and from cloud or you will be bound to a certain application? If there is any disaster then is it possible to recover your data? In vendor lock-in the customer is being locked-in the particular vendor's cloud. The solution of vendor lock-in is the Zoho Creator in which you make database applications, download them and upload in Google's cloud. In private cloud software should support interoperability to protect you from lock-in. In vendor-failure situation the service provider is being occupied by another competitor or he is not financially stable.

11 Conclusion

A promising paradigm cloud computing is used for the delivery of IT services by using pay as you model. Cloud is emerging as a promising solution for many organization's scalability and profitability. In this paper we have presented a comprehensive architecture comprising of most basic and important layers of cloud. This paper is a good source of information for the evaluation of the best cloud vendors as well. It also aims to provide necessary information to organizations willing to adapt cloud for their business scalability. Our study raises several pitfalls and issues of clouds. Advantages of clouds are also highlighted. This paper provides the state of the art knowledge domain of cloud to the developers and researchers.

References

1. Buyya, R., Broberg, J., Goscinski, A.M.: Cloud Computing Principles and Paradigms. Wiley, New York (2011)
2. Kaur, K., Rai, A.K.: A comparative analysis: grid, cluster and cloud computing. Int. J. Adv. Res. Comput. Commun. Eng. **3**(3) (2014)
3. Er. Ramanjeet Kaur.: A Review of Computing Technologies: Distributed, Utility, Cluster, Grid and Cloud Computing. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **5** (2015)
4. Vaquero, L.M., Rodero-Merino, L., Caceres, J., Lindner, M.: A break in the clouds: towards a cloud definition. SIGCOMM Comput. Commun. Rev. **39**(1), 50–55 (2009)
5. Grama, A., Gupta, A., Karypis, G.: Introduction to Parallel Computing: Design and Analysis of Algorithms. Benjamin, Cummings Publishing Company, Redwood City (1994)
6. Buyya, R.: High Performance Cluster Computing: Architecture and Systems, vol. 1. Prentice Hall, Upper Saddle River (1999)
7. Goyal, S.: Public vs private vs hybrid vs community cloud computing: a critical review. Int. J. Comput. Network Inf. Secur. (2014)
8. Armbrust, M. et al.: Above the clouds: a berkeley view of cloud computing. Technical Report No. UCB/EECS-2009-28. Electrical Engineering and Computer Sciences University of California at Berkeley (2009)
9. Lewis, G.: Basics About Cloud Computing. Software Engineering Institute Carniege Mellon University, Pittsburgh (2010)
10. Marks, E.A., Lozano, B.: Executive's Guide to Cloud Computing. Wiley, New York (2010)
11. Plummer, D.C., Smith, D., Bittman, T.J., Cearley, D.W., Cappuccio, D.J., Scott, D., Kumar, R., Robertson, B.: Gartner highlights five attributes of cloud computing. Gartner Rep. **G00167182**, 1–5 (2009)
12. Rimal, B.P., Choi, E., Lumb, I.: A taxonomy and survey of cloud computing systems. In: Proceedings of 5th IEEE International Joint Conference on INC, IMS and IDC, pp. 44–51 (2009)
13. Buyya, R., Yeo, C.S., Venugopal, S.: Market oriented cloud computing: vision, hype, and reality for delivering IT services as computing utilities. In: Proceedings of the 10th IEEE International Conference on High Performance Computing and Communications, Dalian, China (2008)
14. Cisco Cloud Computing—Data Center Strategy, Architecture, and Solutions Point of View White Paper for U.S. Public Sector, 1st edn. Cisco Systems (2009)
15. Youseff, L. Butrico, M. Da Silva, D.: Toward a unified ontology of cloud computing. In: Grid Computing Environments Workshop, 2008. GCE'08, pp. 1–10 (2008)

16. Lenk, A., Klems, M., Nimis, J., Tai, S., Sandholm, T.: What's inside the cloud? An architectural map of the cloud landscape. In: Proceedings of the 2009 ICSE Workshop on Software Engineering Challenges of Cloud Computing, pp. 23–31 (2009)
17. Oracle Infrastructure as a service, [online] available at http://www.oracle.com/us/technologies/cloud/oracle-infrastructure-as-a-service-409028.html
18. Essential characteristics, [online] available at http://www.cloudadvocates.com/Cloud-Computing-Essential-Characteristics
19. Choubey, R., Dubey, R., Bhattacharjee, J.: A survey on cloud computing security, challenges and threats. Int. J. **3** (2011)
20. Marston, S., Li, Z., Bandyopadhyay, S., Zhang, J., Ghalsasi, A.: Cloud computing—the business perspective. Decis. Support Syst. **51**(1), 176–189 (2011)
21. Alger, D.: Build the Best Data Center Facility for Your Business. Cisco Press, Indianapolis (2005)

Neural Network Emulation of Spatio-temporal Data Using Linear and Nonlinear Dimensionality Reduction

V. Triantafyllidis, W. Xing, A.A. Shah and P.B. Nair

Abstract A statistical emulator of a high-fidelity computer model is based on the application of machine learning algorithms to input-output data generated by the model at selected design points. Applications include real-time control, design optimization and inverse parameter estimation. In many of these applications, the outputs are spatial or spatio-temporal fields. In such cases, standard emulation methods are computationally impractical due to the curse of dimensionality, or are limited in their applicability by simplifying assumptions in relation to the correlation structure. In this work, we combine linear and nonlinear dimensionality reduction with artificial neural networks to develop an efficient approach to emulating high-dimensional spatio-temporal models, without making ad hoc assumptions regarding correlations. The approach is tested on models of electromagnetic wave propagation. The necessity of nonlinear dimensionality reduction is highlighted.

Keywords Emulator · Artificial neural network · Machine learning · Manifold learning · Electromagnetic wave propagation

1 Introduction

An emulator [1] is a computationally efficient approximation of a high-fidelity computer model for applications in which repeated calls to the model are not practical or even feasible. This can occur, e.g., in uncertainty analysis, design optimization, control and inverse parameter estimation. Emulators are based on

V. Triantafyllidis · W. Xing · A.A. Shah (✉)
Warwick Centre for Predictive Modelling, University of Warwick, Coventry CV4 7AL, UK
e-mail: Akeel.Shah@warwick.ac.uk

P.B. Nair
University of Toronto Institute for Aerospace Studies, 4925 Dufferin Street, Toronto, ON M3H 5T6, Canada

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_86

machine learning algorithms (e.g., neural networks, Bayesian Gaussian process (GP) models) applied to input-output data generated by the model. A common approach is to assume that the outputs are generated by a GP, most typically for a scalar output that is dependent on one or more model parameters (also termed inputs) [2].

Emulating parameterized spatio-temporal data from computer models poses enormous challenges as a consequence of the exponential growth in the number of hyperparameters. The simplest approach was developed by Kennedy and O'Hagan [3], who treated the output index (spatio-temporal location) as an additional input. This method is impractical for emulating entire spatio-temporal models. Conti and O'Hagan [4] extended GP emulation (GPE) to multiple outputs by placing a multi-dimensional GP prior over the computer model and assuming a stationary, separable covariance structure. This simplification leads to an efficient emulation strategy but is limited in its applicability. The same method was employed by Rougier [5], who took advantage of factorizations of the covariance matrix to improve computational efficiency. Konomi et al. [6] extended the approach to non-stationary GPs. Non-separable covariance structures were considered by Fricker et al. [7] by using the linear model of coregionalization [8]. Due to the exponential growth in the number of hyperparameters, this approach is restricted to low-dimensional problems, e.g., categorical data.

An alternative approach was taken by Higdon et al. [9], who used principal component analysis (PCA) for dimensionality reduction (DR) of the output space. The coordinates in a PCA basis are uncorrelated and can be treated as independent (but not identically distributed) GPs. This method is only applicable to problems in which the manifold on which the data resides can be represented by a linear subspace of the ambient space. An ad hoc dimension reduction was also employed by Bayarri et al. [10], who used a wavelet decomposition and a thresholding procedure to restrict the dimensionality of the output space.

In this work, artificial neural networks (ANN) are combined with DR to develop a method for learning mappings between general input spaces and high-dimensional output spaces, with applications to parameterized spatial and spatio-temporal data sets from computer models. The advantages of ANNs are that they are extremely versatile and learn rapidly. They can also be used to learn multiple coefficients in a reduced basis simultaneously, in contrast to GPE. This has particular advantages in terms of learning multiple spatio-temporal outputs from a model, accounting naturally and efficiently for correlations between the different outputs. The method for multiple fields is outlined in Algorithm 1 in Sect. 2.2.

ANNs can also be placed in a Bayesian framework, which furnishes predictions of the uncertainty in the results. In our proposed linear dimension reduction method for ANN, these uncertainties (in the coefficients of a new basis expansion) can be translated into uncertainties in physical space. Using the nonlinear dimension reduction methods, however, such uncertainties cannot be determined explicitly (an approximate inverse map of the uncertainties in the coordinates of the new basis is not realizable), which represents a potential drawback. Details are provided in the next section.

2 Emulation of Spatio-Temporal Data Sets

An emulator provides a probability distribution for the outputs of a computer model. The computer model is represented as a function $\boldsymbol{\psi} : \chi \to \mathbb{R}^d$, taking as inputs (or parameters) $\boldsymbol{x} \in \chi \subset \mathbb{R}^l$ and generating outputs $y = \boldsymbol{\psi}(x) \in \mathcal{O} \subset \mathbb{R}^d$. The emulator is trained using m computer model outputs $\boldsymbol{y}^{(i)} = \boldsymbol{\psi}(\boldsymbol{x}^{(i)})$, referred to as *training points*, at selected *design points* $\boldsymbol{x}^{(i)} \in \chi \subset \mathbb{R}^l$.

Spatio-temporal data sets from parametrized PDE models

Consider a generic parameterized nonlinear computer model (e.g., a system of parameterized partial differential equations (PDEs)) with input parameters $\boldsymbol{x} \in \mathbb{R}^l$ and outputs $y(\boldsymbol{q}, t; \boldsymbol{x})$ computed at different points in a spatial domain $\Omega \subset \mathbb{R}^2$ (for the purposes of illustration). In this notation, $\boldsymbol{q} = (\xi, \eta)$ denotes the spatial variable and t represents time. The computer model is executed at design points $\boldsymbol{x}^{(k)}$, $k = 1, \ldots, m$. In steady-state problems this yields the values of $y(\boldsymbol{q}; \boldsymbol{x})$ at locations (ξ_i, η_j), $i = 1, \ldots, N_\xi$, $j = 1, \ldots, N_\eta$, on a spatial grid. These values $y_{i,j}^{(k)} := y^{(k)}(\xi_i, \eta_j; \boldsymbol{x}^{(k)})$ can be vectorized as follows:

$$\boldsymbol{y}^{(k)} := (y_{1,1}^{(k)}, \ldots, y_{1,N_\eta}^{(k)}, y_{2,1}^{(k)}, \ldots, y_{2,N_\eta}^{(k)}, \ldots, y_{N_\xi,1}^{(k)}, \ldots, y_{N_\xi,N_\eta}^{(k)})^T \in \mathbb{R}^d \tag{1}$$

for $k = 1, \ldots, m$, and where $d = N_\xi \times N_\eta$. For dynamic problems, $\boldsymbol{y}^{(k)}$ can be defined in a similar manner, with $d = N_\xi \times N_\eta \times N_t$, where N_t is the number of time steps. The extension to 3D problems is also straightforward. The method developed below can be applied to a single field of interest or to the emulation of multiple fields (PDE systems) as explained later.

Clearly, an ANN with d (as defined above) output neurons will not be computationally practical in many cases. For highly complex problems involving, e.g., interface tracking or phase change, the number of grid points required to fully resolve phenomena at all scales can lead to d values in excess of 10^6. To overcome this issue, DR (of the output space) is employed.

2.1 Dimensionality Reduction

Linear methods for dimension reduction

PCA provides a basis $\boldsymbol{w}_i$, $i = 1, \ldots, d$, for $\mathbb{R}^d$ (and therefore $\mathcal{O}$) that is defined by the eigenvectors of the sample covariance matrix. For each input $\boldsymbol{x}^{(j)}$, the corresponding output has a unique representation $\boldsymbol{y}^{(j)} = \sum_{i=1}^{d} z_i^{(j)} \boldsymbol{w}_i$, in which the uncorrelated coefficients in this basis, $z_i^{(j)}$, are naturally ordered in a non-increasing manner with i. The data can be projected (potentially) onto a low-dimensional subspace of $\mathbb{R}^d$ by using the r most dominant basis vectors $\boldsymbol{w}_i$: $\boldsymbol{z}_r^{(j)} := (z_1^{(j)}, \ldots, z_r^{(j)})^T \in \mathbb{R}^r$. ANN is then performed on the input-output pairs $(\boldsymbol{x}^{(j)}, \boldsymbol{z}_r^{(j)})$, $j = 1, \ldots, m$, to yield a mean value for the random vector $\boldsymbol{z}_r = (z_1, \ldots, z_r)^T$ corresponding to a test input $\boldsymbol{x}$. The predicted output at the test input x is obtained as the linear combination

$$\boldsymbol{y} = \boldsymbol{\psi}(\boldsymbol{x}) \approx \sum_{i=1}^{r} z_i \boldsymbol{w}_i \tag{2}$$

or $y = W_r \boldsymbol{z}_r$, where $W_r := [\boldsymbol{w}_1 \ldots \boldsymbol{w}_r]$.

Multidimensional scaling (MDS) is a mapping $\boldsymbol{\chi}_r : \mathcal{O} \to S_r \subset \mathbb{R}^r$ that relates the Euclidean distances δ_{ij} between data points $\boldsymbol{\chi}_r(\boldsymbol{y}^{(i)})$ and $\boldsymbol{\chi}_r(\boldsymbol{y}^{(j)})$ in the mapped space S_r to 'dissimilarities' d_{ij} between $\boldsymbol{y}^{(i)}$ and $\boldsymbol{y}^{(j)}$ in data space $\mathcal{O}$. Let $D := [d_{ij}]$ denote the dissimilarity matrix. Classical scaling [11] is an isometry in which dissimilarities are defined as Euclidean distances:

$$\delta_{ij} = d_{ij} = ||z_r^{(i)} - z_r^{(j)}||$$

for points $\boldsymbol{z}_r^{(i)}$ and $\boldsymbol{z}_r^{(j)}$ in S_r, $i, j = 1, \ldots, m$. Let $\widehat{Z}_d := [\boldsymbol{z}_d^{(1)}, \ldots, \boldsymbol{z}_d^{(m)}]^T$, or in centred form, $Z_d = H\widehat{Z}_d$. It can be shown that

$$-(1/2)H(D \circ D)H = Z_d Z_d^T = K$$

where K is a centred kernel matrix and $\circ$ denotes a Hadamard product. Spectral decomposition yields $K = V_d \Lambda_d V_d^T$, where $\Lambda_d := diag\,(\lambda_1, \lambda_d) \in \mathbb{R}^{d \times d}$ and $V_d := [v_1, v_d] \in \mathbb{R}^{m \times d}$. The non-zero eigenvalues λ_i, $i = 1, \ldots, d$, are arranged in a non-increasing order and the corresponding eigenvectors $v_i \in \mathbb{R}^m$ are normalized. The data can be represented as $Z_d = V_d \Lambda_d^{1/2} \in \mathbb{R}^{m \times d}$, and embedded in an r-dimensional linear subspace S_r of $\mathbb{R}^d$ by setting $V_r := [v_1, v_r]$ and $\Lambda_r := diag\,(\lambda_1, \lambda_r)$ to yield

$$Z_r = V_r \Lambda_r^{1/2}$$

The rows $z_r^{(i)} \in S_r$ of Z_r are the low-dimensional representations of the data points. MDS is equivalent to PCA (the coordinates are identical) when Euclidean distances are used. Both of these linear methods will fail when no linear subspace of $\mathbb{R}^d$ can accurately describe the output space $\mathcal{O}$. In such cases, nonlinear DR (or manifold learning) can be employed.

Nonlinear methods for dimension reduction

In contrast to MDS, Isomap uses geodesic distances for the dissimilarities between points on the manifold $\mathcal{O}$ [12]. Neighbourhood points on the manifold can be determined by using either of the following methods: (i) all points lying within an ε ball; or (ii) the n (neighbourhood number) closest points. A dissimilarity matrix $D := [d_{ij}]$ is then constructed by: (i) using Euclidean distances between neighbours as the geodesic distances; (ii) for non-neighbouring points, using the shortest path distances through neighbouring points. Classical scaling on the kernel matrix $K = -(1/2)H(D \circ D)H$ subsequently yields a representation of the data in $\mathbb{R}^r$ (an r-dimensional parameterization of $\mathcal{O}$).

In kPCA, the training data is mapped to a higher dimensional *feature space* in which it is possible to perform linear PCA [13]. The mapping $\boldsymbol{\phi}: \mathcal{O} \to \mathcal{F}$ is implicitly defined using a kernel function

$$k(i,j) := \boldsymbol{\phi}(y^{(i)})^T \boldsymbol{\phi}(y^{(j)})$$

i.e., the dot product between two data points in feature space. The kernel values define a kernel matrix K with entries $K_{ij} = k(i,j)$. A common kernel is the Gaussian:

$$k(i,j) = \exp\left(-\frac{1}{2s^2}||\boldsymbol{y}^{(i)} - \boldsymbol{y}^{(j)}||^2\right)$$

where s is a shape parameter [14] that controls the flexibility of the kernel. One choice for s (adopted in this work) is the average minimum distance between two observations in the original space [14]:

$$s^2 = \frac{g}{m}\sum_{i=1}^{m} \min_{j \neq i} ||\boldsymbol{y}^{(i)} - \boldsymbol{y}^{(j)}||^2; \quad j = 1, \ldots, m \tag{3}$$

in which g is an adjustable parameter. Centering the points in feature space by defining

$$\widetilde{\boldsymbol{\phi}}(\boldsymbol{y}^{(i)}) = \boldsymbol{\phi}(\boldsymbol{y}^{(i)}) - \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{\phi}(\boldsymbol{y}^{(j)})$$

yields centred kernel values $\widetilde{k}(i,j) := \widetilde{\boldsymbol{\phi}}(\boldsymbol{y}^{(i)})^T\widetilde{\boldsymbol{\phi}}(\boldsymbol{y}^{(j)})$. These kernel values define a centred kernel matrix $\widetilde{K}$ with entries $\widetilde{K}_{ij} = \widetilde{k}(i,j)$. The eigenvectors $\boldsymbol{f}_i$ (of which there are $\min\{\dim(\mathcal{F}), m\}$) of the sample centred covariance matrix S in feature space can be written in terms of the eigenvectors v_i, $i = 1,\ldots,m$, of $\widetilde{K}$.]] >

The eigenvectors of S can be normalized to yield a new set of vectors $\widetilde{\boldsymbol{f}}_i, i = 1,\ldots,\dim(\mathscr{F})$, given by $\widetilde{\boldsymbol{f}}_i = \sum_{l=1}^{m} v_{li}\widetilde{\boldsymbol{\phi}}(\boldsymbol{y}^{(l)})/\sqrt{\lambda_i}$, where v_{li} is the lth component of the eigenvector v_i and λ_i is the ith eigenvector of S. A point $\widetilde{\boldsymbol{\phi}}(\boldsymbol{y}^{(j)}) \in \mathcal{F}$ can be projected onto $\widetilde{\boldsymbol{f}}_i$ as follows:

$$z_i^{(j)} = \widetilde{\boldsymbol{f}}_i^T\widetilde{\boldsymbol{\phi}}(y^{(j)}) = \sum_{l=1}^{m} v_{li}\widetilde{K}_{lj}/\sqrt{\lambda_i} \quad i = 1,\ldots,m \tag{4}$$

As in PCA, these projections are arranged in order of non-increasing variance. The projection of a point $\widetilde{\boldsymbol{\phi}}(\boldsymbol{y}^{(j)})$ onto span $(\widetilde{\boldsymbol{f}}_1,\ldots,\widetilde{\boldsymbol{f}}_r)$ is given by:

$$\sum_{i=1}^{r} z_i^{(j)}\widetilde{\boldsymbol{f}}_i + \frac{1}{m}\sum_{j=1}^{m}\boldsymbol{\phi}(\boldsymbol{y}^{(j)}) = \sum_{i=1}^{r} z_i^{(j)}\Phi H\frac{v_i}{\sqrt{\lambda_i}} + \Phi\mathbf{1} = \Phi\left(H\widetilde{U}_r z_r^{(j)} + \mathbf{1}\right) \tag{5}$$

where $H = I - m\mathbf{1}\mathbf{1}^T$ is the centering matrix, $\mathbf{1} \in \mathbb{R}^m$ is a vector of ones and:

$$\begin{aligned} \boldsymbol{z}_r^{(j)} &= (z_1^{(j)},\ldots,z_r^{(j)})^T \\ \widetilde{U}_r &= \left[\frac{v_1}{\sqrt{\lambda_1}}\ldots,\frac{v_r}{\sqrt{\lambda_r}}\right] \\ \Phi &= [\boldsymbol{\phi}(\boldsymbol{y}^{(1)}),\ldots,\boldsymbol{\phi}(\boldsymbol{y}^{(m)})] \end{aligned} \tag{6}$$

2.2 Main Algorithm

The emulation algorithm employing DR on the output space is now described in the pseudo code below, including for multiple spatio-temporal data sets. The last step relates to reconstruction of the predicted point in physical space (in $\mathcal{O} \subset \mathbb{R}^d$) and is described in the sequel.

Algorithm 1: ANN learning of spatio-temporal models using DR

1. Select design points $\boldsymbol{x}^{(i)} \in \chi \subset \mathbb{R}^l$, $i = 1,\ldots,m$, using DOE and construct outputs $\boldsymbol{y}^{(i)} = \boldsymbol{\psi}(\boldsymbol{x}^{(i)}) \in \mathcal{O} \subset \mathbb{R}^d$, $i = 1,\ldots,m$, from the computer model.
2. Perform DR (PCA, Isomap or kPCA) on $\boldsymbol{y}^{(i)}$, $i = 1,\ldots,m$, to obtain coordinates in a low-dimensional representation: $\boldsymbol{z}_r^{(i)} = (z_1^{(i)},\ldots,z_r^{(i)})^T$, $i = 1,\ldots,m$, with $r \ll d$ (for multiple fields $\boldsymbol{y}_b^{(i)}$, $b = 1,\ldots,B$, this would lead to B sets of coefficients $\boldsymbol{z}_{r,b}^{(i)} = (z_{1,b}^{(i)},\ldots,z_{r,b}^{(i)})^T$.

3. Select a test point $\boldsymbol{x}$ for prediction and perform ANN on the training set $(\boldsymbol{z}_r^{(i)}, \boldsymbol{x}^{(i)})$, $i = 1, \ldots, m$, to obtain $\boldsymbol{z}_r = (z_1, \ldots, z_r)^T$. For multiple fields the training set is $((\boldsymbol{z}_{r,1}^{(i)}, \ldots, \boldsymbol{z}_{r,B}^{(i)}), \boldsymbol{x}^{(i)})$, $i = 1, , m$, which yields $\boldsymbol{z}_r = (z_{1,1}, \ldots, z_{r,1}, \ldots, z_{1,B}, \ldots, z_{r,B})^T \in \mathbb{R}^{rB}$ for a test point $\boldsymbol{x}$.
4. Using z_r approximate the computer model output $\boldsymbol{y} = \boldsymbol{\psi}(\boldsymbol{x})$ by solving the pre-image problem (see below).

2.3 Pre-Image Problem (Inverse Mapping)

When using PCA, the reconstruction of the point in physical space $\mathcal{O}$ is given by the linear combination (2). In Isomap and kPCA only the predicted coordinates, $z_1, \ldots, z_r$, of a point $\boldsymbol{y} = \boldsymbol{\psi}(\boldsymbol{x})$ in the reduced space are available. In Isomap, the Euclidean distances between points in the reduced space are equal to geodesic distances $d_{i,*}$ between a predicted point $\boldsymbol{y}$ and points $\boldsymbol{y}^{(i)} \in \mathcal{O}$ in physical space. Local linear interpolation can be used to approximate the coordinates of $\boldsymbol{y}$ by using these geodesic distances as weights [15]:

$$y = \boldsymbol{\psi}(\boldsymbol{x}) \approx \sum_{i=1}^{N_n} \frac{w_i}{\sum_{i=1}^{N_n} w_i} \boldsymbol{y}^{(i)} \tag{7}$$

where $w_i = 1/d_{i,*}$ and N_n is the number of neighbours selected for the reconstruction.

The same method can be used for kPCA. ANN on the first r kPCA coefficients (in the basis $\widetilde{\boldsymbol{f}}_j$) yields an approximation $\boldsymbol{\phi}(\boldsymbol{y})$ of the point $\boldsymbol{y}$ in feature space. The distance $\widetilde{d}_{i,*}$ between $\boldsymbol{\phi}(\boldsymbol{y}^{(i)})$ and $\boldsymbol{\phi}(\boldsymbol{y})$ in feature space is given by:

$$\widetilde{d}_{i,*}^2 = \boldsymbol{\phi}(\boldsymbol{y})^T \boldsymbol{\phi}(\boldsymbol{y}) + \boldsymbol{\phi}(\boldsymbol{y}^{(i)})^T \boldsymbol{\phi}(\boldsymbol{y}^{(i)}) - 2\boldsymbol{\phi}(\boldsymbol{y}^{(i)})^T \boldsymbol{\phi}(\boldsymbol{y}) \tag{8}$$

Substituting Eq. (5) into Eq. (8) for Gaussian kernel gives:

$$\widetilde{d}_{i,*}^2 = \boldsymbol{\tau}^T \Phi^T \Phi \boldsymbol{\tau} + k(i,i) - 2\boldsymbol{\tau}^T \Phi^T \boldsymbol{\phi}(\boldsymbol{y}^{(i)}) = \boldsymbol{\tau}^T K \boldsymbol{\tau} + 1 - 2\boldsymbol{\tau}^T \boldsymbol{k}_i \tag{9}$$

where

$$\begin{aligned} \boldsymbol{\tau} &= H \widetilde{U}_r \boldsymbol{z}_r^{(j)} + \mathbf{1} \\ \boldsymbol{k}_i &= (k(1,i), \ldots, k(m,i))^T \end{aligned} \tag{10}$$

For an isotropic kernel ($k(\boldsymbol{y}, \boldsymbol{y}^{(i)}) = k(||\boldsymbol{y} - \boldsymbol{y}^{(i)}||^2)$), the relationship $\widetilde{d}_{i,*}^2 = 2 - 2k(d_{i,*}^2)$ follows from Eq. (8). In the case of a Gaussian kernel:

$$d_{i,*}^2 = -2s^2 \log\left(2 - \widetilde{d}_{i,*}^2/2\right) \tag{11}$$

which is combined with (9) to yield $d_{i,*}$. For other kernel functions [16], similar relationships can be derived.

We note that it is also possible to reconstruct $\boldsymbol{y}$ from the predicted coefficients using a least-squares approximation. This method is, however, prone to instability, as is the fixed point algorithm of Mika et al. [17].

2.4 Bayesian Regularization

In order to improve generalization (and avoid cross-validation), Bayesian regularization [18] is used. A prior (zero-mean, Gaussian) distribution is placed on the network weights (for a fixed architecture), which leads to the minimization of $F(\boldsymbol{a}) = \beta E_D/2 + \alpha E_W/2$, where E_D is the network square error, α is the inverse variance of the zero-mean (assumed) Gaussian noise, β is the inverse variance of the weights, and $E_W = ||\boldsymbol{a}||^2$, where $\boldsymbol{a}$ is the vector of network weights. The posterior density of the weights is given by:

$$P(\boldsymbol{a}|D, \alpha, \beta, \mathcal{M}) = \frac{P(D|\boldsymbol{a}, \beta, \mathcal{M})P(\boldsymbol{a}|\alpha, \mathcal{M})}{P(D|\alpha, \beta, \mathcal{M})} \tag{12}$$

where $D = \{\boldsymbol{y}^{(i)}\}_{i=1}^m$ is the data set, $\mathcal{M}$ indicates the ANN model, $P(\boldsymbol{a}|\alpha, \mathcal{M})$ is the prior density and $P(D|\boldsymbol{a}, \beta, \mathcal{M})$ is the likelihood function. The optimal weights should maximize the posterior likelihood $P(\boldsymbol{a}|D, \alpha, \beta, \mathcal{M})$. A uniform prior density $P(\alpha, \beta, \mathcal{M})$ for the parameters α, β gives:

$$P(D|\alpha, \beta, M) = \frac{Z_F(\alpha, \beta)}{Z_D(\beta)Z_W(\alpha)} \frac{\exp(-\beta E_D - \alpha E_W)}{\exp(-F(\boldsymbol{a}))} = \frac{Z_F(\alpha, \beta)}{Z_D(\beta)Z_W(\alpha)} \tag{13}$$

in which $Z_D(\beta) = (\pi/\beta)^{(m/2)}$ and $Z_W(\alpha) = (\pi/\alpha)^{(N/2)}$, where N is the total number of parameters in the model. The unknown normalization factor $Z_F(\alpha, \beta)$ can be approximated in terms of the Hessian matrix H^{MP} of $F(\boldsymbol{a})$ by a quadratic Taylor series expansion of $F(\boldsymbol{a})$ around its minimum, at $\boldsymbol{a} = \boldsymbol{a}^{MP}$. Placing the result in (13) and differentiating yields

$$\alpha^{MP} = \frac{\gamma}{2E_W(\boldsymbol{a}^{MP})} \qquad \beta^{MP} = \frac{m - \gamma}{2E_D(\boldsymbol{a}^{MP})}$$

where $\gamma = N - 2\alpha^{MP}/tr(H^{MP})$. To optimize α and β, the Hessian matrix H^{MP} is required. Using a Gauss-Newton approximation to the Hessian matrix and the Levenberg-Marquardt algorithm, these hyperparameters are calculated using an iterative procedure detailed in [19] (implemented in the trainbr function in Matlab).

3 Results and Discussion

3.1 *Details of Training and Testing*

In both examples below, the data set consisted of 500 points ($\boldsymbol{y}^{(i)} = \boldsymbol{\psi}(\boldsymbol{x}^{(i)})$), with inputs $\boldsymbol{x}^{(i)}$ selected using a Sobol sequence (uniform sampling) design-of-experiment (DOE). This was found to be adequate for the examples presented below. It must be noted that the DOE is, in general, a vital aspect of any emulation strategy. An appropriate sampling of the input space is paramount for generating training samples that are representative of the region output space O that is of interest. In the present case, the training samples must generate a basis (either in physical space or in a feature space) that accurately captures the output space. Since these issues are encountered in all emulation methods, we do not focus on them here.

400 points were reserved for testing and up to 100 points were used for training ($m \leq 100$). The relative square errors (total square error divided by the number of grid points and the magnitudes of the average values of the test points) were used to assess the generalization error. Results are shown for different numbers of components r in the DR methods. In the case of PCA (kPCA), the first r 'components' are the r principal components corresponding to the r largest eigenvalues of the (feature space) covariance matrix. For Isomap the first r 'components' are the r Isomap coordinates corresponding to the r largest eigenvalues of the kernel matrix.

The neighbourhood number method (10 neighbours) was used for Isomap. For kPCA, a Gaussian kernel was used, with a shape parameter dependent on the data set. For reconstruction, $N_n = 10$ points were used for both Isomap and kPCA. The ANN architecture in all cases contained a single hidden layer and the ANN was trained using Bayesian regularization [18]. The number of neurons for each example was selected using sequential network construction [19]. In general, an arbitrary ANN architecture can be used within the framework.

Example 1: 2D h-bend waveguide

This model examines a transversal electric (TE) wave in a h-bend waveguide with a 90 degree bend. The frequencies f are restricted so that TE_{10} is the single propagating mode. The electric field has only one nonzero component E_z in the transversal direction z. The model computes the electromagnetic field by solving Hemholtz equation:

$$-\nabla^2 E_z - n^2 k_0^2 E_z = 0$$

in which n is the refractive index, k_0 is the free space wave number, and ξ and η are the in-plane directions.

On the domain walls, the tangential component of the electric field is zero. The input wave is determined by the boundary conditions for Maxwell's equations: $\widehat{n} \times \boldsymbol{E} = 0$, where $\widehat{n}$ is the unit normal. The incident field has the form:

$$\boldsymbol{E} = (0, 0, \sin(\pi(b/2 - \xi)/b)) = \Re(\boldsymbol{E} e^{i\omega t})$$

in which b is the width of the rectangular sections of the waveguide and ω is the angular frequency of the incident wave. The model was solved ('H-Bend Waveguide 2D' in the 'RF Module' of COMSOL Multiphysics 5.0) for 500 frequency values f between 4 and 6 GHz ($\boldsymbol{x}^{(i)} = f^{(i)}$, $i = 1, \ldots, 500$). For each simulation, the magnitude of the electric field E was recorded on a 100 × 100 regular grid in (ξ, η). The $d = 10^4$ values of $|\boldsymbol{E}(\xi, \eta)|$ for each $\boldsymbol{x}^{(i)}$ at locations (ξ_l, η_j), $l, j = 1, \ldots, 100$, were vectorized (see Eq. 1) to give 500 data points $\boldsymbol{y}^{(i)}$ in $\mathbb{R}^d$. Up to 100 were used for training and the remainder for testing.

Results

All three dimension reduction methods using ANN gave excellent results for at least 20 training points (m = 20). In the case of PCA, box plots of the relative errors for different numbers of principal components (on the horizontal axis) are shown in Fig. 1a for 60 training points. The other methods gave similar results, so to conserve

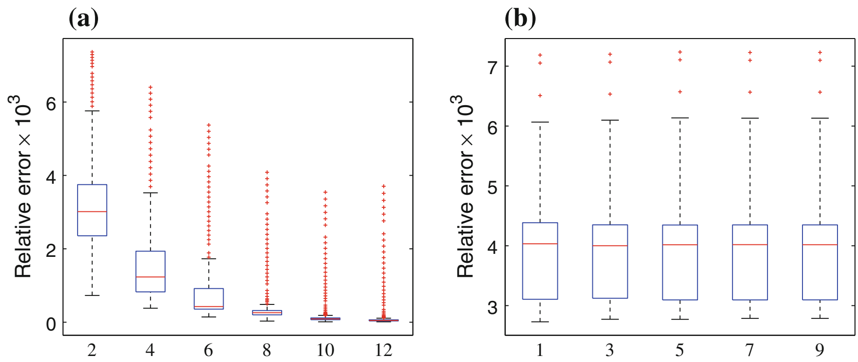


Fig. 1 Boxplots of the relative errors for different numbers of components (r) using ANN with PCA (M = 60) and GPE with PCA (M = 100) in the waveguide example

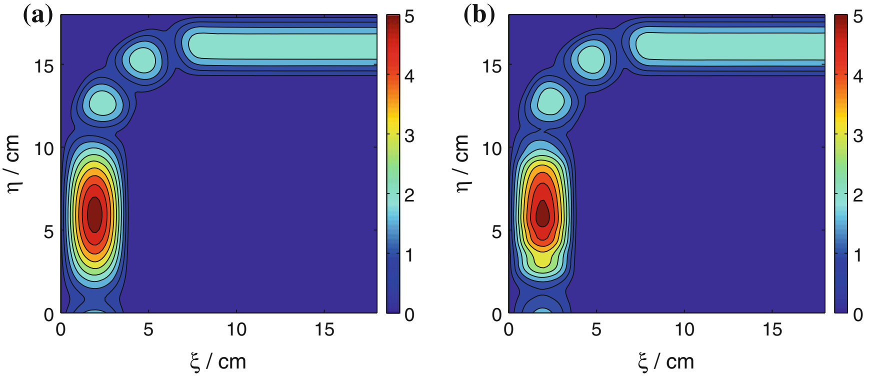


Fig. 2 Representative examples of prediction using ANN with SVD (r = 5, M = 60) for the 2D waveguide

space they are omitted. The standard method of Higdon et al. [9] using a maximum likelihood estimate (MLE) for the hyperparameters failed to provide meaningful results, as demonstrated in Fig. 1b for m = 100. An example of the worst predictions for m = 60 using ANN with PCA (r = 12) is shown in Fig. 2. The relative error is 2.3×10^{-3}.

Example 2: 2D radar interaction with a boat (radar cross section)
The interaction between a boat and the incident field from a radar transmitter is simulated. The transmitter is distant enough that the field can be treated as a plane wave (only the boat and its immediate surroundings are considered). The background field is swept over a range of angles of incidence and the far-field and radar cross section (RCS) are computed. The 2D geometry consists of an inner circle containing the boat and the surrounding air, together with an outer circle representing a perfectly matched layer (PML). The background electromagnetic field from the radar is described by its out-of-plane electric field component:

$$\boldsymbol{E}_b = \exp(ik_0(\xi\cos\theta + \eta\sin\theta))\boldsymbol{e}_z$$

where $k_0 = 2\pi f/c$ is the wave number in vacuum, c is the speed of light, f is the frequency and θ is the angle of incidence. The time-harmonic wave equation is then solved for the relative field, $\boldsymbol{E}_{rel} = \boldsymbol{E} - \boldsymbol{E}_b$, where $\boldsymbol{E}$ is the total field:

$$\nabla \times \left(\mu_r^{-1}\nabla \times \boldsymbol{E}_{rel}\right) - \left(\varepsilon_r - \frac{i\sigma}{2\pi f\varepsilon_0}\right)k_0^2\boldsymbol{E}_{rel} = 0$$

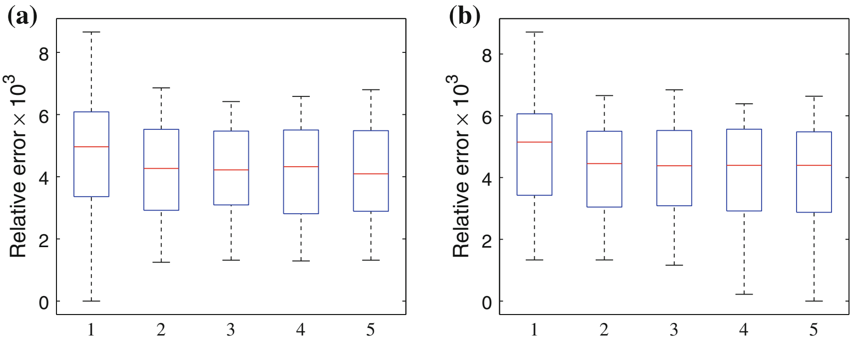


Fig. 3 Boxplots of the relative errors for different numbers of components (r) using ANN with Isomap (M = 80 and 100) in the RCS example

in which $\in$, μ and σ denote the permittivity, permeability, and conductivity of air, respectively (subscripts r denote a 'relative' quantity). The RCS per unit length is defined as

$$\sigma_{2D} = \lim_{r \to \infty} 2\pi r \frac{|\boldsymbol{E}_{rel}|^2}{|\boldsymbol{E}|^2}$$

The model was solved ('Radar Cross Section' under the Radio Frequency module in COMSOL Multiphysics 5.0) for 500 combinations of f and θ as input values; that is $\boldsymbol{x}^{(i)} = (f^{(i)}, \theta^{(i)})^T$, $i = 1, \ldots, 500$. The magnitude of the electric field $\boldsymbol{E}$ was recorded on a regular 500 × 500 square spatial grid in (ξ, η). The d = 2.5 × 10^5 values of $|\boldsymbol{E}(\xi, \eta)|$ for each $\boldsymbol{x}^{(i)}$ at locations (ξ_l, η_j), $l, j = 1, \ldots, 500$, were vectorized to form the data points $\boldsymbol{y}^{(i)} \in \mathbb{R}^d$ used for testing and training.

Results

PCA with both ANN and GPE (method of Higdon et al. [9]) failed to provide meaningful results for any number of training points m or components r. ANN with Isomap and kPCA, on the other hand, exhibited good results, especially in the case of Isomap for $m > 60$, which captured the trends precisely. Box plots of the relative square errors are shown in Fig. 3a, b, up to 5 components (beyond which no improvements were visible). Figure 4 shows two representative examples of the predictions using ANN with Isomap (r = 5 and m = 100). In the first case, the relative error is 6.4 × 10^{-3} (near the maximum) and in the second case the relative error is 2.2 × 10^{-3}.

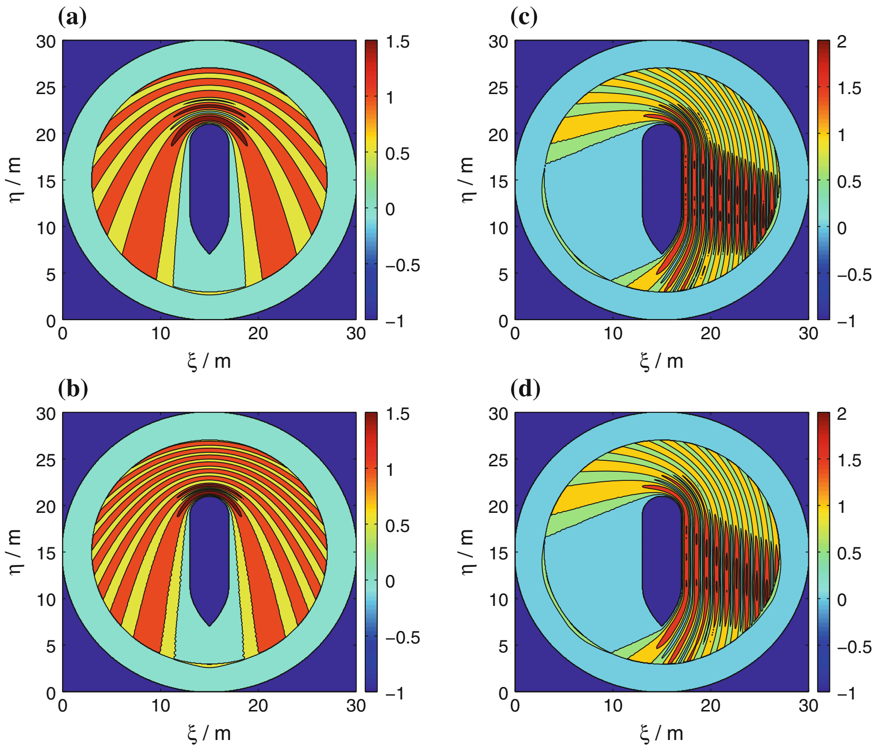


Fig. 4 Representative examples of the predictions using ANN with Isomap (r = 5) and 100 training points in the RCS example. **a** Test 1, **b** ANN−Isomap prediction Test 1, **c** Test 2, **d** ANN −Isomap prediction Test 2

4 Concluding Remarks

An efficient approach for constructing emulators of parameter-dependent spatial and spatio-temporal fields was developed by reducing the dimensionality of the output space. The total learning and prediction time for all methods is on the order of a few minutes, which is a dramatic reduction in computational effort compared to the emulation of d outputs simultaneously—in many cases this would not be computationally feasible. Target applications include uncertainty quantification, design optimization, real-time control and inverse parameter estimation. Results were compared in test cases to a standard method [9]. In both cases (and in many other data sets tested) this standard method fails, while the linear and/or nonlinear methods developed in this paper exhibit accurate results. The second example and numerous others that we have used to test the methods show that nonlinear DR is necessary for many complex data sets.

The lack of an uncertainty measure in the basic ANN framework (in contrast to GPE) can be overcome with Bayesian neural networks. A fully Bayesian approach with Markov Chain Monte Carlo sampling can be used without modification to the general framework developed here. In the linear dimension reduction method, the predicted output is given by $\boldsymbol{y} = W_r \boldsymbol{z}_r$. By the properties of PCA, the coordinates of $\boldsymbol{z}_r$ are uncorrelated. Denoting the calculated variances of these coefficients by $\chi_1, \ldots, \chi_r$ (for $z_1, \ldots, z_r$, respectively), we obtain

$$cov(\boldsymbol{y}) = W_r cov(\boldsymbol{z}_r) W_r^T = W_r diag(\chi_1, \ldots, \chi_r) W_r^T$$

from which the variances in the coordinates of the predicted output can be determined. This procedure is, however, not possible with the nonlinear dimension reduction techniques, even thought the variances in the coordinates in a feature space basis are available. Work is ongoing to overcome this drawback.

Acknowledgements This work was funded by EU FP7 (Grant Number 314159; NECOBAUT). The authors would also like to acknowledge the Chinese Scholarship Council for a scholarship provided to Wei Xing to fund his doctoral studies.

References

1. Santner, T.J., Williams, B.J., Notz, W.I.: The Design and Analysis of Computer Experiments. Springer, New York (2003)
2. Oakley, J.E., O'Hagan, A.: Bayesian inference for the uncertainty distribution of computer model outputs. Biometrika **89**, 769–784 (2002)
3. Kennedy, M., O'Hagan, A.: Bayesian calibration of computer codes. J. R. Stat. Soc. Ser. B Stat. Methodol. **63**, 425–464 (2001)
4. Conti, S., O'Hagan, A.: Bayesian emulation of complex multi-output and dynamic computer models. J. Stat. Plan. Inference **140**, 640–651 (2010)
5. Rougier, J.: Efficient emulators for multivariate deterministic functions. J. Comput. Gr. Stat. **17**, 827–843 (2008)
6. Konomi, B., Karagiannis, G., Sarkar, A., Sun, X., Lin, G.: Bayesian treed multivariate gaussian process with adaptive design: application to a carbon capture unit. Technometrics **56**, 145–158 (2014)
7. Fricker, T.E., Oakley, J.E., Urban, N.M.: Multivariate gaussian process emulators with nonseparable covariance structures. Technometrics **55**, 47–56 (2013)
8. Gelfand, A.E., Schmidt, A.M., Banerjee, S., Sirmans, C.F.: Nonstationary multivariate process modeling through spatially varying coregionalization. TEST **13**, 263–312 (2004)
9. Higdon, D., Gattiker, J., Williams, B., Rightley, M.: Computer model calibration using high-dimensional output. J. Am. Statist. Assoc. **103**(482), 570–583 (2008)
10. Bayarri, M.J., Berger, J.O., Cafeo, J., Garcia-Donato, G., Liu, F., Palomo, J., Parthasarathy, R. J., Paulo, R., Sacks, J., Walsh, D.: Computer model validation with functional output. Ann. Stat. **35**, 1874–1906 (2007)
11. Torgerson, W.S.: Multidimensional scaling: I. Theory and method. Psychometrika **17**(4), 401–419 (1952)
12. Tenenbaum, J.B., De Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290**, 2319–2323 (2000)

13. Scholkopf, B., Smola, A., Muller, K.-R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput. **10**, 1299–1319 (1998)
14. Rathi, Y., Dambreville, S., Tannenbaum, A.: Statistical shape analysis using kernel PCA. In: Proceedings of SPIE, Image Processing: Algorithms and Systems, Neural Networks, and Machine Learning, vol. 6064, 60641B–60641B-8 (2006)
15. Ganapathysubramanian, B., Zabaras, N.: A non-linear dimension reduction methodology for generating data-driven stochastic input models. J. Comput. Phys. **227**, 6612–6637 (2008)
16. Williams, C.K.I.: On a connection between kernel PCA and metric multidimensional scaling. Mach. Learn. **46**, 11–19 (2002)
17. Mika, S., Scholkopf, B., Smola, A., Muller, K.-R., Scholz, M., Ratsch, G.: Kernel PCA and de-noising in feature spaces. Adv. Neural Inf. Process. Syst. **II**(11), 536–542 (1999)
18. Foresee, D., Hagan, M.T.: Gauss-newton approximation to bayesian learning. Int. Conf. Neural Netw. **3**, 1930–1935 (1997)
19. Moody, J.: Prediction risk and architecture selection for neural networks. In: Cherkassky, V., Friedman, J., Wechsler, H. (eds.) From Statistics to Neural Networks, volume 136 of NATO ASI Series, pp. 147–165. Springer, Berlin (1994)

Agent-Based Simulation for Identifying the Key Advantages of Intelligent Environments for Inhabitants with Special Needs

Karel Mls, Richard Cimler and Peter Mikulecky

Abstract In the paper, simulation of the interaction between an intelligent house containing smart sensors and its inhabitants is introduced. The simulation model is mainly focused on monitoring of different inhabitants' needs and their health statuses. Different situations affecting inhabitant's health status are simulated. The proposed simulation model has an ability to test different arrangements of sensors in the environment without the necessity of its real construction. The most critical situation—heart attack occurrence based on the selected attributes—is studied in a practical example. The monitoring and processing system can recognize a person who needs an urgent medical assistance. In this case, the Emergency Medical Responders (EMR) are called immediately. The simulation tool AnyLogic has been used and its usability for modeled cases seemed to be proven.

Keywords Sensors · Simulated environment · Ubiquitous computing · Avatar model · Welfare evaluation · AnyLogic

1 Introduction

Human activity is a highly complex process in which individuals not only decide about frequency of activities, but also about sequencing, timing and duration of activities. Most of the published studies on simulators of human activities are based on behaviorism models using psychological approaches. This paper presents a simulation of human activities based on a mathematical approach using the mul-

K. Mls · R. Cimler (✉) · P. Mikulecky
University of Hradec Kralove, 500 03 Hradec Kralove, Czech Republic
e-mail: richard.cimler@uhk.cz

K. Mls
e-mail: karel.mls@uhk.cz
URL: http://www.uhk.cz

P. Mikulecky
e-mail: peter.mikulecky@uhk.cz

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_87

tiagent model. Experiments on real systems face many crucial problems—take too much time and face many technical problems such as limitations of sensors, loss of data during logouts, and more. In this context, building a simulator for generating data with the use of ideal sensors, configurable profiles of the inhabitants and flexible duration of experiments is very practical.

On the other hand, there are also complex systems proposals such as the Framework of Ubiquitous Healthcare System Based on Cloud Computing for Elderly Living, which are based on the smart phone capabilities and several sensors located within the environment (see [1] or [2]). The sensors transmit data into the computational unit which processes the data and executes actions on the actuators based on the built-in algorithms. Such complex systems can use information from smart phone sensors and combine it with sensors located within the environment [3].

With the increasing number of external sensors and different devices capable of measuring various biometric data, a large number of particular applications are coming into the market. New approaches and methods are studied and experimentally validated. Watch Dog system, for example, uses multicriteria Analytic hierarchy model to combine data from sensors with expert knowledge [4]. The Watch Dog application has been modified to be prepared to communicate with such sensors, process the data and send the results to the appropriate destination. Result of this process can be for example calling an emergency or just notifying relatives about non-standard situation.

External sensors capable of measuring various biometric data can be connected using the wireless technology such as Bluetooth. These sensors can measure, inter alia, blood pressure, pulse, temperature, moisture or the position of the limbs. All the information from sensors is processed in the device or in a case of more complex computation on the server application. The aim of the system is to evaluate the actual state of the monitored person according to the measured values and to make appropriate actions. Some of the measured values can be normal, some of them can be suspicious, and some can be strictly alerting. Suspicious values can signify danger only when are observed together with suspicious values from some other sensors. The importance of each sensor value can vary according to the context and so different types of dangerous situations can be expressed by a combination of measured values.

In the following, second part of the paper, ambient intelligence utilization for home health care is described. Third part contains information about usage of software AnyLogic for creation of an intelligent environment simulation. For better understanding of the dependencies in the simulation a concept map is created. Detailed description of simulated intelligent systems, residents and the simulation results are in the end of the third part.

2 Ambient Intelligence for Home Health Care

Recent intensive research of Ambient Intelligence approaches and applications has resulted in significant results in several specific areas oriented mainly on independent living support of seniors and handicapped persons. One of these—the Ambient Assisted Living (AAL)—has very close relations also with the eHealth area, and certainly with already matured technology of Smart Homes as well. According to a recent review [5], the technology of Smart Homes can be considered to be an instance of ambient assisted living technologies that are designed to assist the homes residents accomplishing their daily-living activities while preserving their privacy. The smart home technologies and solutions have already proven to be a good and acceptable alternative to formal care in hospitals and care homes.

Nehmer with his colleagues [6] emphasized, that Ambient Intelligence systems can be subdivided into three domains: emergency treatment and assistance, autonomy enhancement, and comfort features. According to them, emergency treatment is considered to be the core functionality of every AmI system, where it aims at early prediction and recovery from critical situations, like heart attacks, injuries or sudden falls.

The Smart Home technologies are considered to be a useful way to reduce living and care costs and to improve the quality of life for people with care needs. They have been applied already for many purposes like energy saving, security and safety, fall detection, light management, smoke and fire detection, etc. using various solutions such as video monitoring, alarms, smart planners and calendars, or reminders [5]. Equipped with sensors, actuators and cameras to collect different types of data on about the home and the residents, Smart Homes can enable automatic systems or caregivers to control the environment on behalf of the residents, monitor and even predict their actions and ubiquitously track their health condition (see also [7] or [8]).

Although the Smart Home technologies are a good way to reduce namely care costs, making use from the patient's home environment, it is evident that the design and deployment of the appropriate sensors and other necessary equipment could be quite costly. Sometimes the results need not to be adequate for the particular usage, mostly because this kind of systems is extremely difficult to test and verify (cf. e.g., [9] or [10]). Therefore, various simulation tools and techniques are considered to be useful and applicable [11, 12], or [13]. A general architecture for testing, validating and verifying AmI environments, called AmISim, has been developed by Garcia-Valverde et al. [14]. A methodology for the validation of ubiquitous computing applications focusing on the use of artificial societies has been developed recently by Serrano and Botia [15]. A visionary paper introducing a number of new ideas in this direction was published by Ishida and Hattori [16].

3 Simulation

The aim of the project is to simulate an environment containing ambient intelligent features and to simulate behavior of residents. The simulation is mainly focused on an intelligent house or office equipped with different sensors and actuators. Various layouts of the house and location of the sensors can be created in the simulation model. Information from the sensors about residents is evaluated by the control system of the house and relevant actions are executed. Several relevant projects and studies can be found in [13, 14] or [15]. Advantage of our approach is in using a free multimethod simulation modeling tool AnyLogic (see, e.g., [17]). AnyLogic uses in-build elements as well as own Java code, which can be easily implemented in the project. Usage of Java enables to program own types of sensors, actuators and algorithms for evaluating the data. Sensors can be merged into the more complex systems. Currently, five types of such sensor systems are implemented and more will be added. It is worth telling that up to now we do not know about any paper focused on using AnyLogic for agent-based simulation of intelligent environments for inhabitants with special needs.

3.1 Example Study

This study is mainly focused on the monitoring inhabitants needs and their health statuses. Different situations affecting the inhabitants' health are simulated. The most important among them is the situation of a heart attack occurring by some of the inhabitants. Based on the selected attributes which are health status, blood pressure, cholesterol level and smoking habits, there is the possibility of a heart attack. The reaction of the environment and EMR (Emergency Medical Responders) arrival in the case of a heart attack is simulated.

3.2 Concept Map of the Simulation

The problem of model identification is highly complex and only slightly structured task. During numerous interviews with experts, many more or less significant and sometimes even contradicting opinions have been proved. To be able to devise any functional and relevant model, some method of generalization of the initial ideas and knowledge should be applied. One well-developed approach, based on the specific domain conceptualization, can be supported by the use of several concept mapping tools.

Concept mapping was introduced in 1972 to meet the ability of people to acquire science concepts [18]. Underlying the research program and the development of the concept mapping tool was an explicit cognitive psychology of learning and an explicit constructivist epistemology. In 1987, collaboration began between Novak and Caas and others at the Florida Institute for Human and Machine Cognition, then part of the University of West Florida. Extending the use of concept mapping to other applications such as the integration of concept mapping with the World Wide Web (WWW) led towards the development of software that enhanced the potential of concept mapping, evolving into the current version of CmapTools now used worldwide in schools, universities, corporations, and governmental and non-governmental agencies [19].

The development of new products, services or processes involves the creation, re-creation and integration of conceptual models from the related scientific and technical domains [20]. Simplified conceptual model of interactions between the inhabitants and their environment, together with the structure of the environment is shown at Fig. 1.

3.3 Model Prerequisites

Residents live in a house equipped with ambient intelligent features. There are five types of systems. First and second are more complex systems. Last three are basic systems. New simple and complex systems will be added during extension of this simulation.

Each of the following systems can be switched on or off before each simulation run and subsequent impact on the residents' life can be tested. Systems are independent to each other.

1. Health emergency recognition system (HERS)
2. Smart heating system
3. Medicine reminder
4. Time to sleep reminder
5. Intelligent fridge

3.3.1 Health Emergency Recognition System

Health emergency recognition system (HERS) simulates a complex system composed of many sensors, which can recognize a critical situation—a person needs an urgent medical assistance. In such case, an EMR is called immediately. HERS is based on Watch Dog [4] and Hausy projects [21]. The Watch Dog is a mobile application for detection of life-threatening situations which is currently under development. Research is focused on a connection with the HAUSY (Home AUtomation SYstem). HAUSY is a system for a complex management of an intelligent house. The current implementation in the AnyLogic does not simulate each sensor of the HAUSY and Watch Dog separately. However the model simulates the behavior of such a system as a whole, and it is called HERS—Health emergency recognition system.

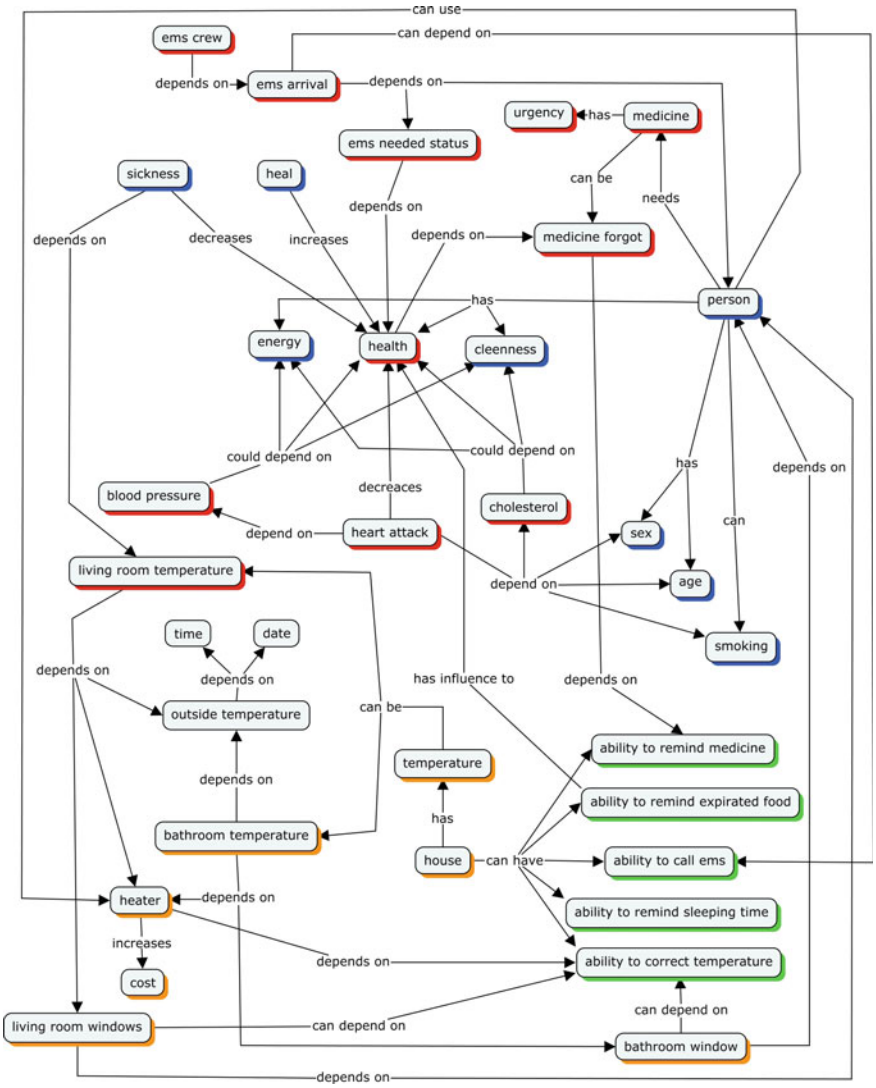


Fig. 1 Concept map of basic interactions within the simulation model

3.3.2 Smart Heating System

Smart heating system is a system based on temperature sensors located within the environment and actuators with the ability to close the windows. The temperature is set due to the preferences of the residents if are at home. During a day, when residents are at work, the temperature is lowered. Before arrival of the residents, the temperature is set to a preferred level again. If residents forget to close windows and leave the house, actuators can close the windows automatically.

3.3.3 Medicine Reminder

It is quite easy to forget to take a medicine or not to remember if a medicine has been already taken today. The Medical reminder simulates a simple system for reminding to take the medicine. More sophisticated systems for more serious situations when it is critical to take a medicine at the right time can communicate with the home system, and contact selected persons via the internet or SMS if medicine has not been taken.

3.3.4 Time to Sleep Reminder

Second of simple systems is Time to sleep reminder. The system reminds resident about time to go to bed every half an hour after the preferred time when a person should go to sleep.

3.3.5 Intelligent Fridge

Intelligent fridge enables to notify about an ending expiration date of the groceries. There is only a minimal probability of fall sick from the expired food if this system is activated in the simulation.

3.4 Residents

Residents are modeled as autonomous agents. Each resident has its attributes, needs and day cycle. Needs such as hunger gradually rise and if any of the needs is on a high level, residents do actions to fulfill it.

3.4.1 Residents' Needs

There are basic needs: hunger, toilet, cleanness, energy and free time in this model. Important is also current health status of the residents. If health status is under set level (90 %) there is a bigger probability that resident becomes ill. During the illness, the energy drops faster and the resident does not go to work.

3.4.2 Residents Day Cycle

Residents get up in the morning, do the hygiene and make a breakfast. Then they drive or go to a work. The temperature inside the house is automatically set to a lower values. After return from the work residents do free time activities such as reading a book, going outside for a walk or watching TV. If a residents' energy is low, they can also take a nap. After a dinner, residents can go to sleep or continue to do some free-time activities based on their energy. During weekends residents can do free time activities or make a whole day travels.

3.5 Simulation Results

In the Table 1 are results from the 15 repetitions of the model run simulating 30 years. Small family house with two inhabitants has been created and all sensors systems described in Sect. 3.3 has been added. Layout of the house can be seen on Fig. 2. Two residents with age starting at 40 and normal day cycle live in the simulated house. Desired temperature inside the house is 21 °C respectively 18.5 °C if residents are not at home. If a health level is below 70 %, resident stays at home and cures. Probability of getting sick increases when the health level is under 90 % Factors such as eating of the spoiled food or low temperature affects residents' health status.

In the simulations with all intelligent features activated, there were significant savings on the heating costs. Concurrently, a number of spoiled food dropped to zero and average value of residents' health was higher because of positive effect of smart systems. System eliminated (1) low temperatures inside the house caused by unwillingly opened windows, (2) eating spoiled food (3) reminded residents to use pills and (4) in some cases prevented from going late to sleep. Due to the higher health level, the number of sick days also decreased significantly.

3.5.1 Simulation of HERS System

After a serious heart attack occurrence, a medical treatment has to be provided within 20 min [22]. Without the system HERS there is a small probability that resident will be able to call an EMR himself. A roommate has to notice this critical

Table 1 One year average for period of 30 years

Parameter	All features OFF	All features ON	Difference	Difference %
Health	96.33 %	98.52 %	2.19 %	2.27
Heating costs	6580	4921	−1659	−25.21
Sick days	53.2	29.7	−23.5	−44.17
Spoiled food	29.6	0	−29.6	−100.00

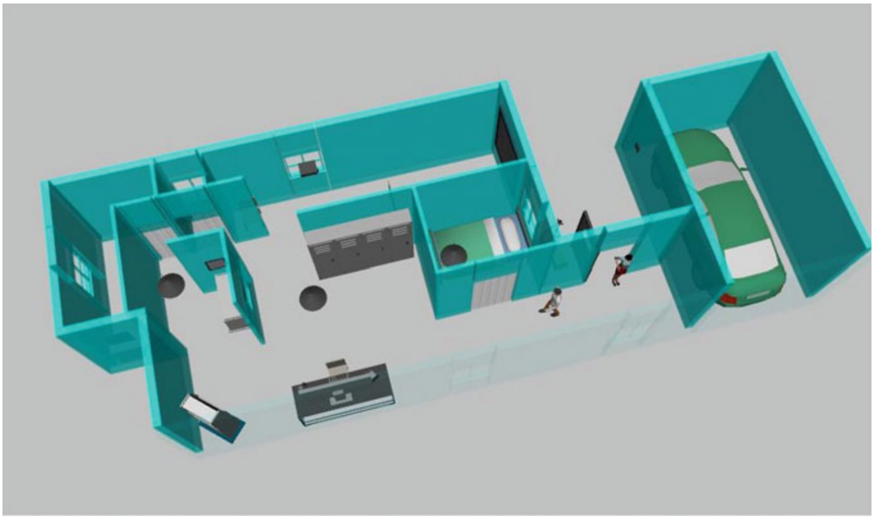


Fig. 2 3D view of the simulated environment

Table 2 Count of death occurrence, AmI features ON (sum of 15 simulation runs)

	HERS OFF	HERS ON
Heart attacks	8	5
Inevitable death	4	2
Preventable death	3	0

situation and call an EMR. HERS should be able to recognize automatically critical situation—fall of a person and notify roommate, relatives or EMR. In the Table 2, it can be seen a number of deaths after the heart attack with and without the system HERS activated. There are two types of heart attacks. Inevitable death means that death comes immediately after the heart attack. Preventable means that resident had chance to be rescued.

4 Conclusion

The tool programmed in AnyLogic for creation and simulation of an intelligent environment had been introduced. Substantial costs can be saved by modeling intelligent environment before its realization. Different house layouts and various sensors can be tested. AnyLogic brings the possibility to add algorithms created in the Java and test different setups of control systems. Residents are modeled as autonomous agents with individual needs and behavior. Many scenarios of different situations can be tested. As an example of a simulation, a house with intelligent systems focused on residents with serious health issues was created and positive impact of the systems on the residents' health status was shown.

New emerging technologies can be tested using the proposed model. It is also possible to test technologies which are still under development, without technical barriers. Results from the simulation runs then can substantiate further adjustments of physical prototypes, which are part of the simulation. Finally, new technological solutions can be inspired by the results of the simulated scenarios.

Acknowledgment The support of Czech Science Foundation GAČR #15-11724S is gratefully acknowledged. Special thanks to programmer Daniel Sacky.

References

1. Babakura, A., Sulaiman, M.N., Mustapha, N., Perumal, T.: Hmm-based decision model for smart home environment. Int. J. Smart Home **8**(1), 129–138 (2014)
2. Hamernik, P., Tanuska, P.: The proposal of ensuring the protection of handicapped inhabitants of smart house. AMM **256–259**, 2962–2966 (2012)
3. Rashidi, P., Mihailidis, A.: A survey on ambient-assisted living tools for older adults. IEEE J. Biomed. Health Inform. **17**(3), 579–590 (2013)
4. Cimler, R., Matyska, J., Sobeslav, V.: Cloud based solution for mobile healthcare application. In: Proceedings of the 18th International Database Engineering and Applications Symposium on—IDEAS'14 (2014)
5. Amiribesheli, M., Benmansour, A., Bouchachia, A.: A review of smart homes in healthcare. J Ambient Intell. Hum. Comput.1–23 (2015)
6. Nehmer, J., Becker, M., Karshmer, A., Lamm, R.: Living assistance systems: an ambient intelligence approach. In: Proceedings of the 28th International Conference on Software Engineering, pp. 43–50. ACM (2006)
7. Kim, S.H., Chung, K.: Emergency situation monitoring service using context motion tracking of chronic disease patients. Cluster Comput. 1–13 (2015)
8. Li, K.F.: Smart home technology for telemedicine and emergency management. J. Ambient Intell. Humaniz. Comput. **4**(5), 535–546 (2013)
9. Bergmann, J., McGregor, A.: Body-worn sensor design: what do patients and clinicians want? Ann. Biomed. Eng. **39**(9), 2299–2312 (2011)
10. Mosa, A.S.M., Yoo, I., Sheets, L.: A systematic review of healthcare applications for smartphones. BMC Med. Inform. Decis. Mak. **12**(1), 67 (2012)
11. Baig, M.M., Gholamhosseini, H.: Smart health monitoring systems: an overview of design and modeling. J. Med. Syst. **37**(2), 1–14 (2013)
12. Lazovik, E., Den Dulk, P., de Groote, M., Lazovik, A., Aiello, M.: Services inside the smart home: a simulation and visualization tool. In: Service-Oriented Computing, pp. 651–652. Springer Berlin (2009)
13. Noury, N., Hadidi, T.: Computer simulation of the activity of the elderly person living independently in a health smart home. Comput. Methods Programs Biomed. **108**(3), 1216–1228 (2012)
14. Garcia-Valverde, T., Serrano, E., Botia, J.A.: Combining the real world with simulations for a robust testing of ambient intelligence services. Artif. Intell. Rev. **42**(4), 723–746 (2014)
15. Serrano, E., Botia, J.: Validating ambient intelligence based ubiquitous computing systems by means of artificial societies. Inf. Sci. **222**, 3–24 (2013)
16. Ishida, T., Hattori, H.: Participatory technologies for designing ambient intelligence systems. J. Ambient Intell. Smart Environ. **1**(1), 43–49 (2009)
17. Borshchev, A.: The big book of simulation modeling: multimethod modeling with AnyLogic 6. AnyLogic North America (2013)

18. Novak, J.D., Canas, A.J.: The origins of the concept mapping tool and the continuing evolution of the tool? Inf Vis **5**(3), 175–184 (2006)
19. Moon, B.M.: Applied concept mapping. CRC Press (2011)
20. Pereira, C., Sousa, C., Soares, A.L.: Supporting conceptualisation processes in collaborative networks: a case study on an R&D project. Int. J. Comput. Integr. Manuf. **26**(11), 1066–1086 (2013)
21. Horalek, J., Matyska, J., Stepan, J., Vancl, M., Cimler, R., Sobeslav, V.: Lower layers of a cloud driven smart home system. Stud. Comput. Intell.219–228 (2015)
22. Stejfa, M.: Kardiologie. Grada (1998)

Design of Mixed Bow-Tie and Ice-Cream Cone Antenna at 120 GHz for Millimeter Wave Application

M.A. Othman, S. Ruthrabathi, M.Z.A. Abd Aziz, H.A. Sulaiman, A.N. Che Pee, M.F.I. Othman and M.R. Kamaruddin

Abstract In this paper, an analysis was done by combining two separate antenna design into one which is the Bow-Tie design and the Ice-Cream cone design. This antenna was designed to operate at 120 GHz for millimetre wave application via CST Microwave Studio Simulation Software. An investigation was done on the antenna return loss and radiation efficiency to the new combined design. This antenna was designed on an Alumina substrate with a dielectric constant 8.8. The proposed antenna shows improvement in the radiation efficiency and the ability to operate at high frequency. Results of the simulation and the antenna design are presented as well.

M.A. Othman (✉) · S. Ruthrabathi · M.Z.A. Abd Aziz
Microwave Research Group, Faculty of Electronics and Computer Engineering,
Universiti Teknikal Malaysia Melaka (UTeM), 76100 Durian Tunggal, Melaka, Malaysia
e-mail: azlishah@utem.edu.my

S. Ruthrabathi
e-mail: ruthra1511@gmail.com

M.Z.A. Abd Aziz
e-mail: mohamadzoinol@utem.edu.my

H.A. Sulaiman ·
A.N.Che Pee · M.F.I. Othman
Communication Network and System Research Group, Faculty of Information Technology and Communication, Universiti Teknikal Malaysia Melaka (UTeM),
76100 Durian Tunggal, Melaka, Malaysia
e-mail: asyrani@utem.edu.my

A.N.Che Pee
e-mail: naim@utem.edu.my

M.F.I. Othman
e-mail: mohdfairuz@utem.edu.my

M.R. Kamaruddin
Wireless Communication Centre (WCC), Fakulti Kej. Elektrik, Universiti Teknologi Malaysia (UTM), 81310 Johor Bahru, Johor, Malaysia
e-mail: ramlee@fke.utm.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_88

Keywords Bow-Tie · Ice-Cream cone · Millimeter wave · Return loss · Radiation efficiency

1 Introduction

In recent years, many types of antenna are designed for millimeter wave applications due to its precision and capability to operate at higher frequencies. Millimeter wave basically linked to the Extremely High Frequency (EHF) radio spectrum in the range of 30–300 GHz with wavelength between one and ten millimeters [1]. The propagation characteristics of millimetre waves through the atmosphere are usually depends on atmospheric oxygen [2], humidity [3], fog [4], and rain [5]. Due to its characteristics as a millimetre wave and high frequency, the antenna architecture will be very small. The types of material used as a substrate and conductor as well as the dielectric characteristics plays an important role in order to determine its performance efficiency [6].

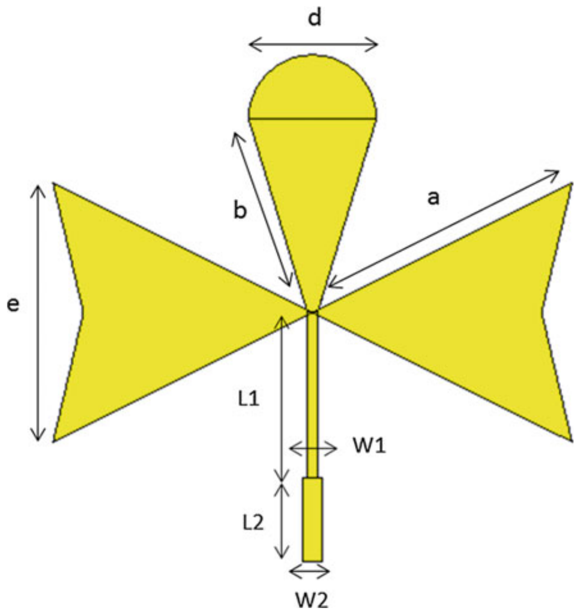
The Bow-tie patch design is based on the design of the triangular micro strip antenna which is the combination of two triangular patches which are mirrored perpendicular. While the design of the Ice-cream cone structure resembles a scoop of ice-cream on a wafer cone [7, 8]. It can be a combination of a triangle and a semicircle, or combination of a rectangle and semicircle [9, 10] . These designs have become an attraction in the present day communication field due to their compact nature. Nowadays the demand for compact wireless communication devices has increased rapidly. This creates an impact to the researches to investigate more on these designs for various applications.

2 Antenna Design and Architecture

The antenna was design on an alumina substrate with the thickness of 0.127 mm and the dielectric constant of 8.8. The dimension of the antenna is 1.6 mm × 1.2 mm. The main structure of the proposed antenna comprises of two main sections which is the Bow-tie structure and the Ice-cream cone structure. These structures are combined and were fed by referring to the 50 ohm characteristic impedance. This conductor structure was designed using gold as the material as well as for the grounding. This antenna was designed to operate at the frequency of 120 GHz. Figure 1 shows the geometrical configuration of the proposed antenna.

The parameters for the antenna design which is the side length of the bow-tie *a,* length of the feeder (L1, L2), width of the feeder (W1, W2), are calculated by using the following formulas, and other parameters are determined by using parametric study [2, 3].

Fig. 1 Geometrical configuration of the antenna



Side length:

$$a = \frac{2c}{2f_r\sqrt{\varepsilon_r}} \tag{1}$$

Efficient value of side length:

$$a_{eff} = a + \frac{h}{\sqrt{\varepsilon_r}} \tag{2}$$

Efficient dielectric constant:

$$\varepsilon_{eff} = \frac{\varepsilon_r + 1}{2} + \frac{\varepsilon_r - 1}{4\sqrt{1 + \frac{12h}{a}}} \tag{3}$$

Wavelength of the antenna:

$$\lambda_g = \frac{\lambda_0}{\sqrt{\varepsilon_{eff}}} \tag{4}$$

The CST Microwave Studio software displays the view of the designed antenna for the millimetre wave application as shown in Fig. 2, as it also reflects the perspective view of the antenna design and the dimension of the substrate.

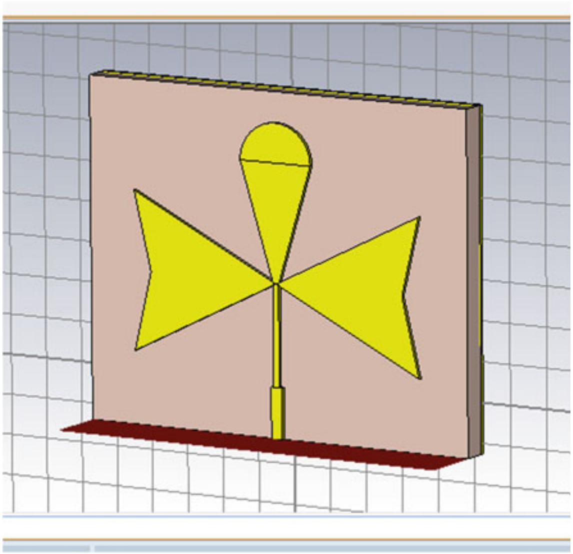


Fig. 2 Perspective view of the proposed antenna

3 Simulated Result and Discussion

The proposed mixed Bow-tie and Ice-cream cone antenna design are simulated and the performance was tested accordingly. The antenna functionality and the efficiency was analysed and observed in the return loss S-parameter graph and the radiation pattern. After certain changes in the parameter of the antenna had been made, the optimized values of these parameters were obtained. Figure 3 presents the return loss graph of the proposed antenna.

From the S-parameter graph, we can observe that the antenna resonates at 121.1 and 137.27 GHz. The return loss for the both resonant frequency are −46.158 and

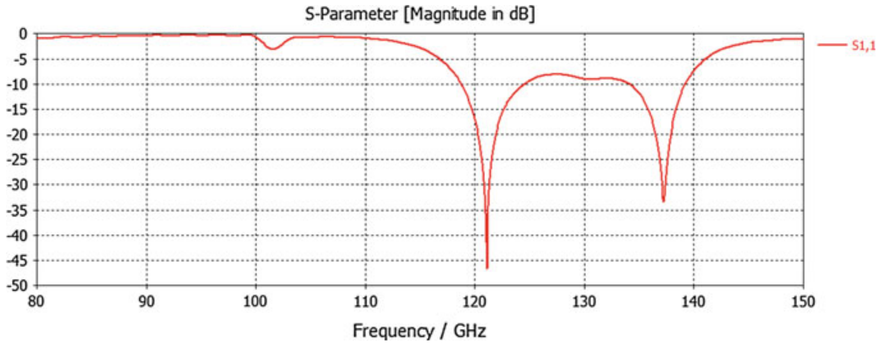


Fig. 3 Return loss of the proposed antenna

−33.238 dB respectively. The designed antenna resonates at two frequency points thus making it suitable for dual band applications (Fig. 4).

Figure 5 presents the far-field plot of the antenna. From the simulated result, we can clearly see the high radiation intensity part of the antenna. The red coloured contour from the simulation describes the region of powerful radiation. The gain obtained for this proposed design was 6.118 dB, while the radiation efficiency and the total efficiency are −1.600 and −1.687 dB respectively. This design shows a stable result in terms of gain and radiation efficiency (Figs. 6, 7 and 8).

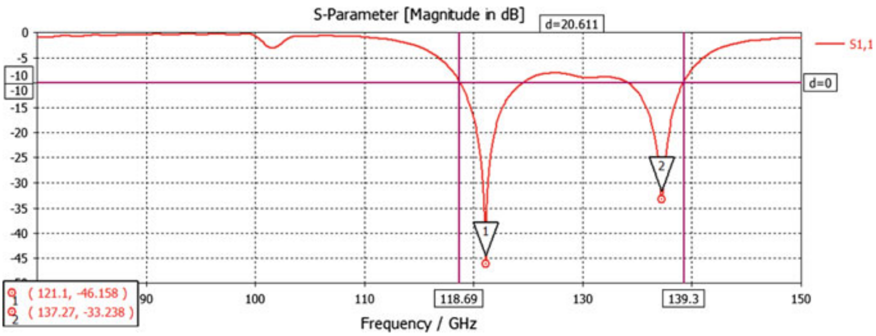


Fig. 4 Return loss marking of the proposed antenna

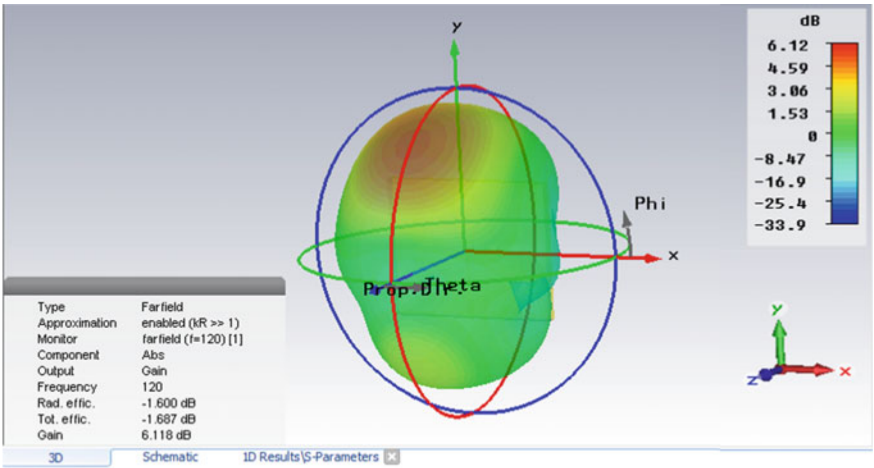


Fig. 5 Simulated radiation pattern in 3D

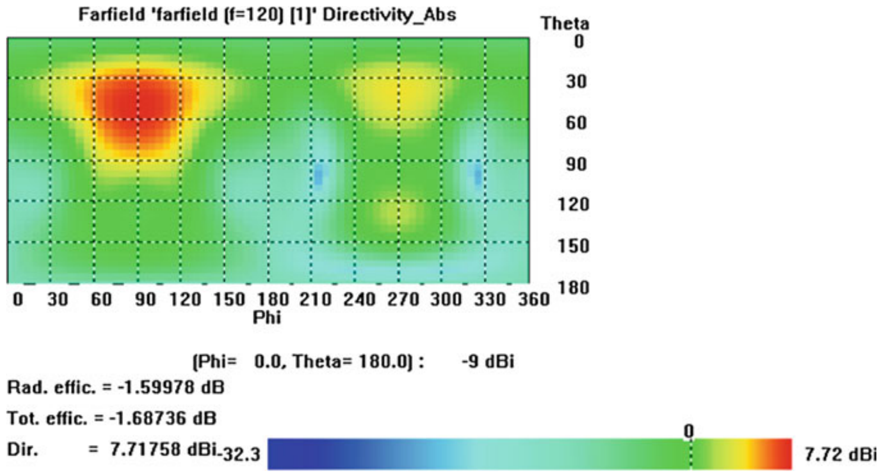


Fig. 6 Simulated farfield in 2D

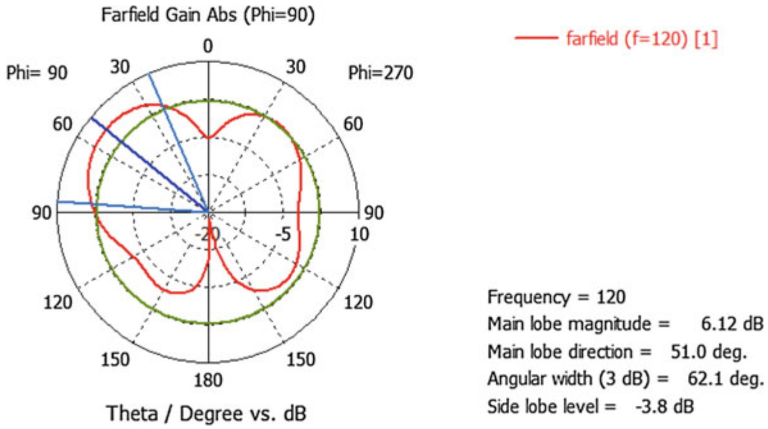


Fig. 7 Simulated farfield in polar

Table 1 shows the performance evaluation of the proposed antenna design. From the simulation, we can see the antenna has a centre frequency at 129.185 GHz. The directivity of the antenna is 7.72 dBi with a main lobe direction of 51.0 deg and side lobe level at −3.8 dB. This antenna shows a satisfying result with an acceptable radiation efficiency and gain. This antenna can be analysed further in future by using different types of material for the substrate and the radiating patch on how these materials affect the radiation pattern. Array type of this antenna can be also studied for further development.

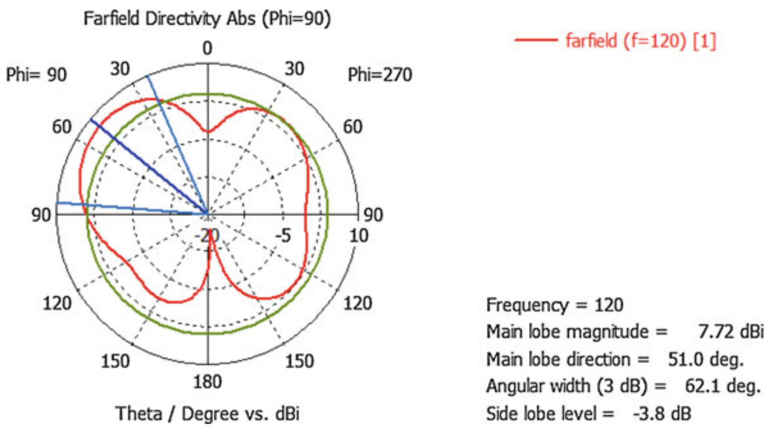


Fig. 8 Simulated directivity in polar

Table 1 Performance evaluation of the proposed design

No.	Parameters	Proposed design
1	Centre frequency	129.185 GHz
2	Gain	6.118 dB
3	Directivity	7.72 dBi
4	Return loss	−46.158/−33.238 dB
5	Radiation efficiency	−1.600 dB

4 Conclusion

An analysis on a mixed bow-tie and ice-cream cone antenna for millimetre wave application was presented in this paper. The proposed antenna was designed and successfully simulated via CST Microwave Studio Software and the results shows that this design can be used for millimetre wave application and satisfies the radiating requirements with a gain of 6.118 dB and a radiation efficiency of −1.6 dB. This antenna design can also be used for dual band application due its two resonant frequencies with a bandwidth of 5.9283 GHz for the first resonant frequency and 5.1716 GHz for the second resonant frequency. The return loss of −46.158 and −33.238 dB achieved also validates the maximum power is transferred to the load. This proposed antenna design will be fabricated in the near future for further analysis.

Acknowledgement I would like to thank first and foremost, the Malaysian Ministry of Education, Universiti Teknikal Malaysia Melaka (UTeM) for financial assistance under project grant FRGS (RACE)/2013/FKEKK/TK3/3 F00194. A special thanks goes to Universiti Teknologi Malaysia (UTM) Research University Grant Vot 00M22 and Vot 00M23 for supporting this paper too. Exceptional dedication to kind-hearted project coordinating supervises lectures, student, parents, public and those involved directly and indirectly for their moral supports, contribution and cooperation in completing this paper. I would like to express my sincere gratitude and respect for those who being helpful when needed whether directly or indirectly.

References

1. Othman. M.A.: Ultra-wide band (UWB) ice cream cone antenna for communication system. Aust. J. Basic Appl. Sci. **7**(3), 10–17 (2013)
2. Pratishthan, V.: Design of bow-tie microstrip antenna with fractal shape for W-Lan application. Int. J. Electron. Commun. Technol. IJECT **3**(4), 445–449 (2012)
3. Rahim, M.K.A., Abdul Aziz, M.Z.A., Goh, C.S.: Bow-Tie microstrip antenna design. In: IEEE 1-4244-0000-7/05/17-20 (2005)
4. Sayidmarie, K.H., Fadhel, Y.A.: Progress in electromagnetics research C, vol. 35, pp. 253–267 (2013)
5. Masud, M.M., Irteza, S.: Design of an inductively coupled miniature UHF bowtie RFID antenna. In: Proceedings of International Bhurban Conference on Applied Sciences and Technology 19–22 January, Islamabad, Pakistan. IEEE, pp. 90–92 (2009)
6. Colin, A., Febvre, P.: Design of a wideband slot bow-tie antenna excited by a microstrip to CPW transition for applications in the millimeter wave band. In: Proceedings of Progress in Electromagnetics Research Symposium 18–21 August, Moscow, Russia, pp. 1421–1425 (2009)
7. Akkermans, J.A.G., Dijk, R.V., Herben, M.H.A.J.: Millimeter-Wave Antenna Measurement. Radiocommunications Group, Eindhoven. University of Technology Eindhoven, The Netherlands (2005)
8. Singh, A.P.: A ice cream shaped microstrip patch antenna. Int. J. Adv. Res. Electron. Commun. Eng. (IJARECE) **4**(2), 498–501 (2013)
9. Kuldip, N.M., Barrie, G.H., Gill, H., Harrison, I.: Low loss ice cream cone ultra wideband antenna. Department of Electrical/Electronic Engineering. The University of Nottingham, Nottingham NG7 2RD, UK. A Compact, pp. 165–168 (2005)
10. Balanis, C.A.: Handbook of Microstrip Antenna. Wiley, New York (1982)

Possibility of Hybrid Multilayered Perceptron Neural Network Realisation on FPGA and Its Challenges

Lee Yee Ann, Phaklen Ehkan and M.Y. Mashor

Abstract This paper reviewed the artificial neural network (ANN), a type of ANN called Hybrid Multilayered Perceptron (HMLP) and existing implementation of ANN on FPGA hardware. The structure of ANN and HMLP is discussed thoroughly. Past works involving HMLP had been reviewed and the HMLP had shown promising improvement over classic MLP. FPGA had seen increasing use for implementing various ANN, however ANN implementations on FPGA had encountered many challenges as discussed in this paper. After the review, it was found that, currently, no implementation of HMLP on FPGA was ever reported. Therefore a novel approach to implement the HMLP directly on FPGA is proposed at the end of the paper. The performance of the proposed FPGA-HMLP is expected to be better due to the characteristic similarity of ANN and FPGA.

Keywords Field programmable gate array · Hybrid multilayered perceptron · Artificial neural network

L.Y. Ann (✉) · P. Ehkan
School of Computer and Communication Engineering, Universiti Malaysia Perlis, Pauh Putra Campus, 02600 Arau, Perlis, Malaysia
e-mail: leeyee4nn@gmail.com

P. Ehkan
e-mail: phaklen@unimap.edu.my

M.Y. Mashor
School of Mechatronic Engineering, Universiti Malaysia Perlis, Pauh Putra Campus, 02600 Arau, Perlis, Malaysia
e-mail: yusoff@unimap.edu.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_89

1 Introduction

Researchers had been trying to imitate the functionalities of human brain for some time. A major breakthrough in this field of research was when, in 1943, McCulloch and Pitts introduced the idea for artificial neuron that model the function of biological neuron cell, which laid the foundation of future ANN researches [1–3]. Currently, there are many kinds of ANN being researched, including HMLP [4]. Researches had also been expended to implement the ANN directly on hardware. Early efforts on this may not been very successful, mostly due to the reasons that most early implementations are ASIC-based which are not competitive enough for large-scale adoption, and performance of FPGA technology at the time are not very capable for ANN implementations [5]. Subsequent improvements in FPGA technology later showed that it can be a viable alternative for hardware implementation of ANN.

2 Artificial Neural Network

Basic interpretation of ANN can be referred to as a network of interconnected artificial neurons, which are simple processing units or nodes [1]. The ANN's processing capability lies on the inter-node connection strength or weight which can be adjusted by allowing the ANN to learn from training data sets [1, 5]. The ANN was introduced with the vision to mimic the operations performed by the biological nervous system and its learning capabilities. The ANN, with its ability to learn, adapt and generalise [6], had shown its capability of successfully solving problems of which conventional processing methods are unable to solve efficiently, such as pattern recognition, classification, function approximation, forecasting, and control [7, 8].

The artificial neurons or nodes in ANN were modelled from a biological neuron cell. These artificial neurons are the basic processing unit of the ANN. The inputs of artificial neuron are similar to the dendrites of biological neuron, the central processing node is the approximation of the soma, and the output correspond the axon. The weights associated to each input represent the properties exhibited by the synapses that bridge the axon of a neuron cell to the dendrites of next connecting neuron cell [1, 8, 9]. The organisation of these artificial neuron nodes into a network makes the ANN a massively parallel processing system [10]. Each artificial neurons in an ANN typically perform a same function, whereby it calculates the weighted sum of its inputs which are then transformed into the output through the activation function. The commonly used activation functions include linear function, threshold function, sigmoid function, and Gaussian function [5, 8].

The ability to learn is fundamental to ANN, which is also an interesting characteristic of ANN. This provides the ANN with the capability to adapt and optimise itself accordingly depending on changes to its inputs [11]. The process of learning for ANN can be viewed as the process of adjusting synaptic weights or the strengths of all connections between every artificial neurons in an ANN [5, 8]. The values of these weights are adjusted repeatedly until gradually the weights of the ANN reached the optimal value for the specified application. The adjustment to these weights is done according to the steps laid-out by specific learning algorithms. The example of popular learning algorithms are back propagation, Levenberg-Marquardt, Recursive Least Square, Recursive Prediction Error, and including any modification done to any of the aforementioned algorithms [12]. Through this learning process, ANN is trained to eventually attain their operational characteristic. Training or learning process may be done in supervised manner where the ANN will be provided with training data sets which specified the desired output for a given input data and the ANN will adjust the weights of its internal connections accordingly to achieve the desired outcome [2, 5] or in unsupervised manner where output data or any information of how to process the input data is not provided and the ANN had to organise itself to classify the input data into clusters of similar pattern [5, 10]. Learning for ANN can be done either online or offline. Offline training is when the ANN is trained on separate location, usually by using a PC or simulator, and the resulting structure and connection weights of the ANN is hard-coded into the ANN for final application, whereas online learning is when the ANN is designed to be able to adjust its connection weights while being deployed for its intended application.

There are many types of ANN, differentiated by the organisation of the nodes in the network and how the nodes are connected to one another, the activation function implemented in the nodes, or the learning algorithms used to train each ANN. Several ANN types are multilayered perceptron (MLP), radial basis function neural network (RBF), Kohonen's self-organising maps neural network (SOM), Hopfield's neural network, Adaptive resonance theory neural network (ART), Probabilistic neural network (PNN), Recurrent neural network (RNN) and HMLP [1, 2, 8, 10].

3 Hybrid Multilayered Perceptron Neural Network

The HMLP is the focus in this paper. It was proposed to improve the performance of MLP [4]. Being the most widely used ANN, the MLP had been modified in various ways to fit it into different range of application. It was noted that classic MLP is a non-linear model, thus a linear system had to be approximated into the non-linear MLP model; nevertheless, modelling a linear system using a linear model is the best approach. A solution to this issue is to improvise the MLP by introducing additional weighted links that connect the input layer directly to the output layer as represented by the dotted lines in Fig. 1. This extra weighted links allows the formation of linear system in parallel with the non-linear MLP topology, thus the name *hybrid multilayered* perceptron. At the same time, modified recursive prediction error (MRPE) algorithm was introduced to train the HMLP [4].

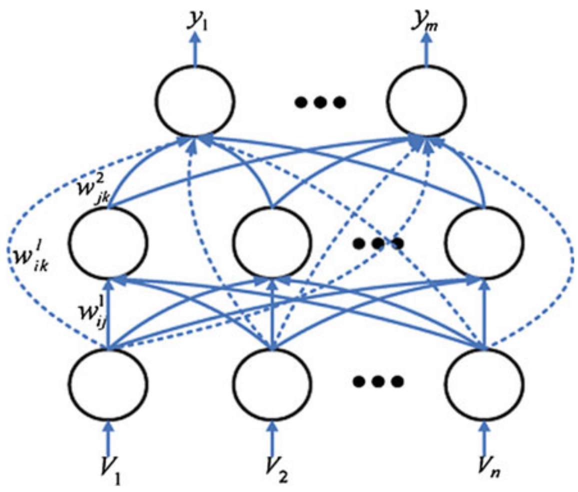


Fig. 1 Hybrid multilayered perceptron network (Reproduced from [13])

The output of the k-th neuron, y_k, in the output layer for HMLP network with one hidden layer can be expressed as

$$\hat{y}_k(t) = \sum_{j=1}^{n_h} w^2_{jk} F\left(\sum_{i=1}^{n_i} w^1_{ij} v^0_i(t) + b^1_j\right) + \sum_{i=0}^{n_i} w^\ell_{ik} v^0_i(t);\ \text{for } 1 \le k \le m. \tag{1}$$

where w^2_{jk} and w^1_{ij} denote the weights of the connections from hidden to output layer and the weights that connect the input to the hidden layer respectively; v^0_i and b^1_j denote the inputs that are supplied to the input layer and thresholds in hidden nodes respectively; n_i, n_h and m are the number of input nodes, hidden nodes and output node respectively. $F(\cdot)$ is the activation function which is normally chosen as sigmoidal function. Notice that, the superscripts indicate the number of layer, for instance v^0_i, w^1_{ij} and w^2_{jk} denote the inputs at 0-th layer (or input layer), weights in first layer (or hidden layer) and weights in the second layer (or output layer) respectively.

The first term on the right-hand side of (1) is similar to the MLP network, of which the HMLP is based on. The second term on the right-hand side of (1) represent the HMLP's extra connections between the input and output layer. w^l_{ik} denotes the weights of the extra connections between input and output layer.

The weights w^2_{jk}, w^1_{ij}, w^l_{ik} and threshold b^1_j are unknown and should be selected to minimise the prediction error defined as

$$\varepsilon_k(t) = y_k(t) - \hat{y}_k(t). \tag{2}$$

where $y_k(t)$ and $\hat{y}_k(t)$ are the expected and network outputs respectively.

3.1 Implementations and Applications of HMLP

To observe and investigate the behaviour and operation of ANN, it needs to be implemented for certain applications. Since its introduction, the HMLP had been implemented for various application such as pattern recognition, classification, and control. It had been employed on fields such as classification of several heart disorders based on ECG input data [13–16], classification of cervical cells [17–19] and breast cells [20, 21] to aid cervical and breast cancer diagnosis, detection of Tuberculosis (TB) Bacilli [11, 22], car speed forecasting [23], fault diagnosis of power system transformer [24], classification of the shape of aggregate [25–27], and satellite attitude controller [28, 29].

Over the years, the HMLP itself had been modified and enhanced by incorporating new or different training algorithms and modifying the structure or topology of the original HMLP. These modifications were done to improve the performance of the HMLP for specific applications such as faster training, better classification and recognition. In [11] and [22], the HMLP used to detect the TB Bacilli in tissue slide images stained using Ziehl-Neelsen method was enhanced by integrating the Extreme Learning Machine training algorithm which resulted in shorter learning duration. Al-Batah et al. [30] proposed a new learning algorithm, the Modified Recursive Least Square (MRLS), for HMLP which is based on Recursive Least Square (RLS) algorithm for MLP. This MRLS for HMLP was tested and compared to HMLP trained with MRPE and MLP trained with RLS using six benchmark datasets. The outcome of the comparison revealed that HMLP-MRLS outperform both HMLP-MRPE and MLP-RLS in term of accuracy, rate of convergence and mean square error (MSE).

Hashim et al. [12] proposed the Multi-Classify HMLP (MCHMLP) for pattern recognition application. MCHMLP was formed by cascading 2 HMLPs into 2 stages where the outputs of the first HMLP stage become the inputs to the second HMLP stage, effectively repeating the classification on the second stage. The outcome revealed that MCHMLP performed better than conventional HMLP while the extra overhead of extra nodes and longer training time is not significant to overall system performance. Another modification to the HMLP, the Clustered-HMLP network and Clustered-MRPE training algorithm was proposed by Mat Isa and Mamat [31]. The performance of this modified HMLP with addition of a cluster layer between the input layer and hidden layer was tested using seven benchmark datasets and the result showed that Clustered-HMLP produce significant improvement compared to HMLP.

4 ANN Implementations on FPGA

In recent times, the ANN had been seen more implementations on hardware, specifically the FPGA. Much can be gained by implement the ANN directly on hardware, especially to exploit the inherent parallelism nature of the ANN [7, 32–34].

Still, majority of ANN implementations nowadays are software-based which run on top of a sequential machine or PCs [5]. Some researchers had used the terms Hardware Neural Network (HNN) [35, 36] or Neuron Machine [37] to refer ANN implementation on hardware, and Field Programmable Neural Array (FPNA) [5, 38] to refer the concept of ANN implementations of artificial neurons with reconfigurable interconnections much like the concept of FPGAs' logic blocks and reconfigurable interconnections. Depending on the application, hardware implementation of ANN is preferable for real-time application that require high processing power [36]; inversely, for certain tasks that do not require real-time computation or intensive processing, software implementation is still appropriate [39].

An FPGA is a type of semiconductor device that was designed to be able programmed, configured, reprogrammed and reconfigured by the user even after manufactured. FPGA contains programmable logic components called logic blocks and reconfigurable interconnections that connect the logic blocks together [36]. These logic blocks and its interconnections can be configured to function as simple gates or as complex combinational circuits. The advantages of FPGA includes the ability to be reprogrammed and reconfigured. Thus, the use of FPGA will significantly reduce the risk of designing unusable system, offering software-like flexibility and allows different designs to be evaluated in a short time [35]. This advantages make the FPGA very suitable for research and prototyping purposes before the design is submitted for fabrication on ASICs. However, the FPGA is relatively much slower than other hard-wired devices and the circuit density of FPGA is comparatively lower that hinders implementation of large systems [35]. Prior to programming the FPGA, the system to be implemented had to be fully described either through schematic capture or by hardware description languages such as VHDL or Verilog.

Interest in the use of FPGA to implement ANN has been increasing over year. The FPGA had been the platform of choice for hardware implementation of ANN for researchers, each using their own design model. RBF networks with online learning had been implemented on Altera Cyclone II FPGA [6] and Xilinx Virtex 6 FPGA [40]. A feed forward ANN for pattern recognition with network training and testing done offline on PC was described using VHDL for Xilinx Zynq-7000 FPGA [41]. SOM had been coded in VHDL for FPGA implementation [42]. Other types of ANN like Spiking NN, Cellular NN, Stochastic NN, Recurrent NN, Pulse Stream NN, and Associative Memory NN to name a few, had too seen FPGA implementations [33]. On the other hand, other researchers focus on specific aspect of ANN implementation as in the FPGA hardware approximation and implementation of various activation functions [43–45], and implementation of learning algorithms like particle swarm optimisation [46] and extreme learning machine [47].

4.1 Challenges for FPGA Implementation

Most of the researches mentioned previously had discussed some of the consideration for implementing their ANN design on FPGA. A major consideration is the limited resources available on the FPGA device, for example, the number of logic elements, resulting in the possibility for the size of the ANN to be implemented being too large that the FPGA is unable to implement [41]. ANN nodes' processing involve heavy use of multiplication and addition due to the large number of inter-node connection in the network, and non-linear activation functions [36, 48]. Direct implementation of multipliers and activation function consume a lot of FPGA resources [49]. Another important aspect to be considered is the data format to be used for the input value, weight value, and activation function calculation as this will affect the precision and overall accuracy of the ANN [50]. ANNs operate with real numbers, floating point data format which offers good precision is preferred but requires more FPGA resources [40]. A compromise had to be made to find right balance between precision and FPGA resource usage. Some quick fixes for these issues are to use fixed point data format instead of floating point at the expense of sacrificing the accuracy [39], to use lookup table (LUT) [51] or LUT with interpolation scheme [43] for multiplication and activation function calculation, and to approximate the activation function through piecewise linear approximation [49], or non-linear approximation [45].

5 Discussion

Through extensive review, software implementations of HMLP are found to be very prevalent. However, hardware implementation of the HMLP is very limited in number. The HMLP had been implemented on 8051 microcontroller in [18], but implementation on FPGA is not available at the moment. Therefore a novel approach to implement the HMLP on FPGA hardware is proposed.

The FPGA architecture of this proposed HMLP implementation will be described using VHDL and the overall system block diagram for this implementation is as in Fig. 2. Research had been ongoing to determine the suitable number of nodes in HMLP and the suitable learning algorithm to be implemented on the FPGA for initial application of classic XOR classification. Once the XOR classification application prove to be successful, the FPGA-HMLP will be applied for other application to compare the performance of this FPGA-HMLP with other existing researches.

Taking into consideration of the limited resources available on the FPGA, the HMLP might had a head start due to the extra links in HMLP that can reduce the number of the hidden nodes compared to MLP [4]. Furthermore, constant improvement of FPGA had gradually increased the capacity of FPGA, increasing the possibility of implementing a large, high precision, fully parallel ANN, with online learning, on a single FPGA device [52].

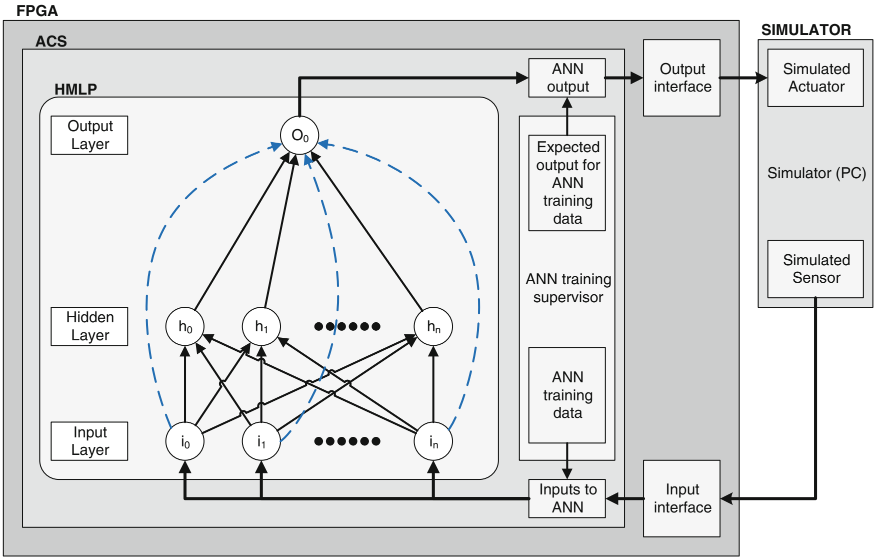


Fig. 2 Proposed block diagram for HMLP implementation on FPGA

6 Conclusion

In this paper, several past researches relating to HMLP had been reviewed, and their implementations and applications had been discussed. The challenges faced by previous works in FPGA implementation of ANNs should be taken into consideration. This proposed FPGA-based implementation of HMLP is expected to perform significantly better due to the characteristic similarity of ANN and FPGA.

References

1. Gurney, K.: An introduction to neural networks. UCL press (1997)
2. Kröse, B., van der Smagt, P.: An introduction to neural networks. University of Amsterdam (1993)
3. McCulloch, W., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. **5**(4), 115–133 (1943)
4. Mashor, M.Y.: Hybrid multilayered perceptron networks. Int. J. Syst. Sci. **31**(6), 771–785 (2000)
5. Omondi, A.R., Rajapakse, J.C. (eds.): FPGA Implementations of Neural Networks. Springer, US (2006)
6. Kim, J.S., Jung, S.: Implementation of the RBF neural chip with the back-propagation algorithm for on-line learning. Appl. Soft Comput. J. **29**, 233–244 (2015)
7. Liu, J., Liang, D.: A survey of FPGA-Based hardware implementation of ANNs. In: International Conference on Neural Networks and Brain (ICNNB) **2**, pp. 915–918 (2005)
8. Jain, A.K., Mao, J., Mohiuddin, K.M.: Artificial neural networks: A tutorial. Computer **29**(3), 31–44 (1996)

9. Gershenson, C.: Artificial neural networks for beginners. CoRR cs.NE/0308031, Cornell University Library (2003)
10. Illingworth, W.T.: Beginner's guide to neural networks. Aerospace and Electronic Systems Magazine, IEEE **4**(9), 44–49 (1989)
11. Osman, M.K., Mashor, M.Y., Jaafar, H.: Hybrid multilayered perceptron network trained by modified recursive prediction error-extreme learning machine for tuberculosis bacilli detection. In: Osman, N., Abas, W., Wahab, A., Ting, H.-N., (eds.) 5th Kuala Lumpur International Conference on Biomedical Engineering 2011 **35**, pp. 667–673, Springer, Berlin Heidelberg (2011)
12. Hashim, F.R.B., Soraghan, J.J., Petropoulakis, L.: Multi-classify hybrid multilayered perceptron (HMLP) network for pattern recognition applications. In: Iliadis, L., Maglogiannis, I., Papadopoulos, H., (eds.) Artificial Intelligence Applications and Innovations **381**, 19–27, Springer, Berlin Heidelberg (2012)
13. Megat Ali, M.S.A., Jahidin, A.H., Norali, A.N.: Hybrid multilayered perceptron network for classification of bundle branch blocks. In: International Conference on Biomedical Engineering (ICoBE), Penang, pp. 149–154 (2012)
14. Megat Ali, M.S.A., Shaari, N.F., Julai, N., Jahidin, A.H., Amiruddin, A.I., Noor, M.Z.H., Saaid, M.F.: Robust arrhythmia classifier using hybrid multilayered perceptron network. In: IEEE 9th International Colloquium on Signal Processing and its Applications (CSPA), Kuala Lumpur, pp. 304–309 (2013)
15. Megat Ali, M.S.A., Rani, M.F., Jahidin, A.H., Saaid, M.F., Noor, M.Z.H.: Identification of cardiomyopathy disease using hybrid multilayered perceptron network. In: IEEE International Conference on Control System, Computing and Engineering (ICCSCE), Penang, pp. 23–27 (2013)
16. Amiruddin, A.I., Megat Ali, M.S.A., Saaid, M.F., Jahidin, A.H., Noor, M.Z.H.: Feature reduction and arrhythmia classification via hybrid multilayered perceptron network. In: IEEE 3rd International Conference on System Engineering and Technology (ICSET), Shah Alam, pp. 290–294 (2013)
17. Jusman, Y., Mat Isa, N.A., Adnan, R., Othman, N.H.: Intelligent classification of cervical pre-cancerous cells based on the FTIR spectra. Ain Shams Eng. J. **3**(1), 61–70 (2012)
18. Yahaya, S.Z., Mat Isa, N.A.: Implementation of HMLP network with different activation function for cervical cells classification. In: IEEE 7th International Colloquium on Signal Processing and its Applications (CSPA), Penang, pp. 266–271 (2011)
19. Mat-Isa, N.A., Mashor, M.Y., Othman, N.H.: An automated cervical pre-cancerous diagnostic system. Artif. Intell. Med. **42**(1), 1–11 (2008)
20. Sakim, H.A.M., Salleh, N.M., Arshad, M.R., Othman, N.H.: Evaluation of morphological features for breast cells classification using neural networks. Stud. Comput. Intell. **166**, 1–9 (2009)
21. Sakim, H.A.M., Salleh, N.M., Othman, N.H.: Neural network inputs selection for breast cancer cells classification. Stud. Comput. Intell. **199**, 1–11 (2009)
22. Osman, M.K., Mashor, M.Y., Jaafar, H.: Detection of tuberculosis bacilli in tissue slide images using HMLP network trained by extreme learning machine. Elektron. ir Elektrotechnika. **120** (4), 69–74 (2012)
23. Saad, Z., Osman, M.K., Mashor, M.Y.: Modelling and forecasting of car speed using detrended inputs and hybrid multilayered percepton network. Contemp. Eng. Sci. **7**(13–16), 603–610 (2014)
24. Mamat, W.M.F.W., Isa, N.A.M., Zamli, K.Z., Mamat, W.M.F.W.: Hybrid version of MLP neural network for transformer fault diagnosis system. In: International Symposium on Information Technology (ITSim), Kuala Lumpur, vol. 2 (2008)
25. Isa, N.A.M., Joret, A., Ali, A.N., Zamli, K.Z., Azizli, K.A.: Application of artificial neural networks to classify the shape of aggregate. WSEAS Trans. Syst. **4**(6), 846–853 (2005)
26. Isa, N.A.M., Al-Batah, M.S., Zamli, K.Z., Azizli, K.A., Joret, A., Noor, N.R.M.: Suitable features selection for the HMLP and MLP networks to identify the shape of aggregate. Constr. Build. Mater. **22**(3), 402–410 (2008)

27. Al-Batah, M.S., Mat Isa, N.A., Zamli, K.Z., Sani, Z.M., Azizli, K.A.: A novel aggregate classification technique using moment invariants and cascaded multilayered perceptron network. Int. J. Min. Process. **92**(1–2), 92–102 (2009)
28. Sharun, S.M., Mashor, M.Y., Norhayati, M.N., Yaacob, S., Yaakob, M., Jaafar, W.N.W.: Adaptive neuro-controller based on HMLP network for InnoSAT attitude control. In: International Conference on Electrical, Control and Computer Engineering (INECCE), Kuantan, pp. 355–360 (2011)
29. Sharun, S.M., Mashor, M.Y., Nazid, N.M., Yaacob, S., Jaafar, W.N.W.: Innosat attitude control system based on adaptive neuro-controller. J. Inf. Commun. Technol. **10**, 45–65 (2011)
30. Al-Batah, M.S., Mat Isa, N.A., Zamli, K.Z., Azizli, K.A.: Modified recursive least squares algorithm to train the hybrid multilayered perceptron (HMLP) network. Appl. Soft Comput. J. **10**(1), 236–244 (2010)
31. Mat Isa, N.A., Mamat, W.M.F.W.: Clustered-hybrid multilayer perceptron network for pattern recognition application. Appl. Soft Comput. J. **11**(1), 1457–1466 (2011)
32. Zhu, J., Sutton, P.: FPGA implementations of neural networks—a survey of a decade of progress. In: Cheung, P.Y.K., Constantinides, G., (eds.) Field Programmable Logic and Application **2778**, pp. 1062–1066. Springer, Berlin Heidelberg (2003)
33. Ehkan, P., Zakaria, F.F., Razalli, M.S., Fuad, F.F.A., Warip, M.N.M.,Yuen, L.C.: Hardware implementation of modeling frequency coded serial communication for Eurobalise using ASK module. Appl. Mech. Mater. **548**–**549**, 785–797 (2014)
34. Ehkan, P., Ann, L., Zakaria, F., Warip, M.N.: Artificial neural network for character recognition on embedded-based FPGA. In: Park, J., Pan, Y., Kim, C.S., Yang, Y. (eds.) Future Information Technology 309, pp. 281–287. Springer, Berlin Heidelberg (2014)
35. Misra, J., Saha, I.: Artificial neural networks in hardware: A survey of two decades of progress. Neurocomputing. **74**(1–3), 239–255 (2010)
36. Lotrič, U., Bulić, P.: Applicability of approximate multipliers in hardware neural networks. Neurocomputing **96**, 57–65 (2012)
37. Ahn, J.B.: Computation of backpropagation learning algorithm using neuron machine architecture. In: 5th International Conference on Computational Intelligence, Modelling and Simulation (CIMSim), Seoul, pp. 23–28 (2013)
38. Bohrn, M., Fujcik, L., Vrba, R.: Field programmable neural array for feed-forward neural networks. In: 36th International Conference on Telecommunications and Signal Processing (TSP), Rome, pp. 727–731 (2013)
39. Nedjah, N., Da Silva, R.M., De MacEdo Mourelle, L.: Compact yet efficient hardware implementation of artificial neural networks with customized topology. Expert Syst. Appl. **39** (10), 9191–9206 (2012)
40. de Souza, A.C.D., Fernandes, M.A.C.: Parallel fixed point implementation of a radial basis function network in an FPGA. Sens. **14**(10), 18223–18243 (2014)
41. Dondon, P., Carvalho, J., Gardere, R., Lahalle, P., Tsenov, G., Mladenov, V.: Implementation of a feed-forward Artificial Neural Network in VHDL on FPGA. In: 12th Symposium on Neural Network Applications in Electrical Engineering (NEUREL), pp. 37–40 (2015)
42. Brassai, S.T.: FPGA based hardware implementation of a self-organizing map. In: 18th International Conference on Intelligent Engineering Systems, pp. 101–104 (2014)
43. Ortega-Zamorano, F., Jerez, J.M., Juarez, G., Perez, J.O., Franco, L.: High precision FPGA implementation of neural network activation functions. In: IEEE Intelligent Embedded Systems (IES), pp. 55–60 (2014)
44. Jeyanthi, S., Subadra, M.: Implementation of single neuron using various activation functions with FPGA. In: International Conference on Advanced Communication Control and Computing Technologies, pp. 1126–1131 (2015)
45. Xie, Z.: A non-linear approximation of the sigmoid function based on FPGA. In: IEEE 5th International Conference on Advanced Computational Intelligence, Nanjing, pp. 221–223 (2012)

46. Bezborah, A.: A hardware architecture for training of artificial neural networks using particle swarm optimization. In: 3rd International Conference on Intelligent Systems, Modelling and Simulation, Kota Kinabalu, pp. 67–70 (2012)
47. Decherchi, S., Gastaldo, P., Leoncini, A., Zunino, R.: Efficient digital implementation of extreme learning machines for classification. IEEE Trans. Circuits Syst. II Express Briefs. **59** (8), 496–500 (2012)
48. Youssef, A., Mohammed, K., Nasar, A.: A Reconfigurable, Generic and Programmable Feed Forward Neural Network Implementation in FPGA. In: 14th International Conference on Computer Modelling and Simulation (UKSim), Cambridge, Cambridgeshire, pp. 9–13 (2012)
49. Hariprasath, S., Prabakar, T.N.: FPGA Implementation of Multilayer Feed Forward Neural Network Architecture Using VHDL. In: International Conference on Computing, Communication and Applications (ICCCA), Dindigul, Tamilnadu (2012)
50. Çavuşlu, M.A., Karakuzu, C., Şahin, S., Yakut, M.: Neural network training based on FPGA with floating point number format and it's performance. Neural Comput. Appl. **20**(2), 195–202 (2011)
51. Lozito, G.-M., Laudani, A., Riganti-Fulginei, F., Salvini, A.: FPGA implementations of feed forward neural network by using floating point hardware accelerators. Adv. Electr. Electron. Eng. **12**(1), 30–39 (2014)
52. Bahoura, M., Park, C.W.: FPGA-implementation of high-speed MLP neural network. In: 18th IEEE International Conference on Electronics, Circuits and Systems (ICECS), Beirut, pp. 426–429 (2011)

Assessment of Statistical and Empirical Conversion Methods of Integration Time for Rainfall Rate in Malaysia

Nuurul Hudaa Mohd Sobli, Ahmad Fadzil Ismail, Ani Liza Asnawi, Mimi Aminah Wan Nordin and Muhamad Haziq Khairolanuar

Abstract This paper presents some preliminary observations of assessments regarding the precipitation intensity conversion methods from 60- to 1-min integration time. There are two conversion methods were identified and used for this findings; Rice-Holmberg and Khairolanuar et al. conversion methods. Different type of precipitation were analyzed and implemented in analytical method of Rice-Holmberg and characteristics of rainfall rate distributions were evaluated from the empirical method of Khairolanuar et al. Rainfall intensity data were acquired from Malaysia Meteorological Department (MMD). The rainfall rate data consist of twelve consecutive months from January to December 2009. The evaluations for the acquired data were carried out to produce annual rainfall rate cumulative distribution and as well as its cumulative distribution at 1-min integration time utilizing the aforementioned conversion methods. The comparisons of the performance of these two methods were also examined. Based on the evaluation, it can be observed that Rice-Holmberg and Khairolanuar et al. produced percentage errors around 19 % as compared to ITU-R P.837 at 0.01 % of time exceedance.

N.H.M. Sobli (✉) · A.F. Ismail · A.L. Asnawi · M.A.W. Nordin · M.H. Khairolanuar
Department of Electrical and Computer Engineering, Kulliyyah of Engineering, International Islamic University Malaysia (IIUM), Jln. Gombak, 53000 Selangor, Malaysia
e-mail: hudaa@iium.edu.my

A.F. Ismail
e-mail: af_ismail@iium.edu.my

A.L. Asnawi
e-mail: aniliza@iium.edu.my

M.A.W. Nordin
e-mail: mimie@iium.edu.my

M.H. Khairolanuar
e-mail: haziq.khairolanuar@yahoo.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_90

1 Introduction

Distribution of precipitation intensity is inhomogeneous in space and time. Rainfall rate or rainfall intensity is defined as “a measure of the intensity of rainfall expressed by the increase in the height of water reaching the ground per unit time” [1]. It is universally conveyed in millimeter per hour and denoted as mm/hr. There are several equipments being used to obtain rainfall intensity measurements and depend on their locations. Some studies had been steered to compare the evaluation of rainfall rates logged using different type of precipitation measuring devices. Previous studies had indicated that different types of these equipments that are co-located at the location of interest would produce different rainfall rates [2].

It is well known that radio wave propagation of satellite links encounters severe attenuation at 10 GHz and above in rainy condition. It becomes a major concern for communication system engineers especially for locations in tropical region. Rain attenuation occurs due to the absorption and scattering of energy by raindrops that degrade the reliability and performance of the communication link. Prediction of rain attenuation for estimation of unavailable time percentage of the communication link above 10 GHz requires precise information and characteristics about rain rate distribution at the location of interest. The accurate knowledge of rainfall rate statistics is used as input for attenuation prediction method that evaluates the statistical behavior of the attenuation caused by rain.

Therefore, the International Telecommunication Union (ITU) recommends a standard that uses rain rate cumulative distribution functions CDFs with 1-min integration time for derivation of attenuation CDFs. The 1-min integration time was selected as a compromise between experimental accuracy and the amount of available rainfall data. Furthermore, it has been confirmed in literature that 1-min integration time interval promises a good level of precision in regenerating the time inconsistency of attenuation [3, 4]. However, 1-min integration time is not the standard time interval that is utilized in meteorological applications because of no significance of rapid changes of precipitation intensity. They most preferable in average quantities i.e. hourly, daily, monthly, or yearly accumulated rain. Thus, most of the data acquired from meteorological departments have longer integration time.

The non-availability of such integration time of rainfall rate data had motivated researchers to develop procedures that focusing on achieving CDFs of rainfall rate with 1-min time resolution from CDFs of rainfall rate of longer integration time or from the information of general local meteorological conditions. There are various prediction methods proposed by researchers and can be classified into three different sets, namely meteorology, analytical, and empirical based methods.

Meteorology based methods utilizes general climatic information as input (i.e. average monthly or yearly rainfall rate, number of rainy days per year, peak annual rain rate, etc.) such as proposed by Dutton et al. [5], ITU-R [6], Rice and Holmberg [7], and Crane [8] methods. On the other hand, analytical methods assumed that by changing the integration time resolution would not affect the analytical form of

CDFs of rainfall intensity. Different integration time has different CDFs due to the different parameters of CDFs equation. These methods give the prediction behavior of the parameters of the distribution change according to the integration time as proposed by Moupfouma and Martin [9], Karasawa and Matsudo [10]. Finally, empirical methods that are most widely proposed by researchers around the world. These methods provide conversion factors between the known CDF and the one to be estimated as a function of probability level. This type of method were utilized in the proposed methods by Segal and Allnutt [11], Chebil and Rahman [12], Burgueno et al. [13], Joo et al. [14], Watson et al. [15], Ismail et al. [16] etc. Comparison analyses using empirical conversion methods of integration time can be found in [17].

2 Conversion Methods from 60- to 1-Minute Integration Time

In 1973, Rice and Holmberg developed a global surface rain rate model from extensive long-term rain rate statistics from over 150 locations throughout the world [7]. Rice-Holmberg model is also known as R-H model provides a statistical rainfall rate distribution by assuming that the rain structure can be categorized into two types, or modes. These modes or types are "thunderstorm rain" and "all other rain". Each mode is determined by exponential functions and the sum of these two modes yields the total rainfall rate distribution. The percentage of an average year for which the rain rate exceeds R mm/hr at a medium location is as follows [7]:

$$P(R) = M\left\{0.03\beta e^{-0.03R} + 0.2(1-\beta)\left[e^{-0.258R} + 1.86e^{-1.63R}\right]\right\}. \tag{1}$$

where M is the average annual rainfall accumulation (mm), M_1 is annual accumulation of thunderstorm rain (mm), $\beta = M_1/M$ thunderstorm ratio, and R is clock minute rate (mm/hr).

The model provides global maps for β, M_1, and M. However, directly measured data can be used when available. Besides, R-H model has shown to produce very good agreement with directly measured long-term rain rate data for locations in the United States.

In 1994, a recommendation ITU-R P.837 had been proposed as a reference on evaluation of the characteristics of precipitation for propagation modeling. It suggests a global map with different climatic zones to represent general characteristics of the rainfall rate at the location of interest. Further development and improvement had been performed in order to produce a reliable conversion method for ITU-R. In 2009, a new-revised method proposed by Capsoni and Luini that exploits EXCELL Rainfall Statistics Conversion (EXCELL RSC) model [18]. It has the capability of facilitating users to generate statistics known as $P(R)$ of the local rainfall intensity, R (mm/hr) at 1-min integration time [18]. Their proposed method provides options for users to input either from local measurements statistic of rainfall intensity at integration times up to 60 min or from the global digital maps of rainfall parameters derived from numerical weather prediction data. The statistics of compilation data using locally measured rainfall intensity and integrated conversion model are predicted to provide the best approximation.

New empirical conversion method proposed by Khairolanuar et al. [19]. This empirical conversion method utilizes polynomial relationship using rainfall intensity data acquired in Malaysia [19]. Optical rain gauge was used for data collection for the duration of almost two years period. The estimation of 1-minute integration time of rainfall intensity statistics is summarized as follows [19]:

$$R_1(P) = aR_\tau(P)^4 + bR_\tau(P)^3 + cR_\tau(P)^2 + dR_\tau(P) + e. \quad (2)$$

where $R_\tau(P)$ is the input of rainfall intensity data for τ integration time and a, b, c, d, and e are the empirical constants [19].

3 System Set-up for Data Collection and Measurements

The ground truth measurement data were collected from the Malaysia Meteorological Department (MMD) rain gauge station located at station Kuala Lumpur International Airport (KLIA) with 2° 44′N and 101° 42′E, about 5 km from the airport and 16.3 m above mean sea level [20]. The location of this study was selected at KLIA so that further analyses concerning the weather surroundings airport can be carried out in the future works.

The rain gauge used by MMD to collect the measured rainfall values consists of standard tipping bucket. The tipping bucket rain gauge follows the standard by World Meteorological Organization (WMO). This tipping bucket collects rainfall rate data every 60 min. The integration time for the collected data for the collected data is 60 min [20].

The tipping bucket comprises of two components; funnel-shaped at the top supported by a cylindrical-shaped at the bottom. This funnel has a water filter at the end of the funnel opening. The water drops of rainfall land in the funnel of the tipping bucket rain gauge. Water flowing into the funnel will be screened and will be collected by two metal water collectors (tipping buckets). The raindrops is poured into the cylinder when one of the collectors receives the raindrops of amount 0.2 mm and the next rain will then fall to the other metal collectors. This process is repeated and this repetition process is connected to a computing system (counter) that will count the number of times the rain that falls into the water collector metal. The amount of rainfall rate is calculated based on multiplication of the number of times the precipitation that falls on the metal rain collector with 0.2 mm of rain droplets. Maximum rainfall amount that can be obtained is 200 mm/h. Table 1 summarized the specifications of the tipping bucket rain gauge operated by MMD and Fig. 1 shows the tipping bucket that is located at KLIA [20].

Table 1 Specifications of tipping bucket rain gauge

	Specifications
Location (Latitude, Longitude)	KLIA, Sepang (2° 44′N and 101° 42′E)
Distance from KLIA	±5 km
Receiving collector	203 mm ± 0.2 mm
Accuracy	±1 % to 200 mm/h
Bucket capacity	0.2 mm
Dimensions	300 mm height, 230 mm body diameter, 280 mm base diameter
Physical	5.5 kg net weight

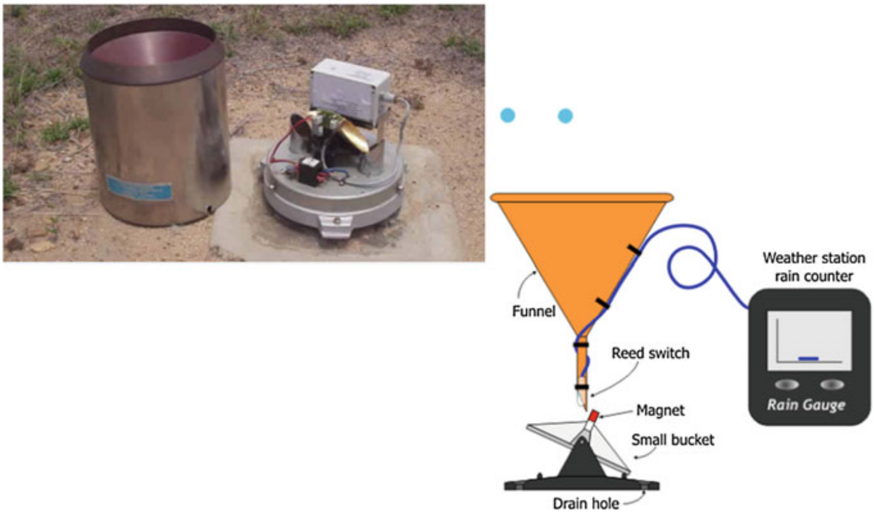


Fig. 1 MMD tipping bucket at KLIA

Figure 2 shows the overall system setup for the study that includes Terminal Doppler Weather Radar that is located at Bukit Tampoi. However, this paper presents only the study of the measured ground truth rainfall rate at the rain gauge station KLIA not inclusive radar data. In addition, radar data consists the information of rainfall rate that will be used for further analysis together with the rain gauge measurement data to find the correction of radar reflectivity to rainfall rate relationships.

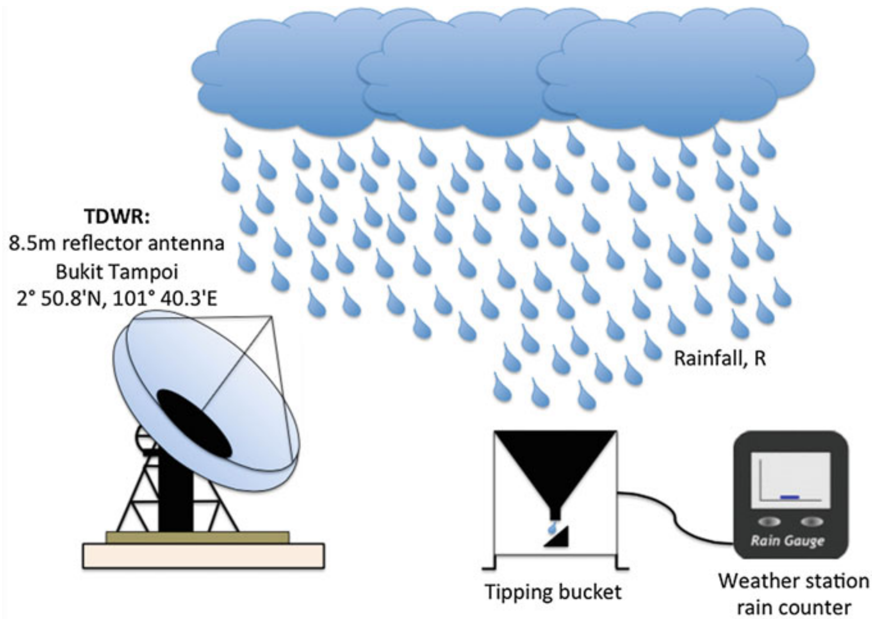


Fig. 2 Overall system set-up for data collection and measurements

4 Results and Discussions

The sampling time for the rainfall data acquired at KLIA is 60-min. The data were measured in real-time quantities for rainfall rate and precipitation accumulation. The measurements of this data were taken from January to December 2009 for period of one year. The highest precipitation intensity value recorded within one-year duration is 71.83 mm/h that occurred in March 2009.

The acquired rainfall data were analyzed in order to obtain its annual cumulative distribution function (CDF). The annual CDF of rainfall rate at KLIA is obtained from the accumulation of monthly CDFs. Figure 3 shows the annual CDF of rainfall data at KLIA for 2009. The figure illustrates that KLIA had experienced rainfall intensity of approximately 54.4 mm/h at 0.01 % time exceedance that is equivalent to 9 h of the year.

The acquired data were then further analyzed by characterizing the rainfall intensity types into stratiform and convective. These two different types of rainfall rate are important in order to implement the Rice-Holmberg's conversion method. The R-H method utilizes the use of two types of rain modes, M_1(mm) and M_2(mm) that represents the thunderstorm rain and all other rain respectively. Total of these two modes gives the average annual rainfall accumulation, M(mm). The ratio of thunderstorm rain to the average annual rainfall accumulation gives the thunderstorm ratio, β. From the collected data, it was found that the average annual rainfall

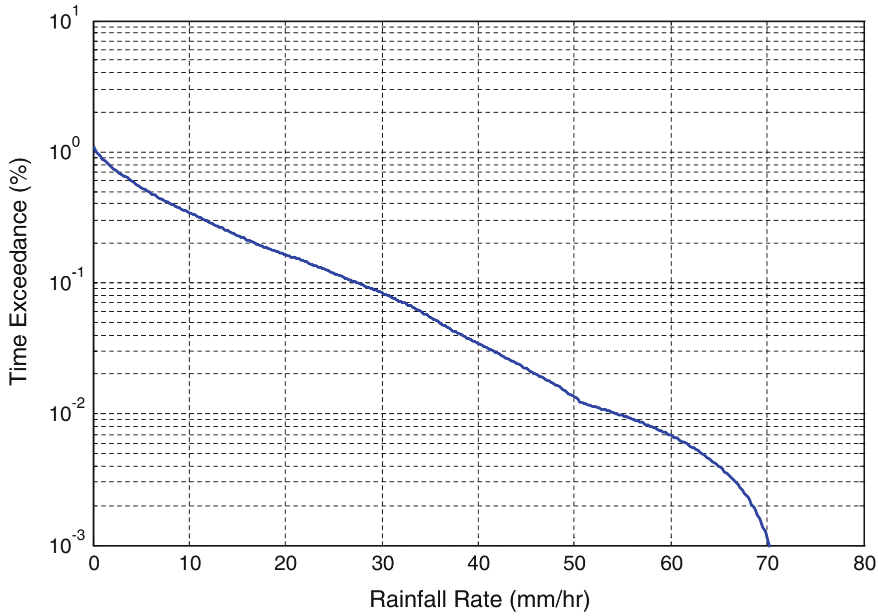


Fig. 3 Annual CDF of rainfall rate at KLIA

M is 2115.6 mm, thunderstorm rain M_1 is 261.4 mm, and thunderstorm ratio β is 0.123583. Then Eq. 1 was used to convert from 60-min integration time to 1-min equivalent integration time of the collected data. Khairolanuar et al. method as described in [19] was used to obtain the equivalent 1-min integration time by using Eq. 2.

Figure 4 depicts the comparison of R-H and Khairolanuar et al. for the collected rainfall data at KLIA. From the figure, it can be deduced that rain intensity value at 0.01 % time exceedance generated by R-H conversion method is 73.0 mm/h and Khairolanuar et al. conversion method is 74.14 mm/h. Table 2 shows summary of rainfall intensity values at 0.01 % time exceedance for the aforementioned converted statistics. The values of rainfall precipitation were chosen at 0.01 % time of exceedance, $R_{0.01}$ because of its reliability for rainfall attenuation prediction as had been used widely in the literature.

Further investigation of the assessments of each method applicability were carried out by determining the root mean square error (RMSE) and the percentage error of rainfall intensity value at 0.01 % of time exceedance. The relevant RMSE for these preliminary findings is defined as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(ECD_i - ITUD_i\right)^2}. \tag{3}$$

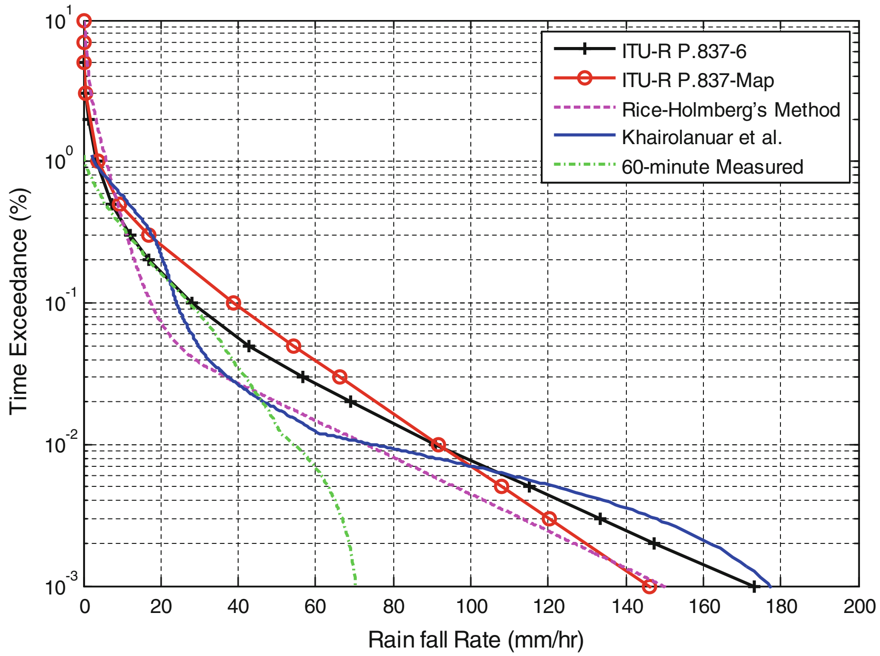


Fig. 4 Comparison of Rice-Holmberg, ITU-R, and Khairolanuar et al. conversion for rainfall rate at KLIA

Table 2 Rainfall intensity values at 0.01 % time exceedance ($R_{0.01}$)

	Rice-Holmberg	ITU-R P.837-Map	ITU-R P.837-6	Khairolanuar et al.
$R_{0.01}$ (mm/h)	73.0	91.75	90.77	74.14

where n is the number of data, ECD is the empirical conversion data and ITUD is the ITU-R data.

Table 3 summarizes the RMSE and percentage difference values at 0.01 % of time exceedance for Rice-Holmberg, ITU-R P.837-6, and Khairolanuar et al. conversion methods of rainfall intensity as compared to ITU-R P.837-Map rainfall intensity of 91.75 mm/h. The table shows that ITU-R P.837-6 physical has the lowest RMSE and percentage difference with 10.958 and 1.068 % respectively. According to the table, it also exhibits that Rice-Holmberg and Khairolanuar et al. yield RMSE of 19.451 and 18.388 % respectively and their percentage difference of 20.341 and 19.193 % respectively.

Table 4 demonstrates the RMSE and percentage difference of the above conversion methods as compared with ITU-R P.837 physical rainfall rate of 90.77 mm/h at 0.01 % of time exceedance. From the table, it can be deduced that

Table 3 Root mean square error and percentage error at 0.01 % time exceedance as compared with ITU-R P.837 rain map

	Rice-Holmberg	ITU-R P.837-6	Khairolanuar et al.
RMSE (%)	19.451	10.958	18.388
Percentage difference (%)	20.341	1.068	19.193

Table 4 Root mean square error and percentage error at 0.01 % time exceedance as compared with ITU-R P.837-6 conversion method

	Rice-Holmberg	ITU-R P.837 Map	Khairolanuar et al.
RMSE (%)	13.657	10.958	12.283
Percentage difference (%)	19.243	1.080	18.321

ITU-R P.837 rain map gives the lowest RMSE and percentage difference values with 10.958 and 1.080 % respectively. The table also reveals that RMSE and percentage difference produced by Rice-Holmberg and Khairolanuar et al.'s had slight difference with values of 13.657, 19.243, and 12.283 %, 18.321 % respectively.

5 Conclusion and Future Works

The scope of this paper is to study the comparison of conversion methods from 60 to 1-min integration time of rainfall rate. The comparison was carried out by selecting ITU-R P.837 proposed equivalent 1-min values at 0.01 % of time exceedance as a benchmark because of its applicability to the local climates. It can be concluded that RMSE and percentage difference values for Rice-Holmberg and Khairolanuar et al. result in around 19 % as compared to ITU-R P.837 at 0.01 % time exceedance are. The comparison of different conversion methods is useful in the analysis of rain attenuation prediction methods due to the fact of the significance of having 1-min integration time in such study. The results presented in this paper are obtained from calculations of rainfall data collected at one location for the duration of one year. More analyses need to be carried out using rainfall intensity data from different locations in Malaysia. Therefore, current evaluations are carried out using rainfall data at different locations with different year and the findings will be disclosed in future works and analyses.

Acknowledgments The authors wish to thank Malaysia Meteorological Department (MMD), Ministry of Science, Technology and Information (MOSTI), and IIUM Endowment Research Grant Funding for their endless support and help for this research.

References

1. Recommendation ITU-R P.310–9.: Definition of Terms Relating to Propagation in Non-Ionised Media. (1994)
2. Brand, E.: A Comparative study of techniques for measuring rainfall rate and accumulation. In: Proceedings AP2000 Millennium Conference on Antennas and Propagation (2000)
3. Segal, B.: The influence of rain gauge integration time on measured rainfall-intensity distribution functions. J. Atmos. Oceanic Technol. **3**, 662–671 (1986)
4. Chebil, J., Rahman, T.A.: Rain rate statistical conversion for the prediction of rain attenuation in Malaysia. IEEE Electron. Lett. **35**, 1019–1021 (1999)
5. Dutton, E.J., Dougherty, H.T., Martin Jr, R.F.: Prediction of European Rainfall and Link Performance Coefficients at 8 to 30 GHz. Institute for Telecommunications Science. U.S. Department Commerce, Washington DC (1974)
6. Recommendation ITU-R pp.837–6.: Characteristics of Precipitation for Propagation Modeling (2012)
7. Rice, P.L., Holmberg, N.R.: Cumulative time statistics of surface-point rainfall rates. IEEE Trans. Commun. **21**, 1131–1136 (1973)
8. Crane, R.K.: Prediction of attenuation by rain. IEEE Trans. Commun. COM-28, 1717–1733 (1980)
9. Moupfouma, F., Martin, L.: Modeling of the rainfall rate cumulative distribution for the design of satellite and terrestrial communication systems. Int. J. Satell. Commun. **13**, 105–115 (1995)
10. Karasawa, Y., Matsudo, T.: One-minute rain rate distribution in Japan derived from AMeDAS one-hour rain rate data. IEEE Trans. Geosci. Remote Sens. **29**, 890–898 (1991)
11. Segal, B., Allnutt, J.E.: On the use of long sampling-time rainfall observation for predicting high-probability attenuation on earth-space links. Proc. Int. Conf. Antennas Propag. **2**, 754–757 (1991)
12. Chebil, J., Rahman, T.A.: Development of 1-minute rain rate contour maps for microwave applications in Malaysia Peninsula. Electron. Lett. **35**, 1772–1774 (1999)
13. Burgueno, A., Puicerver, M., Vilar, E.: Influence of rain gauge integration on the rain rate statistics used in microwave communications. Ann. Telecommun. **43**, 522–527 (1988)
14. Joo, H.W., Kim, Y.S., Kim, J.H., Choi, Y.S.: Empirical conversion process of rain rate distribution for various integration time. In: Asia Pacific Microwave Conference, pp. 1593–1597 (2000)
15. Watson, P.A., Gunes, M., Potter, B.A., Sathiaseelan, V., Leitao, J.: Development of climatic map of rainfall attenuation for Europe. Final Report ESA/ESTEC Contact No.4162/79/NL/DG/(SC), Report 327 (1982)
16. Ismail, A.F., Hashim, W., Abdullah, K., Malek, N.A.: Empirical conversion of rainfall rate distribution for various integration times in Malaysia. In: IEEE International Conference on RF and Microwave, pp. 270–273 (2011)
17. Sobli, N.H.M., Ismail, A.F., Isa, F.N.M., Mansor, H., Saidin, N.A., Abidin, M.S.Z.: Comparison of different empirical conversion methods from 60-minute to 1-minute integration time in Malaysia. Aust. J. Basic Appl. Sci. **8**(24), Special issue IPN Jakarta 2014, 321–327 (2014)
18. Capsoni, C., Luini, L.: A physically based method for the conversion of rainfall statistics from long to short integration time. IEEE Trans. Antennas Propag. **57**, 3692–3696 (2009)
19. Khairolanuar, M.H., Ismail, A.F., Jusoh, A.Z., Sobli, N.H.M., Malek, N.F.A., Zabidi, S.A.: New empirical conversion technique for 1-minute integration time of precipitation intensity in Malaysia. Aust. J. Basic Appl. Sci. **8**(24). Special issue IPN Jakarta, 290–295 (2014)
20. Malaysia Meteorological Department.: KLIA Forecast Office, Meteorological Facilities. http://www.met.gov.my/index.php?option=com_content&task=view&id=92&Itemid=301&limit=1&limitstart=2&lang=english

Comparison of Conversion Methods from 60- to 1-min Integration Time for Rainfall in Malaysia

Nuurul Hudaa Mohd Sobli, Ahmad Fadzil Ismail, Ani Liza Asnawi, Mimi Aminah Wan Nordin and Muhamad Haziq Khairolanuar

Abstract This paper presents some preliminary assessments of the precipitation rate conversion methods from 60-min to 1-min integration time. The conversion methods used in this study are Moupfouma and Khairolanuar et al., and ITU-R P.837. Rainfall rate data of twelve-months duration from January to December 2009 were acquired from Malaysia Meteorological Department (MMD) and exploited for the evaluation. The rainfall data were collected from MMD rain gauge station located at Kuala Lumpur International Airport (KLIA). The investigations comprise of producing annual rainfall rate cumulative distributions of the measured data. The equivalent 1-min annual rainfall rate cumulative distributions using conversion methods as mentioned above. Predicted values of annual 1-min cumulative distribution established by ITU-R P.837 are used as references. The equivalent 1-min equivalent cumulative distribution obtained from Moupfouma's and Khairolanuar et al.'s methods are then compared with that of ITU-R P.837 to validate the applicability and efficiency of each method. According to the results, it can be observed that Khairolanuar et al.'s method capable of generating equivalent 1-min rainfall values with smallest percentage difference as compared to ITU-R P.837 at 0.01 % of time exceedance.

N.H.M. Sobli (✉) · A.F. Ismail · A.L. Asnawi · M.A.W. Nordin · M.H. Khairolanuar
Department of Electrical and Computer Engineering, Kulliyyah of Engineering, International Islamic University Malaysia (IIUM), Jln. Gombak, 53000 Selangor, Malaysia
e-mail: hudaa@iium.edu.my

A.F. Ismail
e-mail: af_ismail@iium.edu.my

A.L. Asnawi
e-mail: aniliza@iium.edu.my

M.A.W. Nordin
e-mail: mimie@iium.edu.my

M.H. Khairolanuar
e-mail: haziq.khairolanuar@yahoo.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_91

1 Introduction

Rainfall or precipitation is very complex and the phenomenon that occurs naturally. Many of its characteristics influence in various ways of different extents such as hydrology, meteorology, water cycle of the earth, remote sensing, radio communication and propagation, etc. Rainfall is generally measured by instrument that is called rain gauge. It provides the total height of water fallen during a given period of time, such as daily, hourly, or minutely [1]. Hence, rainfall rate or rainfall intensity is defined as 'a measure of the intensity of rainfall expressed by the increase in the height of water reaching the ground per unit time' [2]. The information of rainfall rate is considered important for climatologic or general meteorological purposes. However, the time resolution offered by the rain gauge is still insufficient enough when it comes to correlate rain with other fast changing phenomena, such as radar-measured rainfall rate or rain-induced attenuation for microwave telecommunication links.

The effect of attenuation due to rainfall is severe in the countries located in tropical and equatorial regions for microwave system operating at frequency 10 GHz and above because of the average annual cumulative rainfall intensity in tropical and equatorial region is high. Rain attenuation occurs due to the absorption and scattering of energy by raindrops that degrade the reliability and performance of the communication link. Therefore, precise information and characteristics of rainfall rate distribution at the location of interest is vital for prediction of rain attenuation and estimation of unavailable percentage of time at frequency 10 GHz and above in communication link. Accurate knowledge of rainfall rate statistics is used as input for attenuation prediction method that evaluates the attenuation's statistical behavior that caused by rain.

International Telecommunication Union (ITU) recommends the use of rain rate cumulative distribution functions CDFs of 1-min integration time for derivation of attenuation CDFs in order to overcome the problem of different integration time interval [3]. 1-min integration time was selected as a compromise between experimental accuracy and the amount of available rainfall data. However, the standard time interval of 1-min integration time has not been utilized because of the significance of rapid change of precipitation intensity is not a major concern in meteorological applications. The most preferable rainfall rate is in average quantities such as hourly, daily, monthly, or yearly accumulated rain. This is the reason that most of the data acquired from meteorological departments have longer integration time. In addition, literatures also confirmed that lower integration time interval assures a good level of accuracy in regenerating the time variability of attenuation [4, 5].

For such purpose, researchers had came up with different proposed procedures that focusing on obtaining rainfall rate CDFs with 1-min integration time resolution from CDFs with longer integration time that was acquired from the general knowledge of local meteorological surroundings. Various conversion methods had been proposed in literature. These conversion methods can be generally classified into three categories. Firstly, meteorology based methods that utilize general climatic information as input (i.e. average monthly or yearly rainfall rate, number of rainy days per year, peak annual rain rate, etc.) such as proposed by Dutton et al. [6], Rice and Holmberg [7], Crane [8], and ITU-R [9]. Secondly, analytical methods assumed that the analytical form of rainfall intensity CDFs would not be affected by changing the integration time. Different integration time has different CDFs because the parameters of CDFs equation are different. These methods give the prediction behavior of the distribution change's parameters according to the integration time as proposed by Moupfouma [10] and Karasawa [11]. Finally, empirical methods that are most widely proposed by researchers around the world. The empirical methods specify conversion factors to estimate new CDFs from the known CDFs as a function of probability level. This type of method were utilized in the methods proposed by [12–15]. Some of the empirical conversion methods had been studied and compared with ITU-R values using local measurement data in Malaysia [16].

2 Conversion Methods from 60- to 1-min Integration Time

In 1993, Moupfouma had derived a simple empirical model that offers a good description of the global cumulative rainfall intensity distribution above 2 mm/hr, for both high rainfall rates and low rainfall rates [10]. The method was derived using data from USA, Canada, Europe, and India. It is useful for radio system designers. The cumulative distribution of rainfall rate can be expressed as follows [10]:

$$P(R \geq r) = \left(\frac{R_{0.01}+1}{r+1}\right)^{b} \exp\left(u(R_{0.01}-r) - \log\left(10^{4}\right)\right). \tag{1}$$

where r(mm/hr) represents the rainfall rate exceeded for a fraction P of the time. Parameter b is approximated by the following analytical expression [10]:

$$b = \left(\frac{r-R_{0.01}}{R_{0.01}}\right) \log\left(1+\frac{r}{R_{0.01}}\right). \tag{2}$$

The parameter u in Eq. 1 governs the slope of rainfall rate cumulative distributions and depends on the local climatic condition and geographical features. For tropical localities, it is expressed as follows [10]:

$$u = \frac{\log(10^{4})}{R_{0.01}} \exp\left[-\lambda\left(\frac{r}{R_{0.01}}\right)^{\gamma}\right]. \tag{3}$$

where λ and γ are positive constants. Based on the measured 1-min rainfall rate cumulative distribution at several locations in Malaysia, Singapore, and Indonesia, it was found that the best values for the parameters λ and γ are as follows [10]:

$$\lambda = 0.707 \quad \text{and} \quad \gamma = 0.06 \quad M < 3000. \tag{4}$$

where M(mm) is the mean annual rainfall.

In 1994, a recommendation ITU-R P.837 had been proposed by ITU-R as a reference on evaluation of the characteristics of precipitation for propagation modeling. This recommendation suggests a global map with different climatic zones to represent general characteristics of the rainfall rate at the location of interest. The latest version of this recommendation is ITU-R P.837-6.

In 2009, Capsoni and Luini proposed a new-revised method that exploits EXCELL Rainfall Statistics Conversion (EXCELL RSC) model [17]. The method enables users to generate statistics known as $P(R)$ of the local rainfall intensity, R (mm/hr) at 1-min integration time [17]. It also provides options that allow users to input either global digital maps of rainfall parameters derived from numerical weather prediction data or local measurements statistic of rainfall intensity at integration times up to 60 min integration time. The statistics of compilation data using locally measured rainfall intensity and integrated conversion model are predicted to provide the best approximation.

Khairolanuar et al. proposed a new empirical conversion method in 2014. The method was proposed by exploiting polynomial relationship for rainfall intensity data in Malaysia [18]. The data were acquired by using optical rain gauge for the duration of almost two years period. The estimation of 1-minute integration time of rainfall intensity statistics is summarized as follows [18]:

$$R_1(P) = aR_\tau(P)^4 + bR_\tau(P)^3 + cR_\tau(P)^2 + dR_\tau(P) + e. \tag{5}$$

where $R_\tau(P)$ is the input of rainfall intensity data for τ integration time and a, b, c, d, and e are the empirical constants.

3 System Set-up for Data Collection and Measurements

The ground truth measurement data were collected from the Malaysia Meteorological Department (MMD) rain gauge station located at station Kuala Lumpur International Airport (KLIA) with 2° 44′N and 101° 42′E, about 5 km from the airport and 16.3 m above mean sea level [19]. Airport was chosen as the location of interest for the analysis because further analysis regarding rainfall rate and radar estimated rainfall rate could be carried out in the future.

The rain gauge used by MMD to collect the measured rainfall values consists of standard tipping bucket. The tipping bucket rain gauge follows the standard by World Meteorological Organization (WMO). This tipping bucket collects rainfall rate data every 60 min. The integration time for the collected data is 60 min [19].

Table 1 Specifications of MMD tipping bucket rain gauge at KLIA

	Specifications
Location (latitude, longitude)	KLIA, Sepang (2° 44′N and 101° 42′E)
Distance from KLIA	±5 km
Receiving collector	203 ± 0.2 mm
Accuracy	±1 % to 200 mm/hr
Bucket capacity	0.2 mm
Dimensions	300 mm height, 230 mm body diameter, 280 mm base diameter
Physical	5.5 kg net weight

The tipping bucket comprises of two components; funnel-shaped at the top supported by a cylindrical-shaped at the bottom. This funnel has a water filter at the end of the funnel opening. As rain falls it lands in the funnel of the tipping bucket rain gauge. Water flowing into the funnel will be screened and will be collected by two metal water collectors (tipping buckets). The raindrops is poured into the cylinder when one of the collectors receives the raindrops of amount 0.2 mm and the next rain will then fall to the other metal collectors. This process is repeated and this repetition process is connected to a computing system (counter) that will count the number of times the rain that falls into the water collector metal. The amount of rainfall rate is calculated based on multiplication of the number of times the precipitation that falls on the metal rain collector with 0.2 mm of rain droplets. Maximum rainfall amount that can be obtained is 200 mm/hr. Table 1 summarized the specifications of the tipping bucket rain gauge operated by MMD and Fig. 1 shows the tipping bucket that is located at KLIA [19].

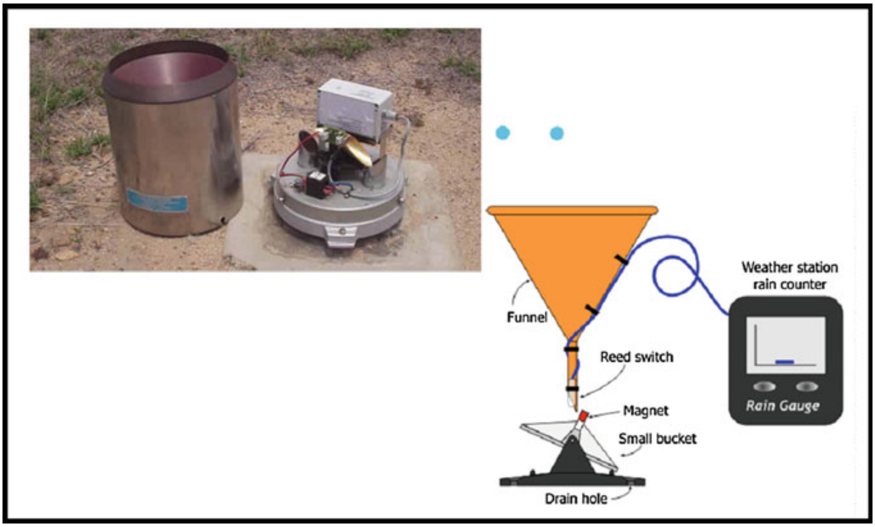


Fig. 1 Rain gauge tipping bucket at KLIA operated by MMD

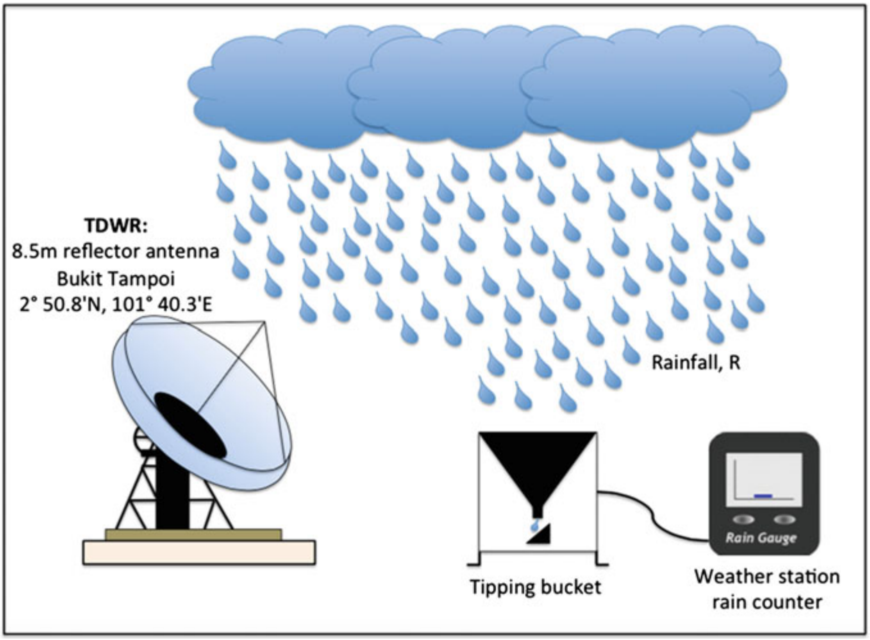


Fig. 2 Simplification of overall system set-up for the study

Figure 2 shows the simplification of overall system setup for the study including Terminal Doppler Weather Radar that is located at Bukit Tampoi. This paper presents only the study of the measured ground truth rainfall rate at the rain gauge station KLIA excluding radar data. However, the information of rainfall rate extracted from radar data will be used for further analysis together with the rain gauge measurement data to find the correction of radar reflectivity to rainfall rate relationships.

4 Results and Discussions

The rainfall intensity data were acquired from MMD rain gauge station located at KLIA. The acquired rainfall rate data were measured in real-time quantities for rainfall rate and precipitation accumulation. 60-min integration time is the standard resolution time employed by MMD for tipping bucket-acquired data at KLIA. The measurements of this data were taken for a period of one year that is from January to December 2009. The highest precipitation intensity value recorded within one-year duration is 71.83 mm/hr.

The acquired rainfall rate data were investigated to produce its annual cumulative distribution function (CDF). The annual CDF of rainfall rate at KLIA is

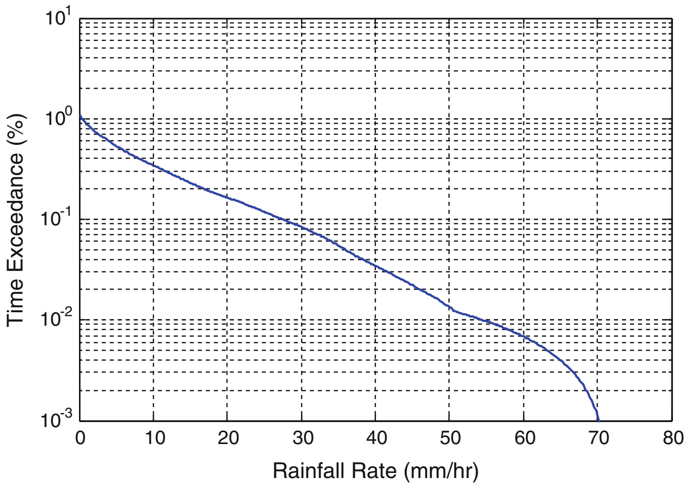


Fig. 3 Annual CDF of rainfall rate at KLIA in 2009

obtained from the accumulation of monthly CDFs. Figure 3 shows the annual CDF of rainfall data at KLIA for 2009. The figure illustrates that KLIA had experienced rainfall intensity of approximately 54.4 mm/hr at 0.01 % time exceedance that is equivalent to 9 h of the year.

Further analyses were carried out by implementing the above conversion methods i.e. Moupfouma, ITU-R P.837, and Khairolanuar et al. The plotted 60-minute precipitation intensity statistics was converted to 1-minute precipitation intensity using eqn. [1] with substitution of the parameters by eqn. [2–4] for Moupfouma's and using eqn. [5] for Khairolanuar et al.'s methods respectively. The conversion method of ITU-R P.837-Map and ITU-R P.837-6 were used as the benchmark for comparison different conversion methods because of its appropriate to the local climate.

Figure 4 illustrates the comparison of the converted 1-minute rainfall intensity statistic using Moupfouma's, Khairolanuar et al. including the proposed values for ITU-R P.837-6, ITU-R P.837-Map for the collected rainfall data at KLIA. From the figure, it can be deduced that the highest rainfall rate value at 0.01 % time exceedance is generated by Moupfouma's conversion method with 119.6 mm/hr, and Khairolanuar et al.'s method produces 74.14 mm/hr. Meanwhile, ITU-R P.837 physical and ITU-R P.837 Rain Map yield 90.77 and 91.75 mm/hr respectively. Table 2 summarizes the rainfall rate values at 0.01 % time exceedance for all conversion methods statistics.

Moupfouma's and Khairolanuar et al.'s conversion methods were compared with the ITU-R P.837 rainfall intensity conversion method as this method has been chosen as subject of reference due to its applicability to the local climate. The values of rainfall rate were selected at 0.01 % of time exceedance, $R_{0.01}$ because of

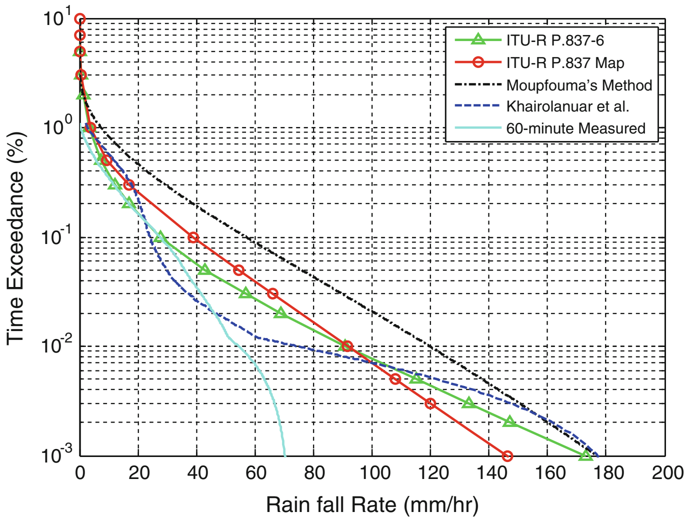


Fig. 4 Comparison of different conversion methods for rainfall rate at KLIA

Table 2 Precipitation rate values at 0.01% time exceedance ($R_{0.01}$) for rainfall rate in KLIA

	Moupfouma	ITU-R P.837 Map	ITU-R P.837-6	Khairolanuar et al.
$R_{0.01}$ (mm/hr)	119.6	91.75	90.77	74.14

its applicability and reliability for further rain attenuation prediction studies as described in literatures.

The assessments of the applicability of each method were explored by determining the root mean square error (RMSE) and the percentage error of precipitation rate value at 0.01 % of time exceedance. The relevant RMSE for these preliminary findings is defined as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(ECD_i - ITUD_i)^2}. \tag{6}$$

where n is the number of data, ECD is the empirical conversion data and ITUD is the respective ITU-R P.837 data under investigations.

Table 3 exhibits the comparison of RMSE and percentage difference values for each aforementioned conversion method rainfall intensity with ITU-R P.837 rain map rainfall intensity of 91.75 mm/hr at 0.01 % of time exceedance. The table shows that ITU-R P.837 physical has the lowest values of RMSE and percentage difference with 10.958 and 1.068 % respectively. However, Moupfouma's conversion method yields the highest RMSE and percentage errors with 68.765 and 75.594 % respectively. It is then followed by Khairolanuar et al.'s conversion method as compared to the ITU-R P.837 rain map with 18.388 and 19.193 % respectively.

Table 3 Comparison of RMSE and percentage error at 0.01% time exceedance with ITU-R P.837 rain map for rainfall rate in KLIA

	Moupfouma	ITU-R P.837-6	Khairolanuar et al.
RMSE (%)	68.765	10.958	18.388
Percentage difference (%)	75.594	1.068	19.193

Table 4 Comparison of RMSE and percentage error at 0.01% time exceedance with ITU-R P.837-6 conversion method for rainfall rate in KLIA

	Moupfouma	ITU-R P.837 Map	Khairolanuar et al.
RMSE (%)	45.863	10.958	12.283
Percentage difference (%)	67.664	1.080	18.321

Table 4 tabulates the RMSE and percentage difference of the discussed conversion methods as compared with ITU-R P.837 physical rainfall rate of 90.77 mm/hr at 0.01 % of time exceedance. From the table, it can be inferred that ITU-R P.837 rain map gives the lowest RMSE and percentage difference values with 10.958 and 1.080 % respectively. The table also demonstrates that the highest values of RMSE and percentage difference were produced by Moupfouma's conversion method, followed by Khairolanuar et al.'s conversion method with 45.863, 67.664 %, and 12.283, 18.321 % respectively.

5 Conclusion and Future Works

The scope of this paper is to study the comparison of different proposed conversion methods of 60- to 1-min integration time of precipitation intensity using measured rainfall rate at KLIA. The comparison was carried out by selecting ITU-R P.837 proposed equivalent 1-min values at 0.01 % of time exceedance as a benchmark for that its applicability to the local climates. It can be concluded that Moupfouma's conversion method produced the highest RMSE and percentage difference values at 0.01 % time exceedance. The analysis in this paper can be useful for rain attenuation prediction methods study because of the significance of using 1-min integration time. The results obtained for this study were the analysis of rainfall data measured and collected at one location for the duration of one year. However, the evaluation of rainfall data at different locations will be discussed and disclosed in the future.

Acknowledgments The authors wish to thank Malaysia Meteorological Department (MMD), Ministry of Science, Technology and Information (MOSTI), and IIUM Endowment Research Grant Funding for their endless support and help for this research.

References

1. Brand, E.: A comparative study of techniques for measuring rainfall rate and accumulation. In: Proceedings AP2000 Millennium Conference on Antennas and Propagation (2000)
2. Recommendation ITU-R P.310-9: Definition of terms relating to propagation in non-ionised media (1994)
3. Recommendation ITU-R P.837-4: Characteristics of precipitation for propagation modeling (2001)
4. Segal, B.: The influence of rain gauge integration time on measured rainfall-intensity distribution functions. J. Atmos. Oceanic Technol. **3**, 662–671 (1986)
5. Segal, B., Allnutt, J.E.: On the use of long sampling-time rainfall observation for predicting high-probability attenuation on earth-space links. In: Proceeding International Conference of Antennas and Propagation, vol. 2, pp. 754–757 (1991)
6. Dutton, E.J., Dougherty, H.T., Martin Jr, R.F.: Prediction of European rainfall and link performance coefficients at 8 to 30 GHz. Institute for Telecommunications Science. U.S. Department Commerce, Washington DC (1974)
7. Rice, P.L., Holmberg, N.R.: Cumulative time statistics of surface-point rainfall rates. IEEE Trans. Commun. **21**, 1131–1136 (1973)
8. Crane, R.K.: Prediction of attenuation by rain. IEEE Trans. Commun. **28**, 1717–1733 (1980)
9. Recommendation ITU-R P.837-6: Characteristics of precipitation for propagation modelling (2012)
10. Moupfouma, F., Martin, L.: Modeling of the rainfall rate cumulative distribution for the design of satellite and terrestrial communication systems. Int. J. Satell. Commun. **13**, 105–115 (1995)
11. Karasawa, Y., Matsudo, T.: One-minute rain rate distribution in Japan derived from AMeDAS one-hour rain rate data. IEEE Trans. Geosci. Remote Sens. **29**, 890–898 (1991)
12. Chebil, J., Rahman, T.A.: Rain rate statistical conversion for the prediction of rain attenuation in Malaysia. IEEE Electron. Lett. **35**, 1019–1021 (1999)
13. Burgueno, A., Puicerver, M., Vilar, E.: Influence of rain gauge integration on the rain rate statistics used in microwave communications. Ann. Telecommun. **43**, 522–527 (1988)
14. Joo, H.W., Kim, Y.S., Kim, J.H., Choi, Y.S.: Empirical conversion process of rain rate distribution for various integration time. In: Proceeding of the Asia Pacific Microwave Conference, pp. 1593–1597 (2000)
15. Ismail, A.F., Hashim, W., Abdullah, K., Malek, N.A.: Empirical conversion of rainfall rate distribution for various integration times in Malaysia. In: IEEE International Conference on RF and Microwave, pp. 270–273 (2011)
16. Sobli, N.H.M., Ismail, A.F., Isa, F.N.M., Mansor, H., Saidin, N.A., Abidin, M.S.Z.: Comparison of different empirical conversion methods from 60-minute to 1-minute integration time in Malaysia. Aust. J. Basic Appl. Sci. **8**(24), 321–327 (2014). (Special issue IPN Jakarta)
17. Capsoni, C., Luini, L.: A physically based method for the conversion of rainfall statistics from long to short integration time. IEEE Trans. Antennas Propag. **57**, 3692–3696 (2009)
18. Khairolanuar, M.H., Ismail, A.F., Jusoh, A.Z., Sobli, N.H.M., Malek, N.F.A., Zabidi, S.A.: New empirical conversion technique for 1-minute integration time of precipitation intensity in Malaysia. Aust. J. Basic Appl. Sci. **8**(24), 290–295 (2014). (Special issue IPN Jakarta)
19. Malaysia Meteorological Department, KLIA Forecast Office, Meteorological Facilities: http://www.met.gov.my/index.php?option=com_content&task=view&id=92&Itemid=301&limit=1&limitstart=2&lang=english

DycScreen—Cross-Platform Dyslexia Screening Test for Malaysian Children Through Hybrid Applications

N. Che Pee, P.H. Leong, M.A. Othman, H.A. Sulaiman, M.F.I. Othman and Y.A. Rahim

Abstract This research is carried out to identify potentially dyslexics among children in Malaysia. Identifying dyslexics among the children in Malaysia has been a difficult task. This is due to lack of comprehensive screening tests in Malaysia and most of the available screening tests are expensive. Indicators of dyslexia are differs for specific age groups, for instance, the identification for children is focusing more in problems learning the alphabets, numbers, shapes and colors. The identification for adults is however more advance in which the focus would be in terms of the written works and achievements in studies. The existing screening test applications are expensive while the free applications are incomprehensive and lacking of useful features. The screening test will be developed using hybrid applications approach as

N. Che Pee (✉) · H.A. Sulaiman · M.F.I. Othman · Y.A. Rahim
Communication, Multimedia, Networks and Systems Research Group, Faculty of Information Technology and Communication, Universiti Teknikal Malaysia Melaka (UTeM), 76100 Durian Tunggal, Melaka, Malaysia
e-mail: naim@utem.edu.my

H.A. Sulaiman
e-mail: asyrani@utem.edu.my

M.F.I. Othman
e-mail: mohdfairuz@utem.edu.my

Y.A. Rahim
e-mail: yahaya@utem.edu.my

P.H. Leong
ICT Innovation and Creativity Research Group, Faculty of Applied Sciences and Computing, Universiti Teknikal Malaysia Melaka (UTeM), 76100 Durian Tunggal, Melaka, Malaysia
e-mail: leongph@acd.tarc.edu.my

M.A. Othman
Microwave Research Group, Faculty of Electronics and Computer Engineering, Universiti Teknikal Malaysia Melaka (UTeM), 76100 Durian Tunggal, Melaka, Malaysia
e-mail: azlishah@utem.edu.my

N. Che Pee · P.H. Leong · M.A. Othman · H.A. Sulaiman · M.F.I. Othman · Y.A. Rahim
Tunku Abdul Rahman University College (TARUC), Jalan Segamat/Labis, 85000 Segamat, Johor, Malaysia

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_92

it provides cross-platform access which allowed users to conduct the screening test according to their preference either through personal computers, desktops, or mobile devices. Initial study shows that currently in Malaysia, there is no screening test which developed using hybrid applications approach. The conventional approach requires the person to perform manual screening test conducted by the expert at the Dyslexia Centres to confirm the disabilities and some fees will be charged. Having a localized screening test is crucial as Malaysian education system is differs from other countries. Hence, it is important for individuals to indicate whether they are dyslexics because many of them suffered in studies without knowing that they are actually dyslexics. Dyslexics can have a better opportunity in life if they were given early remediation, intervention, and support from corresponding society. Identifying dyslexics does not only benefit the dyslexics but also to guardians and countries.

Keywords Dyslexia · Screening test · Cross-platform and hybrid applications

1 Introduction

Dyslexia is known as a specific learning disability (SPD) of neurological origin which affects the ability or capability to process written and sometimes even spoken language [1]. Dyslexia is a lifelong condition for which it is neither a sickness nor disease and therefore cannot be cured. The symptoms of dyslexia are ranging from mild to severe. However, depending on the degree or severity, there are appropriate remedial programs and compensatory strategies available which can assist dyslexics to overcome the difficulties. According to [1], dyslexia is often a cause of school failure and significantly affects the individuals' education, as well as being an important risk factor for the development of more complex psychopathological disorders such as behavior disorder, anxiety, and depression. Dyslexia is generally inherited and not the outcome of race or social background, sensory impairments, or learning a second language and is independent of intelligence, although those reasons can lead the individuals to the risk of inability to read [2].

Likewise, individuals with dyslexia in a multi-cultural country such as Malaysia commonly exhibit more difficulties in reading, spelling and writing since Malaysian are expected to master more than one language. To date, there is no concrete data and research evidence on the prevalence of children dyslexics in Malaysia. However, there was a pilot study conducted by [3] in a primary school of 2000 students revealed that in Standard 2 Malay students, 7 % of the students are identified to be dyslexics. Moreover, according to an article in The Star, there is an estimated of 314,000 dyslexics in Malaysia [4]. This implies that most dyslexics are not getting assistance from corresponding society or given public awareness. Although there are successful individuals with dyslexia, many individuals with dyslexia do not make it to higher studies or acquire the necessary skills to cope with adult life challenges. In fact, some of them contribute to social problems.

This study aims to identify potentially dyslexics among children in Malaysia through hybrid applications approach. Apart from that, this research also intends to capture more data about the prevalence of children dyslexia in Malaysia. The flow of this paper is organized as follows. Section 2 describes digital technologies as a tool to assist dyslexics followed by Sects. 3 and 4 which describe the DycScreen and testing details. Section 5 discusses the results and finally the conclusion of the paper is discussed in Sect. 6.

2 Digital Technologies as a Tool to Assist Dyslexics

In recent years, mobile devices, generally represented by smartphones and tablets are becoming a trend, used by millions of people across the world. Mobile applications, sometimes referred to as mobile apps or apps are application software developed to run on devices such as smartphones and tablets. Typically, the apps are downloaded from the platform to a target device, such as Android phone, Windows Phone, iPhone or BlackBerry.

In the last two decades, digital technology, particularly Personal computers (PC) have played a vital role in supporting remediation and enhancing skills. For instance, the digitization of text with the introduction of e-books has made it possible for dyslexics' students to read books using text-to-speech software [5]. Nevertheless, due to the problems such as high cost, the bulky size on the desk, increase electric power supply and slow startup or powering up tend to make it difficult to use in many situations especially in classrooms. Despite of that, teachers may struggles or face difficulties when using the computer based assistive technologies.

However, depending on preference and needs, some users might prefer to conduct the dyslexia screening test via the personal computers or desktop due to the large screen which make it more interactive. On the other hand, some users might prefer to conduct the screening test via mobile devices particularly because mobile devices are designed to be handheld which allowed the users to conduct the screening at anytime, anywhere.

Hence, in order to satisfy the preference and needs of users, hybrid apps approach is introduced. Hybrid applications is the combination of native apps and web apps. A hybrid apps is written similarly with the web technologies approach such as HTML5, CSS and JavaScript. The hybrid apps are then installed through the app store, or scan through the Quick Response Code (QR code) in order to provide privileged access to run inside a native container on mobile devices. Besides, instead of developing the application using native SDK, hybrid apps work by wrapping the web application through a native web view controller full screen, which indirectly declined normal browser controls and address bar. To simplify, hybrid apps approach indirectly provides cross-platform access which allowed users to conduct the screening test according to their preference either through personal computers, desktops, or mobile devices.

Table 1 DycScreen categories

Categories	Screening area	Performed by
A	Questionnaires	Parents/teachers/guardians
B	Spelling	Child
C	Vision and cognitive skills	Child
D	Direction	Child
E	Mathematics and time	Child

3 DycScreen

DycScreen is a Dyslexia-Screening tool for children age between 9 and 12 years old in Malaysia. The screening test can be conducted anytime, anywhere depending on preference with the availability of internet connection. For instance, the screening test can be performed via personal computer (PC) or mobile devices such as smartphones and tablets. It is divided into five categories pertaining to the dyslexia symptoms. In order to identify whether the particular child has dyslexia indication, the questions in each category are necessary to be answered. The child will be screened in five categories which include questionnaires, spelling, vision and cognitive skills, direction, mathematics and time. The categories of the screening test are portrayed as in Table 1.

The questionnaires category comprises ten questions which are performed by parents, teachers or guardians to examine the child learning history. On the other hand, the remaining categories are performed by the child in assistance of the parents, teachers or guardians to capture the child performance in school. The result of the screening test indicates whether the particular child is potential dyslexic. Below are the screenshots of DycScreen:

Figure 1 shows the multiscreen preview panel of DycScreen. The application will fit automatically to the size of the device used. For instance; the desktop, tablet or mobile view.

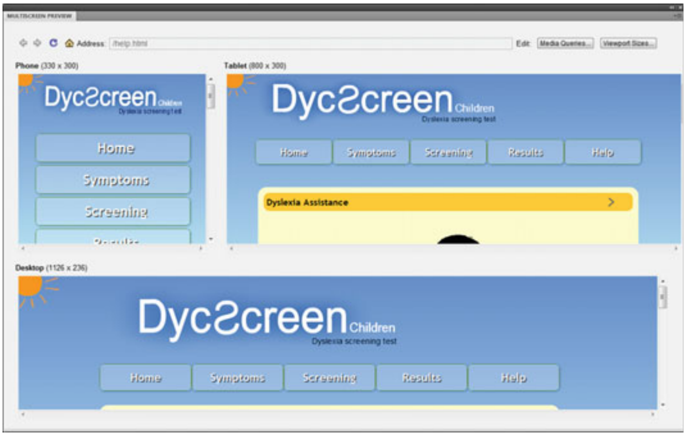


Fig. 1 The multiscreen preview panel

Fig. 2 DycScreen home page (desktop view)

Fig. 3 DycScreen screening categories

Figure 2 shows the home page of DycScreen. Brief information of Dyslexia is portrayed in bullet form.

Figure 3 shows the categories of the screening test. Before the screening test starts, the name of the child is required to be filled in.

Figure 4 shows the questionnaires assessment of the screening test. All of the questions must be answered before proceeding to the next categories.

Figure 5 shows the spelling categories of the screening test. The screening result will be shown after all of the categories are answered.

Fig. 4 Questionnaires assessment

Fig. 5 Spelling categories

4 Testing Details

The screening test was conducted among Malaysian children aged between 9 and 12 years old. The testing procedure is illustrated as below.

4.1 Participants

One Specific Learning Disabilities (SPD), one tutor, three dyslexics' students and fifteen normal students are involved in this testing.

4.2 Procedure

The screening test is divided into five categories pertaining to the dyslexia symptoms. In order to identify whether the particular child has dyslexia indication, all of the questions in each category are necessary to be answered. The categories include questionnaires, spelling, vision and cognitive skills, direction, mathematics and time. The questionnaires category comprises ten questions which are performed by parents, teachers or guardians to examine about the child learning history. On the other hand, the remaining categories are performed by the child in assistance of the parents, teachers or guardians to capture the child performance in school. The result of the screening test indicates whether the particular child is potential dyslexic.

5 Results

Three dyslexic students and fifteen normal students participated in the testing. The results of the analysis are depicted as in Fig. 6.

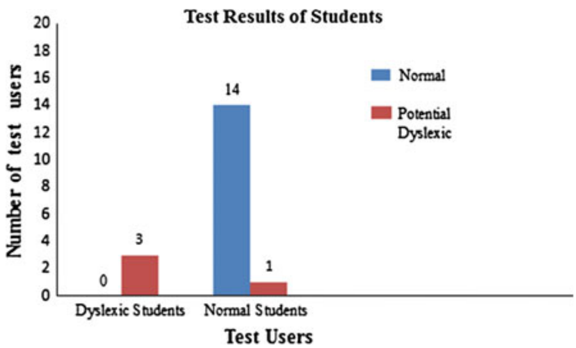


Fig. 6 Test results

5.1 Dyslexic Students

Based on the results in Fig. 6, three of the dyslexic students are identified to be potentially dyslexics. The testing took approximately 1 h and 50 min to complete whereby each of the student undergo the testing for approximately 30 min. The observation indicates that all the three students are reluctant and struggled in answering certain assessments especially in cognitive assessments whereas two students struggled when answering direction assessments. All the students seemed to be normal when answering other categories of the assessments. None of the students managed to answer all the questions correctly. Instead, all the three students tend to make mistakes. Furthermore, according to the specific disabilities learning expert, the students are more engaged in conducting the screening test via digital technologies compared to manual screening test. One of the students preferred to conduct the screening test using laptop whereas two of the students preferred to conduct the screening test using mobile.

5.2 Normal Students

Based on the results in Fig. 6, fourteen students are identified to be normal whereas one student is identified to be potentially dyslexic. The testing took approximately 3 h and 17 min to complete whereby each of the student undergo the testing for approximately 10 min. The observation indicates that fourteen students are able to answers the questions easily without making any mistakes. Despite of that, one of the students tends to make several mistakes. The tutor claimed that the student is normal and achieved good score in exams. However, the student did not pay attention and tend to answer the questions brutally in order to finish the assessment quickly.

Likewise, similar to the specific disabilities learning expert, the tutor claimed that the students are more engaged and paid more attention in learning via digital technologies compared to books. Four of the students preferred to conduct the screening test using laptop whereas eleven of the students preferred to conduct the screening test using mobile.

6 Conclusion

Individuals having dyslexia is increasing in percentage whereby support is crucial to ensure the continuity and quality of their education and better future. With the implementation of dyslexia screening test, it is hoped that these identified dyslexics may seek immediate help from corresponding society and therefore are able to cope with current academic demand in Malaysia. Dyslexia is not a disease and it cannot be cured. But there are remedial programs available which can assist dyslexics, depending on their degree or severity, overcome many of the difficulties they face in reading and writing. In the study conducted in Malaysia, this attempt has revealed that there is no screening test which developed via Hybrid applications approach. In addition to this, it is also observed that children are more engaged in learning using digital technologies such as mobile devices compared to comprehensive learning such as blackboard and book. Likewise, this research indicates that cross-platform applications are suitable for detecting potential dyslexics. It is hope that with the availability of this novel approach, more cases of children dyslexics would be discovered.

The proposition of improvement for this study is to develop a more extensive screening test whereby the other aspects of the symptom dyslexia such as reading and auditory tests are included. There should be a voice capturing tools to capture the fluency of reading and intonation of the users when reading the article as well as webcam to capture users' expression, behaviour and gesture while doing the assessments. Another proposition of improvement is to develop a screening test with multi language in order to overcome the language barrier.

References

1. Gaggi, O., et al.: A serious game for predicting the risk of developmental dyslexia in pre-readers children. In: 2012 21st International Conference on Computer Communications and Networks (ICCCN). IEEE, pp. 1–5 (2012). Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6289249
2. Catherine, E., et al.: Preventing reading difficulties in young children. The National Academies Press (1998). [online] Available at: http://www.nap.edu/catalog.php?record_id=6023. Accessed on 17 Feb 2014
3. Gomez, C.: Dyslexia in Malaysia. International book of dyslexia: A guide to practise and resource. pp. 158–163 (2004). [online] Available at: http://www.wiley.com/legacy/wileychi/dyslexia/supp/Malaysia.pdf. Accessed on 22 Feb 2014
4. Wong, L.Z.: Hope for dyslexicss. The star online (2009). [online] Available at: http://www.thestar.com.my/story.aspx/?file=%2f2009%2f3%2f16%2flifefocus%2f3336348. Accessed on 17 Feb 2014
5. Elkind, J., et al.: Using computer-based readers to improve reading comprehension of students with dyslexia. Ann. Dyslexia **43**(1), 238–259 (1993)

Embedded Packet Logger for Network Monitoring System

Chananankorn Jandaeng

Abstract The network monitoring system (NMS) consist of three parts: data collection, data analysis and visualisation. The sensor agent is implemented on managed device and server in order to send traffic, audit log and network information for analysis. On the other hand, the unmanaged device does not have software to support the management function. Moreover, the end user or client behind L2-switch should be monitored by NMS. This paper implements embedded packet logger (EPL) for packet sniffer in switch rack. As the small switch rack is distributed in campus building, it hards to monitor and support in order to keep it safety. The embedded system can eliminate this problem. The EPL, light weight package sniffer, is implemented and embedded in RaspberryPi board, these hardware installed Linux kernel. We compare the traffic measurement and resource consumption with *tcpdump*(1), the EPL efficiently capture packet like *tcpdump*(1). Moreover, the memory and CPU consumption do not different significantly, with lower storage space than *tcpdump*(1).

1 Introduction

Monitoring a network is importance to network management. A major function of network monitoring is early identification of trend and pattern in both network traffic and device. The network monitoring system (NMS) monitors network service and system, resource capacity plan, statistics and accounting, fault management and performance; such as throughput, latency and round trip time. Further more, the network monitoring supports the network under service level agreement (SLA) and network policy.

C. Jandaeng (✉)
School of Informatics, Walailak University, Thaiburee, Thasala,
Nakorn Sri Thammarat, Thailand
e-mail: cjundang@gmail.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_93

The simple NMS tools are *ping*(1), audit log and Wireshark. These tools provide simple information to diagnose the network, which not enough to fulfil the above requirement. *ping*(1) command is a simple tools but it gives only round trip time, average success and average fail. Audit log needs other tools for log analysis. Wireshark suit for analysis local and small network. However, Simple network management protocol (SNMP) [1] is a standard protocol, which consist of agent and manager. The SNMP agent is embedded in network device, whereas the SNMP manager request and get response from agent via UDP port. The SNMP is able to collect various information but does not have analysis module, beside it can only monitor managed device. Fortunately, PRTG Network Monitor is an example of web based NMS tools that only show traffic and bandwidth usage from SNMP.

Network administrator require network information collection in order to analyse, plan and manage network infrastructure and application. Most efficiency result is require an accuracy information and large collected data. Thus, network administrator inquire network monitoring tools that collect data from network such as network device, server in demilitarised zone (DMZ), traffic in core network and local network.

This research propose Network Monitoring Tools, that consist of data logger, data analysis and visualisation. In data logger phase, all traffic are sent and collected in data centre. All managed device and server are able to send information to data centre, whereas unmanaged device such as L2-switch have no sensor or mechanism to collect and forward packet.

Most NMS analyse traffic in core switch or backbone network but do not analyse data in local network, which result to the accuracy of problem solving; such as ability to solve problem to the right point. For example, NMS process only packet passing from L2-switch then send to core-switch but some end device cannot be connected to Wifi access point or get IP address from DHCP server because of traffic jam. On the other hand, bandwidth chart shows that there are no problem in core switch side.

Although, this problem can be solved, when packet sniffer installed in local network and send traffic information to network monitoring centre. Unfortunately, the packer sniffer installed in computer is not suitable for distributed in physical location. The switch rack has not been installed in server room as core switch and server. Our solution; packet sniffer implementation in embedded device then install in switch rack.

This paper propose the Embedded Packet Logger (EPL), EPL is packet sniffer embedded device. All captured packets are pre-processed and transformed to log message, then forwarded to log server that based on *rsyslogd*(8) service.

First, the literature review is compared and discussed in Sect. 2, after that the Sect. 3 shows and describe the overall system model of network management system. The Embedded Packet Logger will be explained in Sect. 4, then evaluated with traffic measurement and resource consumption in Sect. 5. Finally, the conclusion and future work are mentioned in Sect. 6.

2 Literature Review

2.1 *Freeware Network Monitoring System*

There are many network monitoring tools today, which divided into 2 categories: commercial and free software. The commercial software may either distributed as software system or embedded into the box; called controller. Most software system are implemented in small and non-complex network, while the controller is used in large organisation ex. university, network company, Internet Service Provider, etc. Today commercial controller price is around 4 million bath with maintenance at approximately 800,000 bath per year. Beside network administrator or programmer cannot extend or access raw data for committed their organisations requirement.

Another choice is freeware or open source system. Many freeware consist of 2 editions: free version and commercial version. Most of free version will disable some function and provide some function as a trial version for further selling on commercial vision, which is not applied to the organisation inquiries. In addition, commercial version is expensive. Example of freeware tools are shown in Table 1.

Data source of NMS Tools is packet capture with raw data socket programming, libpcap and HTTP traffic. Moreover some tools analyse data from SNMP and NetFlow Protocol. Port scanning is also a technique that monitor network service.

Regarding to ISO standard, the network management consist of 5 functions [2]. Fault Management (FM) detects up/down of service status, port and link, or monitor for normal network. Performance management (PM) shows bandwidth utilisation, incoming and outgoing traffic. Security management (SM) monitors both normally/anomaly network, and it also detects/prevents the network intrusion. Usually, three functional model are implemented in the traditional NMS Tools, but Accounting management and Configuration management has not been implemented. Therefore, this paper will propose designed and implemented all 5 functional models as a new NMS. Moreover, it will also propose traffic prediction base on machine learning approach, which is applied to new NMS.

Table 1 Freeware network monitoring system

NMS tools	Platform	Data source	FM	PM	SM
Wireshark	All	libpcap		x	x
Xymon	Unix	Traffic	x	x	
NetworkMiner	Unix	Traffic	x	x	
Fiddler	All	HTTP traffic	x	x	x
Total network monitor	Window	TCP 139 and 445	x	x	
PRTG	All	SNMP, Netflow	x	x	
ntopng	Unix	sFlow, NetFlow	x	x	
Angry IP scanner	Window	Scan			x

The details of each NMS Tools available on its official website

2.2 Embedded Packet Sniffer

The libpcap is the traditional library that is selected for implementation of packet sniffer; such as *tcpdump*(1) and Wireshark. *tcpdump*(1) and Wireshark are text mode and GUI mode of packet sniffer, which run on PC or server. PiTap [3] is embedded packet sniffer that implemented in RaspberryPi. PiTap calls *tcpdump*(1) command, then it will export raw data and save to hard disk. However, PiTap does not have analysis process, network administrator has to remote access in order to analyse by themselves.

Wang [4] proposed kernel and protocol analysis tools that embedded on network devices, called KPAT. KPAT uses auto-instrument technique in order to systematically inject software and probe into selected kernel function without recompile kernel. These proposal allows user to observed details of protocol behaviour.

Rahman [5] implemented network monitoring system on embedded Linux platform, named PNtMS. They compared the PNtMS with Wireshark. The result shown that the PNtMS obtain better performance by using limited resources on TS-5400 and shown more statistical analysis result than Wireshark.

3 Proposed System Model

The network monitoring system analyses audit log data. All log message are collected from server, unmanaged devices and network equipment; such as switch and router. The network monitoring system consist of 3 modules: Logger Agent, Log analysis and Visualisation as shown in Fig. 1.

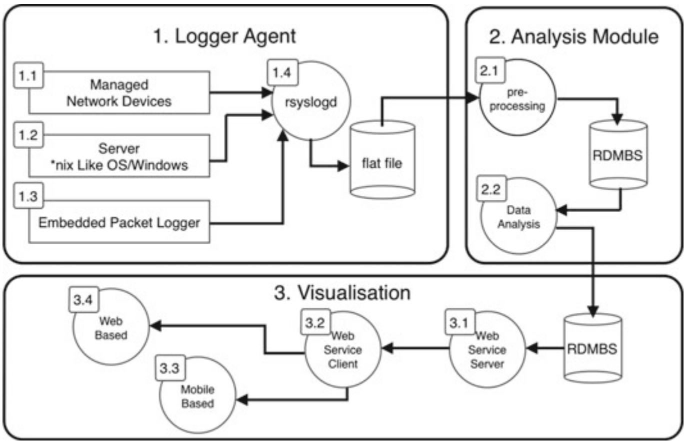


Fig. 1 The overview of system model for network monitoring system

3.1 Logger Agent

Most logger agents are *rsyslogd*(8) client. The managed switch and router are easily enabled these function, in order to forwarded log to data centre via UDP/514 connection. Besides, the *rsyslogd*(8) are installed in all Unix-like OS, network administrator can only just enable the service in order to activate *rsyslogd*(8). For Microsoft Window 2010 or above, the *rsyslogd*(8) can be easily enabled via administrator tools. From the above basic configuration, all log are sent to log server.

On the other hand, the traffic behaviour in LAN cannot be monitored, because most network devices are unmanaged switches. Therefore, network monitoring system trace both number and packet type via uplink of core switch only. Whereas, some broadcast packets are not forward across the unmanaged switch. Thus, network administrator cannot diagnose some problem in local network. Hence, this paper proposes the embedded packet logger (EPL); as describe in Sect. 4, to capture packet in local network and then forward to log server as data centre.

rsyslogd(8) service is shorted from rocket-fast system for log processing. It is enhanced from *rsyslogd*(8) in UNIX that is compatible to most log message from variety of source; such as Unix like OS, managed network equipment and Microsoft Window Server. The *rsyslogd*(8) directly receives log message from process and kernel. In addition, this service accepts the message from remote log service; such as switch or router. EPL captures packet, pre-process and transform to JSON format. Afterward, these messages are sent to log server.

Log server records message in flat file, named /var/log/message by default. Log message consist of 4 parts: (1) Date and time of event, (2) IP Address of remote host/localhost/domain name, (3) Process name with process ID in bracket and (4) Description which depends on OS, application or service.

3.2 Log Analysis

Log analysis is consist of 2 sub modules: pre-processing and log analysis. The pre-processing module encodes standard log message in order to reduce storage, then transform message into suitable data for analysis. The encoded data is stored in RDBMS or NoSQL, which will be in the future [6]. Log analysis sub-module classify and predict the network behaviour from log message in order to plan and detect/prevent intrusion.

There are many techniques that used in log analyser. Data mining is an optimal solution, which composed of data clustering, data association and data classification. Network behaviour is grouped by K-Means [7] or Agglomerating clustering technique, while the behaviour relation is defined by association rules; such as Appiori [8]. In addition, we classify all log messages with classification method; ex. Decision Tree, Naive Bayes and Neural Network. All the analysed results are stored in database for visualisation in next module.

3.3 Visualisation

Once the old said that one picture is worth a thousand words. The network monitoring system is a tool for making decision, planning and judgment on large data. The visualisation is also importance module equivalent to both data collection and log analysis. In visualisation, network administrator access analysed data via mobile device or PC. Mobile/PC application is web service client that request data from web service server via SOAP protocol. The mobile device shows summary result and notify alert message. Whereas all details of event, analysis result or other information are accessed via web based application *ping*(1).

This section explains the whole system model of NMS. However, this paper focus on Logger Agent module only. Network device that supported *rsyslogd*(8) can send their log message to data centre. Because we need to analyse network traffic in local network, so the embedded packet logger is designed and proposed in next section. Whereas log analysis and visualisation do not discuss in this paper.

4 Embedded Packet Logger

Traffic in LAN is broadcast communication and disable to be forwarded via switch or router. In addition, some network and security problem are occurred in LAN i.e. ARP Spoofing, Internet Worm, unnecessary service and flooding traffic. Which causes a problem to end user, there are unable to connect an Internet or connect under low network speed. Thus, network administrator need to monitor traffic behind their switch via packet logger.

The system model of embedded packet logger is presented in Fig. 2, that composed of packet capture with libpcap, pre-processing data and data collection.

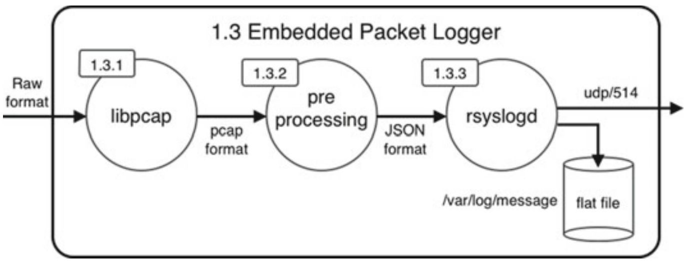


Fig. 2 System model of embedded packet logger is used in unmanaged network devices

4.1 Packet Capture

The packet capture will use libpcap library which is a standard library ex. *tcpdump*(1). These capture mechanism is an event driven. When the process capture raw packet, that packet will be processed with interrupt driven function in pre-processing phase.

4.2 Pre-processing Data

The header raw packet is reassembled into 3 layers as TCP/IP model: data link, network and transport layer. The content or payload of higher layer is not processed because of the privacy, except the security policy defined that all bytes in each packet will be monitored. Thus, header of data link, network and transport layer are transformed to readable data then generated to JSON format. The JSON format is easily access and lightweight more than other document format; such as XML. The example of JSON is shown in Fig. 3.

The end of this phase, JSON message is sent to *rsyslogd*(8) server as log message.

4.3 Data Collection

The data collection store JSON message as log message via library function. The *openlog*(3) is a library function in Unix that establishes a connection to the system


```
Jan 27 21:39:07 localhost emos[1307]: {"ethernet":
    {"dst_addr":"00:e0:b1:66:bd:c0",
    "src_addr":"08:00:27:32:a4:60",
    "ether_type":"IP"},"ip":
    {"version":"4","dst_ip":"192.168.100.11","src_ip"
    :"172.16.74.231","protocol":
    {"name":"ICMP","type":"0", “code":"0"}}}

Jan 28 03:40:26 localhost emos[1307]: {"ethernet":
    {"dst_addr":"ff:ff:ff:ff:ff:ff",
    "src_addr":"20:c9:d0:7f:af:a7",
    "ether_type":"ARP"}, "ARP":
    {“Hardware":"Ethernet","Protocol":"IPv4","Operati
    on":"ARPRequest","SenderMAC":"20:c9:d0:7f:af:a7",
    "SenderIP":"172.16.101.183","TargetMAC":"00:00:00
    :00:00:00","TargetIP":"169.254.255.255"}}

Jan 29 05:38:48 localhost emos[1854]: {"ethernet":
    {"dst_addr":"08:00:27:32:a4:60",
    "src_addr":"40:6c:8f:58:0b:f6",
    "ether_type":"IP"},"ip":
    {"version":"4","dst_ip":"192.168.36.170","src_ip"
    :"192.168.36.115","protocol":
    {"name":"TCP","src_port":"57816",
    "dst_port":"21", "flags":"0x10" }}}
```


Fig. 3 The *rsyslogd*(8) generated form embedded packet logger

logger, while *syslog*(3) send a log message. The example of log message from packet logger is shown in Fig. 3.

This module capture packet from ethernet card. After that EPL extracts packet header then transforms to JSON message. Finally, JSON message is sent to remote log server that will be analyse in log analysis module.

5 Evaluations

Network monitoring system calls EPL to capture packet with libpcap embedded on Rasberry Pi. This section explains the experimental design that is test bed of EPL, which is focus on traffic measurement and resource consumption by comparing with *tcpdump*(1).

5.1 Experimental Design

This test bed consist of 3 modules: embedded packet logger, packet generator and target host. Packet generator and target host are PC that installed Linux, then started Web server and disabled firewall. Whereas EPL is Raspberry Pi, which is installed Linux kernel. As Raspberry Pi has only one Ethernet port, so it will require an additional Ethernet port that connected via USB. The storage is extended with 60 GB external hard disk in order to save log file. The IPv4 forward flag is enabled, while firewall service is disabled.

5.2 Performance Evaluation

Traffic generator send packets to EPL then it will forward packet to target host. All packets are captured, then processed by embedded packet logger. Finally, log message is sent to *rsyslogd*(8) service. The traffic generator send packet with packet rate 1 packet per second for 10 min, after that it will be vary between 5, 10, 50, 100, 500, 1000, and 5000 packet per second. Each generated packet is a TCP packet in three way handshake and connection termination mechanism. Thus, each packet generates 7 TCP segments, that will cause traffic jam. These situation bring to a segment lost or damage. Then the recovery mechanism retransmits these segment repeatedly, which will cause a dense traffic. This experiment focus on traffic measurement and resource consumption, which compare with *tcpdump*(1) as the benchmark.

Traffic Measurement *rsyslogd*(8) is a number of capture packet by EPL and *tcpdump*(1). The statistics result is shown in Table 2 as a comparison of the number of packet captured between the proposed tool, Packet Logger and benchmark

Table 2 Comparison of captured packet between packet logger and *tcpdump*(1)

Data rate (packet per second)	Packet logger	*tcpdump*(1)
1	0.06	0.06
5	0.30	0.30
10	0.60	0.60
50	3.01	3.01
100	6.02	6.03
500	30.15	30.15
1000	60.27	60.26
5000	303.15	302.93

named *tcpdump*(1). Most result indicate that *tcpdump*(1) captures packet more accuracy than EPL. But the EPL captures packet more than *tcpdump*(1) at packet rate 1000 and 5000 packet per second, because kernel drops packet when buffer size is full.

However, these results are statistics compared with paired *t* test method. The statistical analysis show that this difference on the number of captured packets are considered to be not statistically significant. The mean of EPL minus *tcpdump*(1) equals 0.026250, while 95 % confidence interval of this difference is from −0.038784 to 0.091284. These statistical analysis is confirmed with Fig. 4a.

CPU and Memory consumption The EPL program is used to capture packet and store in buffer in order to analyse. Each packet is analysed, extracted header then forwarded to *tcpdump*(1). The application consumes CPU and memory as shown in Fig. 4b.

CPU usage grow as polynomial function along with packet rate, because kernel consume CPU to process and analyse incoming packet. When we compare CPU usage between EPL and *tcpdump* with pair *t* test, we found that the two-tailed *P* value equals to 0.4747. By conventional criteria, this difference is considered to be not statistically significant. The mean of EPL minus *tcpdump*(1) equals to 0.0571 while 95 % confidence interval of this difference between 0.1263 and 0.2405.

Unfortunately, memory usage does not depend on packet rate as shown in Fig. 4b. The *tcpdump*(1) consumes memory more than EPL, because feature of *tcpdump*(1) is complicated more than EPL. In other words, EPL is very simple

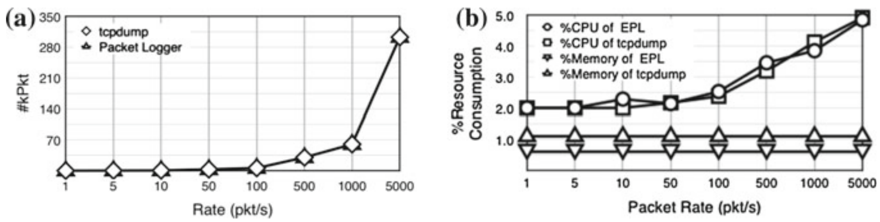


Fig. 4 Comparison: number of captured packet between packet logger and *tcpdump*(1). **a** The number of packet. **b** Resource consumption

packet capture software. It captures and generates JSON message, then JSON message will be sent to *rsyslogd*(8). Thus, packet rate does not effect the memory usage directly.

6 Conclusion and Future Work

This paper proposes embedded packet logger, called EPL, for data collection module in network monitoring system. It captures packet from local network, then generates JSON message and send to *rsyslogd*(8) service. This application is implemented and tested in Linux kernel that embedded on Raspberry Pi board. The embedded packet logger is compared with *tcpdump*(1) in term of traffic measurement and resource consumption, we conclude that both metrics are not difference significantly.

The EPL could be tuned and implemented in kernel mode, in order to reduce memory usage and CPU consumption. Moreover, packet capture in kernel mode will increase robust of mechanism. All information are shown in kernel log that is forwarded to remote log server.

Acknowledgment The authors would like to thank Walailak University to support scholarship for conference participant.

References

1. Case, J., Fedor, M., Schoffstall, M., Davin, J.: A simple network management protocol (SNMP), RFC 1157 (1990)
2. Subramamian, M., Gonsalves, T.A., Usha Rani, N.: Network management: Principles and practice, Dorling Kindersley. Peason Education, Inc. India, pp. 138–139 (2010)
3. Knowless, WW.: PiTap-Automated packet capture on a Raspberry Pi, [Online]. Available: http://williamknowles.co.uk/?p=16, (2014)
4. Wang, M.-H., Yu, C.-M., Lin, C.-L., Yen, L.-H.: KPAT: A kernel and protocol analysis tools for embedded network devices. In: IEEE International Conference on Communication QoS, Reliability, and Modeling Symposium, pp. 1160–1165 (2014)
5. Rahman, M., Khalib, Z.I.A., Ahmad, R.B.: Performance evaluation of PNtMS: A portable network traffic monitoring system on embedded linux platform. In: International Conference on Computer Engineering and Technology (ICCET), pp. 108–113 (2009)
6. Wei-ping, Z., Ming-xin, L., Huan, C.: Using MongoDB to implement textbook management system instead of MySQL. In: IEEE 3rd International Conference on Communication Software and Networks (ICCSN), pp. 303–305, May 27–29 (2011)
7. Wang, T.Z.: The development of web log mining based on improve-K-means clustering analysis. Adv. Comput. Sci. Inf. Eng. **2**, 613–618 (2012)
8. Langhnoja, S.G., Barot, M.P., Mehta, D.B.: Web usage mining using association rule mining on clustered data for pattern discovery. Int. J. Data Min. Tech. Appl. **2**(1), 141–150 (2013)

Performance of MIMO-OFDM System Based on Channel Estimation for IEEE 802.11n

I. Gede Puja Astawa, Yoedy Moegiharto and Hendy Briantoro

Abstract IEEE 802.11n is considered as the current development of standard communication WLAN IEEE 802.11 providing the increase throughput relatively standard to IEEE802.11a/g. The increase of various transmission channel in the wireless communication makes the antenna receiver work harder as the noise and fading happen in the channel. A certain method is needed to predict the channel description between the transmitter and receiver for standard communication WLAN IEEE802.11n. This research does an estimated channel simulation using minimum mean square error (MMSE) method on the MIMO-OFDM system in scheme antennas 2 × 2 and 2 × 4. The rectangular shaping on the time domain is used as channel model approach. The operation of system is stated into impulse responses sent from Tx transmitter antenna to Rx receiver output from the channel estimated output. The computer simulation program shows that the estimated channel works well on the antenna schemes 2 × 2 and 2 × 4.

1 Introduction

In the development of communication system, the need of users in quick access for information is getting higher. To provide the users' demand and increase the system operation, reliable wireless communication system is designed by combining orthogonal frequency-division multiplexing (OFDM) modulation combined with

I.G.P. Astawa (✉) · Y. Moegiharto · H. Briantoro
Electronics Engineering Polytechnic Institute of Surabaya (EEPIS), Wireless Communcation Research Group, Jl. Raya ITS—Kampus PENS, Sukolilo Surabaya 60111, Indonesia
e-mail: puja@pens.ac.id

Y. Moegiharto
e-mail: ymoegiharto@pens.ac.id

H. Briantoro
e-mail: hendy@pens.ac.id

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_94

system multiple-input multiple-output (MIMO) of which it becomes an important engineering development in WLAN communication appropriate to IEEE 802.11n standard [1].

The increase of various wireless communication channel makes the receiver antenna work harder because the noise and fading get high in the channel itself. Therefore, it needs an estimated method which can be used for making prediction and getting information about channel condition between the transmitter and receiver. The channel estimation system using OFDM pilot comb type with algorithm MMSE [2] results in channel impulse response compared between Rician and Rayleigh channel. The use of pilot comb type channel produces better channel response impulse than the use of block pilot type. However the algorithm MMSE for estimating the real channel condition is applied only on the OFDM system of which the signal quality received by OFDM receiver is rather low. Therefore the channel estimation produced is not so good and needs to combine with MMSE algorithm in interpolated channel. The previous research [3] discusses comprehensively about the block and comb pilot types for channel estimators. The channel estimation using block pilot type is done with and without equalizer decision feedback. Meanwhile, the comb pilot type is done by using pilot frequency method and channel intrusion on data frequency. The comparison on each parameter shows that pilot comb type with algorithm minim mean-square (MMSE) combined with low-pass interpolation shows the best result among all parameters being used. When the doppler value is low, the decrease operation can be ignored though the result of estimated channel using comb type with low-pass interpolation gives good influence for increasing doppler frequency. Another paper using estimation on MIMO-OFDM system [4] discusses the intrusion on the pilot channel on the changing time environment. The method uses a number of L-path channel models for estimating the path complex amplitude (CA) and calculate carrier frequency offset, while the data recovery is done by using equalizer QR. When it is compared to conventional method, the output fading shows better Doppler value 0,1. In this research, estimation is done on the channel of MIMO-OFDM System with schemes 2×2, 2×4 according to communication standard WLAN IEEE 802.11n. The system MIMO-OFDM gives more advantages such as the deletion of inter-symbol interference (ISI) and inter-carrier interference (ICI) that are caused by multipath channel. Moreover, it can strengthen the received signal on the OFDM system because the use of more than one antenna is good on both the transmitter and receiver.

2 WLAN IEEE 802.11n

WLAN IEEE 802.11n is the current development of WLAN IEEE 802.11 (a/b/g/n) which continuously gets development and improvement in its throughput. WLAN IEEE 802.11n [1] has been applied into a system having many transmitting and receiving antennas with multiple carrier modulation well known as

MIMO-OFDM. WLAN IEEE 802.11n can support the need of users through qualified video streaming that is good for some users at the same time (video conference in one WLAN network) and consistently produces high throughput (gigabyte). Besides, it can improve its Quality of Service (QoS) that is relatively better than its similarly standard WLAN. According to MIMO-OFDM system, standard WLAN IEEE 802.11n is capable to provide more rate data throughput than the original data rate that is from 54 Mb/s up to 600 Mb/s.

2.1 MIMO-OFDM for WLAN IEEE 802.11n

MIMO-OFDM in this research uses antenna schemes 2 × 2 and 2 × 4, such as follows in Fig. 1. On the transmission side, series of symbols on frequency domain are put in parallel and modulated with 16-QAM. After that, the block IFFT is added with zero padding symbol with IFFT sizes and the series of symbols are penetrated with one pilot symbol on as many as each subcarrier so that the IFFT output produces symbols with time domain the subcarrier being used. On the transmission site, the series of frequency symbols are inserted in parallel and they are modulated with 16-QAM modulation. Then, the IFFT block is added with zero padding symbols according to IFFT size being used and these series of symbols are penetrated with a symbol of pilot on each subcarrier so the IFFT output produces symbols with time domain as many as subcarriers being used. Next, 25 % of IFFT size is inserted with guard interval of cyclic prefix (CP) in the series of symbols to disappear ISI and ICI. The series of symbols are then ordered and transmitted

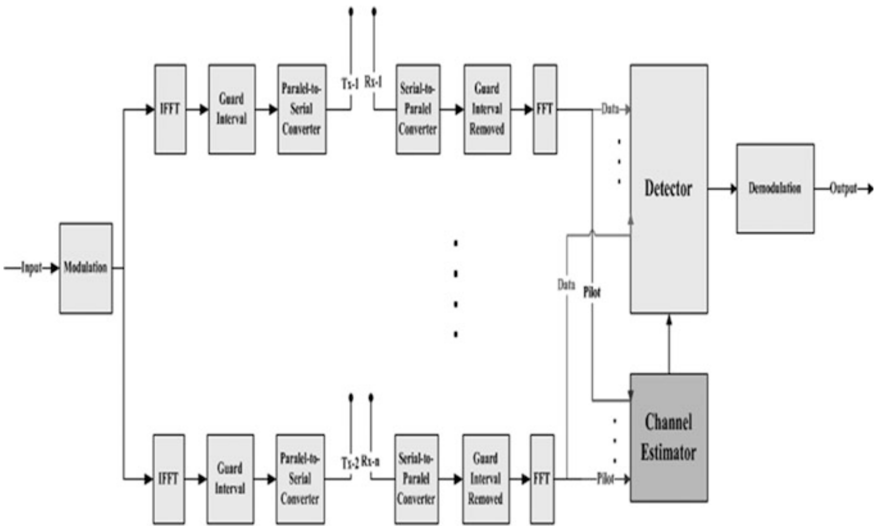


Fig. 1 Block diagram of MIMO-OFDM systems with channel estimation

through the multipath channel. The series of symbols are distributed as Rayleigh and Gaussian and received by an antenna. The series of symbols on the time domain are made parallel and pass through deleting CP block of which the process is done by FFT block. The output of FFT block is the form of pilot and symbol series. The symbol series are accepted by a detector and the series of pilot are managed by an estimator channel block. The output of estimator channel is proceed by a detector altogether with the received symbols. The series of symbols created by the detector are modulated with 16-QAM modulation and symbol output is obtained together with the input on the receiver site. The series of symbols produced by the detector are modulated with 16-QAM and this result in a symbol output that is the same with the input symbol on the receiver side.

2.2 Channel Estimation

On the receiver, the signal is generally made distortion by the channel so that, in order to get back the signal sent, channel estimation and compensation by the receiver must be done. Generally, the signal data can be used to make channel estimation. Some aspects that need consideration in channel estimation are for examples the work expectation, complexion in calculation, the changing of channel time, and so on [5]. In simple way, the channel estimation method together with Minimum Square Error (MMSE) is shown in the block diagram such as on Fig. 2 [7].

Based on Fig. 2, the value of mean square error (MSE) on the estimated channel H is inserted on the following equation.

$$J\left(\widehat{H}\right) = E\left\{\|e\|^2\right\} = E\left\{\left\|H - \widehat{H}\right\|^2\right\} \quad (1)$$

The purpose of MMSE estimation is to get better estimated value; on the other hand, it is intended to choose the most appropriate weight W value. So the Eq. (1) needs to be nominated. And by implementing the orthogonal characteristics on the vector error e = H − H, the equation can be written as follows (2):

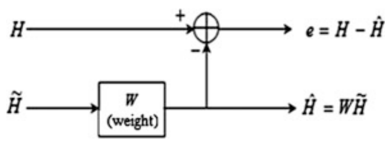


Fig. 2 Channel estimation using MMSE technique

$$\begin{aligned}\left\{e\widehat{H}^H\right\} &= E\left\{\left(H - \widehat{H}\right)\widehat{H}^H\right\} \\ &= E\left\{\left(H - W\widehat{H}\right)\widehat{H}^H\right\}\end{aligned} \tag{2}$$

2.3 The Design of Channel Estimation on MIMO-OFDM System of WLAN IEEE 802.11n

From Fig. 1 channel estimation can be designed we design. This section explains about the approach of channel model that is used so that MMSE method can be processed. This research uses channel model approach that is in rectangular shaping filter for Rayleigh fading channel [6]. The rectangular shaping equation on the time domain is shown as follows (4):

$$p_r(\tau) = \begin{matrix} 1; & 0 \le \tau \le T_1 \\ 0; & otherwise \end{matrix} \tag{4}$$

Note: T1 shows the root-mean-squared delay spread. To get shaping rectangular on frequency domain the equation is derived as follows:

$$\begin{aligned} R_p\left(\Delta_f\right) &= \int_{-\infty}^{\infty} p_r(\tau)e^{-j2\pi\Delta_f\tau}d\tau \\ &= \int_{0}^{T_1} \tau e^{-j2\pi\Delta_f\tau}d\tau \\ &= T_1 e^{-j\pi\Delta_f T_1}\frac{sin\left(\pi\Delta_f T_1\right)}{\pi\Delta_f T_1} \end{aligned} \tag{5}$$

Afterward the Eq. (5) is used to get the matrix covarian such as on the following Eq. (4).

3 Results

This research uses symbols that are transmitted by using 16-QAM modulation. Pilot sequences are insulated by using High Throughtput Long Training Field (HTLTF) basing on IEEE 802.11n on frequency 20 MHz. Other parameters are shown on Table 1.

Table 1 Parameter systems

Part	Parameter	Value
Transmitter	Pilot sequens	HTLTF
	Modulation scheme	16-QAM
	Number of sub-carrier	56
	FFT-size	64
	Dimension of antenna	2 × 2, 2 × 4
	Length of *guard interval*	¼
Channel model	Rayleigh fading	2– *rays*
	Channel model Approximated	Rectangular shaping
Receiver	Channel estimation algorithm	MMSE

3.1 Performance System Testing

3.1.1 Channel Estimation Results

This section shows the results from the output simulation program based on channel estimation design on MIMO-OFDM system using Rayleigh fading channel. The results of simulation is the real and imaginary response impulse values for each antenna schemes that are used in the system as seen from Fig. 3. To save space we only show a few figure. The system with 2 transmitters and 2 receivers gives estimation result that is shown from Fig. 3. The comparison between ideal channel estimation result and ideal channel shows difference error value shown by the error value curve in Fig. 4 for channels h11 in the system with 2 antenna receivers. Each figure states the error value in the real and imaginary subcarrier on each channel. The star curve shows the real error and the other show the error for imaginary. The axis Y shows subcarrier and axis X shows error value on each subcarrier.

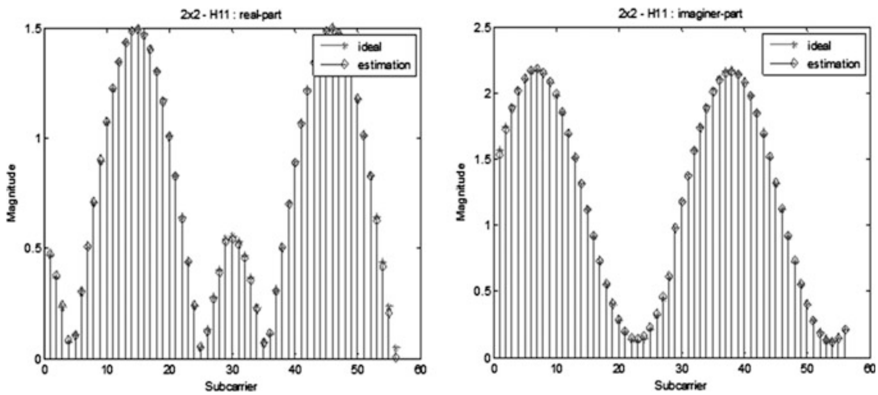


Fig. 3 Estimated channel for real part of MIMO-OFDM 2 × 2 from Tx-1 to Rx-1

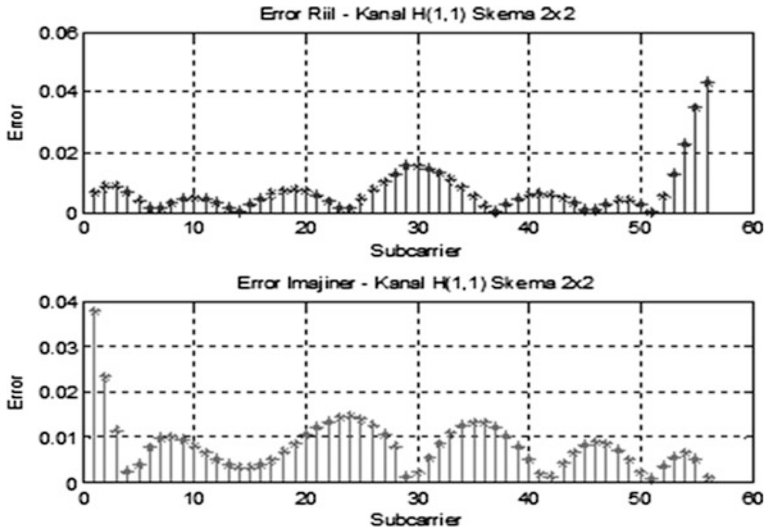


Fig. 4 Deviation/error of real-imaginary parts for channel $h_{1,1}$ of MIMO-OFDM 2 × 2 schemes

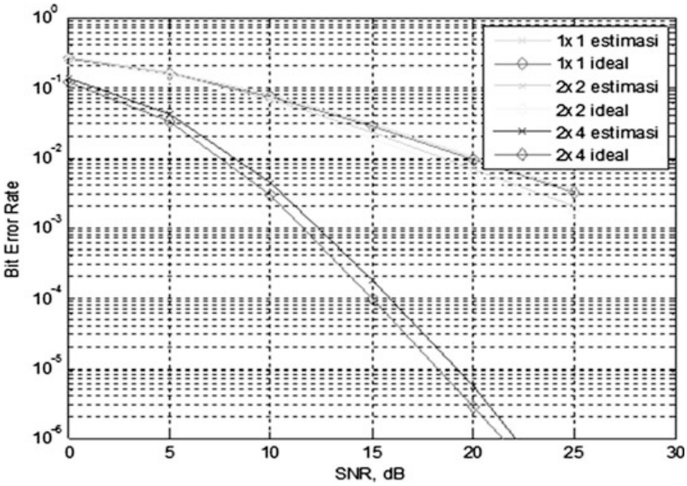


Fig. 5 BER vs SNR of Performance of MIMO-OFDM systems with and without channel estimation for 2 × 2 and 2 × 4 schemes

3.1.2 Bit Error Rate

Figure 5 shows the system SISO-OFDM performance having antenna dimension 1 × 1, system MIMO-OFDM 2 × 2, and system MIMO-OFDM 2 × 4.

4 Discussion

In the wireless communication, signal propagation between the transmitter and receiver passes various different channels. The various channels mark the environment condition of which the multipath continuously changes. The existence of multipath makes the strength of receiving signals different on the receivers. For this reason, MIMO-OFFDM is applied for data transmission on a multipath channel model that is the distribution of Rayleigh fading. Besides, the approach of Additive White Gaussian Noise (AWGN) channel model is also tested by giving disturbance of which the white noise is increased by using Gaussian distribution. On this research, a block estimator channel is added on the MIMO-OFDM system receiver with both antenna schemes 2 × 2 and 2 × 4. The simulation results show that the estimation channel output is shown by series of red dots; while the ideal channel condition is stated with impulse respose in blue color for each real and imagery values. The estimated dot position is not widely different from impulse response ideal channel on both the system scheme with 2 and 4 receivers. The receiver working operation is not influenced by the work of channel estimator however, the operation is still able to estimate exactly the channel condition for each scheme antennas. The different antenna dimension influences on the result of working system on curve BER functioning SNR. The bigger the antenna dimension is, the better system performance it will be. The betterment is marked with the lower value results for BER and SNR shown on Fig. 5.

5 Conclusion

On MIMO-OFDM system with channel estimation technique on Rayleigh channel, it is concluded that estimator channel with antenna schemes 2 × 2 and 2 × 4 can work well because the result shows that the real and imaginary channel response impulse is estimated to have closer response impulse to the ideal channel. When the channel is compared with system MIMO-OFDM having 2 antenna receivers, the MIMO-OFDM system with 4 antenna receivers gives advantage 7.2 dB without channel estimation technique, and 7.5 dB with channel estimation.

References

1. Hsieh, Meng-Han, Wei, Che-Ho: Channel estimation for OFDM systems based on comb-type pilot arrangement in frequency selective fading channels. IEEE Trans, Consum. Electron. **44**, 217–225 (1998)
2. Paul, T., Ogunfunmi, T.: Wireless LAN comes of ages understanding the IEEE 802.11n amendment. IEEE Circ. Sys. Mag. First Quarte (2008)
3. Coleri, Sinem, Ergen, Mustafa, Puri, Anuj, Bahai, Ahmad: Channel estimation techniques based on pilot arrangement in OFDM systems. IEEE Trans. Broadcast. **48**, 223–229 (2002)

4. Simon, E.P., Ros, L., Hijazi, H., Fang, J., Gaillot, D.P., Berbineau, M.: Joint carrier frequency offset and fast time-varying channel estimation for MIMO-OFDM Systems. IEEE Trans. Veh. Technol. **60**, 955–965 (2011)
5. Cho, Y.S., Kim, J., Yang, W.Y., Kang, C.-G.: MIMO-OFDM wireless communications with MATLAB (2010)
6. Astawa, I.G.P., Okada, M.: An RF signal processing based diversity scheme for MIMO-OFDM systems. IEICE Trans. Fundam. Electron. Commun. Comput. Sci. **E95-A**(2), 515–524 (2012)
7. IEEE Standard for Information Technology Telecommunications and Information Exchange Between Systems Local and Metropolitan Area Networks—Specific Requirements. IEEE Computer Society (2009)

A Methodology for Live Forensic Acquisition in Secure Domain Based on Domain Separation Technology

Kyung-Soo Lim, Yong-sung Jeon, Jeong-Nye Kim and Deok-Gyu Lee

Abstract Smart devices, including smart phone and tablet, with easy Internet accessibility, mobility and usability are changing various service environment such as mobile office, smart work, mobile finance, e-government and so on. Recently, those mobile services have trusted execution environments to enhance reliability and security for secure application or preserving sensitive information that should be protected against malicious attacks. In order to overcome cyber threats or vulnerability, domain separation technology has emerged. However, the mobile device to which a domain separation technology has been applied, conventional digital evidence collection tools, based on the general domain, cannot be accessed from the isolated secure domain, furthermore, collecting digital evidence may be impossible. Therefore, this paper describes to a methodology of providing a forensic acquisition in a domain separation-based mobile device and, more particularly, to an apparatus and method that, in order for an investigator to collect digital evidence in the secure domain of a target mobile device.

Keywords Mobile security platform · Domain separation technology · Forensic acquisition

K.-S. Lim (✉) · Y. Jeon · J.-N. Kim
Electronics and Telecommunications Research Institute, 218 Gajeong-ro, Yuseong-gu, Daejeon, South Korea
e-mail: lukelim@etri.re.kr

Y. Jeon
e-mail: ysjeon@etri.re.kr

J.-N. Kim
e-mail: jnkim@etri.re.kr

D.-G. Lee
Seowon University, 377-3 Musimseoro, Heungdeok-gu, Cheongju, Chungbuk, South Korea
e-mail: deokgyulee@gmail.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_95

1 Introduction

Recently, the various and convenient features of smart phones have brought about a rapid growth in the market of mobile phones. The increased usage of smart devices including smart phones and tablets are no more surprising news, because it is an essential item in our living life. These smart devices with easy Internet accessibility, mobility and usability are changing various service environment such as mobile office, smart work, mobile finance, e-government and so on. Meanwhile, a smart device, as a commercial-off-the-shelf device, is applied to military or government services will be increased around the world. Especially, the US Army has installed a Samsung Galaxy smartphone series into the Nett Warrior (NW) system in an effort to enhance situational awareness in the battlefield. The NW system, carried out by part of the army's focus on the tactical network modernization, is provided by the installation as the chest-mounted end-user device. Moreover, the US Army's Joint Battle Command—Platform(JBC-P), which is a networked battle command information system with connected by various military devices such as the NW system, is the first developed under an US Army effort to devise an Android-based smartphone framework and suite of applications for tactical operations [1].

On the other hand, security incidents attributable to attacks, such as an attack of malware, virus, and smishing, have also increased. Especially, in openness-oriented mobile platforms, especially Android OS, have been big security issues recently. Those mobile services on the smart devices to improve reliability have trusted execution environments for secure application and sensitive information that should be protected against malicious attacks [1, 2]. Likewise, security technology is an essential part of operating smart device services to apply smart device. It is necessary with considering private credentials or data leakage protection, because of growing prevalence of smart device, security concerns are also increasing. Thus, researching security countermeasure for the smart device is urgent including unauthorized access prevention, data leakage protection by malicious attack, etc.

In order to overcome these vulnerability and threats, domain separation technology on mobile devices has emerged nowadays. It separates an existing smart devices to diverse domains, such as normal and secure domain. Especially, secure domain manages a user's sensitive information and operates a trusted execution, likewise finance transaction, on a secure OS environment, respectively. The domain separation technology can be divided into a hardware chipset-based domain separation, a logical domain separation, and a hypervisor-based domain separation.

A hardware chipset-based domain separation technology is an isolation technology that is supported at the level of the processor of a mobile platform, and divides the operating mode of the processor into general mode and secure mode. A logical domain separation technology uses separate an application's access control policies and execution rights based on the each domain in which the application belongs, and allows minimal communication between each domains to be performed over only an authenticated channel. A hypervisor-based mobile virtualization technology is a technology that isolates a plurality of virtual machines generated by single piece of physical mobile equipment and allows communication between the virtual machines to be performed over only an authenticated channel, thereby ensuring a secure execution environment. As described above, a mobile device to which a domain separation technology has been applied has a general structure, in which a general and a secure domain are isolated from each other.

However, in the mobile device to which a domain separation technology has been applied, general digital evidence collection tools based on the general domain cannot be accessed and acquired from the isolated secure domain, furthermore, collecting digital evidence may be impossible by conventional forensic tools. Thus, a methodology for data collection corresponding to operating environment (OS) in the isolated secure domain is necessary to collect digital evidence in the secure domain.

In this paper describes to a methodology of providing an forensic acquisition in a domain separation-based mobile device and, more particularly, to an apparatus and method that, in order for an investigator to collect digital evidence in the secure domain of a target mobile device, install an evidence collection tool corresponding to each domain separation technology from an entrustment server and then collect a user's sensitive information, thereby obtaining digital evidence for conducting forensic investigations of the target device.

2 Backgrounds

In this chapter, we describe several domain separation technologies. First of all, A logical domain separation technology uses separate an application's access control policies and execution rights based on the each domain in which the application belongs, and allows minimal communication between domains to be performed over only an authenticated channel. Furthermore, an application for each domain is allowed to be downloaded from an app store for the domain, and is then used, respectively. The recent announcement of Samsung Electronics, as KNOX platform, is one of popular opportunity to let peoples knows a domain separating technique in the market and research area. The KNOX smartphone technology separates the security zone against common user area in Android smart phone. Although the domain separation in KNOX technically is not based on mobile virtualization, that is based on logical separation by marking specific area which applications works in, it will be a positive movement to increase market expansion of mobile virtualization in the future.

A hardware chipset-based domain separation technology is an isolation technology that is supported at the level of the processor of a mobile platform, and divides the operating mode of the processor into general mode and secure mode. Furthermore, the hardware chipset-based domain separation technology enables a security application and a general application to be run in two physically separate environments, respectively. TrustZone supporting ARM processor since version 6 architecture provides secure mobile security environment in the level of the processor chipset [3]. It is divided into the two operation mode with normal and secure mode. According to their functions of general or security application, which specific mode operate will be classified. The ARM processor allows that which operation mode change by this classification. It also provides a secure booting and isolated memory for individual data management with blocking feature for data leakage into other operation mode.

Mobile virtualization, as known as Hypervisor, is mainly used to ensure a trusted operating environment out of former computer virtualization technology to increase flexibility by abstracting the existing physical computing resources [2–5]. A hypervisor-based mobile virtualization technology is a technology that isolates plurality of virtual machines generated by single piece of physical mobile equipment and allows communication between the virtual machines to be performed over only an authenticated channel, thereby ensuring a secure execution environment [6]. In this case, different operating systems (OSs) may be installed on the virtual machines. Such as Secure Execution (SE) is a technique to ensure the secure execution environment by making possible to communicate isolated through an authorized channel created by each of the multiple virtual machines. Corresponding to the malicious attacks, SE are separated by isolated domains by virtualization. It used to operate in the security domain to ensure user-sensitive information protection separately, such as contacts, call history, photos. Furthermore, financial transactions, mobile banking are operated in secure domain. Several ongoing techniques are following: General Dynamics Corporation by OKL4, an open source project Xen-ARM [7], TeeMo by the ETRI [2], MVP (Mobile Virtualization Platform) by VMware and others.

3 A Methodology for Live Forensic Acquisition in Secure Domain Based on Domain Separation Technology

This chapter describes to a methodology of providing an forensic acquisition in a domain separation-based mobile device in order for an investigator to collect digital evidence in the secure domain of a target mobile device, install an evidence collection tool corresponding to each domain separation technology from an entrustment server and then collect a user's sensitive information or traces of malware attacks, thereby obtaining digital evidence for conducting forensic investigations of the target device. Our proposed methodology is based on live forensic acquisition on the particular situation, in which an investigator cannot be seizure a target mobile device, because of extent of the writ or rapid incident response.

Figure 1 is a schematic diagram illustrating the collection of digital evidence in a domain separation-based mobile device according to an embodiment of this paper. An investigator who has secured a target mobile device collects various pieces of

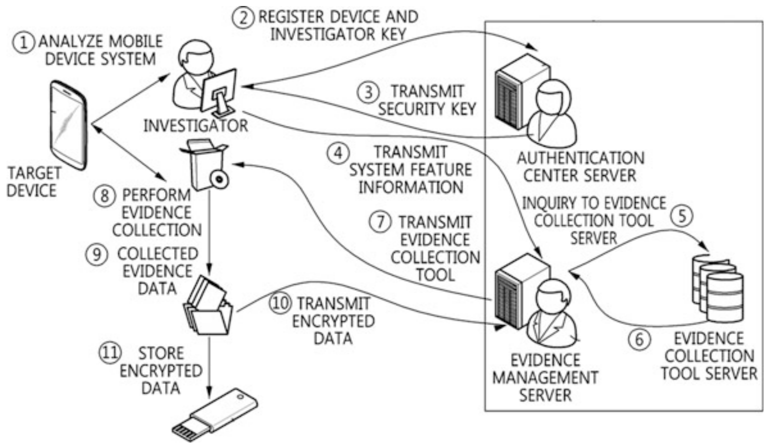


Fig. 1 A schematic diagram illustrating the collection of digital evidence in a domain separation-based mobile device

fundamental information about the corresponding device. In this case, the various pieces of information include system feature information and user identification information. The system feature information may include the model of the corresponding device, OS information, system information, and the type of domain separation technology. The user identification information may refer to unique information, such as the name of a user, a telephone number, and a manufacture serial number. Thereafter, the investigator transfers user identification information and the investigator authentication key value to the authentication center server within the server.

Accordingly, the authentication center server transmits a security key, generated based on the user identification information of the corresponding device, to the corresponding device after authenticating the investigator. The investigator uses the transmitted security key to perform encryption for the secure storage of the collected digital evidence in the future. The investigator transmits the system feature information of the investigation target mobile device to the evidence management server. The evidence management server makes an inquiry to the evidence collection tool server based on the system feature information of the corresponding device, and generates and transfer an evidence collection tool suitable for the corresponding device and to the investigator. Accordingly, the investigator collects data using the received evidence collection tool, and encrypts the collected data using the security key received at an initial authentication step. The collected data encrypted as described above is transferred to the evidence management server over a network and then stored therein, or is stored in a separate digital evidence storage device (e.g., USB memory).

3.1 *Identifying a Domain Separation Techniques*

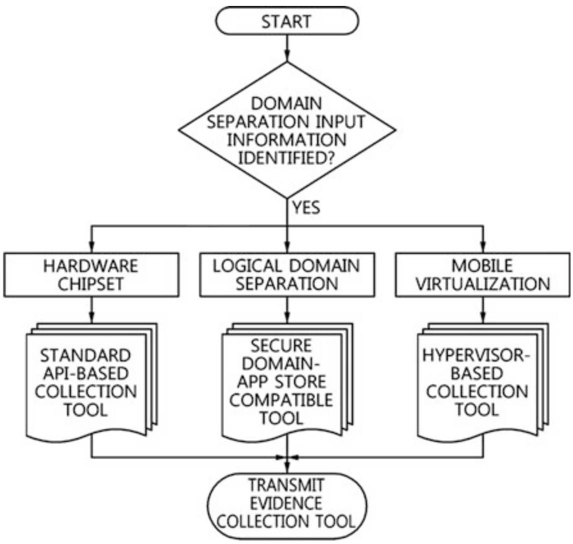
Identifying a domain isolation technology and selecting a suitable evidence collection tool is performed by the evidence collection tool server which is illustrated (⑥) and (⑦) in the Fig. 1. Figure 2 is a flowchart illustrating for identifying a domain isolation technology and selecting a suitable evidence collection tool.

The evidence management server provides the function of identifying a domain separation technology applied to the investigation target mobile device 1 and also selects an evidence collection tool based on received system feature information, and then transmits the selected evidence collection tool to the corresponding investigation target mobile device. Identifying a domain separation technology and selecting a suitable evidence collection tool, the domain separation technology of the corresponding domain separation-based mobile device is identified by detecting the type of domain separation technology from the system feature information at initial domain separation input information.

For example, if, as a result of the determination, the domain separation technology of the corresponding the target device is a hardware chipset-based domain separation technology, a standard API-based evidence collection tool is selected. In this case, the hardware chipset-based domain separation technology supports a standard API, via which access to and collection of a file of a secure domain are enabled.

Meanwhile, for example, if, as a result of the determination, the domain separation technology of the corresponding domain separation-based mobile device is a logical domain separation technology, an evidence collection tool capable of performing data collection by a secure domain compatible tool. In this case, the

Fig. 2 A flowchart illustrating for identifying a domain isolation technology and selecting a suitable evidence collection tool



execution environment of logical domain separation manages access control of app depending on whether it belongs to certain domain. The evidence collection tool corresponding to secure domain should be compatible to app store of secure domain. Only in this case, it could be collect digital evidence in secure domain of the target device. Thus, the collection tools, which is provided by the evidence collection tool server, are already registered or authenticated from app store of secure domain.

Meanwhile, for example, if, as a result of the determination of domain separation is a hypervisor-based mobile virtualization, a hypervisor-based evidence collection tool is selected. In this case, a hypervisor-based evidence collection tool supports various collection methods according to a particular hypervisors. It should provide access and collect digital evidence in the secure domain using the domain communication channel (or driver) based on the hypervisor.

3.2 Apparatus for Collecting Digital Evidence in a Domain Separation-Based Mobile Device

The apparatus for collecting digital evidence for live forensics in the domain separation-based target device, illustrated in Fig. 3, which is determined and selected by the evidence collection tool server. The apparatus for collecting digital evidence includes a target device information collection module, a control module, a collection module, and a transmission module.

The target device information collection module collects the target device information (i.e., system feature information, user identification information) of the corresponding domain separation-based mobile device for conducting forensic investigation. The target device information includes user identification information (a user name, a telephone number, IMEI, and a communication service provider,

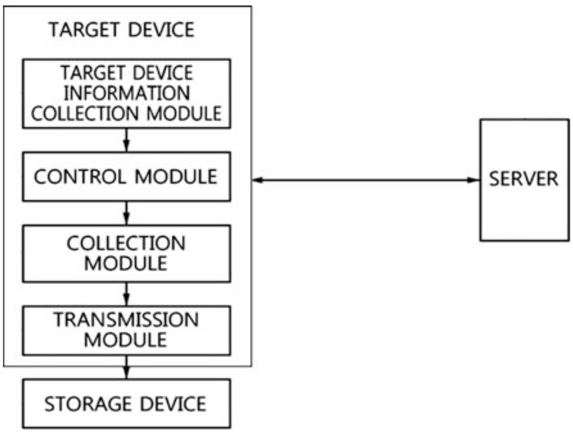


Fig. 3 A configuration diagram illustrating a forensic acquisition tool for collecting digital evidence

etc.) and system feature information (a manufacture name, model name, OS version, manufacture serial number, etc.). The target device information collection module transfers the collected these information to the control module.

The control module temporarily stores the target device information, which is transmitted from the target device information collection module, and then transfers the user identification information and the previously inputted investigator authentication key value to the authentication center server. In this case, the authentication center server performs investigator authentication, generates a security key based on the user identification information of the corresponding domain separation-based mobile device, and transmits the security key to the corresponding domain separation-based mobile device. The communication between the target device and authentication center server is applied to public key infrastructure according to conventional mechanism using server's public key and certification verification. The control module receives and stores the security key, and provides the security key when the transmission module performs encryption later.

Meanwhile, the control module transfers the system feature information of the corresponding domain separation-based mobile device to the evidence management server. In this case, the evidence management server makes an inquiry to the evidence collection tool server based on the received system feature information, and the evidence collection tool server generates an evidence collection tool suitable for the corresponding domain separation-based mobile device, and transfers the evidence collection tool to the evidence management server. As a result, the evidence management server transmits the received evidence collection tool to the corresponding the target mobile device. Furthermore, the control module transfers the received evidence collection tool to the collection module.

The collection module may collect digital evidence (i.e., a file and related data) for requiring forensic analysis in the corresponding domain separation-based mobile device using the received evidence collection tool. The collection module further comprises following: a file duplication unit configured to collect an identical file corresponding to an original file by performing duplicating physical file data, in which the data of the file has been stored, based on metadata of the filesystem; a memory dump unit configured to provide a memory dump unit configured to provide a memory dump function when the memory analysis, used in secure domain of the domain separation-based mobile device, is required; and a deleted file recovery unit configured to recover a deleted file based on metadata of the deleted file based on a processing result of the filesystem analysis unit. The collection module transfers the collected digital evidence to the transmission module.

The transmission module may encrypt the received digital evidence using the security key, which is received from the control module, and may transmit the encrypted digital evidence to the separate storage device or the evidence management server. In this case, when it is possible to store the digital evidence from the corresponding domain separation-based mobile device to the separate storage device, the transmission module encrypts the collected digital evidence and then stores the encrypted digital evidence in the separate storage device. In contrast, when it is impossible to store the digital evidence from the corresponding domain separation-based mobile device to the separate storage device, the transmission module encrypts the collected digital evidence and then transfers the encrypted collected digital evidence to the server of the evidence management server.

4 Test Case

The proposed methodology has applied to military-grade security mobile platform based on domain separation technology (hereinafter MISP, MIlitary-grade mobile Security Platform). The MISP is a researching project to develop military-grade security mobile platform based on mobile hypervisor with CC EAL4 certification and mobile security API for military purpose as final products of our ongoing research [1] (Fig. 4).

The MISP platform has been developing for widely used smart devices in order to ensure essential security applications such as military system, m-banking service using two divergent domains with separated execution environment, as secure domain for operating security service with a trusted execution environment apart from normal domain corresponding to existing mobile operating system particularly Android OS. It provides domain separation, secure filesystem, security service API, secure middleware and access control. And we have a plan to deploy our products to various mobile services to improve security such as military service, smart work, e-government, and BYOD (Bring Your Own Device) service. Furthermore, the prototype of MISP, as we called Trusted Military Zone (TMZ) solution, provides military security service (including secure SMS and MMS, contacts, camera, gallery, secure voice communication, emergency situation notice, etc.), which is based on

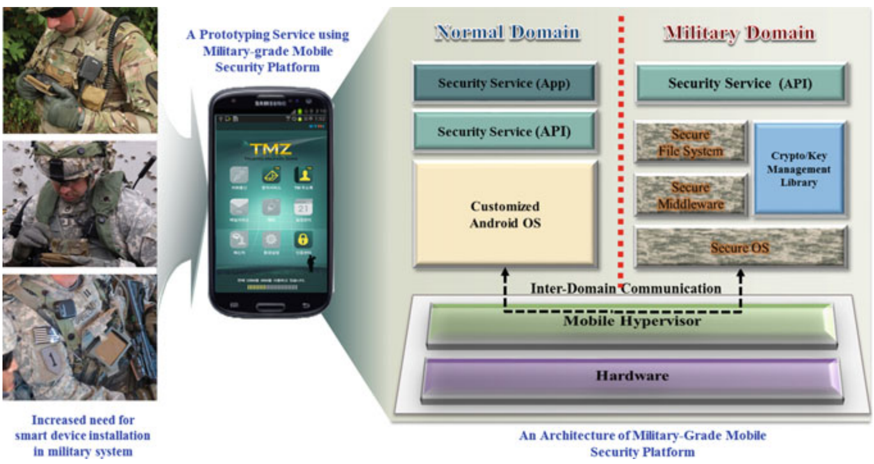


Fig. 4 A prototyping service using military-grade mobile security platform

Table 1 A table for file acquisition time using the forensic collection tool

Type	File size (bytes)	Acquisition time (averages 100 times) (s)
Text messages	3504	0.031
Contact lists	520	0.017
Image files	654,532	4.602

mobile hypervisor and middleware, secure filesystem, cryptography library, and certification authentication service comply with FIPS 196. And it has been installing and testing to a commercial smart device, currently Samsung Galaxy S3 LTE [1].

The proposed methodology can be applied to MISP platform device. Currently, our proposed research has been implemented including user authentication process, control module, collection module and transmission module, partially. If we assumed that TMZ including MISP platform is installed in the target device (Samsung Galaxy S3 LTE) for conducting forensic investigation, the proposed methodology will be determined that corresponding domain separation technology is a mobile hypervisor. The forensic acquisition tool is selected and transferred by the evidence collection tool server as mentioned above. This forensic tool collects digital evidence including particular files such as text messages, contact list and photo images. Above Table 1 shows these acquisition time acquisition time according to file type and size. The reason, we describe file acquisition time in Table 1, have an effect on total time for live forensic acquisition of target device based on mobile hypervisor. In case of large-scale investigation for MISP devices in the field hereafter, the forensic acquisition time is essential part for rapid response, when some device has compromised unidentified threat or malfunction.

5 Conclusion and Future Works

Smart devices with easy Internet accessibility, mobility and usability are changing various service environment such as mobile office, smart work, mobile finance, e-government and so on. Those mobile services on the smart devices to improve reliability have trusted execution environments for secure application and sensitive information that should be protected against malicious attacks. For enhancing security of smart device, domain separation technology has emerged recently. However, in the mobile device to which a domain separation technology has been applied, conventional digital evidence collection tools based on the general domain cannot be accessed and acquired from the isolated secure domain, furthermore, collecting digital evidence may be impossible by conventional forensic tools. Therefore, this paper describes to a methodology of providing a forensic acquisition in a domain separation-based mobile device and, more particularly, to an apparatus and method that, in order for an investigator to collect digital evidence in the secure domain of a target mobile device. We also have been researching a project to develop military-grade security mobile solution based on security platform with mobile hypervisor. The suggested methodology

Acknowledgments Foundation item: This work was supported by the ICT R&D program of MSIP/IITP, Korea. [R0101-15-0195(10043959), Development of EAL 4 level military fusion security solution for protecting against unauthorized accesses and ensuring a trusted execution environment in mobile devices].

References

1. Lim, K., Park, S., Kim, J., Lee, D.: Functional considerations in military-grade security platform using a mobile hypervisor. In: Computer Science and Its Applications. Lecture Notes in Electrical Engineering, vol. 330, pp. 1413–1418 (2015)
2. Lee, Y.K., TeeMo, J.N.: A Generic Trusted Execution Framework for Mobile Devices, International Conference on Computer, Networks, Systems, and Industrial Applications(CNSI), pp. 579–583 (2012)
3. Frenzel, T., Lackorzynski, A., Warg, A., Härtig, H.: ARM TrustZone as a virtualization technique in embedded systems. In: 12th Real-Time Linux Workshop (2010)
4. Colp, P., Nanavati, M., Zhu, J., Aiello, W., Coker, G., Deegan, T., Loscocco, P., Warfield, A.: Breaking up is hard to do: security and functionality in a commodity hypervisor. In: Proceedings of 23rd ACM Symposium on Operating Systems Principles (2011)
5. Andrus, J., Dall, C., Van't Hof, A., Laadan, O., Nieh, J.: Cells: A virtual mobile smartphone architecture. In: Proceedings of the 23th ACM Symposium on Operating Systems Principles (2011)
6. Lange, M., Liebergeld, S., Lackorzynski, A., Warg, A., Peter, M.: L4Android: a generic operating system framework for secure smartphones. In: Workshop on Security and Privacy in Smartphones and Mobile Devices (2011)
7. Kim, K., Kim, C., Jung, S. Shin, H., Kim, J.: Inter-domain socket communications supporting high performance and full binary compatibility on Xen. In: Proceedings of the Fourth ACM SIGPLAN/SIGOPS, pp. 11–20 (2008)

Performance Evaluation of Extended EXACT Algorithm for Analyzing Breast Cancer Using Fractal Dimension

B. Monica Jenefer and V. Cyrilraj

Abstract In the human body, very tiny particles named nano-particles can spoil the living organisms. Analyzing the growth pattern of breast cancer is more important theoretically as well as clinically. The ability of the nano-particles starts with their small size, and spoil and travel in an irregular circular shape in a body. The cancer growth forms a cluster by a group of nano particles, aggregates in a high level and form fractals. In this paper, the cancer growth is simulated and analyzed using various existing mathematical models. These algorithms are taking more time and searches all the points linearly within a distance take costlier. To provide more efficiency in terms of time and cost, a new mathematical model extended from EXACT algorithm is proposed for modeling a cancer growth model. EEXACT algorithm generates a model in terms of time, and number of points searching dynamically in an efficient manner. The performance of the EEXACT algorithm is evaluated by comparing its performance metrics with the performance metrics of existing approaches.

Keywords Cancer growth model · Fractal · Breast cancer · Mathematical cancer growth model · EEXACT growth model

1 Introduction

Cancer is a disease in which the patient is affected by abnormal cell growth which also spreads to other parts of the body. The cancer that affects breast tissues is called as breast cancer. The breast cancer develops in the milk duct linings and the lobules

B. Monica Jenefer (✉)
Department of CSE, Sathyabama University, Chennai, India
e-mail: bmonicajenefer@gmail.com

V. Cyrilraj
CSE, Dr. M.G.R. Educational and Research Institute, Chennai, India
e-mail: cyrilraj@hotmail.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_96

that supply milk to the ducts. The cancer that forms in ducts are known as ductal carcinomas and the cancer developed in lobules is known as lobular carcinomas. There are also some other sub types of breast cancer. The breast cancer diagnosis is carried out by taking a biopsy of the damaged tissues. After the process of diagnosis the tests are conducted to find out the areas affected by cancer and to determine the type of treatment that will suit the type of cancer. The symptoms of breast cancer are lump in the breast, change in breast shape, skin dimpling, ared, scaly patch of skin or fluid coming from the nipple. Also, there will be bone pain, swollen lymph nodes and yellow skin around the area spread with the disease. 5–10 % of breast cancer is formed due to inherited genes. The tamoxifen or raloxifene medications are used in preventing breast cancer from spreading, but the risk of developing those medications is high. Removing the breasts surgically can also be done as a preventive measure which also involves high risk. After diagnosis, surgery, radiation therapy, chemotherapy and targeted therapy treatments are carried out in order to remove the affected tissues completely. The surgery may be a breast conserving surgery or a mastectomy according to the disease. If the cancer has spread to other parts of the body the treatment for improving quality of life and comfort and the eradication of cancer cells is not possible.

There various methods and techniques are available for analyzing the breast cancer data. But one of the main assessing methods for understanding the breast cancer growth is the fractal analysis. Fractal analysis is a process of assessing fractal characteristics of a data [1]. The term fractal refers to the mathematical set that has a repeating pattern in each process or iteration of computing. It is also called as self-similar pattern. The analysis of similar patterns occurring in a data is called as fractal analysis. There are several ways of assigning a particular fractal characteristic to a dataset depending on a pattern or signal extracted from dataset. The fractal analysis of data has wide application in the field of modern science. Even though there is fractal dimension obtained as a result of fractal analysis of a data the pattern cannot be declared as fractal without considering the essential characteristics.

2 Background Study

In this paper the fractal design is chosen for the analysis of breast cancer. Magnetic Resonance Imaging—[MRI] images generally have a degree of randomness in relation with natural random nature of structure. So the Fractal analysis is the best method for MRI image analysis [2]. The FIFs (Fractal Interpolation Functions) are useful in reconstruction of 3d tumor data from 2d slices obtained using MRI [3]. Implementing fractal analysis has advantages such as extraction of fractal dimension spectra for the region of interest and usage of classifier that automatically identifies the kind of tumor [4]. The Breast tumors may be classified based on their shapes using FFT based methods namely benign and malignant [5]. Fractal analysis is useful in digital image processing for the process of assessing the shape and complexity [6]. The Breast tumor masses provide shape and complexity characteristics in mammograms [7]. Brú et al. [8], discussed about various mathematical models providing stochastic equations for tumor-host information and they are referred here in the radial symmetry case. The cells in the breast image are divided into cancer cell and cancer supporting cell. Tumor growth is explained as a multidisciplinary problem and some of the challenges are given in the form of mathematical representation in [9]. While analyzing the tumor growth, cancer cells are always treated as symmetric and cancer supporting cells are treated as symmetric as well as asymmetric. The evidence for asymmetric division for CSC is reported in [10–12]. Farzaneh Keivanfard et al. in their work, have applied feature selection and classification methods based on ANN to classify breast cancer on dynamic Magnetic Resonance Imaging (MRI) [13].

3 Problem Statement

Death rate is increasing due to breast cancer since it grows very fast and it is not easy to detect in the very beginning stage of the cancer. In order to reduce and avoid human death due to breast cancer, it is necessary to identify the cancer symptoms and cancer occurrence in the initial stage itself. By developing a computer based software, the breast image condition, fastness and growth of the cell pattern are analyzed for providing immediate medical treatment to the patients. In this paper, it is aimed to develop a mathematical model based approach to pre-screen the breast image as well as analyzing the growth pattern of the breast tumor. Our approach Extended EXACT algorithm is utilized to analyze the growth pattern of the breast tumor in an accurate manner than the existing approaches.

It is essential to design and develop a mathematical model for breast cancer growth analysis. The main objective of the paper is to provide an efficient mathematical model to verify cancer growth using a mathematical model. Also the proposed model is evaluated by comparing the existing approaches. All the existing approaches are concentrating in verifying the cancer and cancer size. EEXACT approach verifies the cells in the tumor as good cell or bad cells, number of cells belong to tumor, and the size of the tumor in terms of time. It is proved that the EEXACT algorithm is better than the existing EXACT algorithm in terms of time and tumor cell detection rate is given in Figs. 1 and 2. From the figures, it is clear that the number of fractal generation obtained by EEXACT is better than the EXACT in terms of time and fractal analysis.

4 Existing Studies

There are various mathematical models are available for fractal dimension calculation and analysis of breast cancer growth. In general, the growth of the fractals increases, in terms of time and number of bad cells existing already. The process of angiogenesis is very fast after tumor forms in the initial stage. In this paper, it is motivated to analyze the tumor growth, according to the matching between the fractal model data with the clinical-data. The growth of the breast cancer is modeled using some of the specific methods, which are given below:

4.1 Exponential Growth Model

Exponential growth is one of the models for analyzing the tumor growth in breast images. In this model the number of cells are $\boldsymbol{N}$, and $N(0)$ is the starting size of the cell and the time of the tumor growth is t. Let us assume a and b are constants, then the following formula

$$N(t) = N(0) \cdot \exp(bt) \tag{1}$$

is used to compute the growth of the cancer within the time interval $\boldsymbol{t}$. The time $\boldsymbol{t_d}$ needs for the growth $N(t_1)$, becomes double, that is,

$$N(t_1 + t_d)/N(t_1) = 2 \tag{2}$$

is constant all the time at

$$\log_e(2)/b \tag{3}$$

When the actual data fit with exponential pattern growth data, then this data is useful. When population increases, the time $\boldsymbol{t_d}$ is also increases and it is not a constant.

4.2 Gompertzian Model

So that, Gompertzian model [14] is applied. In Gompertzian based growth model $N(t)$ is a function of $N(0)$, t and b also limiting the size of the cells from $N(\infty)$, using the following equation:

$$N(t) = N(0) \cdot \exp\{k \cdot [1 - \exp(-bt)]\} \tag{4}$$

where

$$k = \log_e[N(\infty)/N(0)] \tag{5}$$

This formal fits the experimental data and clinical data, and exposes the growth-regulatory mechanisms in animal [15, 16] and human [14] cancers.

4.3 ODE Tumor Growth Model

There are five different ODE growth models [17] are available such as: exponential, power law, logistic, Gompertzian and von Bertalanffy. Assume P denotes an arbitrary population and t denotes the time. Now the exponential growth models is the first and the simplest ODE growth model, described by

$$\frac{dP}{dt} = rP \tag{6}$$

The growth rate is represented as a constant r. The Exponential growth is derived from power law growth, and it denoted by,

$$\frac{dP}{dt} = rP^{a} \tag{7}$$

where r and a are variables [parameters] should fit with the data. The Logistic growth is another growth model incorporating the population and the capacity carrying by the population, and it can be given as:

$$\frac{dP}{dt} = rP\left(1 - \frac{P}{K}\right) \tag{8}$$

where r denotes the intrinsic growth rate, K denotes carrying capacity. The Logistic growth model is more similar to the exponential growth model. And the von Bertalanffy model is also incorporating the population with its carrying capacity, and it can be defined as:

$$\frac{dP}{dt} = r(K - P) \tag{9}$$

Finally, the Gompertz growth model can be defined as:

$$\frac{dP}{dt} = r\ \log\left(\frac{K}{P}\right)P \tag{10}$$

From the above definitions and mathematical formulas, it is clear that the Gompertzian model is more accurate than the other models and these models are commonly explained and experimented in the earlier researches.

4.4 Power Law Growth Model

In an earlier work, tumor growth was explored using a *CA* model, in the beginning presented in [4] for telling rising tissues. Numerical simulations of this model were conducted to attain the macroscopic dynamics of tumor growth. For the two-dimensional (2D) CA, it is observed that the time course of total number of cells in the unsaturated stage of growth (before space/nutrient limitations are imposed) could be well approximated by a parabola, i.e., it is proportional to the square of time, suggesting power law tumor growth [9]. We also simulated a one-dimensional (1D) CA, which showed linear growth of the total number of cells. Thus, we observed that both in 1D and 2D cases, tumor diameter grows linearly in time.

The following global dynamics, observed in the simulation results of the CA model, now comprise the underlying assumptions of a new macroscopic model for the unsaturated stage of growth. It is important to note that these assumptions are in agreement with the experimental observations of spheroid growth [9]. The assumptions are as follows:

- The cell colony has a spherically symmetric form with diameter $D(t)$;
- For some constant `, in the inner area of depth larger than ` (inner sphere of radius $0{:}5D(t)$¡`), the cell population is homogeneous and in equilibrium (i.e., the net growth rate is zero);
- In the rim between radius $0{:}5D(t)$¡` and $0{:}5D(t)$ the cell densities are dependent on the relative position in that rim, but not on time; cells in this rim have a net growth rate which depends on the depth (distance from the boundary) only;
- The cell mass expands due to cell population growth, but the properties of the inner mass and the outer rim remain invariant.

These assumptions imply that the diameter of the cell mass grows at a constant rate, regardless of the dimension (to see that, note that a unit surface area produces a constant number of cells per time unit, thus increasing the outer radius by a constant length). Therefore, we obtain the following equation

$$\frac{dD(t)}{dt} = k \tag{11}$$

where k is the constant rate of increase in diameter. From our macroscopic model, it directly follows that the diameter of the occupied area will grow linearly in time during the unsaturated stage. In the 2D case we can write the following expression for the total cell number.

$$n = \frac{\pi D^2 \rho}{4} \tag{12}$$

where ρ is cell density. Substituting D(t) from above equation yields

$$n = \left(\frac{k\sqrt{\pi\rho t}}{2} + \sqrt{n_0}\right)^2 \tag{13}$$

where n is the initial cell number. In the same way, cubic growth equation can be derived for the three-dimensional (3D) case. The last equation represents parabolic cell population growth, as indeed appears in the simulations of the original CA model. We subsequently calculated k for different combinations of parameter values as an estimation of the macroscopic growth rate. Importantly, the relative densities of different cell types in the inner part of the cell population during unsaturated growth phase were indeed identical to those found in the steady state phase, enabling calculation of the parameter ½ (data not shown). This proves that the assumptions of the macroscopic model presented here are coherent with simulation results of the cellular automata model in [18].

Some of the user defined models also considered and verified in this paper, to decide about the proposed growth model in order to improve the accuracy and they are discussed.

4.5 HYBRID Algorithm Based Growth Model

The main aim of this method is to compute the perimeter estimation and the hardcore portion of the tumor occurrence within a perimeter. The starting point $\boldsymbol{X_s}$ is taken randomly, and then assume that the perimeter $P = 0$. The current point $X_c(i)$ is compared with the $\boldsymbol{X_s}$ to verify both points are same or dissimilar. To obtain the next point within the perimeter or on the contour, increase the current point $X_c(i)$ index $X_c(i+1)$. Set a point as running point $\boldsymbol{X_r}$ and compare with the next current point $X_c(i+1)$. If, both the point are same connect the points, increase the perimeter P = P + 1 and move ahead within the perimeter or on the contour. This process is repeated and applied to all the points within a distance and within the perimeter.

From the starting point SP, the HYBRID algorithm searches the next pivotal point. To find the next point, the starting point is copied as the next point and the mutual distance ε is computed to collect all the similarity points. The similarity points running through the border are taken into account as running points those who are having the distance from CP which is very close to the distance ε.

The procedure carries on until the SP is reached. Obviously it is likely that after a complete walking the starting point *SP* may be reached before having hit the following current point *CP*. In other words, there may not be a multiple of step size ε so that the final incomplete step length r is added to the others stored distances, whose sum represents the boundary's perimeter. Since the fixed step length is adapted every time during the perimeter computation, its averaged value is then computed and used in the Richardson's plot.

Estimating Perimeter: [Arandom *SP-Starting Point* (X_s, Y_s) lie on the boundary line is selected and assign into an another variable called CP—[*current point* $(X_c(i), Y_c(i))$]. The index i runs through the total number of coordinate points and is iteratively increased defining the *running point R* with coordinates (*xR*, *yR*).

The distance ***d*** between ***SP*** and ***RP*** is calculated and a check on d when smaller than a fixed step length ε is done. The process is repeated until a boundary point whose distance from (*xS*, *yS*) is larger than the step ε is reached. The next pivot point on the boundary line is determined by choosing between the two points the closest to the step length. The distance is then stored and this point becomes the new starting point in order to calculate the next pivot point and so on, until the initial starting point *S* is reached.

4.6 EXACT Algorithm Based Growth Model

In HYBRID algorithm, each step is compared to the previous point and it takes much time. But, EXACT algorithm overcomes this problem and it connects all the contour points with a series of straight lines (Fig. 3).

The EXACT algorithm developed and proposed by Clark [19, 20], whereas this method will take more computational time by providing a simple solution. The EXACT's procedure is also very similar to the HYBRID algorithm. But in EXACT algorithm, the end of the procedure is not coinciding with the coordinates of the boundary. The EXACT algorithm connects all the points using a straight line. The point connection, finding the location of the next point determination is depicted in Fig. 4.

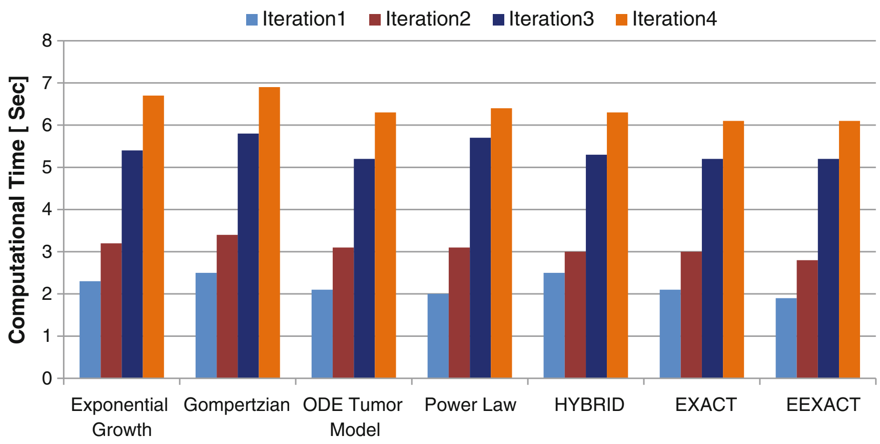


Fig. 1 Comparison of computational time

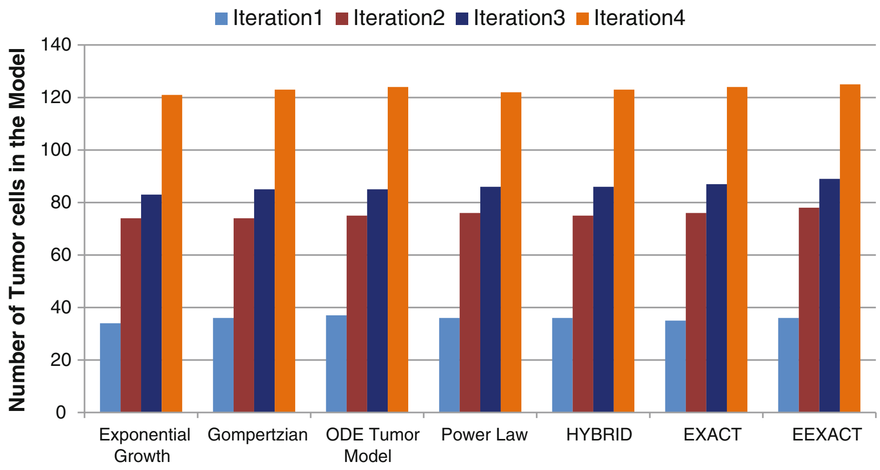


Fig. 2 Number of fractals in different scaling

In EXACT algorithm, starts its functionality from an original starting point (xs, ys), then start searching the next pivot point. Whenever, it finds the next points, the present starting point is copied to current point C[Xc, Yc] obtained within a distance ε. The point searching running though the border, and they are denoted as running point R(XR, YR). The distance between two points, for example, (XR, YR) and (XR + 1, YR + 1) is obtained using the following formula as:

$$\sqrt{(X_R - X_C)^2 + (Y_R - Y_C)^2} < \varepsilon < \sqrt{(X_{R+1} - X_C)^2 + (Y_{R+1} - Y_C)^2} \tag{14}$$

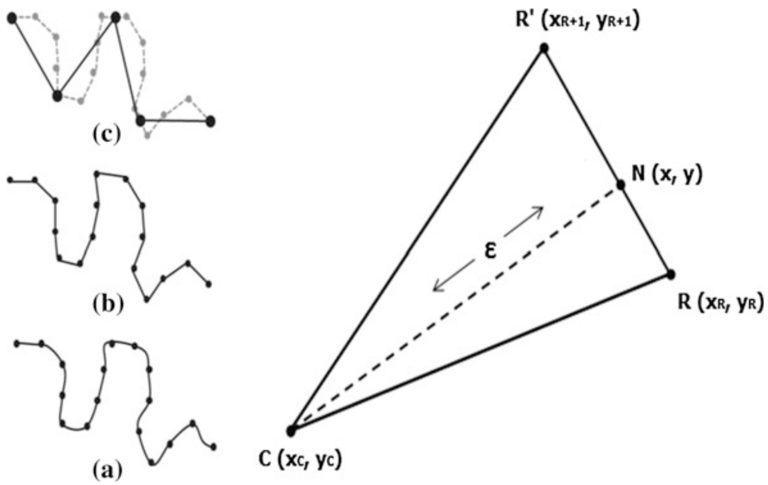


Fig. 3 Point connect by finding a next neighbor point

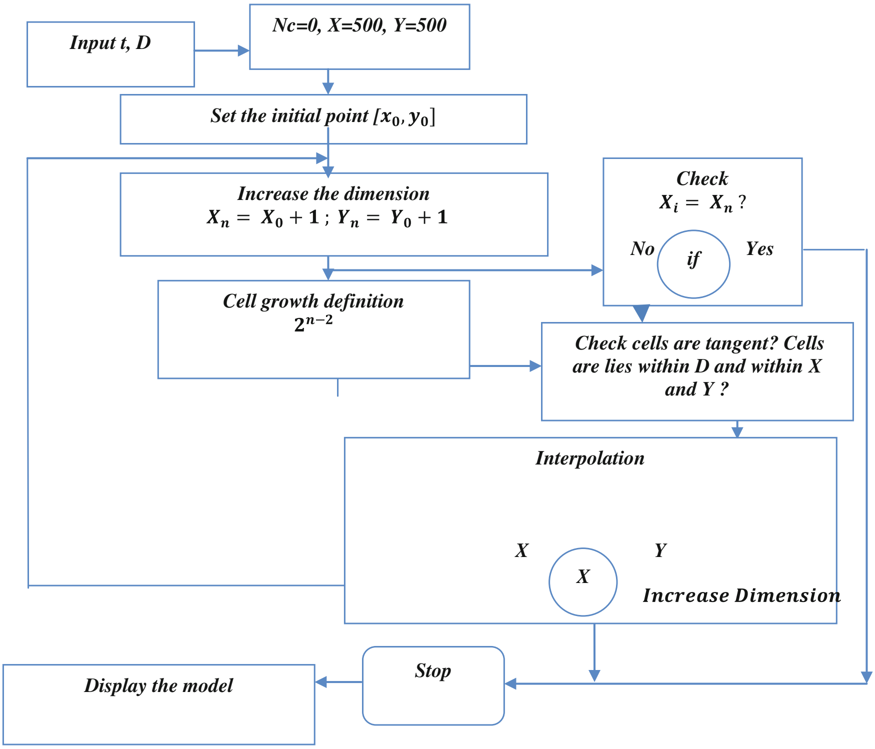


Fig. 4 Flowchart for EEXACT algorithm

And this can provide a new point as the new current point and it can be used to calculate the next boundary point vice versa. This process is stopped the new current point reaches the starting point (Xs, Ys).

The starting point S is chosen in the boundary and copied as current point. Then through the total number of points, iteratively the next points are chosen as running point within a distance d between S and R. Repeats this step until it reached the starting point at the ending point.

5 Proposed Algorithm

The proposed is to obtain the entire tumor cells occur in the breast images. It is important because the existing HYBRID and EXACT approaches are searching the tumor cells within a boundary and it avoid very small number of tumor cells occurs in the breast mammograms. For example, the existing HYBRID and EXACT approaches starts from a starting point S and ending with the same point S by computing nearest neighbours. If, the algorithms cannot match the starting and ending point, then the algorithms takes more time to make the boundary because the neighbours occur in all the directions. So, it is not possible to cover all the tumor cells by drawing a curve or a straight line. In this algorithm, EEXACT approach fetches the entire tumor cells by random comparison in all the directions.

5.1 Extended EXACT Algorithm Based Growth Model

The EXACT algorithm is searching the points linearly within a distance threshold ε on the boundary through the Running points $R(Xr, Yr)$ from the Starting point (Xs, Ys). But, it is difficult to choose the starting point on the boundary; also, developing a simulation of the cancer growth model is having different shapes and different sizes. The model itself developed using assumptions. Flowchart is given in Fig. 2.

Assumption in EEXACT

But in EEXACT algorithm, the simulation of the cancer growth model is initialized with some parameters such as:

$$time - t, size - S, cellsize - c, grid - size[x, y], gridcells - gc.$$

It is assumed that in a defined grid, dimensioned as $[x, y]$, divided into tiny squares representing the pixels [cells] called as $gridcells[gc]$. It is well known that, the cells in the human body are tangent to one another at one point at least. Also, it is assumed that the tumor is growing from a random place in the breast, and grows outside in shapeless shape and it may be an irregular circle. Finally, it is assumed that the size of the tumor growth is increased within the specific time interval $\boldsymbol{ti}$, the number cancerous cells are increased into double the time of available cancerous cells.

In EEXACT, the size of the tumor model can be represented using the following mathematical formation as: At the time t,

$$\frac{ds}{dt} = (t - 1) \cdot 2^{t-2} \tag{15}$$

where the number of cells grows double the time of its availability in a time interval t. Using this, the size of the tumor model is developed in terms of time in the dimension [x, y].

EEXACT Algorithm

- $InputDisthediameter, numberofcellnc = 1.$
- $Input\ the\ time\ interval\ m$
- $For\ I\ =\ 1\ to\ X\quad //\ X\ dimension$
- $For\ J\ =\ 1\ to\ Y\quad //\ Y\ dimension$
- $For\ K\ =\ 1\ to\ D\quad //\ Diameter\ of\ the\ irregular\ circle$
- $nc\ =\ pow(2, nc)\quad //\ number\ of\ cells$
- $If\ (nc\ tangent\ to\ nc{+}1)\ then$
- $Gm\ =\ true\quad //\ growth\ model\ is\ true$
- $End\ if$
- $For\ t\ =\ 0\ to\ m$
- $ds/dt\ =\ (t-1)\ *\ pow(2, (t-2))$
- $End\ t\ \ End\ DEnd\ JEnd\ I$
- $Display\ the\ model\ and\ check\ the\ size$

The numbers of cell grow within the dimension *X*, *Y* and within the diameter *D*. The growth is analyzed in terms of time changes. From an initial time *t*0 to a final time, the size, number of cells and the shape for proper treatment are verified. Figure 2 depicts the flowchart of the EEXACT algorithm proposed in this paper.

6 Simulation Results and Discussion

All the described algorithms are implemented in MALTAB and utilized for biomedical images. The computed results are compared with the theoretical values. The computational time for 2.50 Ghz, Core i5, CPU is shown in Table 1. The values obtained from EEXACT algorithm are time, size, number of cells and are compared with the same values obtained from other methods. This growth modeling is also simulated by scaling the size of the tumor iteratively k number of times (Table 2).

Table 1 Comparison of computational time

Mathematical models	Computational time			
	Iteration1	Iteration2	Iteration3	Iteration4
Exponential growth	2.3	3.2	5.4	6.7
Gompertzian	2.5	3.4	5.8	6.9
ODE tumor model	2.1	3.1	5.2	6.3
Power law	2	3.1	5.7	6.4
HYBRID	2.5	3	5.3	6.3
EXACT	2.1	3	5.2	6.1
EEXACT	1.9	2.8	5.2	6.1

Table 2 Number of fractals in different scaling

Mathematical	Iteration1	Iteration2	Iteration3	Iteration4
Exponential growth	34	74	83	121
Gompertzian	36	74	85	123
ODE tumor model	37	75	85	124
Power law	36	76	86	122
HYBRID	36	75	86	123
EXACT	35	76	87	124
EEXACT	36	78	89	125

From the results and the 2D models generated by all the mathematical models for the cancer growth, the size of the growth [that is growth factor] suggests the patients or the doctor to take the necessary dose of the radiotherapy. The radiotherapy cures the cancer and reduces the size of cancer by increasing the hemoglobin and killing the bad cells. This method avoids the major surgery and reduces the death due to cancer growth.

7 Conclusion

In this paper, an Extended Exact Algorithm is proposed for developing a cancer growth model for fractal dimension. Analyzing the growth helps to identify the severity and further proceed for radiotherapy. The model can be simulated in terms of time. The performance of this EEXACT algorithm is evaluated by comparing its performance with the existing algorithms such as Linear and Exponential mathematical models. All the algorithms are described briefly and implemented in MATLAB. The computational time, cost and model representation are compared among all the algorithms. From the comparison results it is concluded that, our proposed approach EEXACT is better, easier understood, less cost and fast efficient to generate a model in any size. Also, these approaches can be utilized with any benchmark data, any resolution based images and the performances are compared in the future.

References

1. Kimiagar, S.: Fractal shape of cancer tumor growth. Biochem. Biophys. (BAB) **1**(1) (2013)
2. Iftekharuddin, K.M., Jia, W., Marsh, R.: A fractal analysis approach to identification of tumor in brain MR images. In: Proceedings of the 22nd Annual International Conference of the IEEE on Engineering in Medicine and Biology Society, vol. 4, pp. 3064–3066, 23 Jul 2000–28 Jul 2000
3. Craciunescu, O.I., Das, S.K., Samulski, T.V.: Piecewise fractal interpolation models to reconstruct the three-dimensional tumor perfusion [MRI data analysis]. Proceedings of the 22nd Annual International Conference of the IEEE on Engineering in Medicine and Biology Society, pp. 292–305 (2000)

4. George, L.E., Sager, K.H.: Breast cancer diagnosis using multi-fractal dimension spectra. IEEE International Conference on Signal Processing and Communications (ICSPC 2007), pp. 592–595, 24–27 Nov 2007
5. Syed Abdaheer, M., Khan, E.: Shape based classification of breast tumors using fractal analysis. In: Multimedia, Signal Processing and Communication Technologies, pp. 272–275, 14–16 Mar 2009
6. Buzea, C., Pacheco Blandino, I.I., Robbie, K.: Nanomaterials and nanoparticles: Sourcesand toxicity. Biointerphases **2**(4), MR17–MR172
7. Cabral, T., Rangayyan, R.: Fractal analysis of breast masses in mammograms. Synth. Lect. Biomed. Eng. (2012)
8. Brú, A., Casero, D., de Franciscis, S., Herrero, M.A.: Fractal analysis and tumour growth. Math. Comput. Model. **47**, 546–559 (2008)
9. Brú, A., Herrero, M.A.: From the physical laws of tumour growth to modelling cancer processes. Math. Models Methods Appl. Sci. **16**(Suppl. 1), 1199–1218 (2006)
10. Pine, S.R., Ryan, B.M., Varticovski, L., Robles, A.I., Harris, C.C.: Microenvironmental modulation of asymmetric cell division in human lungcancer cells. Proc. Nat. Acad. Sci. USA **107**, 2195–2200 (2010)
11. Knoblich, J.A.: Asymmetric cell division: recent developments and theirimplications for tumour biology. Nat. Rev. Mol. Cell Biol. **11**, 849–860 (2010)
12. Visvader, J.E., Lindeman, G.J.: Cancer stem cells in solid tumours: accumulating evidence and unresolved questions. Nat. Rev. Cancer **8**, 755–768 (2008)
13. Speer, John F., Petrosky, Victor E., Retsky, Michael W., et al.: A stochastic numerical model of breast cancer growth that simulates clinical data. Cancer Res. **44**, 4124–4130 (1984)
14. Norton, L.: A gompertzian model of human breast cancer growth. Cancer Res. **48**, 7067–7071 (1988)
15. Retsky, M.W., Wardwell, R.H., Swartzendruber, D.E., et al.: Prospective computerized simulation of breast cancer: comparison of computer predictions with nine sets ofbiological and clinical data. Cancer Res. **47**, 4982–4987 (1987)
16. Baish, James W., Jain, Rakesh K.: Fractals and cancer. Cancer Res. **60**, 3683–3688 (2000)
17. Sarapataa, E.A., de Pillisa, L.G.: A comparison and catalog of intrinsic tumor growth models. Bull. Math. Biol. **76**(8), 2010–2024 (2014)
18. Monica Jenefer, B., Dr. Cyrilraj, V.: Multi-class abnormal breast tissue segmentation using texture features and analyzing the growth factor using power law. BBRA J. **12**(1) (2015)
19. Clark, N.: Three techiques for implementing digital fractal analysis of particle shape. Powder Technology **46**(45) (1986)
20. Antonio Napolitano, Sara Ungania and Vittorio Cannata.: Fractal Dimension Estimation Methods for Biomedical Images, doi:10.5772/48760

Idea of Emergency Alarm as a Smart Mobile Application with Prediction

Martin Tmej, Jan Dvorak, Ondrej Krejcar, Ali Selamat, Reza Mashinchi and Kamil Kuca

Abstract Mobile telephone is ubiquitous; people carry it around everywhere and all the time. Therefore, there is an effort to get as much information on the mobile phone as possible. It can be weather, latest news, sports and others. This article deals with a problem of obtaining information about unscheduled events and its notification on the mobile phone. An application was created and it acquires information about weather, in particular about the temperature and air humidity, and based on that it ascertains the possibility of rime or black ice being created. The application shares this information with the user through their set alarm clock. If the weather condition in user's locality indicates formation of frost or ice the alarm clock is moved to 15 min earlier. That enables the user to defrost the car or leave house sooner to manage having everything done on time.

Keywords Alarm clock · Notifications · Unscheduled events · Frost · Windows phone

M. Tmej · J. Dvorak · O. Krejcar (✉) · A. Selamat · K. Kuca
Faculty of Informatics and Management, Center for Basic and Applied Research, University of Hradec Kralove, Rokitanskeho 62, 500 03 Hradec Kralove, Czech Republic
e-mail: ondrej@krejcar.org

M. Tmej
e-mail: martin.tmej@uhk.cz

J. Dvorak
e-mail: jan.dvorak2@uhk.cz

A. Selamat
e-mail: aselamat@utm.my

K. Kuca
e-mail: kamil.kuca@uhk.cz

A. Selamat · R. Mashinchi
Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Baharu, Johor, Malaysia
e-mail: r_mashinchi@yahoo.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_97

1 Introduction

Nowadays, in the 21st century, no one can imagine having to give up modern elements of civilization. Over the last few years, a diverse array of electronic devices became part of our lives. Many of us cannot even imagine our lives without them. A representative example of these is a mobile phone that serves not only as a portable phone but also as a calendar, alarm clock, photo camera, internet browser, music player and many others [1]. Owing to the fact that we carry it around all the time, we also demand more functions. One of these functions is a notification of some events. At first, it sufficed for waking up, but later it added notifications for meetings, holidays, birthdays and another infinite number of events that we need to be notified of [2]. Hereby, we get to the problem that this article addresses. That is, how to notify a user of an unscheduled event, for example ice covered road, floods, fire, etc.

Within this article an application on the Windows Phone 8 platform will be created and for that reason methods of notifications that appear on this platform will be described.

The way to inform users about unscheduled events was chosen to be as a notification on the mobile phone. There are three ways to alert the user that the application is informing them about something [3]. The first way is Live Tiles, which are placed on the start screen. The next are Secondary Tiles and the last way is Notification. And that is what this article will be dealing with.

The Live Tiles are supposed to attract the user to open an application. The tiles themselves are the gates to the applications; they are to facilitate opening of the desired application easily. If the tile is “live” it can display additional information and that way it can notify the user of new events in the application.

The Secondary Tiles allow the user to check a function in an application and then to move it to the start screen in the form of a Secondary Tile. The new tile then serves as a link to the chosen function in the application. As a typical example we can take a weather application where the user can create a Secondary Tile for every city and that way he can follow the weather in these cities without having to search for them again in the application. Thanks to this function each user can personalize their application.

The last shortcut is a toast notification. This notification is a great way to inform the user that something is happening. This notification appears in the top part of the display, that way it covers the actual application and it makes a sound. This way the user is securely informed, it is not easy to miss it.

The notifications can be divided also in a different way, according to their starting mechanism [4].

- Scheduled Notification—notification that is switched on at certain time and certain date according to the way the user plans it.
- Period Notification—regular notification that repeats itself constantly. This notification is widely used for Live Tiles where as an example there is an update every hour on the unread emails.

- Push Notification—it allows sending a notification from the external server anytime if the phone is switched on. This type of notification is being used for subscriptions to the news from different servers. For example, we want to be notified when an anticipated movie comes out, or an article we waited for was published. This notification is sent from server and that way it does not require the application to be regularly connected as it is with Periodic notification.

A sample application will be combining first (Scheduled) and last (Push) way of activating the notification. The next section will deal with the way to put in these pieces of information into the mobile phone.

2 Problem Definition

As it was described in the previous section, the topic of this article is informing user of unscheduled events using notifications on mobile phone. Information about events should be acquired from a freely available source. That ensures low expenses on the application and that way its fast distribution can be expected.

The application should be using internet connection for acquiring information from web services, ideally using Wi-Fi. That would ensure another function of the application which is user's localization. If the location if known we can search for information that are happening around him. In the case that there is no Wi-Fi connection around, the user can connect also through mobile internet and thus the location is determined by GPS. In order to save battery of the mobile phone it would be convenient if the user's location would be searched only once and that would be after starting the application. A similar rule should be valid also for internet connection. The unscheduled events control should be optional and chosen by the user and that would be every hour. Shorter time period could deplete the batter very fast.

A similar problem was dealt at the Columbia University in the year 2001 in the USA [5]. The government was providing an effective system of warning the population about unscheduled events. The university looked into the idea of how this system of warning could be made accessible through the mobile phones. They used the SIP (Session Initiation Protocol), which is a signal protocol of application layer that is used for Internet calls (VOIP), instant messaging (chatting online with friends) and for notifications about different events. Using this protocol, users can send requirements of notifications to the server. In case of some event, the server sends a notification and the user can be warned. This system includes two roles—the receiver and the notifier. The receiver represents a user that receives warning reports about unscheduled events. The notifier is a server that sends notifications based on the request of the receiver.

In the article CAP-ONES: an emergency notification system for all [6] the authors are dealing with proposal of a system based on the knowledge base for notifications about danger. For notifications on different devices they use EMS (Emergency Management Systems) [7] that communicates with the user. This system is based on two characteristics. The first one is the knowledge base about the availability, devices, health disabilities, media and unscheduled events so the warning can be adjusted according to many parameters. The second characteristic is protocol CAP (Common Alerting Protocol) [8] that mediates the communication between this and another already existing systems.

Another option how to notify users about unscheduled events is described in the article Personalized Alert Notifications and Evacuation Routes in Indoor Environments [9] that follows the solution described in the previous article. Here the architecture including three elements is used compared to the previous one where only two were proposed. The first one is so called CAP-ONES which has a task to alert to unscheduled events. The next part is NERES that defines the evacuation plan and creates evacuation route for the user. The last element is interface iNERES that gets the message NERES and shows the evacuation plan on the phone.

However, these described solutions do not depict our problem entirely because the basis of our problem is gaining information about unscheduled events from a free source, and then simply and clearly display this information to the user. The solutions mentioned above are focused on the way to use an evacuation route to the user in case of danger or to describe which way should he act in these situations. Our aim is to notify the user about these events. It should be a warning sent to the user. All this information is obviously accessible on the internet through web browsers [10, 11] but the advantage of mobile phones is that we have them with us all the time and that way the user can be informed immediately in the very moment when an unscheduled event originates. The next section will describe the new solution with more details.

3 New Solution

In the previous section we defined a problem that will be investigated and described in detail here.

Our aim is to create an application for mobile phone that will receive information about unscheduled events from some free and accessible source and will show this information to the user in form of a notification.

Nowadays, there is a whole range of OS on the market but the main representatives are only three. Android with the market share 84.4 %, iOS (11.7 %) and Windows Phone 8 (2.9 %) [12]. The rest of the operating systems are only a modification of Android or they have just a minimal market share.

Windows Phone 8, further only as Windows Phone, was chosen as the platform for the mobile phone application. Windows Phone is a mobile operating system developed by Microsoft. It replaced its predecessor Windows Mobile in February 2010. Windows Phone is used in different phones from the companies HTC, Dell, Samsung and LG. In the beginning of the year 2011 it was announced that the system Windows Phone would be used as a principal operating system in the Nokia Phones.

The applications are developed using the platform Silverlight [13] and XNA framework [14]. Both these technologies are built on .Net Framework and the programming can be done using the language C#, Visual Basic .NET and F#. As a programming environment, the most suitable is Visual Studio, which is in the basic version for free. The only requirement is using OS Windows Vista, Windows 7, 8, 8.1. The place where it is possible to download or buy applications for Windows Phone is called Windows Store. If we want to publish our own application it is necessary to sign up as a developer on the Microsoft website and pay a yearly fee 99 dollars. For students there is a free registration within the programme DreamSpark. The finished application is to be sent for a certification process. The application is verified by Microsoft there and if it meets the requirements for publishing it is certified, signed and can be published on the Windows Store. The requirements for publishing an application are quite strict. They control both contents (use of abusive language, trademarks, etc.) and technical part (sources, language localization, etc.).

This way the technology was presented and now onto the description of the problem, obtaining information and informing about unscheduled events. It was quite difficult to find a freely accessible source of information about unscheduled events. Most of them are paid and they are not well updated for the Czech Republic. As sources of information two sources were chosen. The first one Crisis [15], provides information about all unscheduled events, the only problem is that there are very few data for the Czech Republic. On average it publishes one event in three months. The second source is OpenWeatherMap [16], which provides fast and clearly arranged information about the weather all around the world. This information is transmitted using simple API and for the client identification it is enough to have a key that everyone gets after signing up.

The application will consist of two parts—an agent and the application, which will be very similar to a classic organizer or an alarm clock. The user will set a time and whether he wants to be notified in case of an unscheduled event. After having set the preferred functions the second part of the application—Agent, comes to work (Fig. 1). Agent will be gathering and evaluating information in the background from both previously mentioned sources and based on this information it will inform the user. The information gathered by agent is always from the user's location that is determined by GPS or Wi-Fi if it's available.

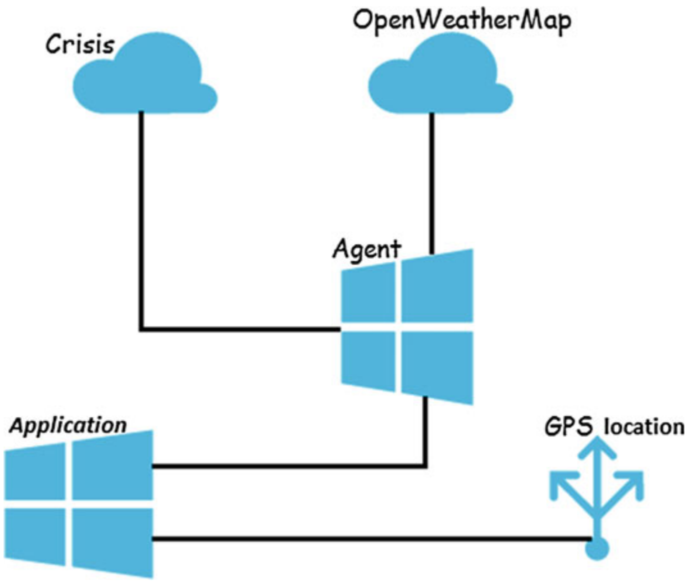


Fig. 1 Scheme of proposed application

4 Implementation

Previous section deals with the new solution that we define. In this section we will analyse the implementation of this solution.

Implementation is in a form of a mobile application on the platform Windows Phone 8. The application will have a similar function such as organiser or alarm clock with the difference that it will be tracking the outside temperature at night or at the time that the user chooses and based on the temperature and humidity of air it will evaluate if there is a possibility of black ice being created. According to this information it will adjust time of the alarm. That way it will save user's time and worry about having to check the weather before going to bed and having to change the alarm clock in order to have time to defrost the car.

The user can simply choose a time to wake up and whether he wants to have the function of the control of unscheduled events available (in this case concerning the creation of black ice). The application then automatically plans an agent that will monitor the temperature and air humidity. If the conditions correspond with the creation of black ice 1 h before the planned activity the agent will automatically move the time of the activity to 20 min earlier.

The whole scheme of the created application is shown in Fig. 2. The application contains two modules Agent and CoreApp. The agent is independent of the application and it cares about the control and downloading data from the web service. The application alone contains a user environment, logic for notifying the user and a component that tracks the location of the user.

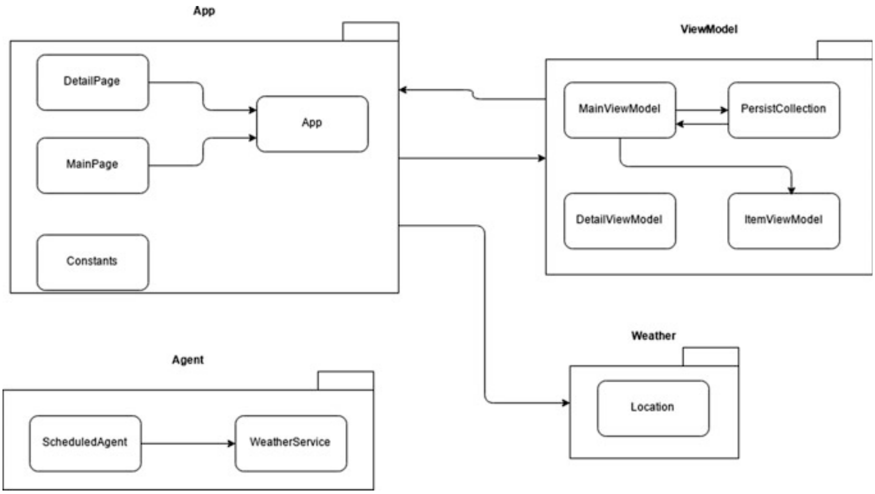


Fig. 2 Scheme of implemented application

As the application connects to the web service, the first step must be signing up at OpenWeatherMap [16]. The registration requires filling up an easy form and in the end a necessary key is obtained. This key will serve as an identification to connect to the services. The second section described that the application would also connect to the service Crisis, however this service was removed from the application because it does not provide good updated data. For the Czech Republic country it contains only one unscheduled event in three years with the only location Prague. For that reason it is unsuitable for our application and we only use the service OpenWeatherMap.

The application is created using the structure MVVM (Model View ViewModel) [17]. Model-View-ViewModel is a designing pattern of WPF [18] application. It offers a solution how to separate the logic of application from user interface. Everything then is much clearer and more organized and possible changes are easier. MVVM separates data, the state of the application and user interface and that way it ensures reusability.

The user interface of the application was designed to be as simple as it can be. The initial page that the user sees after starting the application contains a list of set events, whether the event is active and whether it controls the unscheduled events. In the bottom part of the screen there is a button for adding a new event. On the screen for adding new events we can find a box for inserting the name of an event, the choice of time of the event and a button for turning on or off the control of unscheduled events. As a confirmation of a created event an agent is created and it will be watching and downloading information from the web service. On the initial screen for each event it is possible to activate or deactivate planned event or else a notification for unscheduled events. The sample application should work right in every device that has an operating system Windows Phone 8. It is because

Microsoft determined standards for devices where this operating system is installed. Each one has to meet the requirements: min. 512 MB RAM, GPS, NFC, Wi-Fi and dual core processor. That brings a considerable advantage for developers compared to Android that has a whole array of devices and producing applications so that they would display correctly on all devices is a big problem. During the development of the application a problem in a communication between both modules, Agent and CoreApp, came up. The application uses its isolated memory for saving data. However, a problem can occur if the application runs on more threads and it is necessary that the both threads see the same thing in the mentioned memory. In our case a problem arose when the agent saved information about weather to the memory and the application did not know about this information. The solution of this problem was synchronization of the process of saving data to the memory. That was achieved using class Mutex.

This section contained a description of specific implementation of sample application. The next section is dedicated to the testing of the application.

5 Testing of Developed Application

Implemented application has to be tested. It should be seen in the test if the solution proposed by us was in any way better. The application can be tested based on prediction of unscheduled event, in our case on the creation of black ice. The next test focuses on the success rate of user's localization.

5.1 The Test for Black Ice Prediction

The test was performed with 10 measurements in three different locations with three different settings. As it can be seen in Table 1 the best combination that ensures the biggest frequency of black ice occurrence is within the temperature range 0/−4 °C and air humidity 70 %.

Location can also play a role in the test because in the villages (Opatovice nad Labem—City No. 3) the temperature is lower in general than in the cities (Hradec Kralove—City No. 2, Lanskroun—City No. 1) where more people live, more cars drive and more smog is in the air.

5.2 The Test for Speed and Success Rate of Localization

The second test will test the success rate and the speed of localization of the user. As the application has two options to localize the user, using Wi-Fi or GPS, the test will monitor the speed and the accuracy of found location.

Table 1 The test of black ice prediction

Location	Temperature range (°C)	Air humidity (%)	Frequency of black ice occurrence
City No. 1	0/−4	70	8
City No. 1	−1/−4	75	7
City No. 1	0/−5	80	8
City No. 2	0/−4	70	8
City No. 2	−1/−4	75	6
City No. 2	0/−5	80	9
City No. 3	0/−4	70	10
City No. 3	−1/−4	75	9
City No. 3	0/−5	80	9

The test was performed ten times and it was proved that although GPS is slower it provides 100 % success rate. That is important information in the case of villages. As it can be seen in Table 2 the success rate of Wi-Fi in City No. 3 is only 60 % and that is because the village is close to City No. 2 (3 km). The Wi-Fi gets the location based on IP address and available Wi-Fi networks. In case of the requirement for localization, the data is collected from the surrounding Wi-Fi networks, which means their names, MAC addresses of routers and the strength of the signal. This data is sent to the localization service (Google) that compares them with its Wi-Fi routers database with the known location. Based on that the presumable location is calculated and it is sent back [19]. During the calculation the IP address from which the request was sent is also used in Fig. 3.

This method is not accurate precisely because of the impossibility of getting the exact location and getting only an estimate of the location. In the cities this method of localization works correctly thanks to their big area. The estimate of the user location will always be in that city compared to the village where the area is small. Based on this test I would recommend having GPS always turned on whether Wi-Fi available or not. The application prefers GPS location before Wi-Fi, mostly because of the accuracy. The next criterion that was measured in the test was the speed of obtaining the location. In this criterion Wi-Fi definitely won. For one thing it does not need to look for the position of the satellite and for another it only estimates the position.

Table 2 The test of speed and success rate of localization

Location	Type of connection	Speed (s)	Success rate (%)
City No. 1	Wi-Fi	0.5	100
City No. 1	GPS	1.2	100
City No. 2	Wi-Fi	0.5	100
City No. 2	GPS	1.0	100
City No. 2	Wi-Fi	0.75	60
City No. 3	GPS	2	100

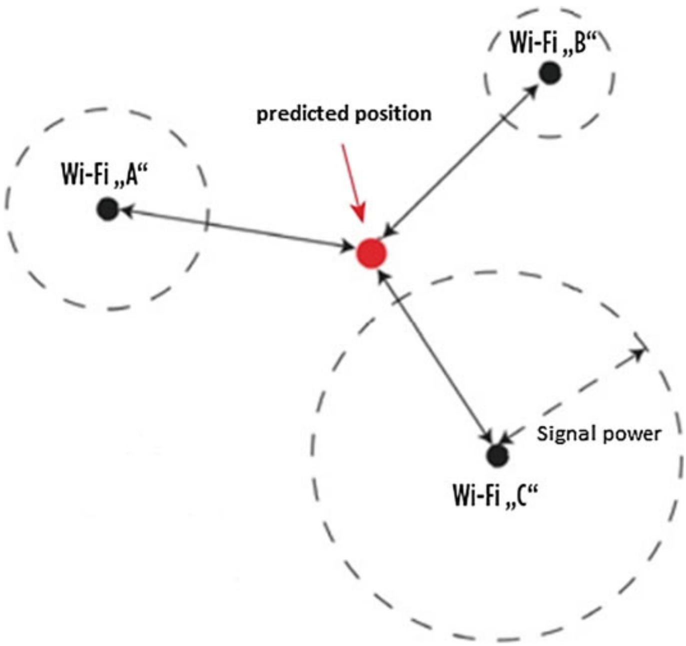


Fig. 3 The scheme of the probable location detection using Wi-Fi [19]

From the results of both tests we can deduce the ideal settings for the application. That is the temperature range 0/−4 °C, air humidity 70 % and switched on GPS [12, 16]. The application set this way should ensure a high probability of successful prediction of black ice. In comparison with other solutions that are described in section number two, this application may not provide the evacuation route but it can reliably notify the user about an unscheduled event in the form of black ice. Our aim was also to find some free source that would provide information about unscheduled events. This aim was not accomplished as the only possible representative was Crisis and it is very rarely updated.

6 Conclusions

In the beginning the aim for creating an application that would inform the user about unscheduled events using the mobile phone was defined. The source of information about the events has to be for free so that a zero price of application would be accomplished. This aim was fulfilled for our territory only partially. The application is working and it is reliable but a source of information that would be for free and would provide updated data about unscheduled events was found only partially. OpenWeatherMap was used as a representative of information about the weather. The source Crisis as a representative of other unscheduled events is unfortunately for the Czech Republic very poorly updated. However, that does not mean that in other bigger countries this source of information could work as reliable as in our case the source OpenWeatherMap. Developed solution can be equipped in various areas of research from biomedicine [20, 21] to measurement and control applications [22–31].

Acknowledgement This work and the contribution were supported by project "SP-103-2015—Smart Solutions for Ubiquitous Computing Environments" Faculty of Informatics and Management, University of Hradec Kralove, Czech Republic. The Universiti Teknologi Malaysia (UTM) and Ministry of Education Malaysia under Research University grants 02G31, and 4F550 are hereby acknowledged for some of the facilities that were utilized during the course of this research work. Last but not least, we acknowledge the technical language assistance provided by Jirina Cancikova.

References

1. Abowd, G.D., Brumitt, B., Shafer, S.A.: Ubicomp 2001: Ubiquitous computing. In: International Conference, Atlanta, Georgia, USA, Sept 30–Oct 2 2001, c2001, xiii, 372 p. Springer, New York. ISBN 35-404-2614-0
2. Kitchens, F.L., Mcguire, J.M., Sharma, S.K.: Emergency preparation and mobile notification through cluster computing. Int. J. Mobile Commun. **4**(1), 82–96 (2006)
3. Liberty, J., Japikse, P., Galloway, J.: Notifications. Pro Windows 8.1 Development with XAML and C#, s. 209. Apress, Berkeley, CA (2014)
4. Lee, H., Chuvyrov, E.: Push Notifications. Beginning Windows Phone App Development, p. 405. Apress, Berkeley, CA (2012)
5. Ishii, N., Lee, R.: 5th IEEE/ACIS International Conference on Computer and Information Science: (ICIS 2006), Honolulu, Hawaii, 10–12 July 2006. Los Alamitos, USA. ISBN 9780769526133
6. Malizia, A., Acuna, P., Onorati, T., Diaz, P., Aedo, I.: CAP-ONES: an emergency notification system for all. Int. J. Emerg. Manage. **6**(3/4), 302 (2009). doi:10.1504/IJEM.2009.031568
7. Hernandez, J.Z., Serrano, J.M.: Knowledge-based models for emergency management systems. Expert Syst. Appl. **20**(2), 173–186 (2001). doi:10.1016/S0957-4174(00)00057-9
8. Saint-andre, P., Fletcher, B.: Common Alerting Protocol (CAP) Over XMPP [online]. 1999–2014 [cit. 2014-12-19]. Available from http://xmpp.org/extensions/xep-0127.html#appendix-legal
9. Aedo, I., Yu, S., Díaz, P., Acuña, P., Onorati, T.: Personalized alert notifications and evacuation routes in indoor environments. Sensors. **12**(12), 7804–7827 (2012). doi:10.3390/s120607804
10. Maresova, P.: Decision making criteria for cloud computing deployment. In: Proceedings of Advanced Multimedia and Ubiquitous Engineering. Future Information Technology, Hanoi, Vietnam. Lecture Notes in Electrical Engineering, vol. 352, pp. 93–98, May 18–19, 2015. Springer, Berlin. ISBN 978-3-662-47486-0
11. Maresova, P., Klimova, B.: Investment evaluation of cloud computing in the European business sector. Appl. Econ. **47**(36), 3907–3920 (2015)
12. Behan, M., Krejcar, O.: Modern smart device-based concept of sensoric networks. EURASIP J. Wirel. Commun. Netw. **1**(155) (2013). doi:10.1186/1687-1499-2013-155 (Received 9 Oct 2012; Accepted 15 May 2013; Published 6 June 2013. ISSN 1687-1499)

13. Behan, M., Krejcar, O.: Adaptive graphical user interface solution for modern user devices. In: 4th Asian Conference on Intelligent Information and Database Systems, ACIIDS 2012, Kaohsiung, Taiwan. Lecture Notes in Computer Science, LNCS, vol. 6592, pp. 411–420, 19–21 Mar 2012
14. Krejcar, O.: Threading possibilities of smart devices platforms for future user adaptive systems. In: 4th Asian Conference on Intelligent Information and Database Systems, ACIIDS 2012, Kaohsiung, Taiwan. Lecture Notes in Computer Science, LNCS, vol. 7197, pp. 458–467, 19–21 Mar 2012
15. Penhaker, M., Kasik, V., Stankus, M., Kijonka, J.: User adaptive system for data management in home care maintenance systems. In: Yeungnam, U. (ed.) LNAI. 3rd International Conference on Intelligent Information and Database Systems, ACIIDS 2011 Daegu, vol. 6592, pp. 492–501 (2011)
16. Mlynka, M., Brida, P., Machaj, J.: Modular positioning system for intelligent transport. In: 5th International Conference on Computational Collective Intelligence Technologies and Applications, Craiova, Romania. Studies in Computational Intelligence 513, vol. 513, pp. 115–124 (2014)
17. Penhaker, M., Darebnikova, M., Cerny, M. (2011) Sensor network for measurement and analysis on medical devices quality control. In: Springer (ed.) CCIS. 1st International Conference on e-Technologies and Networks for Development, ICeND 2011, Dar Es Salaam, vol. 171, pp. 182–196 (2011)
18. Gantulga, E., Krejcar, O.: Smart access to big data storage—Android multi-language offline dictionary application. In: 4th International Conference on Computational Collective Intelligence Technologies and Applications 2012, ICCCI 2012, Ho Chi Minh, Vietnam. Lecture Notes in Computer Science, LNCS, vol. 7653, pp. 375–284, 28–30 Nov 2012
19. Penhaker, M., Krejcar, O., Kasik, V., Snášel, V.: Cloud computing environments for biomedical data services. In: 13th International Conference on Intelligent Data Engineering and Automated Learning, IDEAL 2012, Natal. LNCS, vol. 7435, pp. 336–343 (2012)
20. Cerny, M., Penhaker, M.: Wireless body sensor network in health maintenance systems. Elektronika Ir Elektrotechnika, 113–116 (2011). doi:10.5755/j01.eee.115.9.762, ISSN: 1392-1215
21. Penhaker, M., Darebnikova, M., Cerny, M.: Sensor network for measurement and analysis on medical devices quality control. Commun. Comput. Inf. Sci. **171**, 182–196 (2011)
22. Vanus, J., Novak, T., Koziorek, J., Konecny, J., Hrbac, R.: The proposal model of energy savings of lighting systems in the smart home care. In: Proceedings of IFAC, vol. 12(PART 1), pp. 411–415 (2013)
23. Vanus, J., Koziorek, J., Hercik, R.: Design of a smart building control with view to the senior citizens' needs. In: Proceedings of IFAC (IFAC-PapersOnline) vol. 12(PART 1), pp. 422–427. ISSN: 14746670. ISBN: 9783902823533 (2013)
24. Jancikova, Z., Kostial, P., Bakosova, D., Ruziak, I., Frydrysek, K., Valicek, J., Farkasova, M., Puchky, R.: The study of electrical transport in rubber blends filled by single wall carbon nanotubes. J. Nano Res. **21**(16), 1–6 (2013). ISSN: 16625250
25. Jonsta, Z., Kostial, P., Ruziak, I., Jonsta, P. Jurciova, J., Jancikova, Z., David, J., Kopal, I.: Advanced fillers enhancing thermal and mechanical properties of rubber blends. J. Nano Res. **13**, 27–32 (2011). ISSN: 1662-5250
26. Machacek, Z., Slaby, R., Hercik, R., Koziorek, J.: Advanced system for consumption meters with recognition of video camera signal. Elektronika Ir Elektrotechnika. **18**(10), 57–60 (2012). ISSN: 1392-1215
27. David, J. Svec, P., Frischer, R., Stranavova, M.: Usage of RFID wireless identification technology to support decision making in steel works. In: 21st international conference on metallurgy and materials, pp. 1734–1738 (2012). ISBN: 978-80-87294-31-4
28. Pies, M., Hajovsky, R., Latocha, M., Ozana, S.: Radio telemetry unit for online monitoring system at mining dumps. Appl. Mech. Mater. **548–549**, 736–743 (2014)

29. Ozana S., Pies M., Horacek O., Zidek J.: Concept of automated transportation center and its PIL model. In: World Congress on Engineering and Computer Science (WCECS'13)—International Conference on Modeling, Simulation and Control (ICMSC'13), vol. 2, pp. 992–996 (2013). ISBN 978-988-19253-1-2
30. Horalek, J., Matyska, J., Sobeslav, V.: Communication protocols in substation automation and IEC 61850 based proposal. In: Proceedings of CINTI 2013—14th IEEE International Symposium on Computational Intelligence and Informatics, art. no. 6705214, pp. 321–326 (2013)
31. Cimler, R., Matyska, J., Balík, L., Horalek, J., Sobeslav, V.: Security issues of mobile application using cloud computing. Adv. Intell. Syst. Comput. **334**, 347–357 (2015)

Frequency Modulated Continuous Wave Radar Modeling for Landslide Detection in Malaysia

Nor Fatin Muhd Ariffin, Farah Nadia Mohd Isa, Ahmad Fadzil Ismail and Mohammad Kamrul Hasan

Abstract Landslide is one of natural hazards that is frequently happen after continuous heavy rainfall especially during monsoon season. It effects serious damages on property and also treat human being's live. The objective of this paper is to design and evaluate the performance of a Frequency Modulated Continuous Wave (FMCW) radar for detecting of slow moving landslide movements in Malaysia. The performance of the proposed radar design is tested by using Agilent Advanced Design System (ADS) software. It operates at 2.4 GHz carrier frequency with a resolution range of 0.75 m. Its signal power increases around 38 dBm after de-ramping process. Hence, it increases the sensitivity of the radar to detect the landslide occurrences. The radar has capability to detect landslide movements of 20 m/s speed.

Keywords Landslide · Linear FMCW · Radar · RF de-ramping technique

1 Introduction

Landslide is defined as displacement of earth material (i.e. rock, soil and regolith) from gravity [1]. Over the past few decades, the rising demand for landslide study has given impressive pressure on development of new technologies and methods to

N.F.M. Ariffin (✉) · F.N.M. Isa · A.F. Ismail · M.K. Hasan
Department of Electrical and Computer Engineering, International Islamic University Malaysia, 53100 Kuala Lumpur, Malaysia
e-mail: fatin.nornisa@gmail.com

F.N.M. Isa
e-mail: farahn@iium.edu.my

A.F. Ismail
e-mail: af_ismail@iium.edu.my

M.K. Hasan
e-mail: hasankamrul@ieee.org

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_98

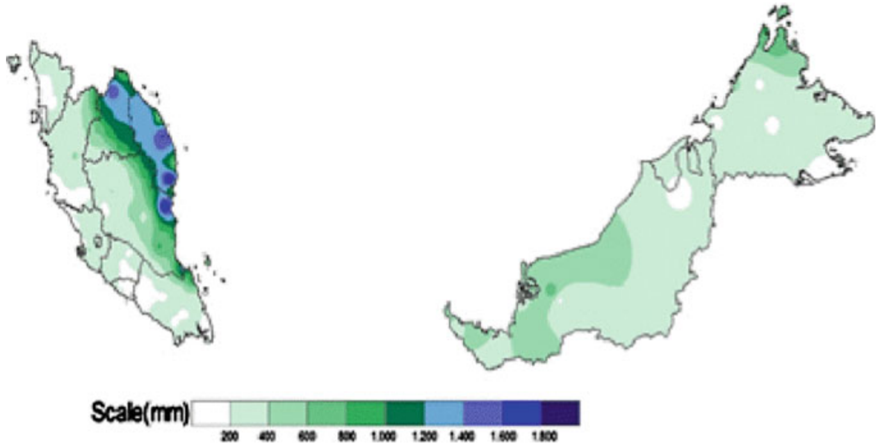


Fig. 1 Malaysia rainfall rate on December 2014 [2]

enhance a better understanding of the studies. Accurate detecting and monitoring of landslide occurrences facilitate authorities to prevent serious damage of infrastructures and loss of lives. In December 2014, East Coast of Malaysia (i.e. Kelantan, Pahang and Terengganu) was hit by severe flood due to excessive rainfall rate, highest recorded by Malaysian Meteorological Department is at 1600 mm [2] as shown in Fig. 1. Several landslides happened during the monsoon season as ground underneath is softened and changed by the excessive amount of rain especially hillside and steep slope area.

There are many different techniques are introduced for landslide studies such as inclinometers, extensometers [3], Global Positioning System (GPS) [4], Synthetic Aperture Radar (SAR) [5, 6], airborne laser scanning [7] and Frequency Modulated Continuous Wave (FMCW) [8, 9]. Each of them has its own advantages and disadvantages in certain categories, yet this paper is only focus on the potentiality of the technique to design and build a low cost tool for the purpose of detecting slow moving landslide in a long range distance.

Landslide study has been done in Japan by using Ground Based Synthetic Aperture Radar (GB-SAR), operated at Ku-band (17.2 GHz) frequency. It moved and scanned repeatedly along 2 m horizontal rail and collected data at every 5 mm [10]. The radar's monitoring was done remotely or not in physical contact with landslide's area as compared to conventional technique such as inclinometers and extensometers. However, the radar has drawback in terms of cost effectiveness as it has complex modulation and circuit design [11] since it used pulse compression technique. Moreover, temporal de-correlation problem raised because the sensitivity or performance of the GB-SAR is easily distracted by the ground vegetation waving [7] which may lead to error data collection since the radar detection performance is closely depends on target's velocity.

Airborne laser scanning is a combination of several techniques such as GPS, SAR and laser scanning methods which is capable to cover and wide landslide area in long range distance. It also provides very high resolution of 3-dimensional (3D) landslide imaging, gives accurate measurements of landslide's characteristics such as velocity, soil moisture, weathering profiles, and etc. [7]. However, cost demanding for the airborne laser scanning to be carried out is very high as it requires flight and very sophisticated instruments. It also needs many experts to handle the project, hence it takes long time to get the landslide studies be done.

Frequency Modulated Continuous Wave (FMCW) method uses continuous waveform as a signal source which is generated at transmitter part. Frequency of the waveforms is varies over time. FMCW radar is popularly used for geophysical studies such as; snow avalanche mapping image [12], landslide displacement measurement [9, 13] and snow thickness measurement [14]. FMCW radar has superior advantages in terms of resolution range, power consumption, simple modulation and circuitry design, less delay of updates data and detect slow moving target [11, 15, 16]. Therefore, FMCW technique is chosen to be used by the radar system in this paper as the technique has simplicity in modulation and design's circuitry which facilitates a portable and low cost FMCW radar system. Unlike GB-SAR, the target detection by FMCW radar is mainly rely upon target size towards resolution cell instead of target's velocity [15]. Hence, the detection of landslide movement can be distinguished from other clutters.

In this paper, radar system is designed and simulated by using Agilent Advanced Design System (ADS) software version 2014. To design and evaluate the performance of the model ADS simulation process is very suitable since it provides essential parameters and specifications to enable hardware radar works well during real experiment. Simulation system is a worthy aid during development process as it capable to minimize the cost and time consumptions.

This paper is organized as: Sect. 2 describes the proposed design of FMCW radar, the evaluation of performance the designed simulation results are depicted and explained in Sect. 3 and finally concluded the paper in Sect. 4.

2 Proposed FMCW Radar Design and Its Specifications

The FMCW radar architecture consists of four main parts which are transmitter, receiver, target and signal processing as shown in Fig. 2. They are connected to each other in a dataflow simulator called Agilent Advanced Design System (ADS) Ptolemy. The software is most relevant for microwave and Radio Frequency (RF) design applications [17]. Details of the FMCW design system is explained in the sub-section of Sect. 2 below.

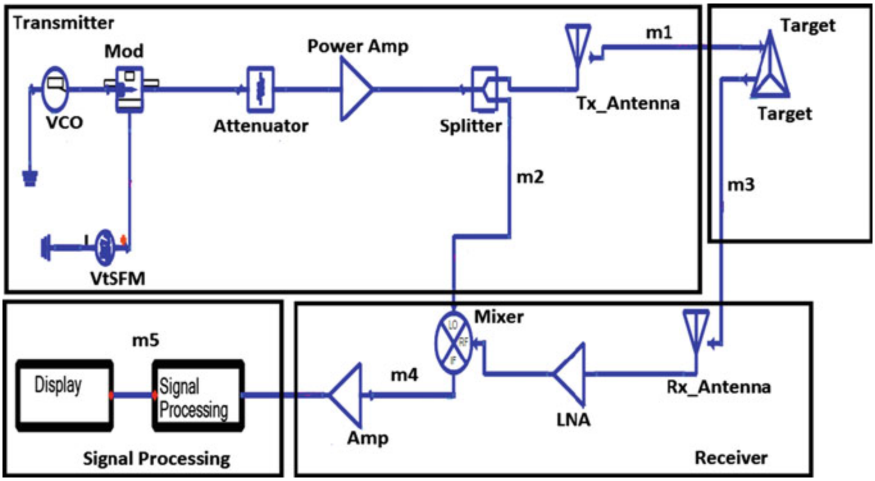


Fig. 2 FMCW radar design block diagram

2.1 *Operating Frequency and Power*

The FMCW radar system operates at 2.4 GHz carrier frequency with approximately 38 dBm transmit power. During operating frequency selection process, there are some parameters need to be considered. One of them is minimum detectable signal (S_{min}) [18] which is required by the radar to get maximum target detection of 1000 m. Theoretically, radar frequency can be used from 300 MHz up to 110 GHz, however very high frequency (i.e. 10 GHz and above) does not penetrate the ground surface well [19]. Modeling and purchasing component processes for real test also would be complex and costly. In addition, the maximum range of target is inversely proportional to the frequency as shown in Eq. (1). Therefore, 2.4 GHz is selected to meet the requirement of 1000 m maximum range with acceptable value of S_{min}. The frequency is categorized in S-band which is known as having impressive advantage when dealing with attenuation constant and atmospheric attenuation as the attenuation increases whenever frequency increases [19].

$$R_{max}^4 = \frac{P_t G^2 \sigma \lambda}{(4\pi)^3 S_{min}} \tag{1}$$

2.2 *FMCW Waveform Generation and Transmitter*

Voltage signal of Frequency Modulated (VtSFM) device is used as waveform generator in time domain. The amplitude of the signal is assumed as 1 V and 0.5

modulation index. In addition, Voltage Control Oscillator (VCO) based system is applied to provide a signal pattern which supports wide FM chirp bandwidth (i.e. GHz frequency) [19]. Next, the modulated signal is amplified and split into two ways. One of them is split towards 20 dB gain transmitter antenna and one more towards a mixer in receiver part, named as reference signal which would be used during de-ramping process. Transmit signal frequency varies from 2.0 GHz until 3.0 GHz over time with 100 MHz step. Attenuator of 3 dB is added after VCO to reduce unnecessary amount of transmit signal during next stages especially during signal amplifying at power amplifier.

2.3 Target Model

It is essential to apply target model parameters as real target's parameters since it gives big impact on the performance of the radar [i.e. backscatter coefficient and Signal to Noise Ratio (SNR)]. Table 1 shows the characteristics of soil. Worst case scenario is selected during simulation process which is wet soil condition. The target model moves at 20 m/s speed with radar cross section of 100 m^2. Dielectric constant (ε_r) is closely depends on dielectric constant of water (ε'') and soil (ε') as shown in Eq. (2).

$$\varepsilon_r = \varepsilon' - j\varepsilon'' \tag{2}$$

2.4 Receiver and Signal Processing Part

Measured return power signal is normally 10–18 times smaller than transmit power signal [20] which leads poor radar performance. It is also decreases when target range increases as shown in Eq. (3). Therefore, Low Noise Amplifier (LNA) device is placed right after the receive signal is collected by receiver antenna to amplify the weak receive signal because it is distorted and attenuated while in the air. Horn antennas with 20 dB gain are used at both transmitter and receiver part. The obstacle also can be solved by mix up the reference signal with the return signal where processing gain is provided to improve the return power and SNR value [12]. Besides, built in Low Pass Filter (LPF) is equipped on mixer to filter out unwanted

Table 1 Soil parameters [19]

Types	Conductivity	Dielectric constant of soil	Dielectric constant of water
Dry soil	0.001 (S/m)	4–7	0.006
Moist soil	0.005 (S/m)	15	0.6
Wet soil	0.02 (S/m)	25–30	32.4

and interference signal before it is collected and analyzed in signal processing part. Frequency difference between transmit and received is used to measure the range and velocity of the target as shown in Eq. (4). Envelope simulator and Harmonic Balance simulator in Agilent Advance Design System is used to analyze complex modulated RF signal in time and frequency domain.

$$P_r = \frac{P_t G^2 \sigma \lambda}{(4\pi)^3 R^4} \tag{3}$$

$$\Delta \mathrm{f} = \alpha \frac{2\mathrm{R}}{\mathrm{c}} = \frac{\mathrm{f_0 2 V_r}}{\mathrm{c}} \tag{4}$$

3 Simulation and Result Analysis

Figure 3 demonstrates power density signal (dBm) at certain frequency (KHz). Representation of each signal at specific part is shown as below:

- m1 is transmit signal
- m2 is reference signal
- m3 is echo signal after transmit signal hits target
- m4 is mixed signal of reference and return signal
- m5 is receive signal at final point which is located at receiver part.

It can be seen in the figure that echo signal power drops drastically from 37.9 to −81 dBm which is certainly gives a negative value of SNR. It is impossible for a radar to detect a long range detection with the negative SNR since noise power is greater than signal power. However, the obstacle is improved by applying de-ramping process as it is capable to work as signal processing gain. The result displays that the echo signal power is improved by approximately 38 dBm after the

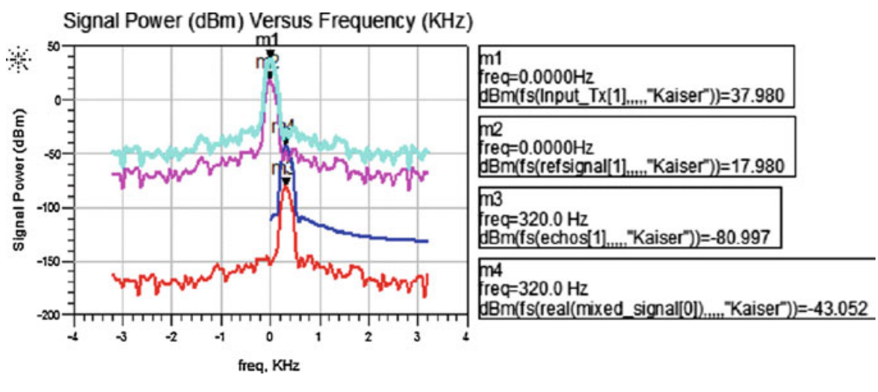


Fig. 3 Signal power at transmitter and receiver parts

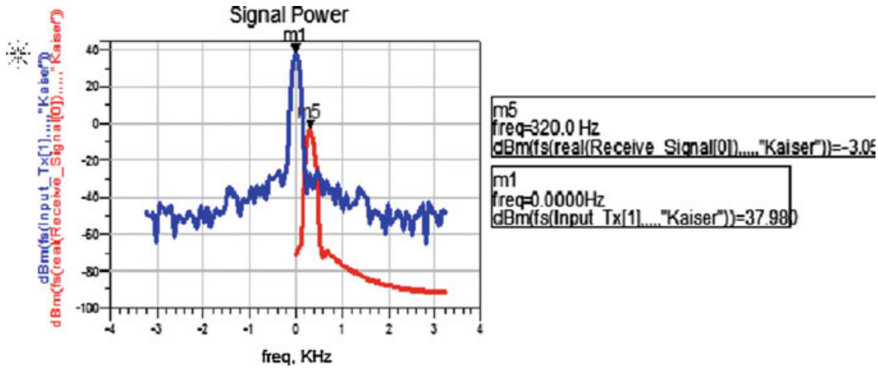


Fig. 4 Shifted frequency

process. It is calculated by subtracting power density of m4 from m3 signal. The process plays important role to increase the sensitivity of the radar especially during detecting slow moving target as it gives very low receive power. Very low speed of target also called blind speed if the movement of the target does not detect by radar system.

The result in Fig. 4 shows that the target frequency shifted is 320 Hz. It is measured by subtracting frequency of m5 from m1 signal. The shifted frequency's information can be used to calculate range and velocity of target by using Eq. (4). The speed of target model is 20 m/s at 1000 m distance from transmitter antenna. The result of shifted frequency from simulation agrees theoretical value from Eq. (4) which is 320 Hz. In addition, it also shows that return signal power, m5 which is measured at last point of receiver part is −3.05 dBm.

Figure 5 represents the value of receive signal power while varying range from 500 to 1500 m with a step of 500 m. Based on Eq. (1), the value of receive power decreases when range of target increases. The value of receive power drops from 8.989 dBm to −3.052 dBm when target's range is increased from 500 m to 1000 m. It drops one sixteenth from original value when the target's range is doubled.

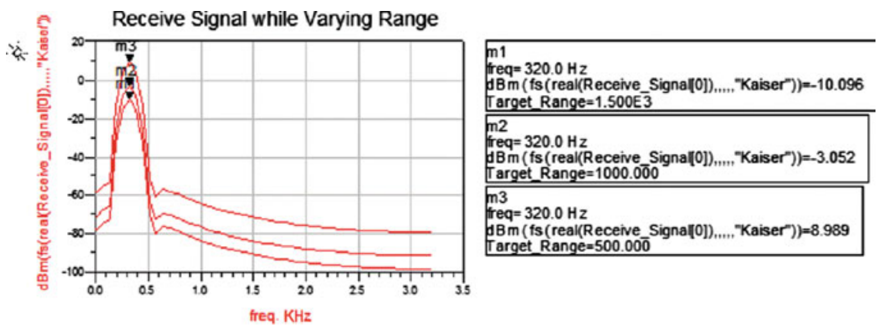


Fig. 5 Return signal power with target range varies

4 Conclusion

A full FMCW radar design for landslide detection has been proposed and simulated by using Agilent ADS software. Simulation is an important process for a development of the radar system in order to meet required specification of the mitigation system.

In this paper, it is shown that poor echo power signal can be improved around 38 dBm by implementing de-ramping process. The radar design is also capable to measure the range and velocity of target by using shifted frequency information. The target model parameters are applied as close as realistic value, so that the radar system is capable to work well during field test. The proposed FMCW radar design capable to detect a slow target with a speed of 20 m/s. It proves that slow moving target can be measured and detected by using low cost FMCW radar.

In future, the FMCW radar design would be modified and improved by implementing better specifications on RF device (i.e. amplifier and filter) and signal processing technique [i.e. Moving Target Indication (MTI)] to be able to detect slower moving target.

Acknowledgment International Islamic University Malaysia (IIUM) and Ministry of Higher Education (MOHE) are highly acknowledged as providing the fund for this project, registered as RAGS 13-027-0090 to be carried out.

References

1. Varnes, C.: Landslide Classification (1996)
2. Monthly Weather Bulletin: Malaysian Meteorological Department (MetMalaysia), 2014. [Online]. Available http://www.met.gov.my/index.php?option=com_content&task=view&id=843&Itemid=158 6. Accessed on 03 Jun 2015
3. Mansour, M.F., Morgenstern, N.R., Martin, C.D.: Expected damage from displacement of slow-moving slides. Landslides **8**(1), 117–131 (2010)
4. Chow, J., Hua, J.J., Wang, W.: Penetrating Radar Study of the Collapse Features in a Nantou Earth Slide, Central Taiwan, vol. 00, no. c, pp. 4143–4146 (2004)
5. Di Martire, D., Iglesias, R., Monells, D., Centolanza, G., Sica, S., Ramondini, M., Pagano, L., Mallorquí, J.J., Calcaterra, D.: Comparison between differential SAR interferometry and ground measurements data in the displacement monitoring of the earth-dam of Conza della Campania (Italy). Remote Sens. Environ. **148**, 58–69 (2014)
6. Monserrat, O., Crosetto, M., Luzi, G.: A review of ground-based SAR interferometry for deformation measurement. ISPRS J. Photogram. Remote Sens. **93**, 40–48 (2014)
7. Razak, K.A.: Airborne Laser Scanning for Forested Landslides Investigation in Temperate and Tropical Environments. Enschede, The Netherlands (2014)
8. Anghel, A., Member, S., Vasile, G., Cacoveanu, R., Ioana, C., Ciochina, S.: Short-range wideband FMCW radar for millimetric displacement measurements. Geosci. Remote Sens. IEEE Trans. **52**(9), 5633–5642 (2014)
9. Chan, Y.K., Ang, C.Y., Koo, V.C., Gan, C.S.: Design and development of a FMCW ground based imaging radar system. PIERS Online **5**(3), 265–268 (2009)

10. Interferometry, G., Takahashi, K., Matsumoto, M., Sato,M.: Monitoring of displacement on a landslide slope by GB-SAR interferometry. In: 2013 Asia-Pacific Conference on Synthetic Aperture Radar, pp. 552–555 (2013)
11. Brooker, G., Hennessey, R., Lobsey, C., Bishop, M., St, R., Building, J., Widzyk-capehart, E.: Seeing through dust and water vapor: millimeter wave radar sensors for mining applications. J. F. Robot. **24**(7), 527–557 (2007)
12. Ash, M.: FMCW phased array radar for imaging snow avalanches (2013)
13. Luo, Y., Song, H., Wang, R., Member, S.: Arc FMCW SAR and applications in ground monitoring. IEEE Trans. Geosci. Remote Sens. **52**(9), 5989–5998 (2014)
14. Galin, N., Worby, A., Markus, T., Leuschen, C., Gogineni, P.: Validation of airborne FMCW radar measurements of snow thickness over sea ice in Antarctica. Geosci. Remote Sens. IEEE Trans. **50**(1), 3–12 (2012)
15. Butler, W., Poitevin, P., Bjomholt, J.: Benefits of wide area intrusion detection systems using FMCW radar. IEEE **07**, 176–182 (2007)
16. Mehdi, G.: Millimeter wave FMCW radar for foreign object debris (FOD) detection at airport runways. In: Proceedings of 2012 9th International Bhurban Conference Applied Science Technology, pp. 407–412. No. 978–1–4577–1929–5 (2012)
17. Malik, V., Malik, T., Nandal, V.: Performance analysis of FMCW sub surface penetrating radar. IJCSMS Int. J. Comput. Sci. Manage. Stud. **12**(June), 124–128 (2012)
18. Charvat, G.L., Fenn, A.J., Perry, B.T.: The MIT IAP radar course : Build a small radar system capable of sensing range, doppler, and synthetic aperture (SAR) imaging. In: IEEE, pp. 138–144. No. 978–1–4673–0658 (2012)
19. Islam, M., Afzal, M.U., Ahmad, M., Taqueer, T.: Simulation and modeling of ground penetrating RADARs. In: IEEE (2012)
20. Ghose, D.: Continuous wave and frequency modulated (Chapter 3). In: Navigation, Guidance and Control, pp. 30–43 (2012)

A Novel FPGA-Based Multi-Channel Multi-Interface Wireless Node: Implementation and Preliminary Test

Witoon Jindamaneepon, Banjerd Rattanalert, Kiattisak Sengchuai, Apidet Booranawong, Hiroshi Saito and Nattha Jindapetch

Abstract The implementation and the preliminary test of a novel FPGA-based Multi-Channel Multi-Interface (MCMI) wireless node is presented in this paper. The MCMI wireless node was developed using a Xilinx Spartan-3E XC3S500E FPGA development board. The CC2500 RF transceivers as low-cost low-power wireless radio modules operating in 2.4 GHz ISM band were connected with the FPGA board via a Serial Peripheral Interface (SPI). The implementation result shows that the designed FPGA-based MCMI wireless node architecture uses small amount of resources of the Xilinx Spartan-3E XC3S500E FPGA. Consequently, we can further include more algorithms or functions on the top of our proposed system. The experimental result from a preliminary test scenario also demonstrates that the FPGA-based MCMI wireless node improves the packet delivery ratio to reach 100 % while varying traffic loads.

Keywords Multi-channel · Multi-interface · FPGA · Wireless node · CC2500 RF transceiver · Packet delivery ratio · Packet inter-arrival time

W. Jindamaneepon · B. Rattanalert · K. Sengchuai · A. Booranawong · N. Jindapetch (✉)
Department of Electrical Engineering, Faculty of Engineering, Prince of Songkla University, Hat Yai, Songkhla 90112, Thailand
e-mail: nattha.s@psu.ac.th

W. Jindamaneepon
e-mail: witoon.jinda@gmail.com

B. Rattanalert
e-mail: banjerdrattanalert@gmail.com

K. Sengchuai
e-mail: ak.kiattisak@hotmail.com

A. Booranawong
e-mail: apidet.boo@gmail.com

H. Saito
Division of Computer Engineering, School of Computer Science and Engineering, The University of Aizu, Aizu, Japan
e-mail: hiroshis@u-aizu.ac.jp

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_99

1 Introduction

Currently, wireless communication networks are widely used for many applications, such as environmental monitoring, structural monitoring, video surveillance, healthcare, home automation, agriculture, and industrial process control [1]. Among these applications, many works specifically require high-throughput of the data transmission. However, in wireless network communications, the communication among sensor nodes over unreliable wireless mediums can cause the occurrence of packet loss and delay that degrades the communication performances. Many effects occurring during the communication, such as multi-path fading, noise, and radio interference can significantly reduce the throughput of the data transmission. Therefore, reliable communication techniques both from hardware and software perspectives should be applied for wireless communications.

According to the research literature, we found that the research problem as described above can be dealt with by developing the wireless sensor nodes to support the MCMI communication. In [2], an adaptive interference-aware multi-channel algorithm with an Acknowledgment/Negative Acknowledgment (ACK/NACK) beacon consideration was proposed for an IEEE 802.15.4 wireless network. This proposed algorithm was implemented on Crossbow's MICAz motes for avoiding the radio interference from IEEE 802.11 Wireless Local Area Networks (WLANs). In [3], a channel switching technique for avoiding the radio interference from the IEEE 802.15.4 network and other wireless networks as using 2.4 GHz ISM band was presented; the theoretical evaluation also shown the performance of the proposed channel switching technique. In [4], the multi-channel communication in a wireless automation system was studied. The work in [4] also investigated how the communication parameters, such as number of channels, packet arrival rates, and packet sizes influence the performance of the wireless automation system by the analytical study. In [5], the MCMI technique for video transmission over IEEE 802.15.4 networks was proposed for avoiding the radio interference of the IEEE 802.15.4 and the IEEE 802.11 devices; the performance evaluations were carried on the simulation. In [6], the MCMI approach as implemented on a microcontroller with the CC2500 wireless radio modules was presented. The experiment also guaranteed the performance of the proposed approach. To the best of our knowledge, we found that there was only the work as presented in [7] implemented the MCMI approach on a Field-Programmable Gate Array (FPGA) with the wireless module as operating in 2.4 GHz frequency band for supporting high throughput applications. The analytical results guaranteed the performance of this wireless sensor platform. Note that the wireless FPGA node as developed in [7] was not evaluated in the real experiment.

In this paper, the implementation and the preliminary experimental evaluation of the FPGA-based MCMI wireless node is presented. The high performance wireless node was implemented on the Xilinx Spartan-3E XC3S500E FPGA development board with the CC2500 wireless modules as the low-cost low-power wireless module operating in 2.4 GHz frequency band. The CC2500 wireless modules were connected with the selected FPGA board via SPI. The implementation result indicates that the MCMI wireless FPGA node uses 26 % of available slices of the Xilinx Spartan-3E XC3S500E FPGA. Hence, we can further include more functions to achieve Programmable System-on-Chip (PSoC) design. For the experimental results, the finding from the given scenario indicates that the FPGA-based MCMI wireless node gives the packet delivery ratio as the ratio of the number of packets received at the receiver and the number of packets sent from the transmitter close to 100 % while varying traffic loads.

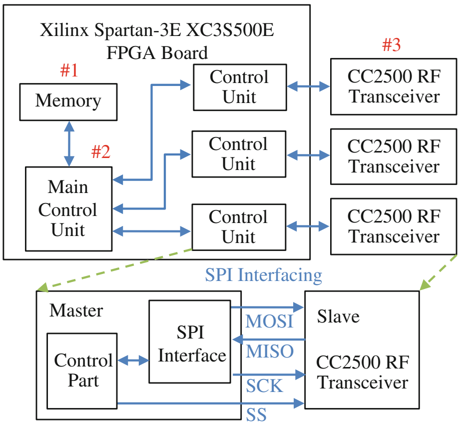
The rest of this paper is organized as follows. Section 2 provides the implementation details of the FPGA-based MCMI wireless node. Section 3 describes the experimental setup. Section 4 provides the experimental results and discussions. Finally, Sect. 5 concludes this paper.

2 FPGA-Based MCMI Wireless Node

2.1 Architecture and Hardware

The architecture of the FPGA-based MCMI wireless node presented in this work is shown in Fig. 1. The MCMI wireless node was implemented using a Xilinx Spartan-3E XC3S500E FPGA development board [8]. This board was selected because it has enough available resource to support control and computing parts, and it also has enough input and output pins for interfacing with three radio transceivers. Note that the number of interfaces can be easily extended more than three depending on the application requirements. The proposed MCMI wireless node architecture consists of three major components: a memory component, a control component and a radio component indicated by block number #1, #2, and #3, respectively. The memory component is employed for collecting both transmitted data and received data. We note that for the transmitted data, they can be

Fig. 1 The FPGA-based MCMI wireless node architecture



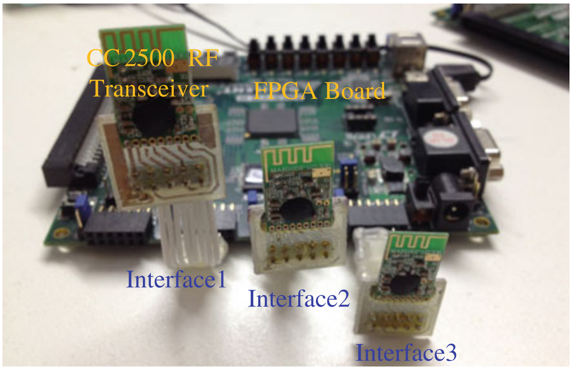
obtained from any sensor devices. However, at the current state of our research, sensor devices are not yet integrated on our wireless node platform. The transmitted data using in the experiment are directly inserted by us. For the control component, it consists of a main control unit and three sub-control units. The main control unit controls all operations of the sub-control units, while each sub-control unit controls the radio transceiver. For the radio component, we employ the CC2500 RF transceiver as the wireless module [9]. The CC2500 is the low-cost 2.4 GHz RF transceiver designed for very low power wireless applications [9]. It also has the maximum data rate of 500 kbps. Consequently, we can utilize it for applications that require a high throughput transmission. In addition, the CC2500 supports the use of the different radio channels as operating in the frequency range of 2.4–2.4835 GHz. Thus, a desired channel can be assigned by users. For this work, we manually configure different radio channels to the three interfaces. The CC2500 RF transceiver connects with the FPGA board via an SPI interface, which consists of four pins including Master Out Slave In (MOSI), Master In Slave Out (MISO), Slave Select (SS), and Serial Clock (SCK). Where the FPGA board is the master, and the CC2500 is the slave, as shown in Fig. 1. For more details of the SPI interface can be further found in [10].

The FPGA-based MCMI wireless node architecture as proposed in Fig. 1 is corresponding to a designed hardware (i.e. a node prototype) as shown in Fig. 2. Three CC2500 RF transceivers are interfaced with the XC3S500E Spartan-3E FPGA board. The CC2500 RF transceivers are placed at the different heights for avoiding the radio interference occurring among them. However, the multi-channel setting is the considered option that can also help to eliminate the radio interference problem.

2.2 Transmitter and Receiver Operations

The flow chart of transmitter and receiver operations of the proposed wireless node is shown in Fig. 3. The designed wireless node supports both transmitter and

Fig. 2 Designed hardware



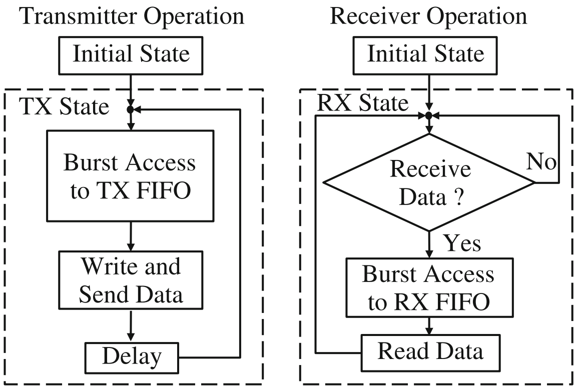


Fig. 3 Flow chart of transmitter and receiver operations

receiver operations, and each operation is automatically run during the communication. Before a node can communicate with others, CC2500 parameters, such as RF output powers, radio channels, data rates, packet formats, data buffers, and modulation techniques are configured using the SPI interface (i.e. initial state). They can be programmed by setting addresses and values to the corresponding registers. When the node wants to transmit its data to others, it performs the transmitter operation by accessing the TX FIFO. Note that the CC2500 contains of 64 byte FIFOs; one for data to be transmitted (i.e. TX FIFO) and one for received data (i.e. RX FIFO). Then, a node writes data in an available TX FIFO and immediately constructs a packet sent to a destination node. After the packet is sent, the node waits with a predefined period of time and repeats the whole procedure if it still require to transmit more data. For the receiver operation, when the node can detect a radio signal in a chosen channel as sending from a transmitter. It also automatically accesses and reads data from the RX FIFO. If there are more data to be received, the node also repeats the whole process, like the transmitter operation.

3 Experimental Setup

The primary objective of our test is to investigate communication performances of the proposed wireless node both in cases of the single-channel and the multi-channel settings. All three interfaces are assigned the same radio channel in the single-channel case, whereas in the multi-channel case, those three interfaces are assigned the different radio frequencies.

In our experiment, there are two wireless nodes; one transmitter and one receiver, as shown in Fig. 4. These nodes operate at 2.4 GHz ISM band with the data rate of 500 kbps. They are placed on the table inside the office room, and the distance between them is set to 15 cm. In this test environment, we guarantee that there is no interference effect from any 2.4 GHz wireless devices. For each test, the transmitter node continuously sends 100 data packets (i.e. via each radio interface)

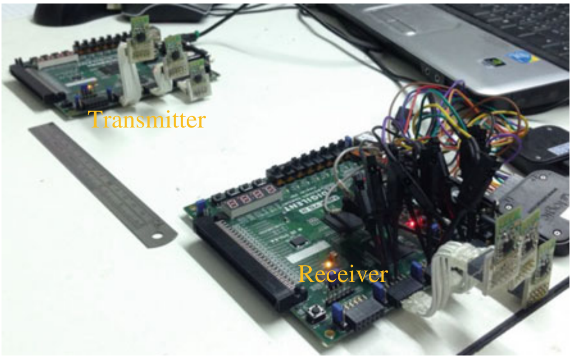


Fig. 4 Experimental scenario

to the receiver node. The data packet sizes are set to 32 and 254 bytes, respectively. Furthermore, at each data packet size, we also test six levels of packet inter-arrival times; 1.27, 2.47, 3.67, 4.87, 6.07, and 7.27 ms, respectively. By this setting, we can study how the traffic rate affects the communication performance and how well the wireless node handles this situation [4, 11]. For a reason why we select the minimum packet inter-arrival time of 1.27 ms and the maximum packet inter-arrival time of 7.27 ms is described here. The packet inter-arrival time of 1.27 ms is corresponding to the required minimum time for the communication process between the FPGA and the CC2500 RF transceiver before transmitting each data packet. Additionally, according to our test results, the packet inter-arrival time of 7.27 ms is enough to reveal the performance of the wireless node both in the single-channel and the multi-channel cases.

The percentage of the packet delivery ratio is selected as the performance metric for evaluating the performance of the proposed wireless nodes. This metric is the ratio of the number of packets received at the receiver and the number of packets sent from the transmitter. It represents the level of the delivered data to the receiver node [12]. The average experimental results of the given performance metric are computed from 30 replications. The 95 % confidence interval is also provided for each average result.

4 Results and Discussions

4.1 Implementation Results

The implementation results from the static timing report indicate that our system can run at the maximum clock speed of 107.23 MHz. However, we divide the clock speed into 16.67 MHz for interfacing with the CC2500 transceiver. The device utilization summary of the proposed wireless FPGA node architecture is shown in Table 1. We can see that the proposed system is a small circuit. It uses 1224 slices or 26 % of available slices of the Xilinx Spartan-3E XC3S500E FPGA. There is

Table 1 Device utilization summary on Xilinx Spartan-3E XC3S500E FPGA for the designed wireless node architecture

Logic utilization	Used	Available	Utilization (%)
Number of slices	1224	4656	26
Number of slice flip flops	553	9312	5
Number of 4 input LUTs	2885	9312	25
Number of bonded IOBs	20	232	8
Number of GCLKs	2	24	8
Number of DCMs	1	4	25

still 74 % remaining slices can be utilized. Therefore, we can further include more communication algorithms or functions on the top of our system, such as a sensor device function, a radio interference detection algorithm, a channel switching scheme and so on.

From the real test, timing diagrams were measured from the Saleae digital logic analyzer [13] when the FPGA communicates with the CC2500 RF transceiver for writing and reading data from the CC2500 buffer (i.e. TX FIFO and RX FIFO) as shown in Fig. 5a and b, respectively. In Fig. 5a, to write data in TX FIFO, the

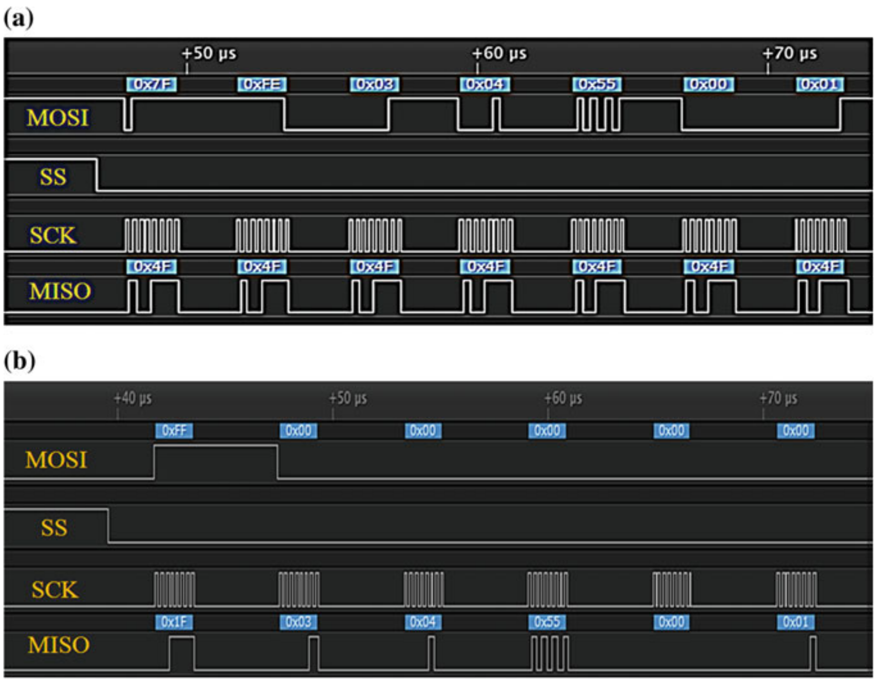


Fig. 5 Timing diagrams; **a** write data in the CC2500 TX FIFO, **b** read data from the CC2500 RX FIFO

FPGA first sends a low logic signal (i.e. logic 0) to the CC2500 via the SS pin. Then, the FPGA sends command signals 0x7F (i.e. a header byte notified burst access to TX FIFO), 0xFE (i.e. data length), 0x03 (i.e. RX address), and 0x04 (i.e. TX address), to the CC2500 via the MOSI pin. For the signals 0x55, 0x00, and 0x01 represent the data signals. Finally, the CC2500 automatically constructs a packet according to a defined packet format, and it immediately sends the packet to the receiver node. Note that each data bit will be sent via the MOSI pin at every positive edge clock signal. In the same time, the CC2500 also responses to the FPGA command via the MISO pin. So data can be communicated in the full duplex manner. In Fig. 5b, to read data in RX FIFO, the FPGA first sends the low logic signal to the CC2500 via the SS pin. Then, the FPGA sends command signals 0xFF (i.e. a header byte notified burst access to RX FIFO) to the CC2500 via the MOSI pin. Finally, the CC2500 transfers received data as collected in its RX FIFO back to the FPGA via the MISO pin; the command signals 0x03 and 0x04 indicate RX address and TX address, while signals 0x55, 0x00 and 0x01 are the data.

4.2 Experimental Results

Figure 6 shows the relationship between the percentage of packet delivery ratio and the packet inter-arrival time at 32 bytes data packet size. The experimental results show that when three radio interfaces use the same radio channel (i.e. single-channel setting) for the communication, the packet delivery ratio is very low at the packet inter-arrival time of 1.27 ms. However, when the packet inter-arrival time is higher, the packet delivery ratio is getting improved. The packet delivery ratio is close to 100 % when the packet inter-arrival time is equal to or greater than 4.87 ms. This result indicates that using the small packet inter-arrival time in the case of single-channel setting is not an appropriate choice for data transmission; more interference occurring among three radio interfaces can cause higher packet loss.

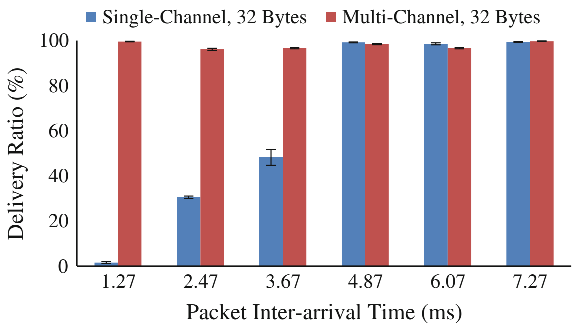


Fig. 6 The relationship between the percentage of packet delivery ratio and the packet inter-arrival time at 32 bytes data packet size

For the multi-channel setting, the packet delivery ratio is close to 100 % for all levels of the packet inter-arrival time. This result demonstrates that the multi-channel multi-interface setting can help to alleviate the effect of the radio interference as well as the possibility of packet loss. We note that if we use the small value of the packet inter-arrival time for data transmission in the case of multi-channel, we are going to utilize the FPGA-based MCMI wireless node for an application that requires a high throughput. This is because both the reliability as indicated by the delivery ratio and the traffic rate as indicated by the packet inter-arrival time are high.

Figure 7 shows the relationship between the percentage of packet delivery ratio and the packet inter-arrival time at 254 bytes data packet size. The results show that the packet delivery ratio in the single-channel case is lower than the multi-channel case for all levels of packet inter-arrival times. For the single-channel setting, the packet delivery ratio at the 254-bytes packet size is also less than the packet delivery ratio at the 32-bytes packet size. This is because transmitting bigger data packet size requires more transmission time; more radio interference as well as the packet collision can be occurred.

Although the data packet size is 254 bytes, the packet delivery ratio in the case of multi-channel setting is still close to 100 % for all levels of packet inter-arrival times, like the case of the data packet size of 32 bytes. This result indicates that the MCMI wireless FPGA node performs well to support bigger data packet sizes.

We summarize the experimental results from Figs. 6 and 7 here again. The multi-channel setting almost gives the better results than the case of the single-channel setting during varying the packet inter-arrival time. However, the single-channel setting is not complex compared with the multi-channel setting. If the data traffic rate is not fast, like the case of the packet inter-arrival time of 4.86–7.27 ms, the single-channel setting can be utilized for the data transmission. However, it may not support an application that requires small delay or high throughput data transmission (i.e. trade-off). For the multi-channel setting, the performance results both in cases of the packet size of 32 and 254 bytes at all levels of packet inter-arrival times are not significantly different. The FPGA-based MCMI wireless node works well although the packet size is higher.

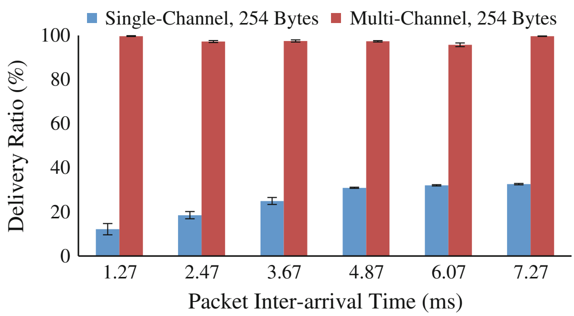


Fig. 7 The relationship between the percentage of packet delivery ratio and the packet inter-arrival time at 254 bytes data packet size

The summary of our findings as described above may not true for other wireless network scenarios. In an unreliable wireless communication, the performance of the FPGA-based MCMI wireless node may be dropped due to the effects of the radio interferences from multiple transmitters, dynamic traffic rates, multi-hop communications, and etc. Consequently, for the future work, we plan to investigate its performance in various scenarios. Additionally, we also plan to further develop the FPGA-based MCMI wireless node; radio interference detection, channel switching algorithms [9] and area optimization are firstly considered for avoiding radio interference from any 2.4 GHz wireless devices and reducing the FPGA resources.

5 Conclusions

In this paper, the implementation and the preliminary test of the FPGA-based MCMI wireless node is presented. The FPGA-based MCMI wireless node was implemented using the Xilinx Spartan-3E XC3S500E FPGA development board. Three CC2500 RF transceivers as the low power wireless radio modules connect with the FPGA via SPI. The implementation result indicates that the proposed architecture requires 26 % of available slices of the Xilinx Spartan-3E XC3S500E FPGA. There is more remaining slices can be utilized. Therefore, more algorithms or functions can be further included in our system to achieve the PSoC design. For the experiment, the preliminary result from the given scenario shows that the proposed wireless FPGA node with the multi-channel setting gives the packet delivery ratio close to 100 % while varying the traffic rates.

Acknowledgments This work was fully supported by Prince of Songkla University funding contact number ENG560014S and partially supported by Center of Excellence in Wireless Sensor Networks (CoE-WSN), Faculty of Engineering, Prince of Songkla University, Hat Yai, Songkhla, Thailand.

References

1. Yick, J., Mukherjee, B., Ghosal, D.: Wireless sensor network survey. Comput. Netw. **52**, 2230–2292 (2008)
2. Kang, M.S., Chong, J.W., Hyun, H., Kim, S.M., Jung, B.H., Sung, D.K.: Adaptive interference-aware multi-channel clustering algorithm in a zigbee network in the presence of WLAN interference. In: 2nd International Symposium on Wireless Pervasive Computing, pp. 200–205 (2007)
3. Yun, J., Lee, B., Li, J., Han, K.: A channel switching scheme for avoiding interference of between IEEE 802.15.4 and other networks. In: International Multisymposiums on Computer and Computational Sciences, pp. 136–139 (2008)
4. Nieminen, J., Björkbom, M., Jäntti, R., Eriksson, L.: Multichannel communications in wireless automation: interdependencies between communication and control parameters. Int. J. Distrib. Sensor Netw. (2012)

5. Zainaldin, A., Lambadaris, I., Nandy, B.: Video over wireless zigbee networks: multi-channel multi-radio approach. In: International Conference on Wireless Communications and Mobile Computing, pp. 882–887 (2008)
6. Yang, L.D.: Implementation of a wireless sensor network with eZ430-RF2500 development tools and MSP430FG4618/F2013 experimenter boards from texas instruments. Master's thesis of Louisiana State University (2011)
7. Kohvakka, M., Arpinen, T., Hännikäinen, M., Hämäläinen, T.D.: High-performance multiradio WSN platform. In: 2nd International Workshop on Multi-hop Ad Hoc Networks: From Theory to Reality, pp. 95–97 (2006)
8. Xilinx: Spartan-3E FPGA Family Data Sheet. http://www.xilinx.com/support/documentation/data_sheets/ds312.pdf (accessed on 19 July 2013)
9. Instruments, T.: CC2500 Datasheet. http://focus.ti.com/lit/ds/symlink/cc2500.pdf (accessed on 6 February 2015)
10. Liu, T., Wang, T.: IP design of universal multiple devices SPI interface. In: The IEEE International Conference on Anti-Counterfeiting, Security and Identification, pp. 169–172 (2011)
11. Booranawong, A., Teerapabkajorndet, W., Limsakul, C.: Energy consumption and control response evaluations of AODV Routing in WSANs for building-temperature control. Sensors **13**, 8303–8330 (2013)
12. Booranawong, A., Teerapabkajorndet, W.: An enhanced AODV routing protocol for wireless sensor and actuator networks. Int. J. Electr. Comput. Electron. Commun. Eng. **7**, 1203–1210 (2013)
13. Saleae: Logic & Logic16 User's Guide. http://downloads.saleae.com/Logic+Guide.pdf (accessed on 5 June 2015)

Multilayer Performance of Green Biomass Coated Pyramidal Hollow Microwave Absorber

Hasnain Abdullah, Linda Mohd Kasim, Mohd Nasir Taib, Norhayati Mohd Noor, Noor Azila Ismail, Azizah Ahmad, Nazirah Mohamat Kasim, Najwa Rawaida Ahmad and Ahmad Rashidy Razali

Abstract This paper investigates the performance of multilayer pyramidal hollow green biomass coated absorbing materials. The designed absorber is simulated using CST Microwave Studio. Several experimental evaluations have been executed to investigate the performance of the tested absorber using free space arch measurement method. The frequency between 8 and 12 GHz was selected for performance evaluation. The results show that, the multilayer pyramidal shaped of microwave absorber produces excellent performance results.

1 Introduction

Communication technology is growing rapidly along with the ease of modern communication requests. It is more closely related to the use of wireless communication equipment. Advances in communication technology for antennas and electro-magnetic shield system based on important development of materials that absorb electromagnetic radiation in a wide range. High absorbance rate of such materials must be assisted by combination with low reflectance from the same frequency range. One of the techniques to solve the problem of interference is to use microwave absorbers. Signal interference absorbed when it travelled through microwave absorbers. Signal reflectivity occurs through several mechanisms of

H. Abdullah (✉) · L.M. Kasim · M.N. Taib · N.M. Noor · N.A. Ismail · A. Ahmad
N.M. Kasim · N.R. Ahmad · A.R. Razali
Faculty of Electrical Engineering, Universiti Teknologi Mara (UiTM), 13500 Permatang Pauh, Pulau Pinang Malaysia
e-mail: hasnain@ppinang.uitm.edu.my

L.M. Kasim
e-mail: linda@ppinang.uitm.edu.my

M.N. Taib
e-mail: dr.nasir@ieee.org

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_100

loss. One of the mechanisms is the power dissipation by dissipative elements of absorbers in microwave. Dissipative elements absorb signal by changing the form of microwave energy into heat energy.

The use of absorbers is to abolish the energy from the microwave radiation. Signal from the transmitter will be reflected across the surface of the absorber in which form this signal be amended by way of cancellation and scattering [1]. There are various forms of absorber such as pyramidal, truncated pyramidal, wedge, convoluted, hybrid absorber, flats absorber, metamaterial type absorber, oblique types and others. Honey comb shape absorber has also been used because of its high strength-to-weight ratio, which also provides support for good reflectivity [2]. Absorber design will distinguish reflectivity performance.

Developments in material science and engineering technology lead to the introduction of several new concepts for the design of the absorber including materials used and a specific frequency selection. The pyramid-shaped absorber is selected due to the potentiality of producing the zig-zag-shaped signal on the surface of the pyramid and dematerializes at the surface of the base. The pyramid is then coated with a material produced from a mixture of palm ash [3]. Parameters that must be taken into vigilant consideration during design stage for this type of absorber are material dielectric constant, angle of incidence, dimension of absorber, and carbon coating thickness. The reflectivity can express the absorbing performance of the material and is a function of the complex permittivity and permeability of the material, and the wide range high frequency [4].

The material chosen in designing the absorber must have the capability to absorb the radio frequency signal through several loss mechanisms. Criteria for loss mechanism are involving loss of power through liberation of heat. Losses of the absorber are seen on the imaginary representing dielectric and magnetic properties of the absorber. The materials consist of carbon such as Carbonyl-iron, barium ferrite and ferroxides are causing the loss of dielectric materials [5, 6].

The reflectivity of a metal-backed microwave absorber when an electromagnetic wave strikes the surface at normal incidence is a function of frequency, thickness, dielectric permittivity and magnetic permeability of the absorbing material. The intrinsic impedance of a metal-backed flat absorber is given by [7].

$$\eta = Z_o \sqrt{\frac{\mu_r}{\varepsilon_r}} \tanh(\gamma t) \tag{1}$$

$$Z_o = \sqrt{\frac{\mu_o}{\varepsilon_o}} = 377\,\Omega \tag{2}$$

where Zo is the intrinsic impedance of free space, t is absorber thickness and γ is the propagation constant in the material.

The earlier study shows that the average reflection by using the paint absorber based on coconut shell for single layer is 3 dB. Meanwhile for double layer absorber is 12 dB and the highest reflection rate is 17 dB at frequency of 10.2 GHz [8].

2 Methodology

Radiation absorbing material can be divided into two types, namely coating and structure design. The coating is intended to absorb the incident wave and reduce the return loss. Meanwhile the structure design is to reduce the reflection [9]. The structure design of single layer and multilayer absorber has also been used in commercial. The reflectivity rate of the use of single layer absorber is only good for narrow frequency range; while for multilayer absorber is good for broad frequency range [10]. The theory of multilayer absorber is based on Jaumann microwave absorber [11].

Throughout the study, absorber was designed in hollow pyramidal shape. Figure 1 shows the sizing and arrangement of 9th pyramidal absorber. Naturally, carbon is a material that has good semiconductor properties thus it is suitable for this study. The proportion of carbon and composite material mixture must be handled carefully. In this paper, to demonstrate the effectiveness of the pyramidal hollow microwave absorber there are three types of design development absorbent introduced namely uncoated, single-sided and double-sided coated absorber.

In the early stage of preparation, modelling and simulation had been carried out by using CST Microwave Studio to get the reflectivity performance or the decision of S_{11}. Analysis of the results and measured absorbent performance has been carried out to determine the overall performance of microwave reflectivity.

Figure 2 shows the set up for this investigation, known as an arch measurement method which applied together with vector network analyzer (VNA) for reflectivity measurements. This method requires a pair of horn antenna, VNA and material measurement software (Agilent 85071E). VNA is also used to measure scattering matrix in the frequency range corresponding to the working condition of the absorber while S-parameter measurement gives required constitutive parameter of the material under test [8].

The measurement data at frequency range from 8 to 12 GHz are then recorded and observed. The solid commercial wedges absorber IP-045C produced by TDK RF solution Inc. at UiTM Pulau Pinang is also tested to study the performance [12].

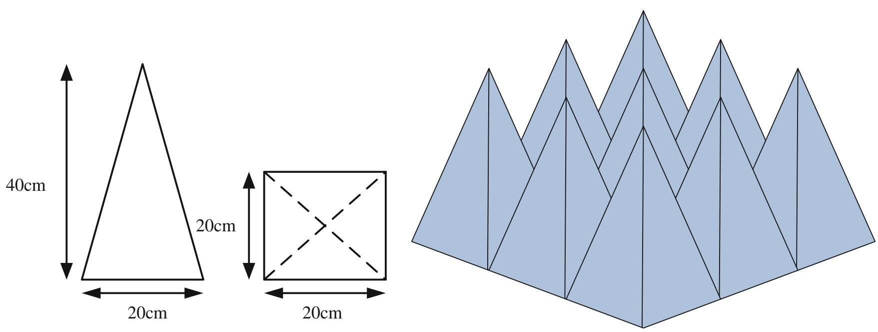
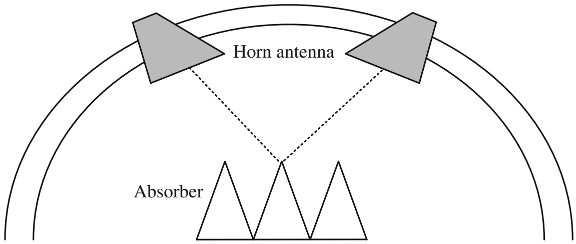


Fig. 1 Size of the pyramid

Fig. 2 Reflective Performance Test



3 Result and Discussion

Figures 3, 4 and 5 shows the simulation result for two, three and six layers absorber respectively using CST.

Table 1 shows the data from the simulation by referring to Figs. 3, 4 and 5. Simulation is performed for two, three and six layers with different thicknesses of 0.5 mm, 1 mm and 1.5 mm respectively. Minimum reflectivity is −7.42 dB for 2 layers, −7.87 dB for 3 layers and −8.55 dB for 6 layers while the maximum reflectivity for 2, 3 and 6 layers are −41.29, −48.20 and −48.96 dB. The average reflectivity of each layer is −17.71, −17.48, and −19.10 dB. The value of reflectivity at the centre frequency of 10 GHz are −12.46, −13.11 and −15.92 dB for each layer. Figures 6, 7 and 8 shows the result for two layers absorber in uncoated, single sided coated and double sided coated absorber using Free Space Measurement Method.

Microwave signal reflectivity is measured through experiments performed using microwave absorbers are classified into uncoated, single sided and double sided coated. The frequencies used for this experiment are from 8 GHz to 12 GHz. The measurement parameters apply for this experiment are the minimum, maximum, average and reflectivity at 10 GHz. Figures 6, 7 and 8 shows the measured performance for two layer absorber and the result summarization can be find in Table 2. The average reflectivity entirely describes the ability of the absorber to

Fig. 3 Simulation result for 2 layer elephant board absorber

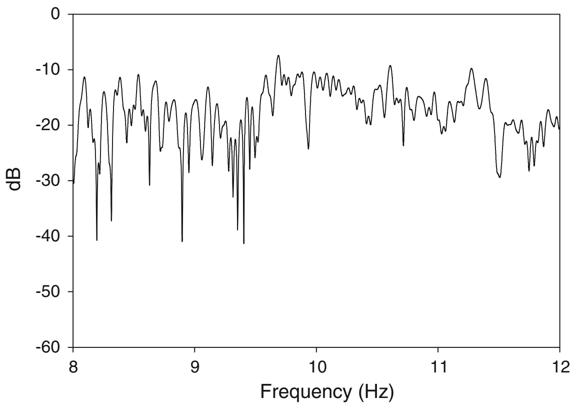


Fig. 4 Simulation result for 3 layer elephant board absorber

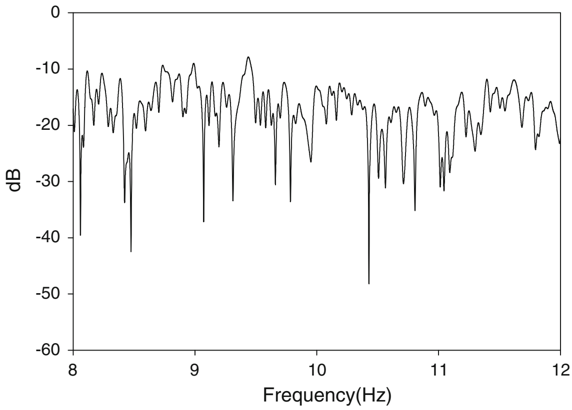


Fig. 5 Simulation result for 6 layer elephant board absorber

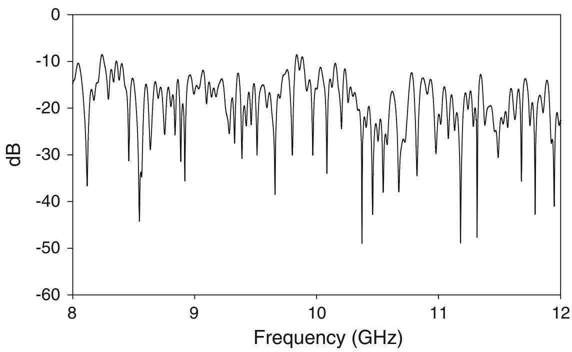
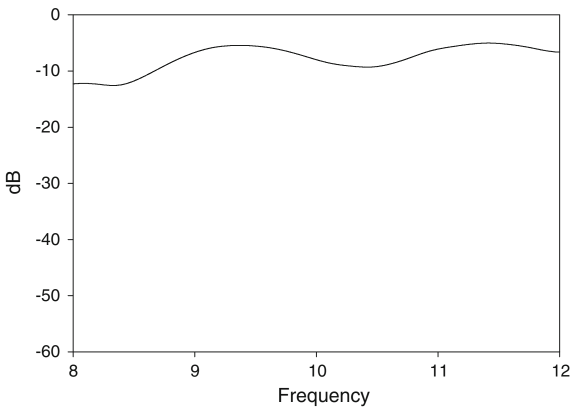


Table 1 Parameter of simulation

Layer	Min	Max	Average	Reflectivity at 10 GHz
2	−7.42	−41.29	−17.71	−12.46
3	−7.87	−48.20	−17.48	−13.11
6	−8.55	−48.96	−19.10	−15.92

Fig. 6 Measurement result for 2 layer uncoated absorber



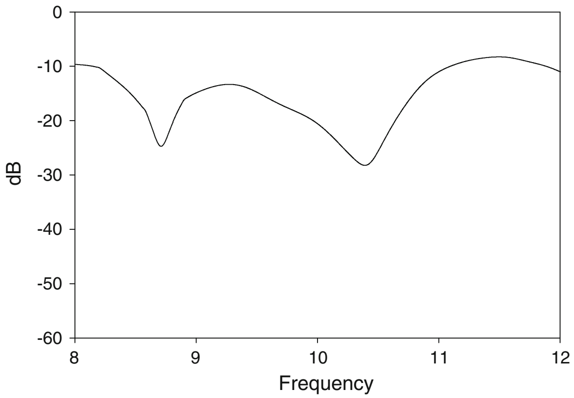


Fig. 7 Measurement result for 2 layer single sided coated absorber

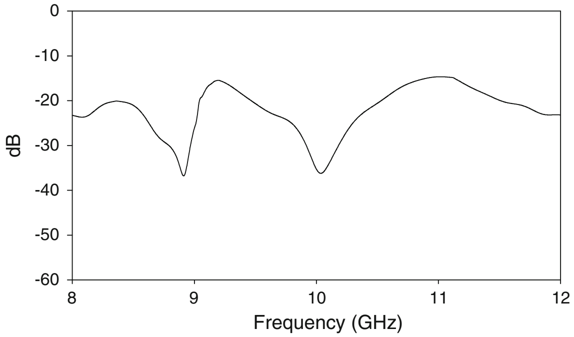


Fig. 8 Measurement result for 2 layer double sided coated absorber

Table 2 Parameter measurement for 2 layers absorbers

Layer	Min	Max	Average	Reflectivity at 10 GHz
Uncoated	−5.03	−12.57	−7.71	−7.98
Single sided coated	−8.26	−28.21	−15.36	−20.55
Double sided coated	−14.66	−36.61	−22.00	−35.3

absorb microwave signals. Experiments of uncoated absorber produce −7.71 dB while the single sided coated absorber absorbs the −15.36 dB. The reflectivity is at maximum with reading of −22 dB for double sided coated absorber. Generally for the three measurements of two layer absorber shows double sided coated absorber has the highest reflection levels.

Table 3 shows the measurements results for three layers absorber. Figures 9, 10 and 11 shows the overall reflectivity image for any circumstances. The three layers absorber shows a good measurement result when compared to two layers absorber. Average reflectivity for single sided coated absorber is increased, −20.09 dB for three layers compared to −15.36 dB for two layers. However there is a decreased in reflectivity at 10 GHz, −31.18 dB for three-layer single sided coated when

Table 3 Parameter measurement for 3 layers absorbers

Layer	Min	Max	Average	Reflectivity at 10 GHz
Uncoated	−1.25	−7.80	−3.87	−3.25
Single sided coated	−5.05	−40.21	−20.09	−31.18
Double sided coated	−18.71	−52.18	−24.68	−21.50

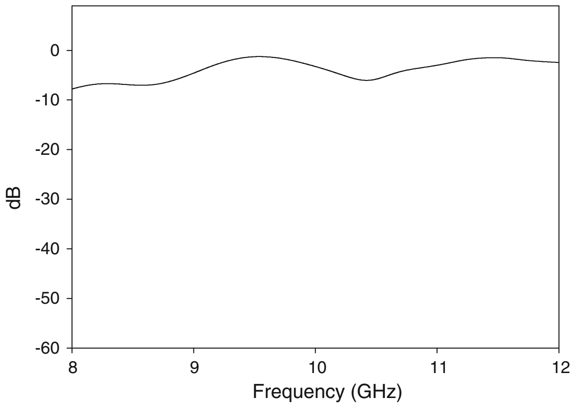


Fig. 9 Measurement result for 3 layer uncoated absorber

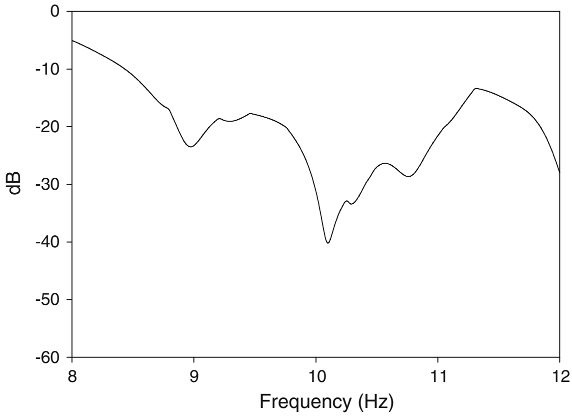


Fig. 10 Measurement result for 3 layer single sided coated absorber

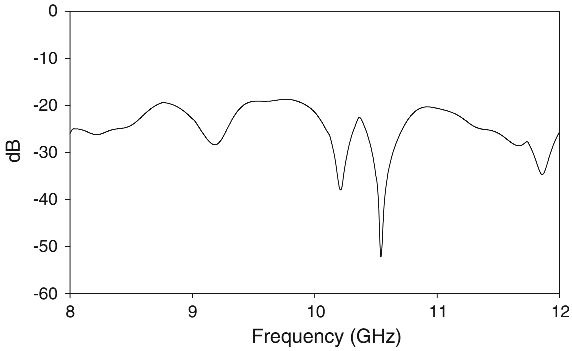


Fig. 11 Measurement result for 3 layer double sided coated absorber

compared to double sided coated with reading of −21.50 dB. Although there are differences in reflectivity at 10 GHz, the average reflectivity for three layers is very good in terms of reflection rates.

Table 4 shows the measurement reading of the six layers absorber. These data are analyzed according to Figs. 12, 13 and 14. An overview of the actual reflection of microwaves can be seen referring to average −29.13 dB for double sided coated in contrast to −17.59 dB for single sided coated absorber. Refers to reflectivity at 10 GHz, six-layer, double sided coated absorber shows maximum reflectivity at −36.38 dB compared to −22.42 dB for single-sided coated absorber. It shows the

Table 4 Parameter measurement for 6 layers absorbers

Layer	Min	Max	Average	Reflectivity at 10 GHz
Uncoated	−8.98	−17.06	−11.57	−12.10
Single sided coated	−10.49	−33.41	−17.59	−22.42
Double sided coated	−20.49	−52.26	−29.13	−36.38

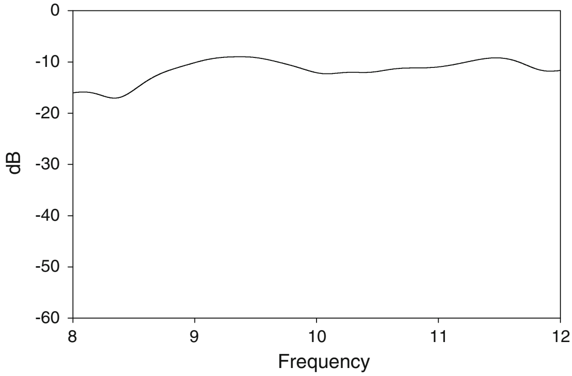


Fig. 12 Measurement result for 6 layer uncoated absorber

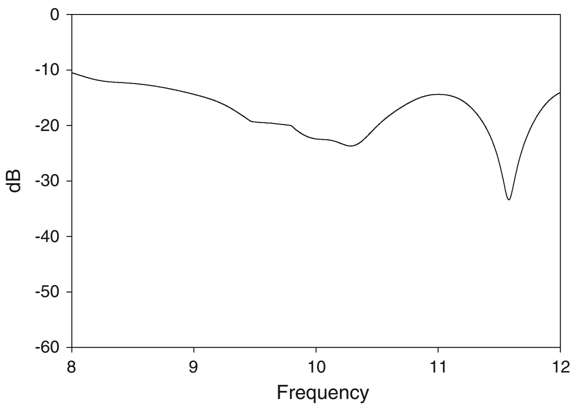


Fig. 13 Measurement result for 6 layer single sided coated absorber

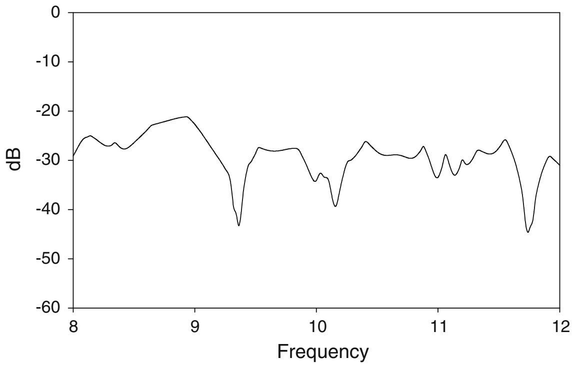


Fig. 14 Measurement result for 6 layer double sided coated absorber

frequency in between 8–12 GHz, six-layer, double sided coated absorber strives to remain on the uptake of between −30 and −40 dB.

Tables 5, 6 and 7 shows the difference between coated uncoated, single sided and double sided coated refers to two, three and six layers absorber. Uncoated absorber shows less reflectivity. Single sided coated absorber shows reflectivity for two layers are better than six layers absorber. Although there was a decrease if appointed at 10 GHz, but the average value shows almost exact reflectivity. Double-sided painted carbon absorber gives a striking impact if analyzed for all of the two-layer, three layers and six layers as shown in Table 7. All parameters measured showed a very good readings. The average reflectivity indicates reflectivity of less distinction

Table 5 Comparison for uncoated absorber

Layer	Min	Max	Average	Reflectivity at 10 GHz
2	−5.03	−12.57	−7.71	−7.98
3	−1.25	−7.80	−3.87	−3.25
6	−8.98	−17.06	−11.57	−12.10

Table 6 Comparison for single sided coated absorber

Layer	Min	Max	Average	Reflectivity at 10 GHz
2	−8.26	−28.21	−15.36	−20.55
3	−5.05	−40.21	−20.09	−31.18
6	−10.49	−33.41	−17.59	−22.42

Table 7 Comparison for double sided coated absorber

Layer	Min	Max	Average	Reflectivity at 10 GHz
2	−14.66	−36.61	−22.00	−35.3
3	−18.71	−52.18	−24.68	−21.50
6	−20.49	−52.26	−29.13	−36.38

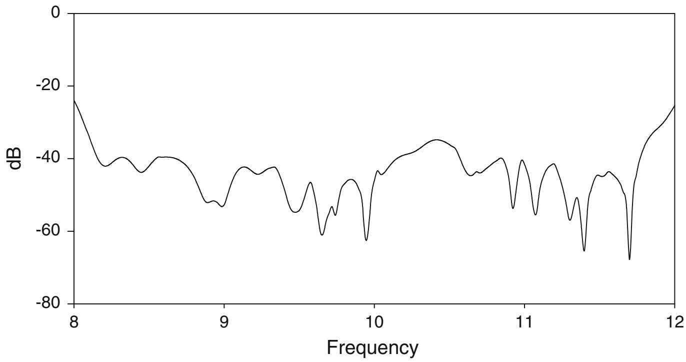


Fig. 15 TDK—Microwave wedges absorber IP-045C

between −22 and −26.9 dB. If referred to the frequency of 10 GHz, there is no significant difference in reflectivity between the three layers. In conclusion, six layers double sided coated absorber is the best among three which can be used as a microwave absorber. Figure 14 represents the reflectivity measurement result for the solid commercial wedges absorber (Fig. 15).

4 Conclusions

This paper presented the performances of multilayer pyramidal hollow green biomass coated absorbing materials. With double sided coated, the measurement results shows an improvement on the reflectivity performance. The average reflectivity performance has improved 14–20 dB with reference to 2, 3 and 6 layers coated absorber. Low reflection coefficients, S11 results from the coconut shell carbon paint strongly suggest that the material can behave as a good absorbing material. Utilizing from natural agriculture sources not only can contribute towards the environmental-friendly absorbing material but economical. Reflectivity can be enhanced if more research is carried out in the future. The use of agricultural sources is in line with the government's call towards green technology.

Acknowledgement The authors would like to acknowledge the financial support from RMI, Universiti Teknologi MARA, Malaysia for 600-RMI/RAGS 5/3(190/2014) for their financial and equipment assistant for the project research.

References

1. Abdullah, H., Damit, D.S.A., Taib, M.N., Noor, N.M., Ali, W.K.W., Jalil, S.Z.A., Abdullah, A.T., Zanal, A.: Green microwave radiation absorbing paint. In: 2012 International Conference on System Engineering and Technology (ICSET), pp. 1–5 (2012)
2. Zhou, P., Huang, L., Xie, J., Liang, D., Lu, H., Deng, L.: Prediction of microwave absorption behavior of grading honeycomb composites based on effective permittivity formulas. IEEE Transactions on Antennas and Propagation, vol. 99, p. 1, doi:10.1109/TAP.2015
3. Noordin, I.R.M., Idris, H.A., Taib, M.N., Sharif, J.M., Rosli, A.D., Zanal, A., Abdullah, A.T.: Investigation of oil palm ash microwave absorber for broadband application. In: 2012 IEEE 8th International Colloquium on Signal Processing and its Applications, pp. 232–235 (2012)
4. Anzai, H., Sailawa, M., Naito, Y.: The equivalent representation of pyramidal Albsorbers and absorber' s characteristics. IEEE (1995)
5. Farias da Silva, E., Silva Rocha, J., Ribeiro Lins Junior, P., Diniz da Nobrega, S., Sampaio de Alencar, M.: Characterization of electromagnetic radiation absorber materials. International Conference on Microwave and Optoelectronics, 2005 SBMO/IEEE MTT-S, pp. 326–329 (2005)
6. Kotsilkova, R.: Polymer nanocomposites for microwave, pp. 4–7 (2004)
7. Tomasi, W.: Electronic communication systems. Arizona, Pearson, Prentice Hall (2014)
8. Abdullah, H., Taib, M.N., Rahayu, I., Noordin, M.: Computational design and performance evaluation of green painting absorbing material, pp. 963–967 (2012)
9. Ke, L., Xin, Z., Xinyu, H., Peng, Z.: Analysis and design of multilayer Jaumann absorbers. In: Proceedings of the IEEE, 2011, pp. 81–84
10. Cao, M., Zhau, J., Yuan, J., Zhang, T.F., Peng, Z., Gao, Z., Xiao, G., Qin, S.: Computation design and performance prediction towards a multilayer microwave absorber. Mater. Des. **23**, 557–564 (2002)
11. Du Toit, L.J.: The design of Jaumann absorbers. IEEE Transection on Antennas and Propagation, vol. 36, no. 6 (1994)
12. Idris, H.A., Khairuddin, W., Ali, W.: Investigation of hollow pyramidal shape microwave absorber coated with green absorbing paint (2012)

Tricuspid Valve Extraction in Transesophageal Echocardiography

Naziffa Raha Md Nasir, Rahmita Wirza O.K. Rahmat, Puteri Suhaiza Sulaiman, Suhaini Kadiman and Mohd Zamrin Dimon

Abstract In the past decades, echocardiography has appeared as an important modality in medical field to assess heart's function and structures as well as for diagnosis and evaluation. Many image processing researches are done to enhance the imaging aspect and produce better quality of image. Numerous research have been conducted on mitral valve, but only a few on the geometry or annular dynamics of the tricuspid valve. Accurateness in measuring and reconstructing tricuspid valve is an important issue, not only for surgical decision-making process but also in deciding the suitable surgical technique on patient such as valve implication or ring placement. In this paper, we will discuss on techniques that have been applied recently in measuring and modelling tricuspid valve and as for experiment, 3DTEE image was used using level set technique discussed in this paper. Our findings will be focusing more on those techniques applied on 3D echocardiography images from different angels and positions.

Keywords Three dimensional echocardiography (3DE) · Transesophageal (TEE) · Transthoracic (TTE) tricuspid valve (TV) · Tricuspid annulus (TA)

N.R.M. Nasir (✉)
College of Information Technology, Universiti Tenaga Nasional, Kajang, Malaysia
e-mail: naziffa@uniten.edu.my

R.W.O.K. Rahmat · P.S. Sulaiman
Faculty of Computer Science and Information Technology, University Putra Malaysia, Serdang, Malaysia

S. Kadiman
Anaesthesia and Intensive Care, National Heart Institute, Kuala Lumpur, Malaysia

M.Z. Dimon
Cardiothoracic Unit, Surgical Cluster, Faculty of Medicine, UiTM, Shah Alam, Malaysia

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_101

1 Introduction

Recently, numerous research have been conducted on mitral valve, but only a few on the geometry or annular dynamics of the tricuspid valve. Positioned on right side of human heart between right ventricle and right atrium, tricuspid valve is an anatomical structure to prevent blood flows from right ventricle. It consists of three leaflets known as anterior, posterior and septal and surrounded by annulus. Accurateness in measuring tricuspid valve annulus is an important issue, not only for surgical decision-making process but also in deciding the suitable surgical technique on patient such as valve plication or ring placement. Appear in oval structure, the normal size of annulus is about 6cm2 and possible to increase up o 12 cm^2 in full dilation [1]. Due to the leaflets being poorly echogenic, the TA has been used to quantify the magnitude of valve remodeling due to chronic volume overload [2].

Automatic segmentation technique is still a challenging subject for image that comes with noise, different regions, incomplete edges and lack of texture contrast between regions and background. A variety of different approaches to segmentation in how much user input is required. Even though the manual segmentation can produce a very accurate result, it still comes with few drawbacks like time consuming and subject to bias. Another approach is called semi-automatic technique. This technique allows users to be involved in the segmentation process. Interaction with user is needed especially to identify the seed point or focused area. For an automatic technique, user input or interaction is not required and it usually uses prior knowledge to come out with segmentation output.

This paper is organized as follows, Sect. 2 provides explains the modalities and techniques used in assessment of tricuspid valve; review on current segmentation method and study on the measurement of tricuspid valve are discussed in Sects. 3 and 4; experiments in Sect. 5 and the paper finishes with discussion and conclusion in Sect. 6.

2 Tricuspid Valve Assessment

Echocardiography is a procedure used to assess the heart's function and structures. Echocardiography uses ultrasound technology to produce two-dimensional, three-dimensional and recently four-dimensional heart images. Compared to other modalities such as MRI and CT, even echocardiography is suffering with lower spatial resolution and imaging artifacts including noise and obscuration, it has gained much popularity for the assessment of heart disease as it is a noninvasive low-cost, non-ionizing, real time and safe imaging technique as well as allows interactive exploration that is available in many clinical scenarios. There are two standards techniques in echocardiogram; transthoracic (TTE) and transesophagel (TEE) where both can produce 2D and 3D echo images.

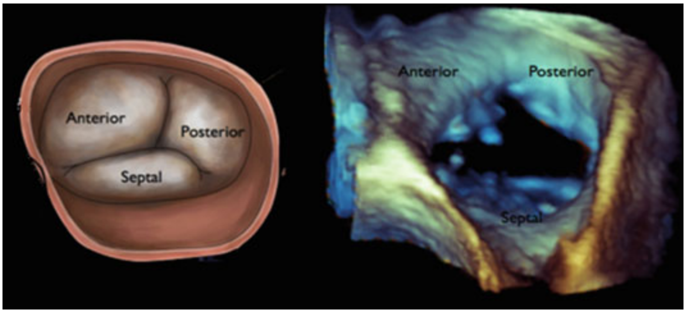


Fig. 1 Tricuspid valve (*left*) and 3D visualization in TEE (*right*)

For tricuspid valve assessment, standard assessment usually is done through 2D echocardiography. Even though this routine promising fast visualization of echo images, however all three leaflets of TV are impossible to be visualized together since only two leaflets can be captured by transducer at the same time. This limitation of visualization can be overcome by the technology of 3D echocardiography (Fig. 1). 3DTEE technique promising more comprehensive assessment and better geometric understanding on valve annulus [3–6] and has capability to obtain an “en face” view of the TV where during the cardiac cycle, all three leaflets can be seen in the same imaging plane together with tricuspid annulus.

As mentioned in [7], with regards to resolution capabilities of current transducer of 3DE, there are three main approaches to be used with 3DE; (1) parasternal short axis: best approach to obtain an “en face” view of the TV from the right atrium in which structures are imaged by the axial and lateral dimensions, (2) parasternal right ventricular inflow: to obtain intermediate result in which the axial and elevation dimensions are used, and (3) apical: worst result is expected to be obtained which uses the lateral and elevation dimensions.

3 Segmentation Techniques

There are numerous techniques recently available for medical image segmentation. One of these techniques is Level set which was introduced by Osher and Sethian on 1988 [8]. It has been efficiently used in computational schemes with its ability to handle topological changes such as merging and splitting of connected components. Shang et al. [9] uses an intensity-based level set method to segment volumetric leaflets of mitral valve from 3 dimensional ultrasound data. In this work, they presented a novel 3D object extracting approach called Region Competition based Active Contour model. Even the extraction is successfully done without oscillation at the object boundaries, it still not able to eliminate valve from tissue and upon coaptation, both leaflets anterior and posterior are not well separated. Burlina et al. [10] used the same technique together with thin tissue detector on a single frame of

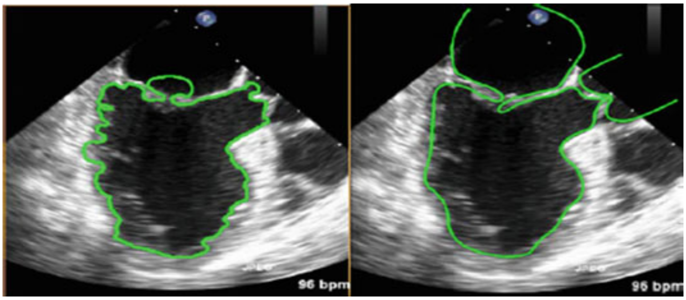


Fig. 2 Segmentation result presented in [13]. *Left* image is obtained using Level set and *right* image is based on subsequent user-refined segmentation

open mitral valve to describe geometry of valve and its surroundings. This work however, not only relies on assumption of planar annulus, but it also needs extensive user intervention in valve modification. Myocardial wall boundaries can be automatically extracted using Level set technique. This has been done by Sprouse et al. [11]. In this work, user interaction is necessary for seeding process in preliminary segmentation and to refine the preliminary segmentation result for better output. Figure 2 shows the segmentation result of their work. In 2011, the work done by Burlina et al. [12] used 3D Level set approach to find endocardial wall boundaries. The result of segmentation process assists in valve illustration and myocardium pathophysiologies. It is also can be extended as model input to patient-specific biomechanical.

Graph cut is another segmentation technique that has been used widely in medical image processing. Using echocardiography images, Graph cut approach has been used by [14] in extended leaflet surface estimation before the refinement on surface is completed using level set approach. Works done by [15, 13], use 4 dimensional graph cuts for image sequences segmentation. Shi et al. [16] adopted the same approach in their work to utilize information from multiple images that come with different spatial resolution. This approach which referred as multi-image graph cut segmentation, is performed concurrently and consistently on multiple images. Connection has been done in two ways, intra-image voxel neighbors based on intensity similarity and distance, and inter-image voxels neighbors according to their spatial overlap. Radford et al. [17] applied graph cut technique in their work on intraoperative images to automatically segment atrium and left ventrical without user interaction. Figure 3 shows the contour delineating segmentation results when valve is closed (a) and opened (b).

The work done by Cheng et al. [19] applied watershed segmentation approach to discover center of left ventricle for automatic boundary detection using echocardiography images (Fig. 5). To enhance accuracy level of watershed segmentation, Amorim et al. [20] proposed an algorithm that combines watershed with image fusion. Image fusion was used to improve preprocessing stage that contribute to better watershed segmentation accuracy. Mahmood et al. [21] used graph cut

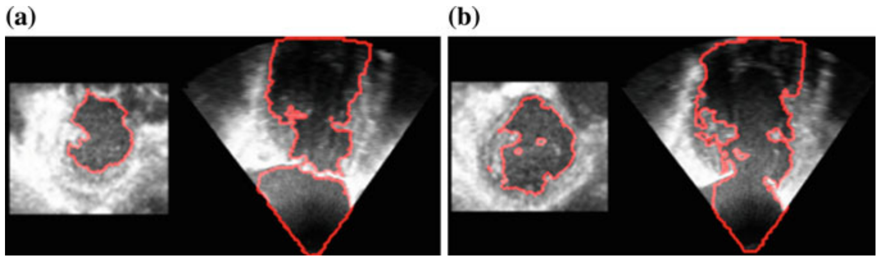


Fig. 3 Segmentation result of closed (**a**) and opened (**b**) valve presented in [18] using graph cut approach

approach in their work to propose algorithm that capable to delineate left ventricular cavity accurately in dynamic 2D-echo sequences. This algorithms combines three stages, insertion of atrioventricular barrier to separate leaft ventricle and atrium, watershed segmentation, and contour correction.

4 Tricuspid Annulus Measurement

In 2DTEE, tricuspid annulus diameter (septolateral) is best measured through mid-esophageal 4-chamber view, while in 3DTEE, the diameter measurement can be done in multiple axes with variations. In this section, the discussion on current work of tricuspid annulus measurement will be focusing more on the 3DTEE technique since compared to 2DE, the measurement of the TA diameters with 2D echocardiography consistently will underestimate the actual annulus size [18].

One of the most popular technique to calculate and modelling tricuspid annulus is through cross sectional planes. Study conducted by [22] used their customized software to trace the tricuspid annulus from real-time 3D echocardiography. Referring to Fig. 4, lateral portion of annulus was marked through manual method (a), and then (in b) the process to mark posterolateral portion (yellow point) was done by rotating the cross-sectional plane around the corner of axis at 45° intervals (red line). In Fig. 4-c, by rotating cross-sectional plane 90° from the original position, posterior portion was then marked (yellow point). Kwan et al. [23] specified the axes through transducer position to define center of the annulus in the volumetric image. The coordinate system axes for anatomical orientation of each annulus were done by (1) septo-medial (SM) axis connecting septal portion (S)'s mid-point of the TA and the medial point (M) on opposite side of the TA and (2) another orthogonal axis also named as AP axis passing through the center of the SM axis in a roughly antero-posterior (AP) direction. To divide the TA into septal and free wall parts, the authors used AP axis that crossed the center of SM axis. As illustrated in Fig. 5, the angel estimation between two vectors, (1) from anterior axis points of annulus to center of SM axis and (2) from posterior axis points of annulus to center of SM axis; produce the non-planarity of the TA.

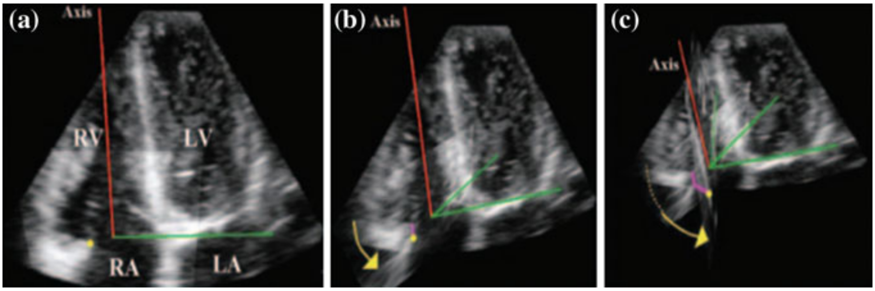


Fig. 4 Tracing tricuspid annulus from RT3DE images by [22]

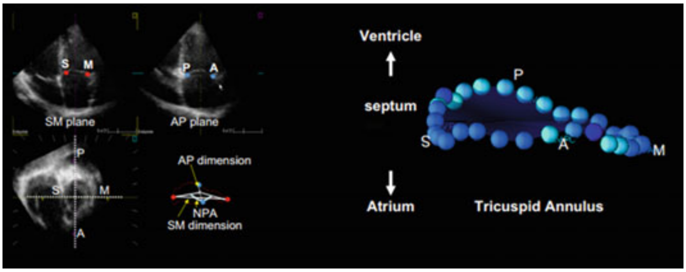


Fig. 5 *Left* Cross sectional volumetric plane and geometrical measurements of tricuspid annulus. *A* anterior point, *P* posterior point, *S* mid-point of septal annulus, *M* medial point of free wall annulus, *AP* antero-posterior, *SM* septo-medial, *NPA* non-planar angle. *Right* Reconstruction of tricuspid annulus

Another study has been conducted in [24] to study the characteristic of tricuspid valve, dynamic changes of tricuspid annulus in vivo and at the same time observe the changes in this structure in patients with dilated right hearts. As shown in [24], using Mitral Valve Quantification software, Philips (MVQ), it involved two stages which are:

1. Optimization of TV view in three orthogonal planes
2. Identification of tricuspid valve hinge point in eight rotational planes

Total parameters of tricuspid annulus inclusive of AP diameter, SL diameter, circumference, area, and height were then provided by the software. Based on these values, assessment process is done and it was repeated up to six time points of cardiac cycle which are early, mid and late of diastole and systole.

Cross-sectional plane technique also has been applied in [25] where tricuspid annulus dimensions were calculated by choosing images from end-diastole and end-systole cycle in apical four-chamber view in 2D Transthoracic echocardiography environment. With aims to find reference values for diameter of TA and to assess the extent of the differences in TA diameter measurements in relation to 2DE view and timing, study done by [26] used images from apical four chamber view

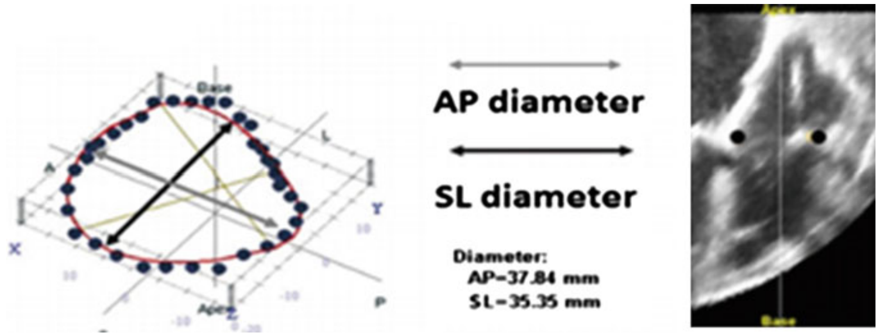


Fig. 6 Reconstruction of tricuspid annulus. Marker points were placed in each plane to mark the annulus (*right*) and all parameters then automatically calculated by the RealView® software (*left*)

together with parasternal long-axis RV inflow (PLAX) and parasternal short-axis at aortic plane (SAX) view in same environment. However, there are few limitation in this study where it is no guarantee that they have measured the same diameter throughout the process due to some factors such as complexity of tricuspid annulus geometry, translational movement of heart as well as limitations of 2D echocardiography.

In [27], the reconstruction of TV was done using (RealView®; YD, Nara, Japan) software with capability to provide views from various directions. In order to get optimum record of full volume images of the right ventricle, angle 0° was selected for the best view of tricuspid valve. As for TV annulus geometry assessment, image at mid-systolic or mid-diastolic frame was selected. Using marker points that were placed on each annulus in each plane, annulus can be recognized by observing the movement of the leaflets in all frames. The point at which the leaflet was attached was marked as the annulus. A reference line was placed at the anteroposterior (AP) axis through the anterior leaflet and at the septolateral (SL) axis from septal to lateral to define the orientation (Fig. 6).

5 Experiment

Preprocessing is important because noise is inherent in some medical images. Dispersion of the electromagnetic waves produced by the transducer make echocardiogram images suffering with speckle noise in registered image. To cater this issue, choosing perfect denoising methods is necessary in order to have an accurate model. For this experiment, we divided the processing stage into two parts. First part, Homomorphic Wavelet Filter was applied on the image and followed with Histogram Equalization (Fig. 7b). Main advantage of wavelet filter is its capability in deconstructing complex signals into basis signals of finite bandwidth, and then reconstructing them again with very little loss of information. This

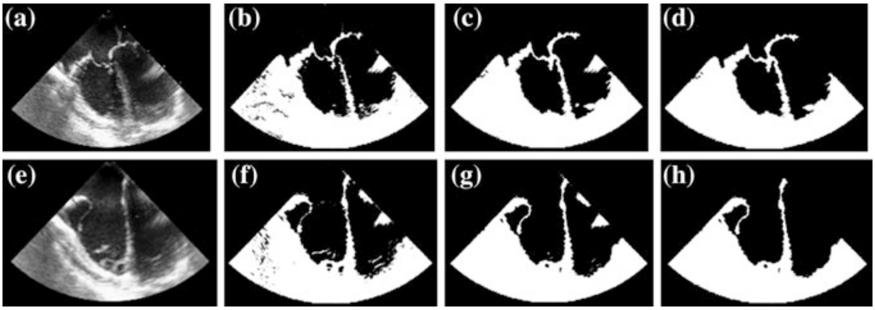


Fig. 7 Segmentation result using Level set approach on both close (**a–d**) and open (**e–h**) valve condition. **a**, **e** Original valve image. **b**, **f** Valve image after Homomorphic Wavelet Filter and Histogram Equalization were applied. **c**, **g** Valve image after morphological functions were applied. **d**, **h** Result of valve segmentation using Level set technique

technique able to denoise the particular signals far better than conventional filters that are based on Fourier transform design. Histogram equalization is a technique for contrast adjustment using histogram of the image. It has been widely used in research due to its simplicity and comparatively effective on almost all types of images. There are two types of histogram equalization, global and local. In this experiment, we applied global histogram equalization on our DICOM image to enhance the contrast of the whole image.

As for the second part of preprocessing stage, the image was then went through smoothing process using morphological functions, erosion and dilation technique (Fig. 7c). Figure 7d shows the result of level set segmentation. Using seeding process of region growing techniques, the edge of wall and leaflets were detected and then segmented to eliminate unwanted part. Both leaflets were clearly displayed without having any over-segmentation issue.

6 Discussion and Conclusion

For echocardiography image segmentation, there are numbers of techniques with high capability to produce better results. Noise, intensity, structure and texture are the main challenging constrains in image processing. We have done experiments using different techniques with the same data set in [28, 29].In this paper, the discussion and experiments are limited to few segmentation approaches that recently are widely applied in medical field. In the future, more techniques will be explored to find the best technique for echocardiography images. Tricuspid annulus measurement plays important role in surgical planning, post-surgical assessment, abnormality detection as well as decision-making procedure. Geometrical knowledge of the annulus such as its shape and size, is crucial in understanding the pathophysiology of valvular heart disease and planning optimal surgical treatment. Formed in saddle-shape, tricuspid annulus is assumed as complicated structures that needs precise and most accurate measurement for diagnosis. The challenge stands tall especially in identify the shape of annulus since it still become topic of debate among researchers. Detection and evaluation of tricuspid annulus are best done through cross-sectional plane. In next paper, we will presenting the experiment on tricuspid annulus based on this technique. Previous works in our lab towards the goal of Cardiac Intervention Environment, Cardiovascular Information System, Heart Diseases Diagnostic Systems, Computer Assisted Medical Research and 3D Medical Visualization has included the develop a new human heart vessel segmentation and 3D reconstruction mechanism under different illumination conditions [30, 31], an automatic coronary arterial tree extraction in angiograms [32], cardiac ultrasound fusion system development [33], reconstruction of Tricuspid valve using TEE echocardiography [28], Wavelet enhancement for x-ray angiography [34], image reconstruction of heart with specular reflection remover [35], 3D Multimodal Cardiac Data Reconstruction in CT Angiography [36], CT Angiography components categorization and coronary artery enhancement [37], and some surveys including review on segmentation approaches [29, 38, 39], and review on registration of cardiac images [40].

References

1. Tei, C., Pilgrim, J.P., Shah, P., Ormiston, J., Wong, M.: The tricuspid valve annulus: study of size and motion in normal subjects and in patients with tricuspid regurgitation. Circulation **66** (3), 665–671 (1982)
2. Maslow, A.D., Schwartz, C., Singh, A.K.: Assessment of the tricuspid valve: a comparison of four transesophageal echocardiographic windows. J. Cardiothorac. Vasc. Anesth. **18**(6), 719–724 (2004)
3. Fattouch, K., Castrovinci, S., Murana, G., Novo, G., Caccamo, G., Bertolino, E.C., Sampognaro, R., Novo, S., Ruvolo, G., Lancellotti, P.: Multiplane two-dimensional versus real time three-dimensional transesophageal echocardiography in ischemic mitral regurgitation. Echocardiography **28**(10), 1125–1132 (2011)
4. Gabriel, V., Kamp, O., Visser, C.A.: Three-dimensional echocardiography in mitral valve disease. Eur. J. Echocardiogr. **6**(6), 443–454 (2005)
5. Noack, T., Mukherjee, C., Kiefer, P., Emrich, F., Vollroth, M., Ionasec, R.I., Voigt, I., Houle, H., Ender, J., Misfeld, M.: Four-dimensional modelling of the mitral valve by real-time 3D transoesophageal echocardiography: proof of concept. Interact. CardioVasc. Thorac. Surg. **20** (2), 200–208 (2015)
6. Qamruddin, S., Naqvi, T.Z.: Advances in 3D echocardiography for mitral valve, 2011
7. Muraru, D., Badano, L.P.: ‘Assessment of tricuspid valve morphology and function’: ‘Textbook of real-time three dimensional echocardiography’, pp. 173–182. Springer, Berlin (2011)
8. Osher, S., & Sethian, J. A.:Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. J. Comput. phys. **79**(1), 12–49 (1988)
9. Shang, Y., Yang, X., Zhu, L., Deklerck, R., Nyssen, E.: Region competition based active contour for medical object extraction. Comput. Med. Imaging Gr. **32**(2), 109–117 (2008)
10. Burlina, P., Sprouse, C., DeMenthon, D., Jorstad, A., Juang, R., Contijoch, F, Abraham, T., Yuh, D., McVeigh, E.: Patient-specific modeling and analysis of the mitral valve using 3D-TEE. In: Book Patient-Specific Modeling and Analysis of the Mitral Valve Using 3D-TEE’, pp. 135–146 (2010)

11. Sprouse, C., Yuh, D., Abraham, T., Burlina, P.: Computational hemodynamic modeling based on transesophageal echocardiographic imaging. In: Book computational hemodynamic modeling based on transesophageal echocardiographic imaging (2009)
12. Burlina, P., Mukherjee, R., Juang, R., Sprouse, C.: Recovering endocardial walls from 3D TEE. In: Book Recovering Endocardial Walls from 3D TEE' (2011)
13. Wolz, R., Heckemann, R.A., Aljabar, P., Hajnal, J.V., Hammers, A., Lötjönen, J., Rueckert, D., and Initiative, A.s.D.N.: Measurement of hippocampal atrophy using 4D graph-cut segmentation: application to ADNI. NeuroImage, **52**(1), 109–118 (2010)
14. Schneider, R.J., Burke, W.C., Marx, G.R., del Nido, P.J., Howe, R.D.: Modeling mitral valve leaflets from three-dimensional ultrasound. In: Book Modeling Mitral Valve Leaflets from Three-Dimensional Ultrasound (2011)
15. Linguraru, M.G., Pura, J.A., Chowdhury, A.S., Summers, R.M.: Multi organ segmentation from multi-phase abdominal CT via 4D graphs using enhancement, shape and location optimization. Lect. Notes Comput. Sci. **6363**, 89–96 (2010)
16. Shi, W., Zhuang, X., Wolz, R., Simon, D., Tung, K., Wang, H., Ourselin, S., Edwards, P., Razavi, R., Rueckert, D.: A multi-image graph cut approach for cardiac image segmentation and uncertainty estimation. Lect. Notes Comput. Sci. **7085**, 178–187 (2012)
17. Juang, R., McVeigh, E.R., Hoffmann, B., Yuh, D., Burlina, P.: Automatic segmentation of the left-ventricular cavity and atrium in 3D ultrasound using graph cuts and the radial symmetry transform. In: Book Automatic segmentation of the left-ventricular cavity and atrium in 3D ultrasound using graph cuts and the radial symmetry transform, pp. 606–609 (2011)
18. Mahmood, F., Kim, H., Chaudary, B., Bergman, R., Matyal, R., Gerstle, J., Gorman, J.H., Gorman, R.C., Khabbaz, K.R.: Tricuspid annular geometry: a three-dimensional transesophageal echocardiographic study. J. Cardiothorac. Vasc. Anesth. **27**(4), 639–646 (2013)
19. Cheng, J., Foo, S.W., Krishnan, S.M.: Automatic detection of region of interest and center point of left ventricle using watershed segmentation. In: Book Automatic detection of region of interest and center point of left ventricle using watershed segmentation, pp. 149–151 (2005)
20. Amorim, J.C., dos Reis, M.D.C., de Carvalho, J.L.A., da Rocha, A.F., Camapum, J.F.: Improved segmentation of echocardiographic images using fusion of images from different cardiac cycles. In: Book Improved segmentation of echocardiographic images using fusion of images from different cardiac cycles, pp. 511–514 (2009)
21. Melo Jr, S.A., Macchiavello, B., Andrade, M.M., Carvalho, J.L., Carvalho, H.S., Vasconcelos, D.F., Berger, P.A., da Rocha, A.F., Nascimento, F.A.O.: Semi-automatic algorithm for construction of the left ventricular area variation curve over a complete cardiac cycle. BioMed. Eng. OnLine, **9**, 1–17 (2010)
22. Fukuda, S., Saracino, G., Matsumura, Y., Daimon, M., Tran, H., Greenberg, N.L., Hozumi, T., Yoshikawa, J., Thomas, J.D., Shiota, T.: Three-dimensional geometry of the tricuspid annulus in healthy subjects and in patients with functional tricuspid regurgitation a real-time, 3-dimensional echocardiographic study. Circulation, **114**(1 suppl), I-492–I-498 (2006)
23. Kwan, J., Kim, G.-C., Jeon, M.-J., Kim, D.-H., Shiota, T., Thomas, J.D., Park, K.-S., Lee, W.-H.: 3D geometry of a normal tricuspid annulus during systole: a comparison study with the mitral annulus using real-time 3D echocardiography (2007)
24. Ring, L., Rana, B.S., Kydd, A., Boyd, J., Parker, K., Rusk, R.A.: Dynamics of the tricuspid valve annulus in normal and dilated right hearts: a three-dimensional transoesophageal echocardiography study. Eur. Heart J.-Cardiovasc. Imaging, jes040 (2012)
25. Dwivedi, G., Mahadevan, G., Jimenez, D., Frenneaux, M., Steeds, R.P.: Reference values for mitral and tricuspid annular dimensions using two-dimensional echocardiography. Echo Res. Pract. **1**(2), 43–50 (2014)
26. Miglioranza, M.H., Mihăilă, S., Muraru, D., Cucchini, U., Iliceto, S., Badano, L.P.: Dynamic changes in tricuspid annular diameter measurement in relation to the echocardiographic view and timing during the cardiac cycle. J. Am. Soc. Echocardiogr. **28**(2), 226–235 (2015)

27. Nishi, H., Toda, K., Miyagawa, S., Yoshikawa, Y., Fukushima, S., Kawamura, M., Yoshioka, D., Saito, T., Ueno, T., Kuratani, T.: Tricuspid annular dynamics before and after tricuspid annuloplasty. Circulation J. **79**, 873–879 (2015)
28. Nasir, N. R. M., Kadiman, S., Rahmat, R. W. O., Dimon, M. Z., & Sulaiman, P. S. (2014, December). 3 Dimensional Reconstruction of Tricuspid Valve Using Transesophagel Echocardiography Images. In Computer Assisted System in Health (CASH), International Conference on 90–95, IEEE (2014)
29. Naziffa Raha Md Nasir, Rahmita Wirza, P. Suhaiza Sulaiman, Suhaini Kadiman, M. Zamrin Dimon.:Image Segmentation Techniques Using Echocardiography Images, The 3rd International Conference on Computer Science and Computational Mathematics, ICCSCM2014, Langkawi, Malaysia (2014)
30. Aqeel Al-Surmi, Rahmita Wirza, Ramlan Mahmod, M. Zamrin Dimon.: A new human heart vessel identification, segmentation and 3D reconstruction mechanism. J. cardiothor. surg. **9** (1):161 (2014)
31. Aqeel Al-Surmi, Rahmita Wirza, M. Zamrin Dimon, Ramlan Mahmod, Fatima Khalid.: Three dimensional reconstruction of human heart surface from single image-view under different illumination conditions. Am. J. Appl. Sci. **10**(7):669 (2013)
32. Moosavi Tayebi, R., Suhaiza Binti Sulaiman, P., Wirza, R., Zamrin Dimon, M., Kadiman, S., Khalid, F., and Mazaheri, S.: A fast and accurate method for automatic coronary arterial tree extraction in angiograms. J. Comput. Sc. **10**(10):2060–2076 (2014)
33. Mazaheri S., Sulaiman P., Wirza R., Dimon Z., Khalid F., and Moosavi Tayebi R., "Hybrid Pixel-based Method for Cardiac Ultrasound Fusion Based on Integration of PCA and DWT", in Computational and Mathematical Methods in Medicine Journal, (MMMI14) (2014)
34. Moosavi Tayebi, R., Wirza, R., Suhaiza Binti Sulaiman, P., Zamrin Dimon, M., Khalid, F., and Mazaheri, S.:Using Wavelet for X-ray Angiography Enhancement, in Proceeding of International Conference on Agricultural, Ecological and Medical Sciences (AEMS) (2015)
35. Rahmita Wirza, Aqeel Al-Surmi, Ramlan Mahmod, Fatima Khalid, M. Zamrin Dimon; Single Image Reconstruction of Human Heart Surface with Specular Reflection Remover; Advanced Computer Science Applications and Technologies (ACSAT) (2012)
36. Moosavi Tayebi, R., Wirza, R., Suhaiza Binti Sulaiman, P., Zamrin Dimon, M., Khalid, F., Al-Surmi, A., and Mazaheri, S.:3D Multimodal Cardiac Data Reconstruction using Angiography and Computerized Tomographic Angiography Registration, published in J. cardiothor. surg. (2015)
37. Moosavi Tayebi, R., Wirza, R., Suhaiza Binti Sulaiman, P., Zamrin Dimon, M., Khalid, F., Al-Surmi, A., and Mazaheri, S.: Cardiac Components Categorization and Coronary Artery Enhancement in CT Angiography, in Scopus Proceeding of International Conference on Computer Assisted System in Health (CASH) (2014)
38. Mazaheri s., Suhaiza P., Wirza R. and Moosavi Tayebi R.:Echocardiography Image Segmentation: A Survey, 2nd International Conference on Advanced Computer Science Applications and Technologies–ACSAT2013, Sarawak, Malaysia (2013)
39. Moosavi Tayebi, R., Suhaiza Binti Sulaiman, P., Wirza, R., Zamrin Dimon, M., Kadiman, S., Nurliyana Binti Abdullah, L., and Mazaheri, S.: Coronary artery segmentation in angiograms with pattern recognition techniques–A survey, in IEEE Proceeding of International Conference on Advanced Computer Science Applications and Technologies, 321–326 (2013)
40. Mazaheri s., Suhaiza P., Wirza R. and Moosavi Tayebi R.: A Review of Ultrasound and Computed Tomography Registration Approaches, The International Conference on Computer Assisted System in Health, CASH2014, Putrajaya, Malaysia (2014)

Parallel GPU-Based Hybrid String Matching Algorithm

Hakim Adil Kadhim and Nuraini Abdul Rashid

Abstract String or pattern matching is an essential part of most computer applications that are widely used in text processors, Internet-based search engines and computer security. A key concept of string matching is identifying the existence of a pattern with m characters in a much longer text string. String matching operations are basic to many algorithms such as indexing algorithms, search algorithms and bioinformatics algorithms that are used in molecular biology, adopted pattern matching concept in their programming structure. Although the mathematical and algorithmic calculations of pattern matching algorithms are simple, they are still the main reason of a large ratio of the computational load in many applications This research proposes to parallelize a hybrid string matching algorithm called Maximum-Shift algorithm, by using CUDA device (GPU) to increase its speed up during the matching process. The parallel Maximum-Shift algorithm shows slight improvement over the sequential version in terms of running time, speed up rate and percentage of performance gain.

Keywords Parallel · GPGPU · CUDA applications · String matching algorithms

1 Introduction

The millennium is the age of Petabytes. Information explosion has becomes the trends as the results of advanced technology in many area of research. The use of a single string matching algorithms in many applications has become infeasible with

H.A. Kadhim
Univerisity of Kufa, Najaf, Iraq
e-mail: hakim.kadhim@uokufa.edu.iq

N.A. Rashid (✉)
Pusat Pengajian Sains Komputer, Universiti Sains Malaysia, Minden 11800, Pulau Pinang, Malaysia
e-mail: nuraini@usm.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_102

this huge data. Therefore, merging two or more string matching algorithms has become popular in order to increase the searching performance [1]. A hybrid string matching algorithm named Maximum-Shift algorithm that exploit the advantages of three existing algorithm (Quick-Search [2], Zuh-Takaoka [3] and Horspool [4], was proposed [5]. Although Maximun-Shift algorithm improved the speed of the comparison, there are still rooms for improvement by using advancement in the parallel techniques.

The parallelism technique via the high-speed processors is used to perform multiple tasks at the same time by. As the fast calculations in the recent computers become important, parallel computing has become the dominant paradigm in computer architecture, mainly in the form of multi-core processors.

Chip-Multiprocessors (CMP) or the single-chip multi-core technology is adopted by chip manufacturers in CPU designs with multithreaded capability. A thread is an active execution sequence of instructions within a process [6]. Recently, the co-processors in multi-core processing have become the mainstream design in personal desktop and handheld computers. The notable increase of processing speed up in the machines using multi-core technique compared to those using single-core makes the multi-core computing as one of the desirable design strategies in the sophisticated machines.

A related concept to the co-processors, GPU (Graphical Processing Unit), also known as video cards, are defined as the large number of adjacent processors with massive parallel computations integrated into a single chip. It was adopted, firstly, to render graphical information of game applications and medical sciences. GPU is characterized with an efficient underlying architecture; a huge number of data-parallel cores and high memory bandwidth; and simple programming handling through sets of flexible frameworks such as NVIDIA's CUDA. CUDA uses a simple programming interface to allow programmers to execute general- purpose logic on NVIDIA GPU) [7]. Consequently, GPU device has offered dramatic speed up for a variety of general-purpose applications better than the contemporary general-purpose processors (CPU) has to offer. The most common general- purpose (non-graphical) algorithms that exploit the full advantages of GPU are Matrix Multiplication [8], Quicksort [9] and string matching applications [10].

In this paper we present a parallel hybrid Maximun-Shift algorithm which was design and implemented on a GPGPU. Section 2 discusses the works that have implemented parallel string matching algorithms followed by the presentation of our proposed algorithms in Sect. 3. Section 4 discusses on the results and analysis of the experiments.

2 Parallel String Matching

The aim of combining two or more algorithms into one is to obtain a new optimum hybrid algorithm that behaves efficiently with complex and large applications by reducing the processing time. Recently, the growing size of data especially produced by the textual applications makes the sequential versions of hybrid string matching algorithms insufficient to accommodate the fast operations with these applications [11, 12]. Therefore, a new technique is the main concern of the researchers who decided that the parallelism is the complement technique in this field. Many of the endeavors have strived to design special hardware and software architectures for efficient parallel implementation of string matching algorithms. The proposed parallel architecture discussed by Yovtis [13] was one of the hardware approach that is called team comparator. Yotvis in 1990 defined team comparator as a mechanism that has the responsibility to examine huge textual data to find certain pattern quickly within a long text. He summarized four techniques for team comparator implementations [13] and listed below:

1. Parallel comparator which consists of many comparator units; each one runs independently and eventually the results are collected together at the end of the process.
2. Associated memory is to compare the memory words with the input words and used instead of the parallel comparator.
3. Cellular array consisted of the array of logical cells; each cell represents the location of single character where all cells implement in the parallel manner.
4. A finite-state automata is composed from a table divided into smaller state tables, where the characters of input text are distributed in a logical ring form.

Designing only hardware architectures for parallel implementation of the string matching algorithms or any other algorithms is insufficient, unless they are supported by suitable parallel designs. Thus, there are many researchers who addressed the fundamental criteria of designing generic parallel algorithms; one of these works was [14]. Bernam and Paul [14] emphasized that the designing of any parallel algorithm is limited to one of the three strategies listed below:

1. Identify the intensive computations parts in the bodies of the sequential algorithms and modify them to be suitable for parallel implementations.
2. Design a purely parallel algorithm without involving any sequential part.
3. Run the similar sequential algorithm on various processors and based on conditional statements, collecting the fastest running to be as a best solution.

In addition, [14] claimed that the implementation speed of the algorithms has tightly restricted with the devices and the hardware specifications, and they could clearly summarize these limitations in five points as listed below:

- The numbers of instructions, i.e. single instruction or multiple instructions.
- Type of used processors and the rate of their availability.
- Memory management, i.e. the shared memory and distributed memory.
- The bandwidth of communication between processing units.
- Read/write mechanisms (I/O).

These assumptions for parallelizing string matching algorithms have led recently to the emergence of many works under different modern parallel architectures such as multi-core CPU and many-core GPU, and a discussed in the next subsections.

2.1 Related Work of CPU-Based String Matching Parallelism

Parallelizing the string matching algorithms by multi-core CPU owns wide popularity in many realms of computer science; one of these fields is the security applications. Hnaif et al. [12] discussed the possibility of increasing the speed of Intrusion Detection System (IDS), which detects the malicious hackers in the network, by parallelizing Quick-Search algorithm to be faster through the filtering process. He used OpenMP directives and Pthread API to test his proposed idea, which was based on studying many factors such as length of pattern and size of dataset, to assign the number of threads for parallel execution. On the other hand, parallel string matching algorithms also have a remarkable role in bio- logical applications. Therefore, Kouzinopoulos [15] presented a hybrid OpenMP/MPI parallel model by exploiting the advantages of shared and distributed memory mechanisms to parallel three types of string matching algorithms. Consequently, they were able to achieve optimum results with specific types of biological databases in their proposed model. The same authors but in another work Kouzinopoulos and Margaritis [16] proved that the way of data partitioning and the type of data are very important factors that govern the parallelization efficiency.

2.2 Related Work of GPU-Based String Matching Parallelism

Very few studies have addressed the parallelization of string matching algorithms in GPU, especially on hybrid string matching algorithm. However, in the single string matching algorithms, Giorgos et al. [17] discussed the possibility of avoiding the over-head that is associated with the data transferring to the GPU by using a special buffer with predetermined size. This buffer is used to collect the limited number of patterns each time and transfer them to the GPU device to perform the actual matching with the existing dataset. He proved that the speed up with his proposed design is 48 times faster than the CPU implementation. Similarly, Schatz and Trapnell [18] succeeded to get 35 times speed up when performing the Cmatch algorithm in parallel by using CUDA. The parallel Cmatch was implemented on sets of biological sequences. The online string matching algorithms such as Naive, Knuth-Morris-Pratt, Boyer-Moore-Horspool and Quick-Search have sequential behavior for accessing the memory. Kouzinopoulos and Margaritis [16] alleviates this behavior in GPU-based implementation by utilizing the low-latency shared memory capability in the graphic device to achieve up to 24 times speed up when compared to the serial implementation. In separate work, Tay [19] has submitted samples of CUDA implementation for three single algorithms, which are Naive, Horspool and Quick-Search algorithms. Some of these samples helped us to understand the way of building GPU kernels in this work and then creating new kernels after modifying the originals to be applicable with the hybrid algorithm. GPU-based implementation with hybrid string matching algorithms is almost non-existent, except one work submitted by Samsudin [20] and that the work is considered as the most related work to this research. Samsudin reduced the hybrid algorithm that is submitted by Abdul Rozaq [21], which consists of Quick- Search, Karp-Rabin and Two-way algorithms, to Quick-Search and Karp-Rabin to be considered as a reduced hybrid algorithm. He implemented these two hybrid algorithms on GPU device and noted that the reduced algorithm was much faster than the original algorithm because the latter suffers threads-warp divergent when it is implemented on the CUDA programming platform.

The two implementations that are submitted by Samsudin were used as the benchmark to evaluate our proposed algorithm. Hasan et al. [22] has proposed a GPU-based multi-patterns string matching algorithms to speed up the intrusion detection system.

3 Parallelizing the Maximum-Shift Hybrid Algorithm

Despite that combining two or more string matching algorithm to create an efficient hybrid algorithm is a good way to solve the string matching problem, the new algorithm is still very slow in searching. The sequential implementation nature of the hybrid algorithm is the main reason why the algorithm is slow especially with huge size datasets [11, 12]. The sequential execution is executed by CPU, which provides only one thread, also known as master thread, to execute the algorithm step by step until the task is completed. Therefore, we parallelize our proposed algorithm, Maximum-Shift algorithm, which is designed to increase its searching speed with large size dataset applications on parallel platform. The recent parallel capabilities of CUDA device in handling the non-graphic applications are considered in this work to parallelize the proposed Maximum-Shift algorithm. The next sections discussed in more detail our proposed parallel design for Maximum-Shift algorithm to be executed on GPU device.

3.1 Maximum-Shift Parallel Design for CUDA Implementation

The proposed Maximum-Shift algorithm consists of two main phases, which are the preprocessing phase and the searching phase. The preprocessing phase involved simple computations unlike the searching phase, which often deals with large size dataset to locate the occurrences of a given pattern. The searching phase is often the main reason of the high workload in the algorithm. Therefore, in this section, we present a proposed parallel design for our algorithm to parallelize the searching phase of the algorithm. It is designed to be executed on a GPU device by exploiting the CUDA standard APIs. The proposed design is virtually depicted in Fig. 1 and consists of 6 stages.

These stages can be summarized as:

- *First stage*: The CUDA program is a serial code implemented on the host processor (CPU). In this stage, the code build the Quick-Search bad character table (qsBc), which keeps the shift values that is used for pattern shifting.
- *Second stage*: In this stage a serial function is used to compute the Zuh-Takaoka bad character table (ztBc), which contains the shift values used in sliding the pattern along the text,
- *Third stage*: This stage represents the Max-Shift function that is proposed in [5]. This function is used to choose the maximum shift value between Quick- Search shift value and Zuh-Takaoka shift value. The chosen value is used to maximize the distance that the pattern is shifted along the text. This function is also implemented as a serial code within the host processor.

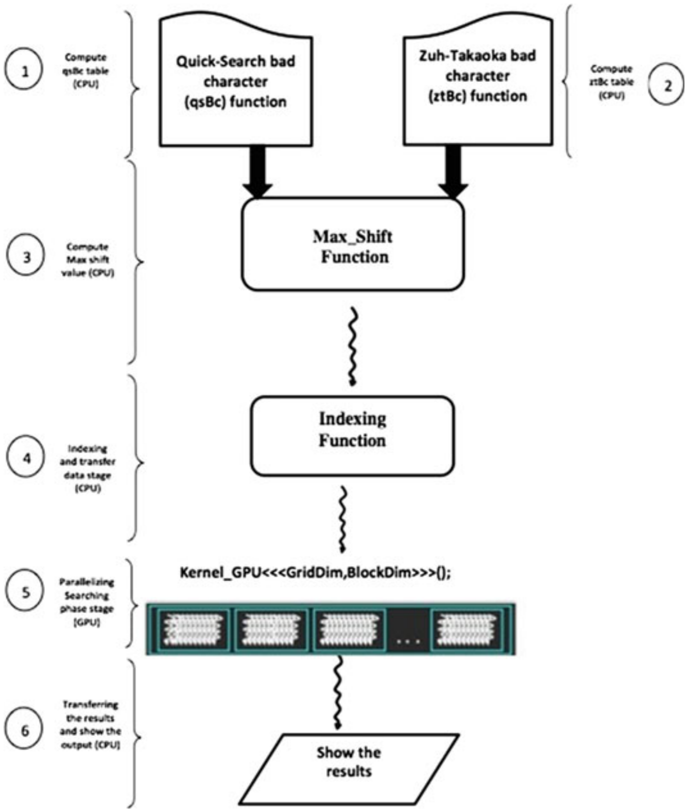


Fig. 1 The proposed design of GPU-based hybrid string matching algorithm

- *Fourth Stage*: The idea of pre-computing the skipping distance, which is discussed by (Tay, 2010), has been adopted in the CUDA-based proposed design. Skipping distance structure, which is a 1D array stores the indexes of skipping characters along the text based on the values of a single bad character table. The proposed idea in this stage is to modify the skipping distance function by considering the shift values of two bad character tables (qsBc and ztBc) instead of using the values of one table. The Max- Shift function, which is addressed in the stage three, is invoked within the new modified function to compute the maximum shift value to be considered as a skipping index. All skipping indexes are stored in a 1D array regardless of a match or a mismatch occurrence, and each valid index is considered as a CUDA thread's ID used in the CUDA kernel function that addresses in the next stage.
- *Fifth Stage*: This stage represents the part of the proposed hybrid algorithm that is executed in the device processor (GPU). As mentioned previously, the high workload part in our hybrid algorithm occurred in the searching phase. Therefore, this phase is considered to be as a kernel function that will be implemented on the processor (GPU) in parallel.
- *Sixth Stage*: This stage is the final step of CUDA-based design of the proposed hybrid algorithm. After the data is processed in the device, the results of the parallel processing are transferred from device memory to the host memory to be used with other operations in the host code.

4 Implementation and Evaluation for Parallel Hybrid String Matching Algorithm

English Text are used to test the parallel algorithm. The data size ranges from 100 to 400 Mb. The performance metric used is the execution time, speed up and percentage of performance gains. The implementation was tested on the GPGPU server in the School of Computer Sciences (biruni.cs.usm.my). GPGPU server consists of AMD Phenom 2.6 GHz Quad-Core processor with 4 GB Random Access Memory (RAM). AMD processor used to test the sequential algorithm and also used to invoke the parallel kernel function to implement in the GPU device. On the other hand, GPU card of GPGPU server is GeForce GTX 260 and consists of:

- CUDA Capability version is 1.3.
- Total global memory is 890 MB (939327488 Byte).
- Multiprocessor count is 27.
- Max threads per block are 512.
- Max thread dimensions are (512, 512, and 64).
- Max grid dimensions are (65535, 65535, and 1).

In the GPGPU server, the operating system is Ubuntu Linux 8.04 of 64-bit with NVIDIA CUDA Toolkit v2.2 and GNU C Compiler (GCC) v4.2.4. We use secure shell (SSH) software to access to the GPGPU server remotely, write the code, compile and run the programs.

5 Results and Discussion

This section addresses the results of the parallel algorithm compared to the sequential version of the algorithm. The length of the pattern used to evaluate the algorithm is 4. To get stable result the experiment was repeated five times with each pattern selected randomly from the database. Figure 2 shows the speedup and performance gain of the parallel algorithm. Unfortunately, the parallel running time slightly outperformed the running time of the sequential algorithm. This is because CUDA device is designed especially for highly intensive and complex arithmetic computations, while the proposed hybrid algorithm has simple arithmetic operations represented by finding the number of attempts and the number of character comparisons. On the other hand, the proposed parallel algorithm incurs high overhead in terms of using CUDA threads, where the number of unused threads is greater than the number of used thread and that, in turn, considers as a reason that causes the poor performance for the proposed parallel algorithm. The underlying structure of the proposed hybrid algorithm is also affected directly on the performance of the parallel algorithm, where the hybrid algorithm composes from many flow control instructions (if, for, while). These instructions cause heavy divergence in the threads of the same warp and then obtaining limited performance for the proposed parallel algorithm.

The high overhead of exchanging the data between host memory and device memory in CUDA program is often deplored with the intensive arithmetic computations and the complex problems, which the CUDA device is expected to process. The simple operations that are part of the proposed parallel algorithm make the overhead of the communication between the host and the device a serious challenge. The communication overhead reduces the algorithm performance.

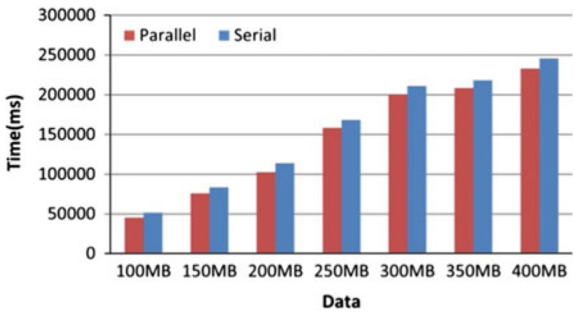


Fig. 2 The elapsed time of parallel and sequential hybrid string matching algorithm

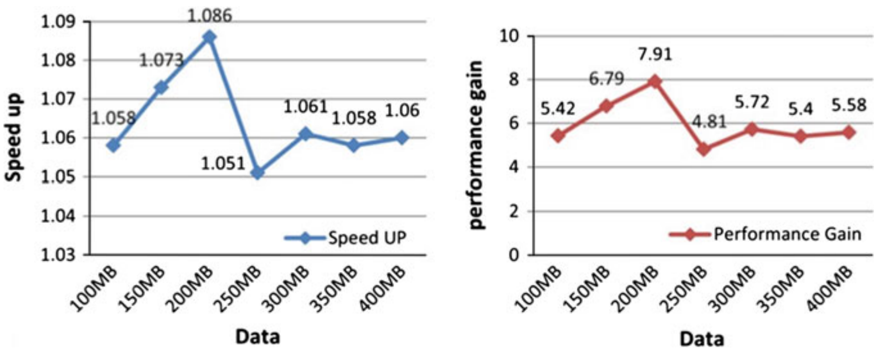


Fig. 3 Speed-up and performance gain of the GPU-based hybrid string matching algorithm

We notice clearly the effect of the mentioned CUDA program limitations on the execution time of different sizes starting from 100 MB until 400 MB, and as explained in Fig. 2. The weakness in the execution time of the proposed parallel algorithm is obviously reflected on the speed up of the algorithm itself. The proposed parallel algorithm achieved 1.086 times as the highest speed up and 7.91 % as the highest percentage of the performance gain for English text data type (see Fig. 3).

6 Conclusion

In this paper, we have presented a parallel hybrid string matching algorithm implemented on GPGPU. The experiments show that there is a improvement in the running time of the algorithm. However, because of the nature of the hybrid algorithm, the improvement is not high. In the future, we proposed that the design of the sequential hybrid algorithm is modified to make it more appropriate for parallel implementation.

Acknowledgement We would like to acknowledge Universiti Sains Malaysia for supporting this publication under the RU grant Pkomp/1001/817065 and School of Computer Sciences, USM for the supporting the publication of this work.

References

1. Chen, Y.: A new algorithm for subset matching problem. J. Comput. Sci. **3**(12): 924–933 (2007). doi:10.3844/jcssp.2007.924.933 (ISSN 1549-3636)
2. Sunday, D.: A very fast substring search algorithm. Commun. ACM **33**(8), 132–142 (1990)
3. Feng, Z.R., Takaoka, T.: On improving the average case of the Boyer- Moore string matching algorithm. J. Inf. Process. **10**(3), 173–177 (1987) (ISSN 0387-6101)

4. Horspool, R.N.: Practical fast searching in strings. Softw. Pract. Exp. **10**(6), 501–506 (1980)
5. Kadhim, H.A., Abdul Rashid, N.A.: Maximum-shift string matching algorithms. In: 2014 International Conference on Computer and Information Sciences (ICCOINS), pp. 1, 6 (3–5 June 2014)
6. Quammen, C., Introduction to Programming Shared-Memory and Distributed-Memory Parallel Computers, ACM Crossroads (2002)
7. Che, S., Boyer, M., Meng, J., Tarjan, D., Sheaffer, J.W., Skadron, K.: A performance study of general-purpose applications on graphics processors using CUDA. J. Parallel Distrib. Comput. **68**(10), 1370–1380 (2008)
8. Ohshima, S., Kise, K., Katagiri, T., Yuba1, T.: Parallel processing of matrix multiplication in a CPU and GPU heterogeneous environment. Graduate School of Information Systems, the University of Electro-Communications 1-5-1, Chofugaoka, Chofu-shi, Tokyo, Japan (2006)
9. Cederman, D., Tsigas, P.: A practical quicksort algorithm for graphics processors. In: ALGORITHMS—ESA 2008 Lecture Notes in Computer Science, vol. 5193/2008, Springer Berlin, pp. 246–258 (2008)
10. Alexander, G.: Research into GPU accelerated pattern matching for applications in computer security. Department of Computer Science and Software Engineering University of Canterbury, Christchurch, New Zealand (2009)
11. Park, J.H., George, K.M.: Parallel string matching algorithms based on dataflow. Computer Science Department, Oklahoma State University, Stillwater, OK 74078, USA (1999)
12. Hnaif, A., Alhalaiqah, M., Abouabdalla, O., Ramadass, S., Kadhum, M.: Parallel quick-search algorithm to speed packet payload filtering in NIDS. J. Eng. Sci. Technol. 220–230 (2009)
13. Yovtis, M.C.: Advances in Computers, vol. 30, United Kingdom edition published by Academic Press (August 11, 1990), Hardcover: 352 pages
14. Bernam, K.A., Pual, J.L.: Algorithms: sequential, parallel and distributed, Thomson Course Technology.(2005)
15. Kouzinopoulos, C.S., Michailidis, P.D., Margaritis, K.G.: Parallel processing of multiple pattern matching algorithms for biological sequences: methods and performance results. Syst. Comput. Biol. Bioinf. Comput. Model. (2011)
16. Kouzinopoulos, C.S., Margaritis, K.: Parallel implementation of exact two dimensional pattern matching algorithms using MPI and OpenMP. In: 9th Hellenic European Research on Computer Mathematics and its Applications Conference (2009)
17. Giorgos, V., Michalis, P., Spiros, A., Evangelos, M., Sotiris, M.: Regular expression matching on graphics hardware for intrusion detection, pp. 265–283. Springer Berlin (2009)
18. Schatz, M., Trapnell, C.: Fast Exact String Matching on the GPU. University of Marylan (2007)
19. Tay, R.: Sample implementations of exact string matching algorithms in CUDA. URL: http://code.google.com/p/exactstrmatchgpu/ (2010)
20. Samsudin, M.N.B.: Adoption of two hybrid string matching algorithms on general purpose graphical processing unit, Masters thesis, Pusat Pengajian Sains Komputer, Universiti Sains Malaysia (2010)
21. Abdul Rozaq, A.A.: Fast hybrid string matching algorithm using message passing programming model, Master thesis, Univesiti Sains Malaysia. Pust Pengajian Sain Komputer, Dissertations. Thesis (M.SC.), Universiti Sains Malaysia (2009)
22. Hasan, A.A., Rashid, N.A.A., Abdulrazzaq, A.A.: GPGPU based hybrid multi-pattern algorithm design for high-speed intrusion detection system. In: 2014 IEEE International Conference on Control System, Computing and Engineering (ICCSCE), pp. 141–146 (28–30 Nov. 2014)

Multi Spatial Resolution for Image Spam Filtering

Nor Azman Mat Ariff, Azizi Abdullah and Mohammad Faidzul Nasrudin

Abstract Image spam filtering is a challenging task because spammers are constantly creating new tricks and making anti-spam filters ineffective. To overcome these problems, many new techniques have been developed. Most of these techniques use a basic bag-of-features representation where global approach is used to extract the feature. This global representation leads to limited descriptive power for the features due to neglecting the spatial information, which can create powerful cues for classification tasks. Spatial Pyramid Representation (SPR) is one of the most effective and widely used image processing techniques that embedding spatial information into a feature. Inspired by this technique, we propose Multi Spatial Resolution (MSR) approach, which transform the image to Base-64 encoding, divided the Base-64 encoding into a sequence of increasingly finer grids on different pyramid level. The n-gram technique is used to extract the features from each grid cell or partition. Frequency histogram for each partition is concatenated to form a single feature vector. The experiments were conducted on Dredze and SpamArchive data sets at four different resolutions using SVM classifier. The results show that MSR increased the classification performance compared to global approach.

N.A.M. Ariff (✉)
Faculty of Information and Communication Technology, Universiti Teknikal Malaysia, Melaka, Malaysia
e-mail: nazman@utem.edu.my

A. Abdullah · M.F. Nasrudin
Faculty of Information Science and Technology, Universiti Kebangsaan, Bangi, Malaysia
e-mail: azizi@ftsm.ukm.my

M.F. Nasrudin
e-mail: mfn@ftsm.ukm.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_103

1 Introduction

Currently, e-mail users are bombarded by messages from spammers promoting products and services to a point where these messages have become a highly irritating nuisance. Spam is information or content that is inappropriate, unsolicited, repeated and irrelevant, and yet spam is hardly avoidable in our mail messages, search results, blogs, forums, social communities and product reviews [1]. Spam causes several problems for e-mail systems and users. Spam not only misuses network resources but also affects computational power [2], storage capacity of servers, the credibility and reliability of e-mail services [3]. Apart from these considerations, spam reduces productivity because significant amount of time are wasted in vetting spam e-mails. Spam e-mails may also include malware and when combined with botnets, spammers are able to launch large scale spam campaigns causing traffic overload that leads to considerable economic loss [4].

In the early days, spam was in the form of text. In response, many text-based anti-spam approaches were appropriately used, such as Bayesian filters and Support Vector Machine (SVM) filters [5, 6]. These types of filters have been effective, and most of the present day text-based spam e-mails still find it difficult to pass these filters. In late 2005, spammers modified their ways by representing all text messages as images known as image spam. Initially, Optical Character Recognition (OCR) was used to analyze text information embedded in images and process this text information as normal text. Spammers then responded by introducing various obscuring techniques like wavy text, random noise, blurring of text outlines, colorful background and animation. OCR is not the best choice for image spam filtering because OCR is vulnerable to spammer tricks as well as a computational cost is quite high [7]. Researchers then proposed content-based filters that based on color, edge, properties, texture, and layout. In addition, a number of header-based filters were proposed to detect e-mails with exploited header field. The method in [8] used Base-64 encoding and process it as normal text processing.

Most of these anti-spam techniques use basic bag-of-features as these features are effective and efficient in filtering spam. However, the use of basic bag-of-features may lead to limited ability to describe features, as spatial information is not included. Thus, the techniques cannot take advantage of the spatial layout of the features, which can improve classification tasks. The features can be represented more precisely if all information about spatial layout of the features is considered [9]. Spatial Pyramid Representation (SPR) is a method that embeds global and local spatial information into a feature [10]. SPR has been used for two-dimensional information (scene and object recognition) rather than one-dimensional information such as text classification. SPR uses a partitioning scheme where the image is repeatedly divided into several partitions at increasingly finer resolutions. Features are extracted locally from each partition and combined to form a single feature vector. Therefore, by applying SPR to existing anti-spam methods, spatial information will be embedded to the features.

In this paper, we apply an extension to the bag-of-features to existing anti-spam techniques. We choose features as suggested by Xu et al. [8] in our experiment. The author uses n-gram analysis over Base-64 encoding and generates a feature vector from character n-grams. Previous research has shown that SPR implemented in two-dimensional image space has successfully improved classification performance. Thus, we want to investigate the performance of features with one-dimensional information when spatial information is considered.

The original aspects of our work is we compare the effectiveness of n-gram representation at multi-resolution levels to the global features. The rest of the paper is organized as follows: related work is discussed in Sect. 2. In Sect. 3, describe our proposed method, MSR. Evaluation of our experiment is contained in Sect. 4, and Sect. 5 concludes the paper.

2 Related Work

2.1 Spatial Pyramid Representation

In image processing community, object recognition is still a challenge and many approaches have been applied to increase the detection rate. One of the key elements that impacts the recognition performance is feature representation. Feature representation can be classified into global and local features. A global feature can be extracted from the whole image. Typically, this feature is represented in equal length feature vector such as edge direction, grayscale, and color histogram [11, 12]. Previous studies have demonstrated that the used of global image features in object recognition have shown promising results. However, global representations are sensitive to image conditions such as occlusions, clutters, image noise or spatial variation of objects in the image [12]. A local feature can be extracted from local regions within an image using saliency-based approach. This feature is claimed more robust and provide a representation that invariant to image transformations such as scale, rotation, viewpoint, and illumination. The main advantage of this representation is it efficiently matching local structures between images which led to reliable detection the same object in different images.

One of the popular approach that used local feature is pyramid match kernel introduced by Grauman and Darrell [11]. This study has proposed the pyramid matching to find approximate matching between two sets of vectors in a highly dimensional feature space. Although the pyramid match kernel allows precise matching between two sets of vectors, one major drawback of this approach is that it abandons spatial information between features [9]. The benefits of embedding spatial information into a local feature have been proven to generate more reliable and increase discriminative power than matching individual features. Thus, Lazebnik et al. [9] introduced the spatial pyramid representation that considers spatial information. Spatial pyramid representation uses a fixed partitioning scheme where the process involves repeated subdivision of an image on each pyramid level.

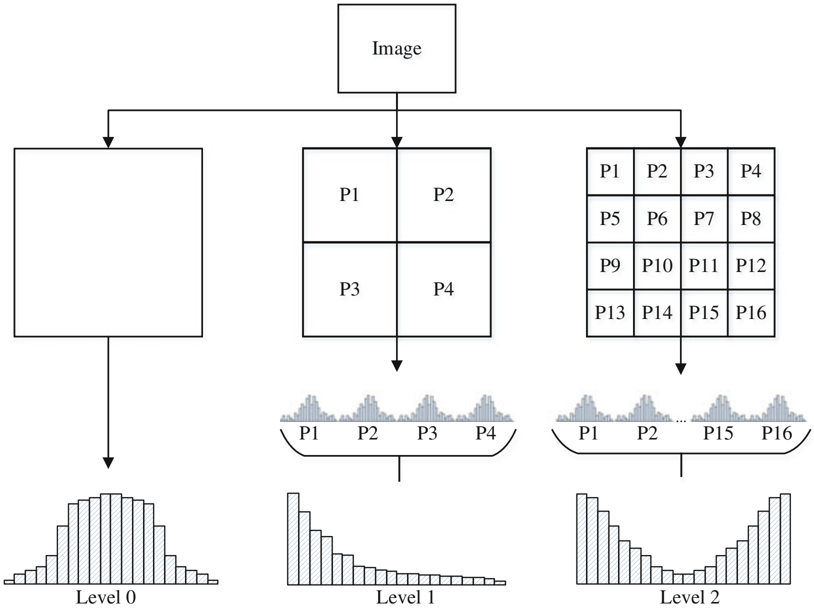


Fig. 1 The process of spatial pyramid representation

Histogram of features are generated from all of the grid cells or partitions as illustrated in Fig. 1. SPR usually processes the image at most at level 2 which produce 16 partitions. Features are extracted from each partition on each pyramid level. Refer to Fig. 1, each pyramid level produces different histogram distributions which generate different representation for the image.

This approach is based on the idea that not all objects can be represented by only global or single level representation. The assumption is that there are cases where the optimal representation of certain objects can be a global scheme, but some objects are better represented using different levels of resolution. For example, a category such as a bicycle is described better at level 0, a car at level 1, a truck at level 2 and a train by a combination of levels 0 and 1. Several studies have revealed that by combining multiple levels, the recognition performance improves compared to a single level [9, 11, 12].

3 Multi Spatial Resolution

A method of image spam filtering inspired by SPR called Multi Spatial Resolution (MSR) is proposed in this section, as shown in Fig. 2. The method mainly includes seven modules: image to Base-64 conversion, multi spatial resolution partition, n-gram feature extraction, feature matching, feature vector generation, normalization, and classification.

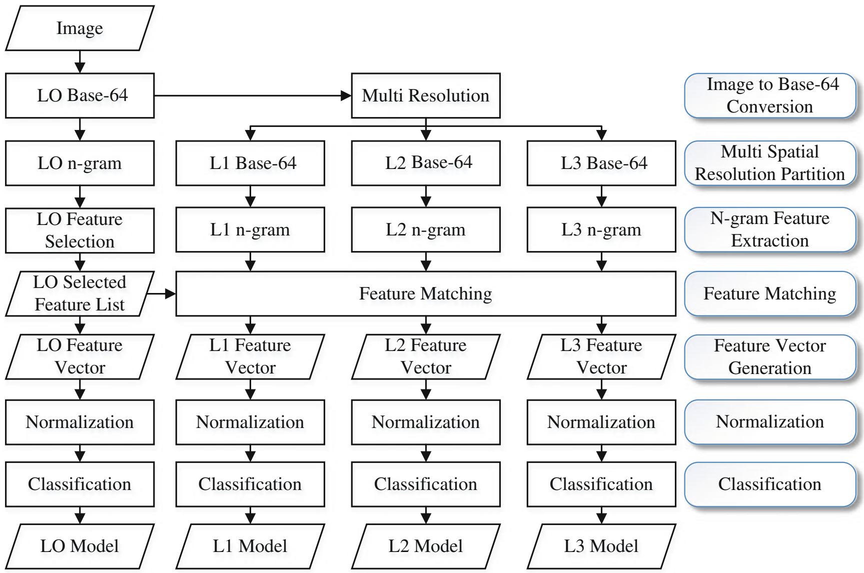


Fig. 2 The framework of multi spatial resolution

3.1 *Image to Base-64 Conversion*

Each image file format has its own file organization structure. For example, the structure of the GIF89a image file has 4 main blocks, header and color table information, extension information, image data, and trailer. Each block has its own field and size. An e-mail containing an image normally uses Base-64 as the *content-transfer-encoding* value as it was designed to represent 8-bit non-text and binary data. Base-64 has 64 characters and converts three octets of binary into four character codes. Before an image is sent over the internet, the image will be coded and the bits for GIF files will be converted to Base-64 encoding [8]. This conversion transformed the image that can be represented using two-dimensional into one-dimensional text representation. This transformation decreases the computational complexity.

3.2 *N-Gram Feature Extraction*

N-gram frequency is globally (L0) and locally (L1-L3) extracted. Base-64 has 65 characters, including one padding character that is "=". Thus, it generates 65^n number of features can be increased drastically with the increase of n. For example, if n is 3, the number of features generated is 274,625 and 17,850,625 if n is 4. It is not practical if all of these n-gram features are used as input features to feed the classifier since many of n-gram features exist in both classes (spam and legitimate) and have less discriminator power. Based on this reason, only a limited set of n-gram features need to be identified before it is use in classification phase.

3.3 Feature Matching

In this phase, Chi-Square feature selection is applied to select the best n-gram features. Chi-Square will rank n-gram features according to their individual relevance. Feature selection is performed at Level 0 only. Occurrence of each n-gram feature at certain partition is recorded if the n-gram feature in certain partition is matched with the ranked features that generated from Chi-Square feature selection.

3.4 Feature Vector Generation

Only a relevant n-gram that identified in the feature matching phase is used to develop a feature vector. Frequency histogram obtained at feature matching phase is used to generate feature vector. The bag-of-words model is applied, where the frequency of each n-gram is used as a feature for training a classifier. Frequency histogram for each partition on each level are concatenated to form a single feature. The size of a feature vector is highly dependent on the number of features. For example, if the combination of 1000 features and level 0 is used, the feature vector size is 1000 as the number of features. In contrast, 1000 features with level 1 will double the vector space size to 2000 as the number of features.

3.5 Normalization

Many researchers have argued that to get better classification performance, the feature vectors must go through the process of normalization. The purpose of feature normalization is to avoid the domination of greater numeric ranges over smaller numeric ranges in feature space. Term frequency (tf) weighting is applied, and after that, tf is normalized to the interval [0, +1] by using min-max normalization.

3.6 Classification

The SVM algorithm is used to train and classify the test image. SVMs will find a hyper plane that separates the two classes of data with as wide a margin as possible.

4 Experimental Evaluation

In this section, we evaluate our proposed method: a performance comparison of global approach (single level or level 0) versus multi-resolution levels. We begin by describing an experimental setup.

4.1 Experimental Setup

Data Set. Building a proper data set is difficult due to the nature of e-mails because they are totally personal, especially legitimate images. Nevertheless, some researchers have built their personal data set and made it available for research communities in their literature. There are a number of openly accessible data sets that are often used such as Dredze and SpamArchive data sets. Among these, to the best of our knowledge, the Dredze data set is the only one that contains both spam and legitimate images. For this reason, many researchers prefer Dredze's data set, while there are some who use legitimate images from their private e-mails, e-mails received by a limited group of people as well as the Google Images engine. Our experiments were conducted on two different data sets. The first data set is Dredze's spam and legitimate data set namely the personal Dredze data set while the second data set is SpamArchive spam combined with Dredze's legitimate data set (also known as SpamArchive and Dredze data set). We obtained the SpamArchive data set from Dredze's website as SpamArchive.org has been shut down. Dredze's data set contains 3297 spam and 2020 legitimate images while SpamArchive consists of 15,090 spam images. However, only 3209 spam and 1828 legitimate images from Dredze's data set and 13,745 spam images from the SpamArchive data set were processed by our system.

Evaluation. The evaluation of our proposed method was performed using 50 repeated random sub-sampling validation. This validation method was chosen because of an unbalanced data set, especially the SpamArchive data set that consists of 13,745 spam images and 1828 legitimate images. We expected that the classifier would learn more spam than legitimate if we used k-fold cross validation. For each iteration, the data set was randomly divided into 2000 training set images (1000 images for each class) and 1000 test set images (500 images for each class). The experiment was repeated 50 times, while the accuracy was calculated by averaging the results from all 50 runs.

The SVM classifier is used to classify the images. Considering that only two classes in the data set were to be tested, the binary classification approach was used. A variety of SVM-based tools can be used to do the classification task. Among these SVM-based tools, one of the most widely used is LIBSVM as introduced by [13]. Although our task was to solve an image classification task, which typically uses visual features, we considered our problem as a text classification problem for the reason that every image was represented by an n-gram that was extracted from Base-64 encoding. Thus, Java Software LibLinear 1.92 was chosen instead of LIBSVM as previous studies have shown that LIBLINEAR is very efficient for training large-scale problems.

4.2 Classification Results

This paper reports results on a 4-gram with 1000 as the number of features. An accuracy measurement was conducted to evaluate the performance of the MSR approach. The best result is illustrated with bold characters. Table 1 shows the classification result on four different levels of the MSR. The results show that accuracy for each level increased in proportion to the increase in resolution. We report the performance using the mean and standard deviation (SD) to verify the significance of the classification results obtained. Table 1 shows that the MSR approach outperformed the single level representation for both data sets.

The highest accuracy was achieved at level 3, with values of 95.04 and 93.56 % for Dredze and SpamArchive data sets, respectively, indicating that MSR can be applied to one-dimensional information. Table 2 presents further analysis for both data sets where naïve approach is used by combining all feature vectors (L0-L3) to form a very large single feature vector. As Table 2 shows, there is a slight increase in accuracy compared to level 3. There is a clear trend of increasing in performance when the number of feature vector is large. Thus, we can conclude that MSR has captured spatial information among Base-64 encoding sequences. The result for the significance test between each level for both data sets was considered statistically significant.

Table 1 Classification accuracy (mean and standard deviation) of single and multiple levels for two different data sets

Corpus	Level 0 (%)	Level 1 (%)	Level 2 (%)	Level 3 (%)
Personal Dredze	92.33 ± 0.97	94.63 ± 0.78	94.75 ± 0.77	**95.04 ± 0.75**
SpamArchive and Dredze	91.93 ± 0.91	92.29 ± 0.83	92.93 ± 0.92	**93.56 ± 0.87**

Table 2 Classification accuracy (mean and standard deviation) for Naïve approach

Corpus	Naïve approach (%)
Personal Dredze	95.20 ± 0.78
SpamArchive and Dredze	93.63 ± 0.80

5 Conclusion

In this paper, the researchers have proposed Multi Spatial Resolution (MSR) approach to improve the classification performance of image spam filtering. MSR, inspired by SPR, can capture spatial information of one-dimensional features (specifically Base-64 encoding) to enrich feature description. Our finding notes that MSR improved the n-gram representation where classification performance at multi spatial resolution levels outperforms global or a single-level features.

Acknowledgments The first author wants to thank to Universiti Teknikal Malaysia Melaka (UTeM) and Ministry of Higher Education Malaysia which funded this Ph.D research.

References

1. Hayati, P., Potdar, V.: Evaluation of spam detection and prevention frameworks for email and image spam. In: Proceedings of the 10th International Conference on Information Integration and Web-based Applications and Services—iiWAS'08, p. 520 (2008)
2. Blanzieri, E., Bryl, A.: A survey of learning-based techniques of email spam filtering. Artif. Intell. Rev. **29**(1), 63–92 (2009)
3. Soranamageswari, M., Meena, C.: Statistical feature extraction for classification of image spam using artificial neural networks. In: 2010 Second International Conference on Machine Learning and Computing, pp. 101–105 (2010)
4. Al-duwairi, B., Khater, I., Al-jarrah, O.: Texture analysis-based image spam filtering. In: International Conference for Internet Technology and Secured Transactions (ICITST), pp. 288–293 (December 2011)
5. Sahami, M., Dumais, S., Heckerman, D., Horvitz, E.: A Bayesian approach to filtering junk e-mail. In: Learning for text categorization: papers from the 1998 workshop. AAAI Technical Report WS-98-05 (1998)
6. Kunlun, L., Houkuan, H.: An architecture of active learning SVMs for spam. In 6th International Conference on Signal Processing, vol. 2, pp. 1247–1250 (2002)
7. Attar, A., Rad, R.M., Atani, R.E.: A survey of image spamming and filtering techniques. Artif. Intell. Rev. **40**(1), 71–105 (2011)
8. Xu, C., Chen, Y., Chiew, K.: An approach to image spam filtering based on Base64 encoding and N-gram feature extraction. In: 2010 22nd IEEE International Conference on Tools with Artificial Intelligence, pp. 171–177 (Oct. 2010)
9. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 2169–2178 (2006)
10. Harada, T., Ushiku, Y., Yamashita, Y., Kuniyoshi, Y.: Discriminative spatial pyramid. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1617–1624 (2011)
11. Grauman, K., Darrell, T.: The pyramid match kernel: discriminative classification with sets of image features. In: Tenth IEEE International Conference on Computer Vision (ICCV'05), vol. 1, 2, pp. 1458–1465 (2005)
12. Abdullah, A., Veltkamp, R.C., Wiering, M.A.: Spatial pyramids and two-layer stacking SVM classifiers for image categorization: a comparative study. In: International Joint Conference on Neural Networks, IJCNN, pp. 5–12 (2009)
13. Chang, C.-C., Lin, C.-J.: LIBSVM. ACM Trans. Intell. Syst. Technol. **2**(3), 1–27 (2011)

Correlation Feature Selection Analysis for Fault Diagnosis of Induction Motors

Thanaporn Likitjarernkul, Kiattisak Sengchaui, Rakkrit Duangsoithong, Kusumal Chalermyanont and Anuwat Prasertsit

Abstract This paper presents a feature selection method for stator winding fault analysis of induction motors by using a Correlation-based Feature Selection (CFS) method. The 14 original motor parameters are selected from the feature selection method with various searching approaches. The classification efficiency of optimal features obtained from the feature selection method is compared with results from the feature extraction method and the original features. In our experiment, we employ a 2.2 kW delta-connected motor which drives a dc generator as a load. The experimental results demonstrate that 4 common selected features for stator winding fault analysis of induction motors are a percent of load *(%Load*), a power factor (*pf*), a negative sequence voltage (V_n), and a negative sequence impedance (Z_n). The accuracy of the classification using this feature subset is higher than using all original features for three classification methods.

Keywords Feature selection · Induction motor · Stator winding faults

T. Likitjarernkul (✉) · K. Sengchaui · R. Duangsoithong · K. Chalermyanont · A. Prasertsit
Faculty of Engineering, Department of Electrical Engineering, Prince of Songkla University, Hat Yai, Songkhla 90112, Thailand
e-mail: l.thanapond@gmail.com

K. Sengchaui
e-mail: ak.kiattisak@hotmail.com

R. Duangsoithong
e-mail: rakkrit.d@psu.ac.th

K. Chalermyanont
e-mail: kusumal.c@psu.ac.th

A. Prasertsit
e-mail: anuwat.p@psu.ac.th

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_104

1 Introduction

Electrical motors are critical equipments for any machine in process industries. Motor deteriorations or any fault occurred since rotors and stators have been forced by electromagnetic field all the time. Moreover, mechanical damages caused by thermal and electrical stresses have an effect to the performance and the lifetime of motors.

About 37 % [1] of induction motor faults are stator winding faults due to the deterioration of the winding insulation from contamination of oil, humidity, and sewage. They impact on opening or shorting one or more circuits of windings.

Current and voltage signals of the induction motor contain information of stator winding faults, and they are widely used to detect and locate stator winding faults in various methods. The Motor Current Signature Analysis (MCSA) method is one of the most frequently used methods to analyze the motor fault by identifying stator current spectrums in abnormal harmonics [2, 3]. In addition, the Extended Park's Vector Approach (EPVA) is applied to analyze EPVA signatures by identifying a spectral component at twice the fundamental supply frequency [4]. Motor sequence components (i.e. negative and zero sequence components of the current) are also used for stator fault diagnosis [5–8]. Moreover, modeling and simulation studies can provide useful information about the electric behavior of the motors, and they relate to the analysis of the presence of the internal fault in the stator windings [9, 10]. Other techniques, such as an instantaneous angular speed technique, temperature monitoring, air-gap torque monitoring, magnetic flux monitoring, noise/acoustic noise, induced voltage monitoring, surge testing, gas analysis, and partial discharge [11] are also used in order to diagnose stator winding faults.

Currently, some or all electrical features of induction motors mentioned previously are used for the motor fault analysis and detection. However, using all or inappropriate electrical features will increase the complexity of the system and the stodgy storages. Moreover, they may not be able to classify the motor faults correctly. Consequently, the feature selection should be required in as preprocessing in order to reduce original features and extract the appropriate features. Basically, there are two methods to reduce the feature dimension. One is the feature selection and the other is feature extraction. The feature selection can be used to choose optimal features from original features to remove irrelevant and redundant of original features and also decreases the complexity of the system. The feature extraction is another method to reduce a number of features by transforming original features to lower dimensional spaces. The Principal Component Analysis (PCA) [12] is one of the example of the feature extraction. The PCA reduces the dimension of features without eliminating the signal information using the principal component. The proposed fault diagnosis system is shown in Fig. 1.

This paper presents the use of the correlation feature selection for selecting electric features. Optimal electric features are obtained by the feature selection for stator winding fault analysis of induction motors. The accuracy of motor fault classification obtained by the optimal features from the feature selection method is

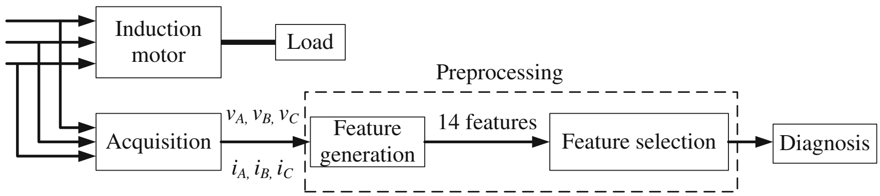


Fig. 1 The proposed fault diagnosis system

compared with one that obtained from the feature extraction method and original features. Various theories related to electrical features calculation are reviewed in Sect. 2. In Sect. 3, the feature selection methods are explained. Section 4, an experimental setup is described and the experimental results of the feature selection analysis are showed and discussed in Sect. 5 and conclude in Sect. 6.

2 Feature Generation

The electrical features were calculated by motor current and voltage signals. These features are popular features for the stator winding fault diagnosis. These signals were fed to the preprocessing block to reduce a number of features, and optimal features are searched and selected for such faults. The original features in this paper can be obtained from the following methods.

2.1 *Symmetrical Components*

The positive and the negative sequence components of the induction motor are normally used to indicate the stator faults. Figure 2 shows a workflow for extracting the positive (I_p) and the negative (I_n) sequence components from three-phase

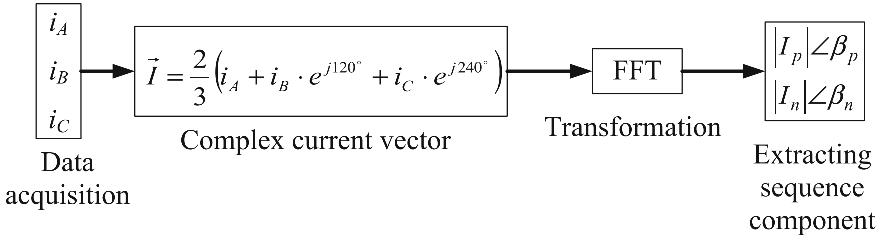


Fig. 2 Signal processing workflow to extract sequence components

current signals. The three-phase currents are used to construct a complex current vector ($\overrightarrow{I}$), as expressed by

$$\overrightarrow{I} = \frac{2}{3}\left(i_A + i_B \cdot e^{j120^\circ} + i_C \cdot e^{j240^\circ}\right) \tag{1}$$

where i_A, i_B, and i_C are the currents in the phase A, B, and C, respectively.

Fast Fourier Transform (FFT) of the complex current vector can automatically separate the positive and the negative sequence currents for all frequency components. An example of an actual current spectrum is shown in Fig. 3. The three-phase voltages are processed in the similar approach.

2.2 Extended Park's Vector Approach (EPVA)

The EPVA method is the observation of the spectrum of the Park's vector module. The motor current Park's vector components (i_D, i_Q) are

$$i_D = \left(\sqrt{\frac{2}{3}}\right) i_A - \left(\sqrt{\frac{1}{6}}\right) i_B - \left(\sqrt{\frac{1}{6}}\right) i_C \tag{2}$$

$$i_Q = \left(\sqrt{\frac{1}{2}}\right) i_B - \left(\sqrt{\frac{1}{2}}\right) i_C \tag{3}$$

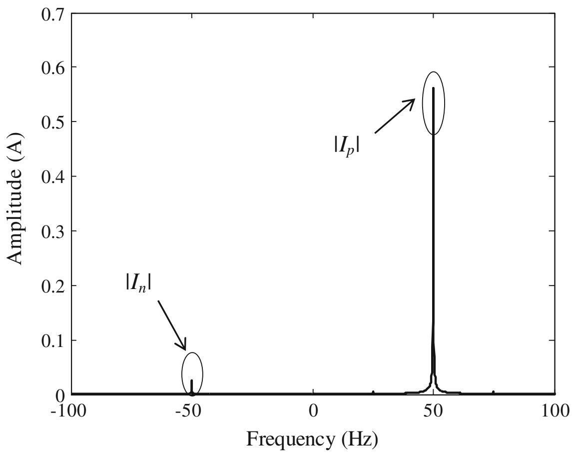


Fig. 3 A magnitude spectrum of the complex current vector

Table 1 The original features

	Features	Description
1	*%Load*	Percent of load
2	*pf*	Power factor
3	I_p	Positive sequence current
4	*Angle* (I_p)	Angle of positive sequence current
5	I_n	Negative sequence current
6	*Angle* (I_n)	Angle of negative sequence current
7	V_p	Positive sequence voltage
8	*Angle* (V_p)	Angle of positive sequence voltage
9	V_n	Negative sequence voltage
10	*Angle* (V_n)	Angle of negative sequence voltage
11	Z_p	Positive sequence impedance
12	Z_n	Negative sequence impedance
13	I_{dc}	Magnitude of square of Park's vector module at DC level
14	$I_{100\mathrm{Hz}}$	Magnitude of square of Park's vector module at twice the supply frequency

The square of the Park's vector module is given by

$$|i_D + ji_Q|^2 = \left(\frac{3}{2}\right)\left(i_d^2 + i_i^2\right) + 3i_d i_i \cos(2\omega t - \alpha_d - \beta_i) \tag{4}$$

where i_d is the maximum value of the direct sequence current, i_i is the maximum value of the reverse sequence current, ω is the angular frequency (rad/s), t is the time variable (s), α_d is the initial phase angle of the direct sequence current (rad), and β_i is the initial phase angle of the reverse sequence current (rad).

The square of the Park's vector module can be used to identify unbalanced three-phase currents. The spectrums of dc level and the component located at twice the supply frequency are obtained by applying the FFT to the square of the Park's vector module [1].

Two feature generation methods contain 12 original features and two extra features including the percent of load (*%Load*) and the power factor (*pf*) are added to be the original features. The 14 original features as shown in Table 1 are used to be the inputs for the next feature selection process.

3 Feature Selection

The Feature selection is a process to select optimal features from original features. It can reduce a number of features by removing irrelevant and redundant features. Basically, it can be divided into four categories: Filter, Wrapper, Hybrid, and Embedded methods [13]. The Filter method is the feature selection that applied independent evaluation criteria without involving any classification algorithms with measurement technique such as CFS [14] and consistency based subset evaluation. The Wrapper method applies a classification algorithm for a feature subset evaluation. This method is better than the Filter method, but it takes longer time for a computation. The Hybrid method combines the advantage of above two approaches. Finally, the Embedded method has built-in the feature selection in classifier.

Generally, the feature selection has four steps [15]: subset generation, subset evaluation, stopping criteria, and result validation. First, it searches optimal features by using searching algorithms. Then, this subset is evaluated by subset evaluator, and it stop by stopping criteria. Finally, it validates selected features.

Searching algorithms for finding the feature set have several methods which are shown below

1. **Best first** is the searching method that selects the feature with the best heuristic value.
2. **Exhaustive search** searches all possible feature subset.
3. **Greedy stepwise** is the searching method. It starts with empty or full feature set. Then, it adds the suitable feature or removes the inappropriate feature.
4. **Linear forward selection** (**LFW selection**) is the searching method that begins with an empty set and successively adding features.
5. **Random search** randomly selects the feature subset from original features.
6. **Rank search** selects the feature subset from ranking of total features.

CFS [14] is a well-known feature selection method that considers the correlation between features and classes and between features and other features. Relevance of the feature subset can be defined by using Pearson's correlation equation [13] which is expressed by (5)

$$Merit_s = \frac{kr_{kc}}{\sqrt{k + (k-1)r_{kk}}} \tag{5}$$

where k is the number of features, c is the number of classes, $Merit_s$ is relevance of the feature subset, r_{kc} is the average linear correlation coefficient between these features and classes, and r_{kk} is the average linear correlation coefficient between different features.

The linear correlation coefficient is defined by

$$r = \frac{\sum_i (x_i - \overline{x_i})(y_i - \overline{y_i})}{\sqrt{\sum_i (x_i - \overline{x_i})^2}\sqrt{\sum_i (y_i - \overline{y_i})^2}} \tag{6}$$

where i is the number of values (x or y), x_i is the x value for observation i, $\overline{x_i}$ is the mean x value, y_i is the y value for observation i, and $\overline{y_i}$ is the mean y value.

4 Experimental Setup

The three-phase four-pole delta-connected induction motor is used in the experimental setup as shown in Fig. 4. The motor parameters and ratings are summarized in Table 2. The motor is modified for interturn stator winding faults in each phase. A shorting resistor is used to limit the short-circuit current in the winding not exceed to 5 A. The induction motor is monitored by three current sensors and three voltage sensors. The measured signals are sent to the computer through a National Instruments (NI) data acquisition device with 6000 Hz sampling rate and ten operate conditions. The operation conditions contain of four cases: healthy motor, and the short- turns motor of 7, 15, and 31 turns in each phase. Each condition is operated under 6, 30, 60, and 90 % rated load of the motor. Note that the 14 original features are obtained from the normalized signals with rated parameters of the testing motor. Classes for the fault classification consist of ‘0’, ‘A’, ‘B’, and ‘C’ as the following meanings

‘0’ is Healthy motor
‘A’ is Interturn fault in phase A
‘B’ is Interturn fault in phase B
‘C’ is Interturn fault in phase C

Optimal features analyzed from the correlation feature selection algorithms are used in three well-known classifiers: k-Nearest Neighbor or kNN (k = 3), Naïve Bayes, and Decision Tree. The accuracy rated of classification will be compared with one obtained from the feature extraction using the PCA by using data set of 1320 samples. The result is validated by using 10 folds cross-validation technique.

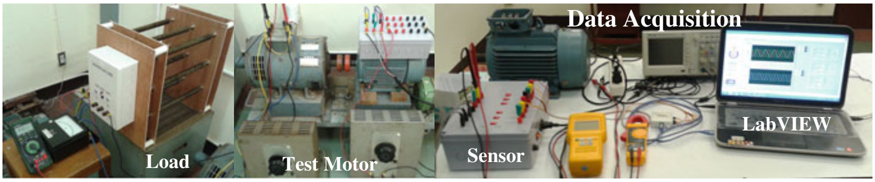


Fig. 4 Experimental setup

Table 2 Parameters and ratings of test machines

V	Hz	r/min	kW	cosØ	A
230Δ/400Y	50	1430	2.2	0.79	8.66/4.98
415Y	50	1435	2.2	0.765	4.94

5 Results and Discussions

The original features are analyzed by CFS for the feature selection method with aforementioned search algorithms. The results in Table 3 show that the correlation feature selection can effectively select 4 or 5 optimal features, while the PCA can reduce to 8 features. The 4 common features including *%Load, pf,* V_n, and Z_n are chosen from each search algorithms. These features are the part of optimal features for stator winding fault analysis. It is shown that these features have high correlation between features and classes, and they have low correlation between features and other features.

According to experimental results, the accuracy rates using original features are 58.5606, 43.0303, and 85.0758 % for kNN, Naïve Bayes, and Decision Tree, respectively. Based on feature selection, the accuracy rates using Decision Tree is higher than using kNN and Naïve Bayes. These classifiers provide the accuracy rates as ranged 85–87, 75–78, and approximately 43 %, respectively. Significantly, it can be explained that using the feature selection to select optimal features can reduce a number of features and increase the accuracy of the classification system. Fault classification provides better performance than using the feature extraction and original features, respectively.

Table 3 Classification correction of the proposed approach

Search algorithms	Subset evaluators	Number of selected features	Selected features	Accuracy rate (%)		
				kNN ($k = 3$)	Naïve Bayes	Decision tree
Original features	–	14	Original feature set	58.5606	43.0303	85.0758
Best first	CFS	5	*%Load, pf,* I_n, V_n, Z_n	75.1515	43.7121	86.0606
Exhaustive search	CFS	5	*%Load, pf,* I_n, V_n, Z_n	75.1515	43.7121	86.0606
Greedy stepwise	CFS	4	*%Load, pf,* V_n, Z_n	76.1364	43.0303	85.6818
LFW selection	CFS	5	*%Load, pf,* I_n, V_n, Z_n	75.1515	43.7121	86.0606
Random search	CFS	4	*%Load, pf,* V_n, Z_n	78.1818	42.803	87.1212
Rank search	CFS	4	*%Load, pf,* V_n, Z_n	77.0455	42.9545	85.6061
Ranker	PCA	8	–	60.5303	43.1061	59.8485

6 Conclusions

This paper presents the correlation feature selection for stator winding fault analysis of the induction motor. According to our experimental results, it can be found that the common selected features for stator winding fault analysis of the induction motor are *%Load, pf,* V_n, and Z_n. These features are good indicators to predict stator winding faults, and they can be applied for any size of motors. Furthermore, using the feature selection and the feature extraction can improve the accuracy of the classification system. For future works of our research, more features will be considered and compared with other feature selection methods.

Acknowledgments This work was fully supported by Prince of Songkla University Graduate Studies Grant Contract number 95000201 and partially supported by Center of Excellence in Wireless Sensor Networks (CoE-WSN), Faculty of Engineering, Prince of Songkla University, Hat Yai, Songkhla, Thailand.

References

1. Bonaldi, E.L., de Oliveira, L.E.L., Borges da Silva, J.G., Lambert-Torres, G., Borges da Silva, L.E.: Predictive maintenance by electrical signature analysis to induction motors. In: Araujo R. E (ed.) Induction Motors Modelling and Control, InTech (2012)
2. Jung, J.-H., Lee, J.-J., Kwon, B.-H.: Online diagnosis of induction motors using MCSA. IEEE Trans. Industr. Electron. **53**(6), 1842–1852 (2006)
3. Thomson, W.T.: On-line MCSA to diagnose shorted turns in low voltage stator windings of 3-phase induction motors prior to failure. In: IEEE International Conference on Electric Machines and Drives Conference, IEMDC 2001, pp. 891–898 (2001)
4. Cruz, S.M.A., Cardoso, A.J.M.: Stator winding fault diagnosis in three-phase synchronous and asynchronous motor, by the extended park's vector approach. IEEE Trans. Industr. Appl. **37** (5), 1227–1233 (2001)
5. Ben Khader Bouzid, M., Champenois, G.: New expressions of symmetrical components of the induction motor under stator faults. IEEE Trans. Industr. Electron. **60**(9), 4093–4102 (2013)
6. Arkan, M., Perovic, D.K., Unsworth, P.: Online stator fault diagnosis in induction motors. Proc. IEE Electr Power Appl. **148**(6), 537–547 (2001)
7. Cheng, S., Habetler, T.G.: A new method to detect stator turn to turn faults in a closed-loop multiple-motor drive system. In: IEEE international symposium on diagnostics for electric machines, power electronics and drives, SDEMPED 2009, pp. 1–6 (2009)
8. Cheng, S., Zhang, P., Habetler, T.G.: An impedance identification approach to sensitive detection and location of stator turn-to-turn Faults in a closed-loop multiple-motor drive. IEEE Trans. Industr. Electron. **58**(5), 1545–1554 (2011)
9. Chang, X., Cocquempot, V., Christophe, C.: A model of asynchronous machines for stator fault detection and isolation. IEEE Trans. Industr. Electron. **50**(3), 578–584 (2003)
10. Bachir, S., Tnani, S., Trigeassou, J.-C., Champenois, G.: Diagnosis by parameter estimation of stator and rotor faults occurring in induction machines. IEEE Trans. Industr. Electron. **53**(3), 963–973 (2006)
11. Siddique, A., Yadava, G.S., Singh, B.: A review of stator fault monitoring techniques of induction motor. IEEE Trans. Energy Convers. **20**(1), 106–114 (2005)
12. Jolliffe, I.T.: Principal Component Analysis. Springer, New York (1986)

13. Duangsoithong, R., Windeatt, T.: Relevance and redundancy analysis for ensemble classifiers. In: P. Perner (ed.) Machine Learning and Data Mining in Pattern Recognition, MLDM 2009, LNAI 5632, pp. 206–220. Springer, Heidelberg (2009)
14. Hall, M.A.: Correlation-based feature selection for discrete and numeric class machine learning. In: Proceeding of the 17th International Conference on Machine Learning, pp. 359–366. Morgan Kaufmann, San Francisco (2000)
15. Liu, H., Yu, L.: Toward integrating feature selection algorithms for classification and clustering. IEEE Trans. Knowl. Data Eng. **17**(4), 491–502 (2005)

Reconfigurable Tri-Colour RGB SD Card Bitmap Image File Writer

Chee Yuen Lam, Voon Siew Soon and Phaklen Ehkan

Abstract Field Programmable Gate Array is a popular choice to implement and test digital image processing algorithm due to its high speed and parallelism capability. The Altera DE2_TV sample project provides a basic real time digital video acquisition function. But this sample project unable to capture a static digital video frame to let the user to perform an analysis. Therefore this paper proposed a tri-colour channel red, green, blue SD card bitmap image file writer to capture the digital video frame and store the image in SD card. The propose architecture in this paper is able to enhance the DE2_TV sample project with the image capturing function.

1 Introduction

A sample project in Altera University program name DE2_TV was developed to interface the TV decoder on the DE2 education board to capture TV input PAL signal and displays on the VGA screen. By making use of this sample project, it can directly interface with the Closed Circuit Television (CCTV) video camera with the PAL video output signal and transform the DE2_TV sample project into a digital video acquisition system.

This paper is to solve and enhance the DE2_TV sample project by developing a frame grabber subsystem and Secure Digital (SD) card bitmap image storage subsystem which did not supported by this project. The capacity of the SD card

C.Y. Lam (✉) · V.S. Soon · P. Ehkan
School of Computer and Communication Engineering, University Malaysia Perlis,
Pauh Putra Campus, 02060 Arau, Perlis, Malaysia
e-mail: cheeyuen_lam@hotmail.com

V.S. Soon
e-mail: svsiew@hotmail.com

P. Ehkan
e-mail: phaklen@unimap.edu.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_105

used is 256 MB from Kingston. Section 2 will explain the hardware architecture of the DE2_TV sample project and also the FAT16 file system which used in SD card. Sections 2.2 and 2.4 will discuss about the frame grabber architecture and the SD card bitmap image writer architecture. Lastly, a result is discussed in Sect. 3.

2 Proposed Architecture of Bitmap Image File Writer

The bitmap image file writer consists of three main modules which include a frame grabber module, a bitmap file construction module and a SD card access controller as shown in Fig. 1. Each of the sub modules in this SD card bitmap image file writer is controlled by its own control unit. SD card is used as the storage media for the image.

Figure 2 shows the general process flow of the RGB bitmap SD card image file writer. When the 'Capture' button is pressed, the system start to capture and save the entire video frame red colour channel 8-bits binary data to the SRAM on board of DE2. The SRAM chip is IS61LV25616. After completed store all the red colour data byte to the SRAM, the data-in of SRAM is read out again and write it to SD card using SPI mode. The data written to the SD card is comply with bitmap file

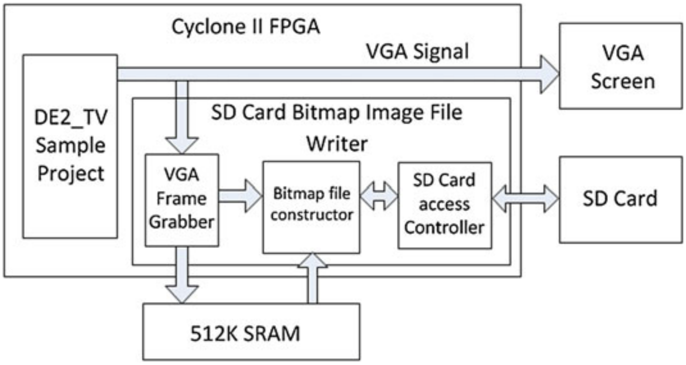


Fig. 1 Architecture of bitmap image file writer

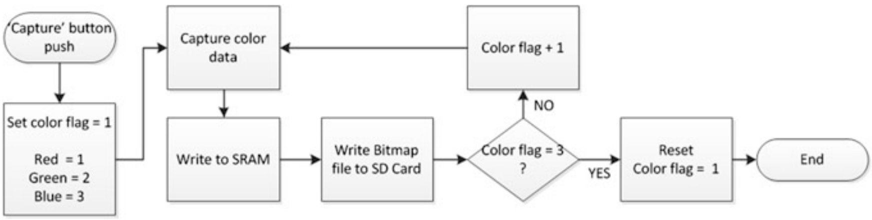


Fig. 2 Process flow of RGB bitmap SD card image file writer

format (.BMP) and organize with FAT16 (File Allocation Table) file system structure. These processes repeat twice followed by the red colour channel for green and blue colour channel to store a complete full colour image.

2.1 FAT16 File System

SD card is just a flash memory chip. In order to store the image files into the SD card, and is able to read directly from a computer, the file needs to be organized in FAT16 file format. This is a standard file system that computer uses to access and read file in the SD card.

In FAT16 file system, a memory space of the flash disk is partitioned into several region, with the sequence starting by master boot sector, FAT, root directory and user data space [1]. However, the starting address of each region is varying due to the total memory capacity of the flash chip. A 256 MB SD card is used in this project. The starting address for the corresponding memory region is listed in Table 1. The bitmap file constructor writes the file to the SD card based on these starting addresses.

SD card have a minimum addressable block write size of 512 bytes (1 sector). All the data has to write in the size of a sector which is 512 bytes. Each of the write command is sent to the SD card and the on-card controller counts 512 bytes of incoming data and stores them into the memory space.

2.2 VGA Frame Grabber Module

The VGA frame grabber module is responsible to capture and store the VGA red, green and blue colour data to the onboard Static Random Access Memory (SRAM). It uses 'print screen' method to capture the image. This module as shown in Fig. 3a captures the VGA output stream to the LCD monitor as the image data. The frame size is a standard VGA frame with resolution of 640 × 480 pixels.

When the 'Capture' push button is pressed, this module begin to synchronize with the input VGA pixel stream through the x and y pixel coordinates. This module waits for the beginning of a new frame in which the x and y pixel coordinates become (0, 0) and start to store the colour data into the SRAM. When the colour data of a frame is completely stored into the SRAM, this module will invoke the bitmap file construction module to start preparing the bitmap file header, FAT and root directory entry for writing to the SD card.

Table 1 256 MB SD card memory region starting address

Memory region	Starting address (hex)
Boot sector	00000000
FAT list	0000CA00
Root directory	0001BE00
User data	0001FE00

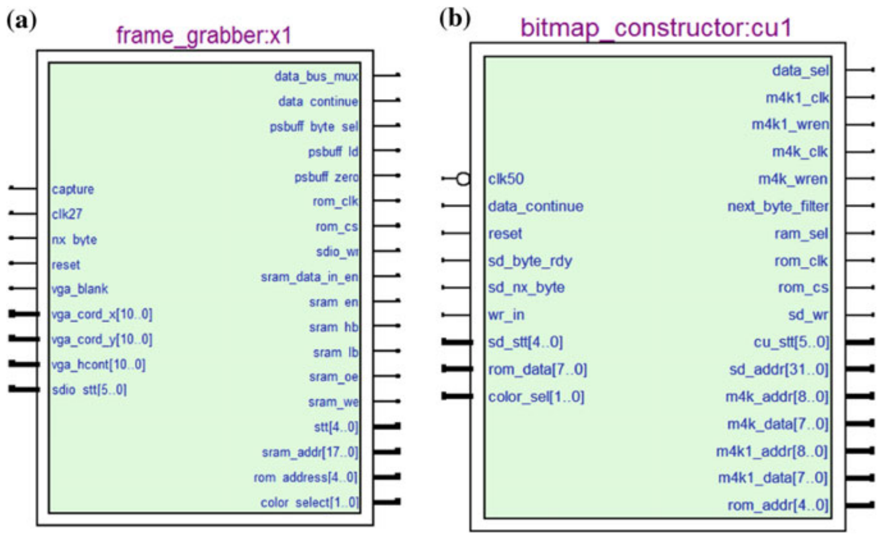


Fig. 3 **a** Frame grabber module. **b** Bitmap file constructor module

$$(H\,x\,W) \times 3 = M \tag{1}$$

In order to store a frame with RGB colour channel, the required memory is calculated by using (1); where H is image height in pixel; W is image width in pixel and M is total image size in bytes. To store a complete VGA frame with RGB colour and resolution 640 × 480 pixels, it requires 921,600 bytes (921 KB) of memory. Since the onboard SRAM contain 524,288 bytes (512 KB) of memory space, it is not sufficient to store the entire RGB colour channel of the frame. Therefore, the RGB colour data of a frame have to store separately and consecutively so that the onboard SRAM is able to fit for the entire operation. Since the SRAM is 16 bits word length, the sequence of data stored into SRAM is from High byte to Low byte.

2.3 *Bitmap File Constructor Module*

This bitmap file constructor module is responsible to construct the bitmap file for the SD card access controller written into SD card. This module is shown in Fig. 3b. It is responsible to generate the FAT, root directory file entry, bitmap file header and also the bitmap file data to be written into the SD card sequentially. It contains two 512 bytes M4 K RAM modules and a ROM. The first and the second M4 K RAM are used as a temporary buffer to store the FAT and the root directory file entry of the SD card. The ROM is used to store the 54 bytes predefine bitmap file header. Since the root directory file entry and the bitmap file header are a fixed set of data, hence these data are predefined in a ROM.

Besides preparing data to be written into SD card, this module also take an important task of calculating and sequencing the data of starting address and address offset for the SD card. All the data are written from the beginning of the starting address. This task is crucial when storing the bitmap file into the SD card because any error of the starting address will cause the bitmap file corrupted and unable to read by the computer.

2.4 SD Card Access Controller

This SD card access controller is responsible to communicate and also perform serial write operation to the SD card. This module accept control signal from both Bitmap file constructor module and SD card. When the write signal of this controller is triggered, it will first latching the memory sector address which provided by the bitmap file constructor module and start sending a BLOCK_WRITE (CMD24) command with 7 bytes long to the SD card. Once the SD card response acknowledge, the controller proceed to request further data from the bitmap file constructor module and transmit the data-in back to the SD card. Total of 512 bytes of data are written into each BLOCK_WRITE command send.

According to the SD specification [2], the SD card has two types of write mode namely, single block and continuous write mode. This project uses a single block write mode and this triggering the state machine of the controller easier to be designed. Figure 4 shows the SD card access controller FSM in single block write mode. The works of [3, 4] provide a comprehensive explanation on the SD card data writing sequence and also the command sequence with operation flow chart. Both covered reading and writing operation of the SD card. The clock signal provided to this controller is 27 MHz, and output SDI clock is half of the driving clock which is 13.5 MHz.

3 Result and Analysis

This project is tested on Altera DE2 education board with Cyclone II EP2C35F672C6 FPGA core. Figure 5 demonstrates the real hardware setting for this research.

A license free software “HxD” is used to verify the binary data written into the SD card by this bitmap image writer. The first 54 bytes of data in Fig. 6 which is surrounded by red line is the bitmap image file header. Whitrow [5] provides a detail description and explanation for the 54 bytes of the bitmap file header. The

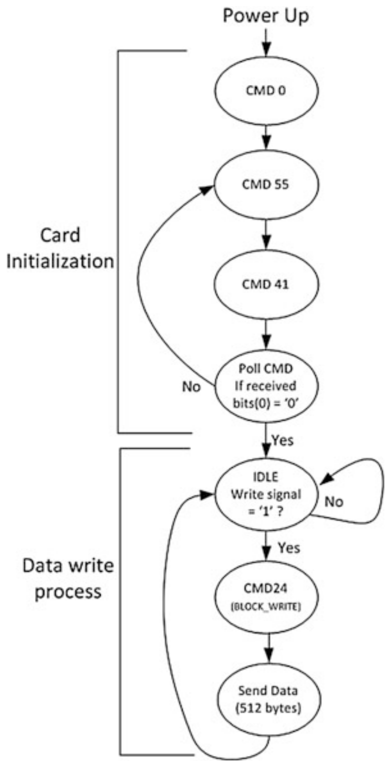


Fig. 4 FSM of SD card access controller

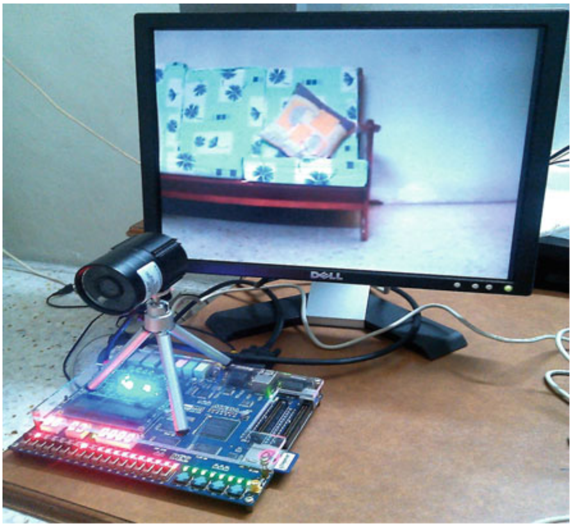


Fig. 5 Real hardware setup

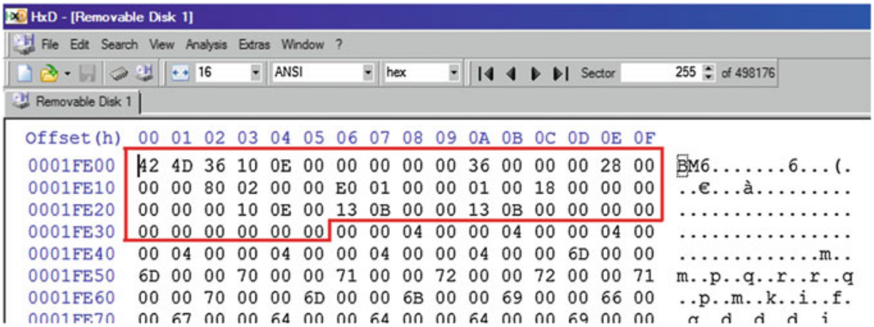
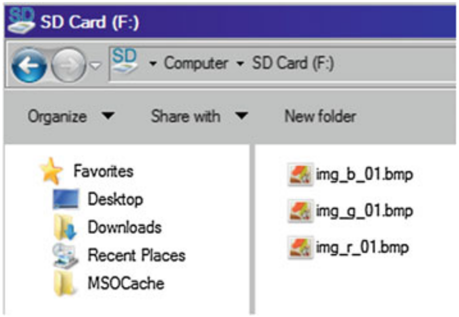


Fig. 6 Bitmap file header data in SD card

Fig. 7 Three bitmap files in SD card



data generated by the bitmap file constructor module can be checked using HxD software. Any error of the data can be verified byte by byte.

After verified the result using HxD software, the SD card is directly access using windows operating system of the computer. When accessing the SD card, it is expected that it contains of three bitmap files with the filename 'img_b_01.bmp', 'img_g_01.bmp' and 'img_r_01.bmp' as shown in Fig. 7. This filename is defined in a filename ROM of bitmap file constructor module. The filename also direct denoted the colour channel of the image. Figure 8a, b and c show the captured single colour channel bitmap image in this project. Figure 8d shows the combined full colour RGB bitmap image using C programming in computer post processing. The recombine image is the same with the image display on the LCD screen of the DE2_TV sample project.

4 Conclusion

From the experimental result of this project, it is proven the proposed architecture for bitmap image file writer is able to function well as its design. This architecture is able to capture separated red, green and blue colour channel images from the VGA

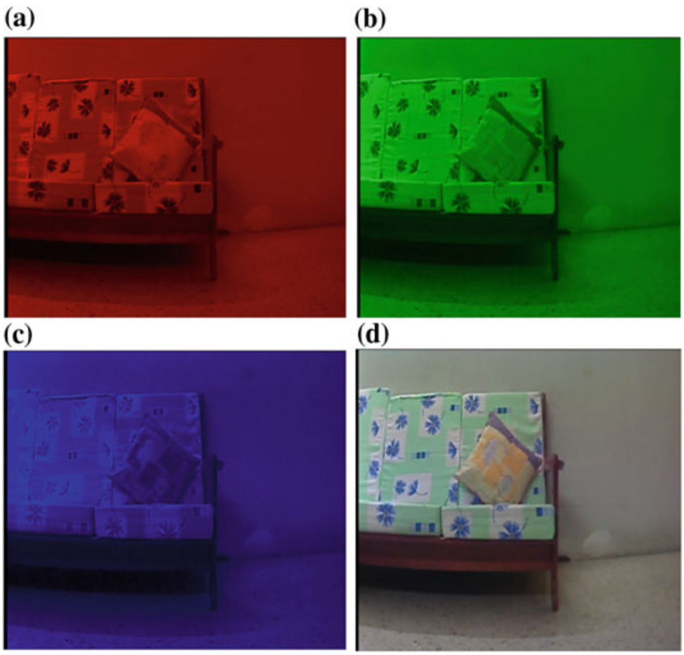


Fig. 8 **a** Red colour channel image. **b** Green colour channel image. **c** Blue colour channel image. **d** Post processed full RGB image

screen and store these images in 24-bit bitmap image file in the SD card with resolution of 640 × 480 pixels. This project has enhanced the existing DE2_TV sample project with image capturing function. The studies of reconfigurable machine vision in FPGA made easy with the function of digital image capturing from this proposed architecture.

References

1. SD Specification—Part 1 Physical layer simplified specification V 4.10. https://www.sdcard.org/downloads/pls/simplified_specs/part1_410.pdf. Accessed 8 Oct 2014
2. FAT file system. http://www.ntfs.com/fat-systems.htm. Accessed 18 Nov 2014
3. Elkeelany, O., Todakar, V.S.: Data archival to SD card via hardware description language. Embed. Syst. Lett. IEEE **3**(4), 105–108 (2011)
4. Zhenlin, Lu., Jingjiao, Li., Yao, Zhang.: The reading/writing SD card system based on FPGA. In: First International Conference on Pervasive Computing Signal Processing and Applications (PCSPA), pp. 419, 422, 17–19 Sept 2010
5. Whitrow, R.: OpenGL Graphics Through Applications, pp. 39–42. Springer, London (2008)

Detection of Privacy Threat by Peculiar Feature Extraction in Malwares to Combat Targeted Cyber Attacks

Farhan Habib Ahmad, Komal Batool and Azhar Javed

Abstract Targeted cyber-threats are topmost concern of organizations and technologies of today. Malwares having similar objectives bear common artifacts. Thus defining a detection mechanism based on such peculiar artifacts will not only help in detecting existing risks but also gives a considerable defense against unknown malicious attacks. About 903 known malware samples related to espionage were analyzed statically and a data set comprising related artifacts is established and also checked against the benign software. Weightage is given to each artifact on the difference of its existence in malicious and benign code and artifact's relation to the expected targeted organization or technology thus catering for targeted attacks. Designed algorithm for detection of espionage attack has given 99.16 % of authentication and 99.33 % of precision. Real time alarm generation is also incorporated by API hooking using Detour library for latter detailed analysis of suspicious program or application by proposed algorithm.

1 Introduction

Life is heading towards the smart living by incorporating smart devices interconnected with internet. Where it brings a lot of comfort and luxury to the life at the same time it is prone to issues like privacy, unavailability and authenticity. Symantec reported 62 % increase in cyber security breaches in 2013 [1] and Kaspersky [2] is detecting 315,000 malicious files per day. Financial impact due to

F.H. Ahmad (✉) · K. Batool · A. Javed
National University of Sciences and Technology, Islamabad, Pakistan
e-mail: farhan@mcs.edu.pk

K. Batool
e-mail: kb10june@gmail.com

A. Javed
e-mail: azharjaved@mcs.edu.pk

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_106

malicious applications and software is up to trillion of US dollars [3]. Therefore there is a dire need to go for mechanism and solutions beyond the existing techniques depending on the signature based and heuristic based detection tools. It is not possible for any vendor to provide the signature of every coming malware especially in case of targeted malware. Trojans for espionage leads in malware community by acquiring a share of 68.84 % as per PANDALABS annual report 2014 [4]. Therefore an effort is made in this research to detect the advanced malwares designed for data espionage by establishing a data base of peculiar artifacts found in known spyware through static analysis. The research is carried out on the analogy that the malwares having common objectives bear similar artifacts thus resemblance of indicators from already identified threats [5] will help in detection of unknown malwares [6]. Weightage is given to each artifact on the difference of its existence in malicious and benign code and on its relation to the expected targeted organization or technology, in our case we have worked for the privacy of telecomm sector. Real time alarm generation mechanism is also incorporated by hooking espionage related APIs using Detour Library for latter detailed analysis by proposed detection algorithm.

This research paper has been organized as follows. Background and related work is discussed in Sect. 2. Section 3 throws light on the research undertaken by the author. Section 4 describes the results and validation of proposed algorithm. Future work and conclusion are described in Sect. 5.

2 Background and Related Work

Malware is the term collectively used for all sort of malicious scripts and codes used with malevolent intentions in cyber domain. Malwares can be divided in different classes like Virus [7], Worm [4], Trojan [8], Botnet [9], Spyware and Rootkit [4], depending upon their propagation and threat ability [10].

Symantec encountered 403 million new malware only in 2011 and a 42 % increase is noted in 2012 [11]. Therefore a thorough malware analysis is required for each category of malware for appropriate defense against each type. Malware analysis are divided in two major categories *Static Malware Analysis* and *Dynamic Malware Analysis* [10]. *PE Explorer, STRING, MD5SUM, PEiD, XORSearch, Bentext* and *Virus Total* are some of good static malware analysis tools [12]. *Anubis, CW Sandbox, Norman Sandbox, Joebox, WiLDCAT* and *LASTLINE* are some example of dynamic malware analysis tools [12]. *Renovo, OmniUnpak, UPX, PolyUnpack* and *Justin* can be used for unpacking packed binaries [10].

Analysis of API calls and API calls sequence can determine the objective and behavior of a particular code. Malwares prefer to use Windows APIs instead of their own dlls as to remain unnoticed by reducing their binary size. API names and their input arguments are used by Salehi et al. [13] for distinguishing benign from malicious applications. In [7] Veeramani and Rai proposed the automated detection mechanism for executable codes based on their relevant API calls. Static Analyzer also uses the API calls sequence to analyze a sample [14]. As the polymorphic malwares keep on changing their signatures but they always follow the same flow pattern, thus analyzing the sequence of API calls will help in detecting the polymorphic malwares.

Analysis revealed that Command and Control communication remains consistent therefore with the detailed analysis of network traffic and recognition of patterns it is not only possible to safeguard against known threats but also can guard against new threats [15]. Similarly *Windows Filtering Platform* (WFP) is a set of APIs and system services that provide a platform for creating network filtering applications at several layers in the networking stack of the operating system [16].

Retrospective detection approach may be a solution for finding the advanced malware-infected computers. It is mechanism which allows the resemblance of indicators from already identified threats [5]. Normally the web related protocols are allowed in most of organizations, thus the logs of http and https requests offer the capability to uncover an attack retrospectively [6].

3 Alarm Generation and Detection of Espionage Threat

3.1 Threat Canvas

Although different operating systems are in use by organizations but the most common is Windows Operating System. Therefore most of the threats found are designed for Windows. All Windows based executable like .exe, .dll and other object codes have a standard format as Portable Executable (PE). As per Security Threat Report 2014 by Symantec Corporation in year 2013, more than 50 % of email attachments are in .exe format and used as a bait for phishing attack [1]. Therefore for detection of malware in this research Windows as operating system and PE as file are analyzed. Moreover applications and information related to telecomm sector are considered to focus for a targeted organization and technology.

3.2 Data Set

Malware dataset was acquired from Georgia Tech Information Security Center [17] and Contagio Malware Dump [18]. These resources are well reputed [19]. A total of 1203 known malwares of data espionage belonging to 51 variants were used out of which 903 are analyzed for establishing of artifacts data base and 300 for the validation of proposed algorithm. Name and number of espionage samples are given in Table 1.

Table 1 Data espionage samples analysed in this paper

Name	Samples	Name	Samples	Name	Samples
Backorfice	22	Prorat	20	Dantom	31
Servu	23	Striker	19	BOFacil	29
DonaldDick	29	RemoteControl	20	InCommander	22
Flammer	38	Win32	21	Wincrash	19
Girlfriend	36	CoreFlood	28	BAckDoor	14
Firehotker	29	Bifrose	17	Spyeye	36
Shamoon	19	NetDevil	33	CyberSpy	18
SpySender	27	OptixPro	24	NetSpyDK	14
Hydraq	24	Bubble	19	Bebloh	31
Brainspy	28	ECC	21	NetControl	13
Devil	21	Gauss	23	Ptakks	37
Doly	23	Sub7	18	Krippled	31
Lamer	25	Agent	13	NetBus	19
Nerte	23	APT	21	Bunker-Hill	14
Spy	17	Delf	19	Aforce	27
Bget	29	Thief	16	Beta	19
Optix	34	Zeus	22	Gift	28

3.3 Extraction of Artifacts

Detection of cyber espionage is a challenge due to sophistication, obfuscation and polymorphism. But it is also a fact that there are specific and peculiar artifacts, like *APIs, Strings, IP Addresses, URLs and Email Addresses* which exist in a malware designed for data espionage and remain constant. 903 out of 1203 malware samples were analyzed for extraction of common artifacts and patterns present in the malwares for data espionage. Process is divided in five steps as shown in Fig. 1 and elaborated below.

Classification of Malware Sample. Samples are classified in the variants as per their tagged identification through online malware detection engines as elaborated in Fig. 2.

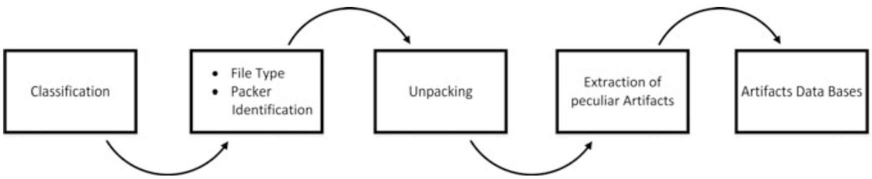


Fig. 1 Extraction of artifacts from PE

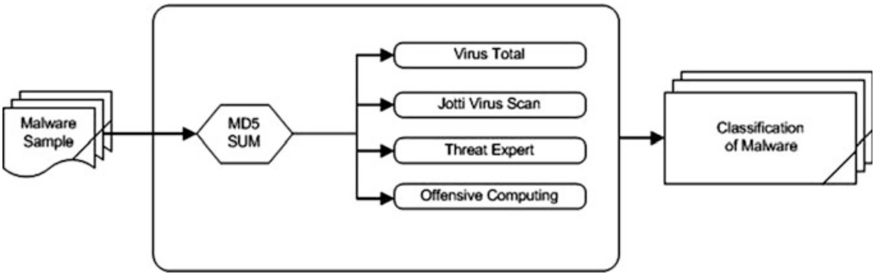


Fig. 2 Classification of malware

File Type and Packer Identification. Attackers hide the type of file by assigning the wrong file extension. TrID utility [20] is used to determine correct file type. Online GUI based PEid [21] facility is used to identify the Packer.

Un-Packing can be performed by making use of utilities as mentioned in Sect. 2. In our case we have used UPX. It is observed that the unpacked version of the file differs in hash as well as in the size from the actual file.

Extraction of Peculiar Features. Utilities as mentioned in Sect. 2 can be used for extraction of DLLs, API and specific strings from PE.

PE Explorer is used to extract the DLLs and APIs from the binary. DLLs and API function calls are carried by the import table of the PE. Thus API calls were extracted by dissecting the PE with PE explorer without in actual executing the code.

Similarly the ASCII strings being embedded in the binary during compilation are extracted with the help of *STRINGS* developed by the Windows SysInternals. Extracted strings include the ASCII strings, IP addresses, URLs, names, classified words, etc. The strings were graded on the basis of difference of their existence in malware and benign software and relation to targeted organization, technology and expected threat.

Artifacts Data Base. Artifacts extracted in previous step are stored separately. 903 malwares analyzed and an “API Data Base” of 1723 and “String Data Base” of 809 artifacts are established. Each artifact in “API Data Base” and “String Data Base” is denoted with a unique ID as “Ài” and “Śi”, where i ranges from 1 to n.

3.4 Artifact Weightage

Weightage to every artifact from previous step is calculated in three steps. In step one a point (+1) is given to every “Ài” and “Śi” artifact for each occurrence in “API Data Base” and “String Data Base” respectively. In Step two occurrences of same artifacts extracted from malicious samples is also checked in benign files. For this a sample set of 600 benign applications being used in telecomm organizations are taken and checked for the artifacts by using same process as mentioned in Sect. 3.3. A negative point (−1) is given to every "Ài" and "Śi" artifact for each occurrence in the "API Data Base" and "String Data Base" for benign files. In step three values calculated in step one and two for each artifact are added to get the artifact weightage of the particular artifact. Higher is the value more is the weightage.

Weightage is represented with "ώi" and "λi" against each Artifact "Ài" "Śi" respectively, where i ranges from 1 to n. However while giving the weightage to artifacts manual effort was also put in specially for "*Strings Data Base*" by giving due weightage to type of organization, extracted or related IP addresses, type of strings being used in the organization, expected target data, expected target file type and other peculiar strings for a superior detection rate and defense mechanism. The artifacts are then sorted in descending order from most malicious to least malicious. The artifacts with zero or negative values are the one whose occurrence in benign samples is either equal or more than malicious samples. Therefore all artifacts having weightage zero or less are dropped from the artifacts list to be used for malware detection except selected manually depending upon their peculiarity. Top 51 artifacts from "API Data Base" and "String Data Base" after calculating the weightage are appended in Tables 2 and 3.

Weightage of Sample. The weightage "¥" for all known malware samples and benign samples is calculated by adding all the weightage of artifact "ώi" and "λi" present in a sample as shown in Eq. 1.

$$¥ = \sum (\acute{\omega}\, i + \lambda\, i) \quad \text{where} \quad i = 1 \rightarrow n \tag{1}$$

Malicious Threshold. Malicious weightage "¥" calculated in previous step is sorted in descending order from most malicious to most benign. Mean of minimum malicious and minimum benign is taken as the threshold bench mark for the detection algorithm.

Table 2 Top weightage API artifacts

Artifacts	Artifacts	Artifacts	Artifacts
RegQueryValueExA	Send	LocalAlloc	SetErrorMode
RegOpenKeyExA	ReadFile	FreeLibrary	LoadLibraryA
DefWindowProcA	WriteFile	TlsSetValue	SetFilePointer
DispatchMessageA	FindClose	TlsGetValue	MessageBoxA
CreateWindowExA	CreatFileA	VirtualAlloc	PeekMessageA
GetDiskFreeSpaceA	RtlUnwind	RegCloseKey	GetLocalInfoA
GetModuleHandleA	VirtualFree	SetEndofFile	FindFirstFileA
GetCommandLineA	GetVersion	ShowWindow	RegisterClassA
WaitForSingleObject	GetFileSize	PostMessageA	RaiseException
GetModuleFileNameA	ExitProcess	SendMessageA	GetVersionExA
WideCharToMultiByte	CreatThread	EnumWindows	DestroyWindow
InitializeCriticalSection	GetFileType	FindWindowsA	CallNextHookEx
SystemParametersInfoA	FindNextFile	FindNextFile	

Table 3 Top weightage string artifacts

Artifacts	Artifacts	Artifacts	Artifacts
UseDockManager	Disconnect	SOCKS	Host
File access denied	REGISTER	Disabled	Timer
Class %s not found	FocusControl	WinSock	Shared
Too many open Files	WindowState	SERVER	Stream
Privileged instruction	HideSelection	Password	Sender
No Address Specified	PasswordChar	Passwords	UrlMon
Read beyond end of file	SHUTDOWN	ScktComp	Owner
Floating point overflow	File not found	CreateKey	Timer
Floating point underflow	Stack Overflow	Username	SysUtils
No argument for format '%S'	Stream read error	LOCKED	Remote
Floating point division by zero	Division by zero	Not Found	WSocket
Deleting all files of current folder	Stream write error	On connect	classified
Copy all files of current folder	IP Addresses	username	

3.5 Alarm Generation

For development of alarm generation mechanism for data exfiltration we have short listed the APIs used for data exfiltration as FindNextFile(), ReadFile(), WriteFile() and Send() functions. DLL for alarm generation is injected in all running processes using Detour library as shown in Fig. 3 only for FindNextFile() function. Injected DLL monitors the actions of the processes by hooking the functions FindNextFile(), ReadFile(), WriteFile() and Send(). Our injected DLL gains control, whenever a process calls any of these functions. FindNextFile(), ReadFile() and WriteFile() functions are hooked in Kernel32.dll and Send() function is hooked in

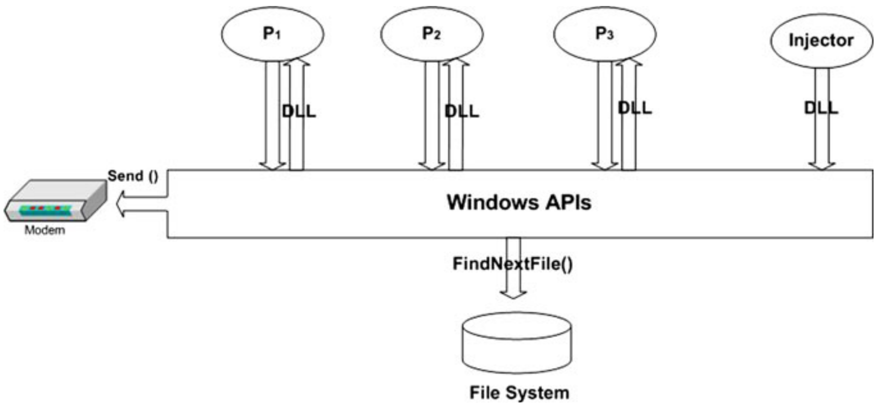


Fig. 3 Injection of monitoring DLL in all processes

Ws2_32.dll. Our DLL maintains the record that how many times these functions are called and also have capable to generate alarm in the form of pop-up or logs. Than latter on the identified suspicious applications or programs are analyzed in detail by the detection algorithm.

3.6 Detection Algorithm

Suspicious programs or applications identified by the alarm generation mechanism by hooking APIs in previous step are checked for quick detailed static malware analysis by the automated proposed detection algorithm. The algorithm is divided in eight steps. The detection algorithm gets the suspicious file as input and separate malicious from benign. Following are the algorithm steps and flow chart is shown in Fig. 4.

- PE is checked and classified using detection engines.
- Correct file type is checked.
- Packer identification and unpacking if required.
- Extraction of peculiar artifacts.
- Comparison with established malicious artifacts.
- Calculation of malicious weightage "¥".
- Comparison with threshold value.
- Malicious or Benign declaration with the artifact report for deep analysis.

4 Results

The proposed algorithm was tested against data set comprising malicious and benign files. Depending on the results authentication, precision and false alerts are calculated. Validation data set comprising 300 known malicious malwares from authentic source as mentioned in Sect. 3.2 for data espionage and 300 benign files related to telecomm sector were tested. Following experimentation results were achieved.

- Total No Samples 600
- Malicious 300
- Benign 300
- True Positives 297
- True Negative 298
- False Positive 2
- False Negative 3

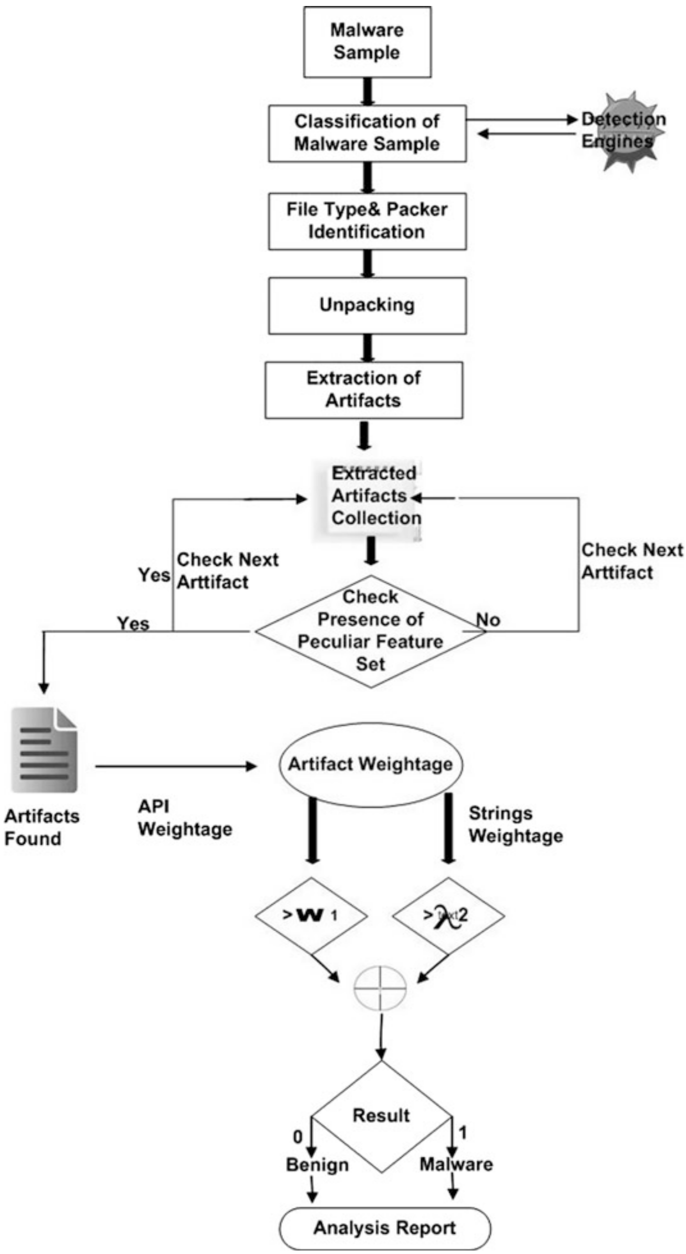


Fig. 4 Detection algorithm for malware detection

Table 4 Comparison of Results with Existing Work

Metric	IMDS (%)	Detection of malware on mining API (%)	Proposed method (%)
Authentication	93.07	98.31	99.16
Precision	80.13	98.5	99.33
False alerts	19.86	1.51	0.833

On the basis of above results authentication, precision and false alerts of the algorithm are calculated as below.

$$\begin{array}{lll} \text{Authentication} & = ((TP+TN))/((TP+TN+FP+FN)) & = 0.9916666 \\ \text{Precision} & = TP/((TP+FP)) & = 0.9933110 \\ \text{False Alert} & = (FP+FN)/(TP+FP+TN+FN) & = 0.0083333 \end{array}$$

Alarm generation is performed on real time basis. It is observed that CPU and system memory utilization by our Hook Handlers used for alarm generation is 5 and 4.5 % respectively. Thus it's not heavy on the system resources.

Similarly the refinement of weightage value of artifacts in relation to expected targeted organization, technology and expected threat for the detection mechanism provides improved results than the previously suggested work like "Detection of Malware on Mining API" by Sami and Hamze [22] and "Intelligence Malware Detection System (IMDS)" by Ye et al. [23] as shown in Table 4.

5 Conclusion and Future Work

In this research, we proposed a novel espionage detection mechanism which analyze the suspicious programs and applications identified by the alarm generation mechanism. We focused on the refinement of artifacts in relation to the targeted organization, technology and the expected threat which has improved the detection probability as compared to previous research. The paper contributes by providing a step towards the defense against targeted cyber threats for data espionage.

The present research is limited to Windows as operating system and PE as format for malicious files. Therefore the future work can benchmark on other operating systems like Linux, Apple Mac or Android along with related file formats like ECOFF, ELF for Linux, .DMG, .APP for Mac and .dex, .apk for Android. Similarly the research was directed towards the malwares designed for data espionage. Therefore future work can be done on similar lines for availability or authentication aspects of security. Furthermore the file packed with unknown application is a challenge in the field of malware analysis and is an area for future work.

References

1. Symantec Corporation Internet Security Threat Report, vol. 19 (2014). http://www.symantec.com/content/en/us/enterprise/other_resources/b-istr_main_report_v19_21291018.en-us.pdf. Accessed 12 Apr 2014
2. Kaspersky Lab: Available http://www.kaspersky.com/about/news/virus/2013/number-of-the-year. Accessed 12 Apr 2015
3. de Vries, J.A.: Towards a roadmap for development of intelligent data analysis based cyber attack detection systems. Dissertation, Delft University of Technology (2012)
4. Denning, P.J.: The ARPANET after twenty years. Am. Sci. **77**, 530–535 (1989)
5. Oberheide, J., Cooke, E., Jahanian, F., CloudAV : N-version antivirus in the network cloud. Electrical Engineering and Computer Science Department, University of Michigan
6. Hung,, S.L.Y.C.H.: N-victims : an approach to determine N victims for APT investigations. Paper Presented in Information Security Application 13th International Workshop WISA 2012
7. Veeramani, R., Rai, N.: Windows API based malware detection and framework analysis. Int. J. Scientif. Eng. Res. **3**(3), 1–6 (2012)
8. Spafford, E.H.: The internet worm incident (1989). Available http://docs.lib.purdue.edu/cstech/793. Accessed 13 Apr 2015
9. Botnet Research Survey: Available http://www.computer.org/csdl/proceedings/compsac/2008/3262/00/3262a967.pdf. Accessed 22 Mar 2015
10. Egele, M.: A Survey on Automated Dynamic Malware Analysis Techniques and Tools, vol. V, pp. 1–49. Vienna University of Technology
11. Alam, S., Horspool, R.N., Traore, I.: MARD: a framework for metamorphic malware analysis and real-time detection. In: Proceedings of International Conference on Advanced Information Networking and Applications (AINA), pp. 480–489 (2014)
12. Javed, A., Akhlaq, M.: On the approach of static feature extraction in trojans to combat against zero-day threats. In: International conference on Paper Presented in IT Convergence and Security (ICITCS), pp. 1–5, 28–30 Oct 2014
13. Salehi, Z., Ghiasi, M., Sami, A.: A miner for malware detection based on API function calls and their arguments. In: 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012), pp. 563–568, May 2012
14. Burji, S., Liszka, K.J., Chan, C.: Malware analysis using reverse engineering and data mining tools. In: International Conference on System Science and Engineering (ICSSE), pp. 619–624 (2010)
15. T.M. Incorporated: Detecting APT Activity with Network Traffic Analysis (2012). Available http://www.trendmicro.com/cloud-content/us/pdfs/security-intelligence/white-papers/wp-detecting-apt-activity-with-network-traffic-analysis.pdf
16. Windows Filtering Platform (Windows). Available http://msdn.microsoft.com/en-us/library/aa366510(VS.85).aspx. Accessed 21 Jan 2013
17. Open Malware. Available http://oc.gtisc.gatech.edu:8080/. Accessed 22 Mar 2015
18. Contagio. Available http://contagiodump.blogspot.com/. Accessed 22 Mar 2015
19. Malware Sample Sources for Researchers. Available https://zeltser.com/malware-sample-sources/. Accessed 22 Mar 2015
20. Marco Pontello's Home—Software—TrID. Available http://mark0.net/soft-trid-e.html. Accessed 22 Mar 2015
21. PEid—aldeid. Available http://www.aldeid.com/wiki/PEid#PEid. Accessed 22 Mar 2015
22. Sami, A., Hamze, A.: Malware Detection Based on Mining API Calls, Categories and Subject Descriptors. ACM, pp. 1020–1025 (2010). 978-1-60558-638-0/10/03
23. Ye, Y.F., Wang, D.D., Li, T., Ye, D.Y.: IMDS: intelligent malware detection system, KDD-2007. In: Proceedings of Thirteenth ACM SIGKDD International Conference on Knowledge Discovery of Data Mining, pp. 1043–1047 (2007)

Design and Analysis of Reconfigurable Antenna Using RF MEMS and Fractal Geometry

Yogita Nafde and Rajesh Pande

Abstract RF MEMS technology offers superior performance and has edge over the conventional solid state devices. The reconfigurability aspect in an antenna system is a desired characteristic and has been the focus area of research in recent years. The performance of reconfigurable antenna can be improved by optimization of the location of the switches and also by identifying the suitable material of RF MEMS structure. The design of Reconfigurable Micro strip Antenna using Koch fractal geometry is analysed in this paper. The simulation of design is carried out using HFSS. The fractal Antenna can provide Multiband and Wideband operation but the reconfiguration aspect can further help to improve the performance of antenna. This antenna design can support the bands of several applications including WiFi, 3G, WiMax as well as UWB range.

Keywords RF MEMS · Reconfigurable antenna · Fractal geometry · Koch curve

1 Introduction

Reconfigurable multi-band antennas are useful for many commercial and military applications where it is required to have a single antenna that has the capability of dynamic reconfiguration to transmit or receive on multiple frequency bands. The paper focuses on the optimisation of the RF MEM equivalent switches along with fractal geometry to come up with a new antenna design suitable for several wireless applications.

Y. Nafde (✉)
Priyadarshini College of Engineering, Nagpur, India
e-mail: yogita.nafde@gmail.com

R. Pande
Ramdeobaba College of Engineering and Management, Nagpur, India
e-mail: panderaj@yahoo.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_107

WLAN and WiMax wireless standards use different operation bands. This necessitates the need for terminal antenna that are multiband as well as wideband. To address the constraints like cost, dimensions, radiation patterns, gain and ease of integration with communications devices, printed monopole with microstrip line feed antenna is used [1]. To operate the antenna in multiband and wideband operation a U-slotted rectangular patch and partial ground plane flushed with feed line have been used.

2 Koch Curve

The Koch snowflake is one of the mathematical fractal curve. The Koch curve has an infinite length since with each iteration the total length of the curve increases by one third [2]. All the iterations create four times the line segments as compared to the previous iteration and length of every segment is one-third the length of the segment of the previous stage. The final length of the curve after n iterations will be $(4/3)^n$ times the original triangle perimeter, which tends to infinity. Thus Koch curve fractal dimension should be near about (log 4)/(log 3). The Koch curve is continuous everywhere but is differentiable nowhere.

A fractal geometry is self repetitive and is generated using an iterative process. The parts of fractal geometry have the same shape as the whole geometry but with different scales. The fractal based radiators keep similar radiation parameters over several bands. The space—filling property is the important aspect of the fractal geometry.

Koch curve is the good example of self similar space—filling fractals which have been used for wideband/multiband/miniaturized antennas [3]. The self similar fractals affect the electromagnetic properties of the antennas that are created on the basis of these geometries leading to multiband structures. The different iterations of the Koch curve are as given in Fig. 1.

Reconfigurability is the property of antenna to change its characteristics as per the requirements. Reconfigurability applied in coordination with a self-similar fractal structure results in significant improvement in antenna performance. The reason is that not only a wider selection of frequencies is obtained, but also similar radiation properties for all desired frequency bands are achieved [4].

Mechanical, Electronic or optical switching may be used with reconfigurable antennas. However, electronic switching is more frequently used because of its reliability and efficiency especially in dynamic bandwidth allocation. Electronic reconfigurability is generally obtained using lumped components such as FET transistors, PIN diodes, or RF MEMS switches. As compared to PIN diodes and FET transistors, RF MEMS switches usually have better performance in terms of insertion loss, linearity, isolation and power consumption [2].

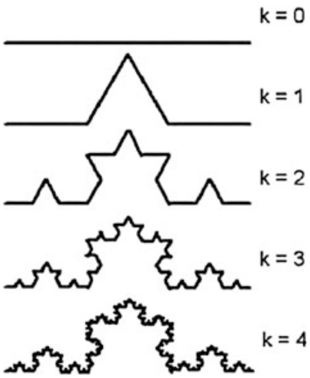


Fig. 1 The Koch curve iterations

3 Design of Antenna With and Without Koch Curve

3.1 Design Methodology

The critical requirement of the Antenna while dealing with several wireless standards such as WLAN, Wi-Fi, Wi-Max and UWB etc. is its Multiband and Wideband Characteristics. For meeting the same, the methodology is adopted which makes the best combination of the Koch curve geometry and the RF MEMS switch equivalents in the Antenna structure. The optimisation of the switch position leads to the multiple resonant frequencies.

3.2 Antenna Design-I

The Antenna Design-I uses Monopole antenna using the substrate, that is 1.6 mm thick with FR-4 Epoxy material having Dielectric constant εr = 4.4. The substrate dimensions are 4 cm by 4.5 cm. Here the feed line is Trapezoidal in shape which is connecting to the patch. The copper patch is rectangular and upper section is cut stepwise as shown in Fig. 2. The ground plane is partial and is flushed with the feedline. The Fig. 3 shows the Return loss characteristics. The wide band and Multiband Characteristics are observed which will be useful for many applications such as UWB, Wi-Max and Wi-Fi etc. The Koch fractal curve in the patch and U-slot can further improve the results as observed in Antenna Design-II.

3.3 Antenna Design-II

In this case the U-Koch fractal curve is used along with rectangular patch is shown in Fig. (1).

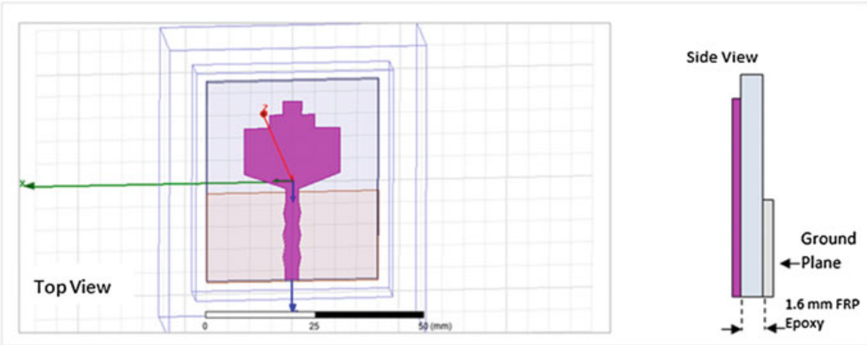


Fig. 2 Partial ground plane patch antenna

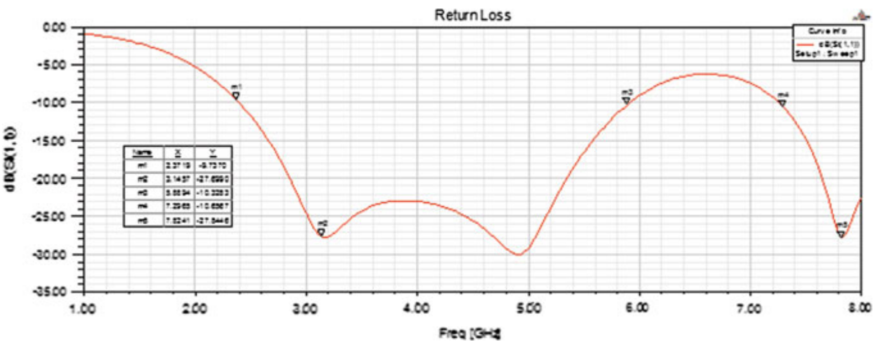


Fig. 3 Return loss characteristics of Design-I

In this antenna structure the partial ground plane with the zig-zag edges feed line is used. The FR4-epoxy material (εr = 4.4) is used for the substrate and its dimensions are 4 cm (*width*) × 4.5 cm (*Height*). The first iteration of Koch fractal geometry in the patch and the U-slot is incorporated in the design. This increases the electrical length of the antenna for operating it at lower frequency band. The matching section is of trapezoidal shape which connect the feed line to a U-Koch slotted rectangular patch [4] (Figs. 4, 5 and Table 1).

Figure 6 shows the antenna is resonating at three different frequencies with first resonance at 1.95 GHz having −20 dB Return Loss and the next two at 3.05 GHz with −34.3 dB, 6.2 GHz with Return Loss of −33 dB respectively over required frequency of operation.

The U-Koch Antenna is shown in Fig. 7 with first and second iteration. The results shows good Multiband and Wideband Characteristics from 2 to 6.2 GHz frequency range in Fig. 8.

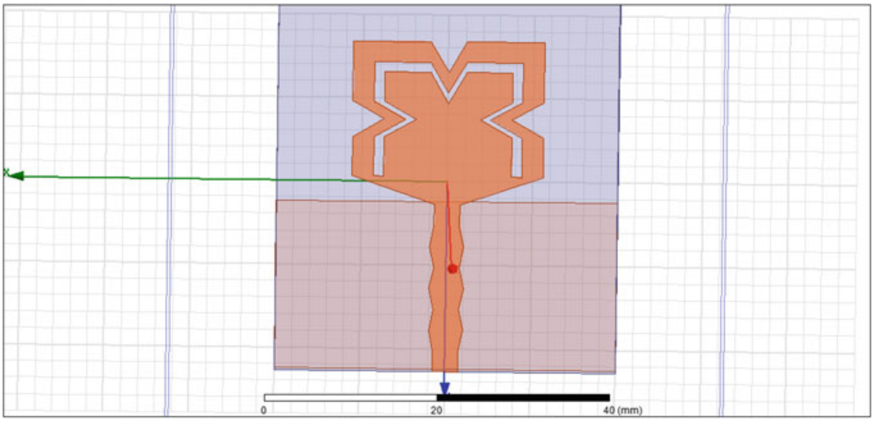


Fig. 4 U-Koch antenna structure without switches

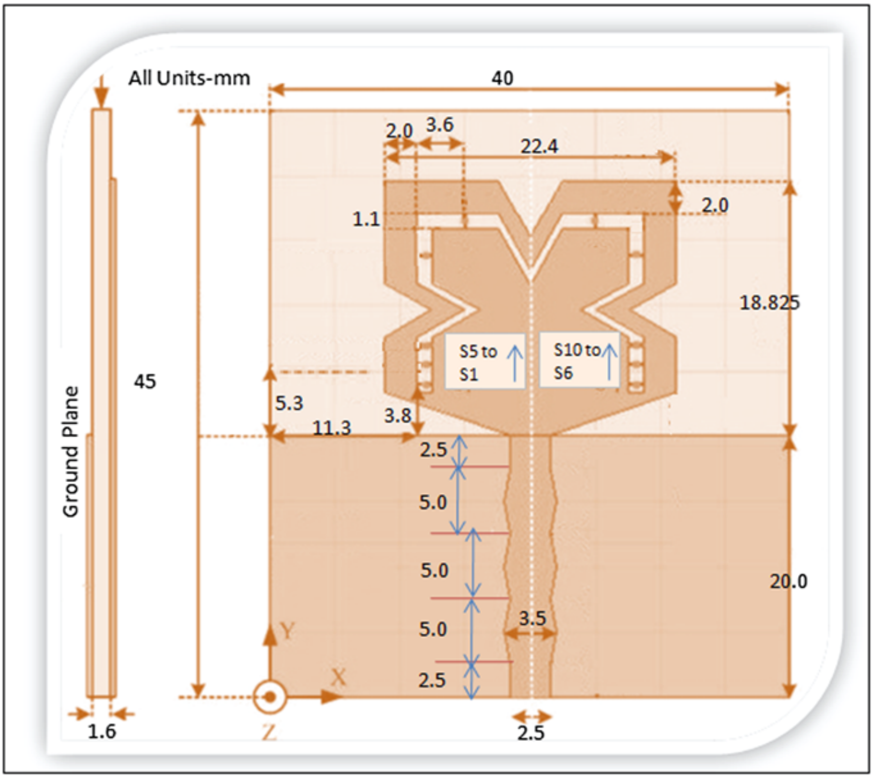


Fig. 5 U-Kotch antenna structure with first iteration and switches

Table 1 Design parameters

Design parameters	Material used	Dimensions
Substrate	FR-4-Epoxy (εr = 4.4)	(4 cm × 4.5 cm), 1.6 mm thickness
Patch	Copper	(22.4 mm × 18.825 mm)
Ground plane	Partial ground plane	(4 cm × 2 cm)
Slot (fractal dimensions)	–	1.1 mm
Feed line (trapezoidal matching section) microstrip line fed	–	(2.5 mm × 20 mm)
Switches	Copper rectangular strip	(400 μm × 200 μm)

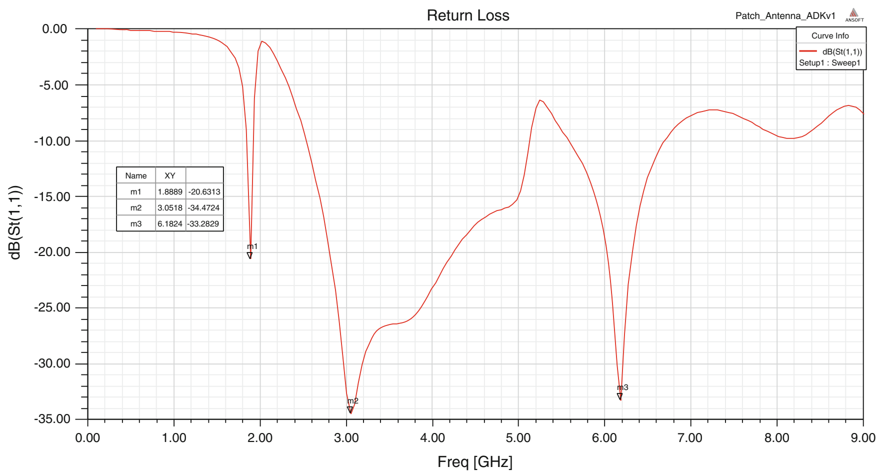


Fig. 6 Return loss characteristics of U-Koch antenna structure without switches

3.4 Antenna Structure with the Switches

The fractal antenna design shows the multiband and wide band characteristics but the performance of the antenna will further be enhanced by the reconfigurability.

3.4.1 RF MEM Switches

RF MEMS Switches have experienced increased use for Telecommunication application in last decade due to their high performance such as excellent isolation, Low insertion loss, linearity and low power consumption compared to other RF switches like PIN diode. These switches are classified by the following

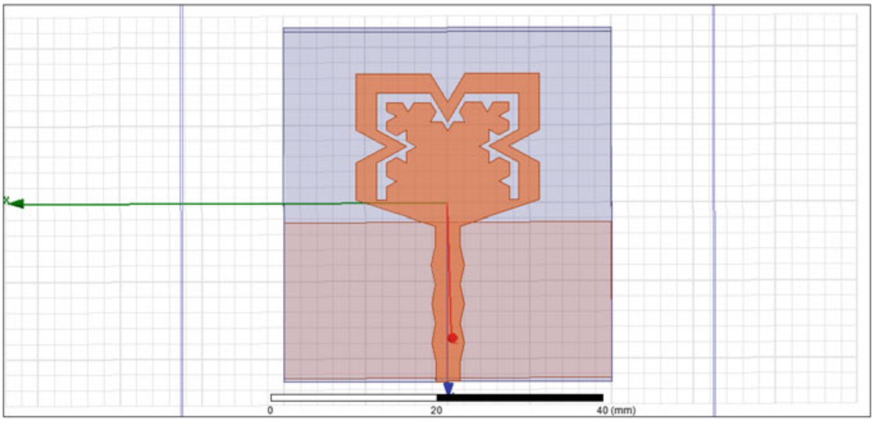


Fig. 7 U-Koch antenna structure with first and second iteration

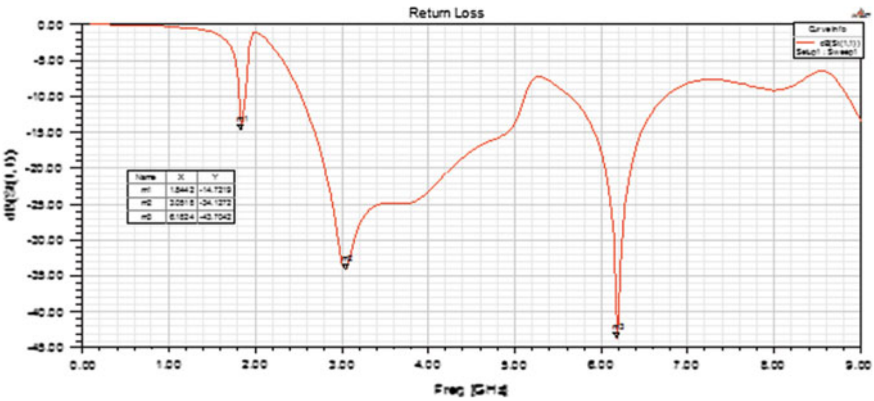


Fig. 8 Return loss characteristics of U-Koch antenna structure with first and second iteration

characteristics: (1) Mechanical structure (2) RF Circuit configuration (3) Form of contact. Spasos et al. [5] suggest that the Ohmic RF MEMS switch with gold cantilever and contact surfaces are best suited for Patch based Reconfigurable Antenna Design.

3.4.2 Switch Integration and DC Biasing Considerations

One of the main challenge of Reconfigurable Antenna development is the integration of the switches. Commonly, two major types of switches PIN diode and RF MEMS are used in developing Reconfigurable Antenna for Wireless applications. The advantages of PIN diodes are very a relatively high power handling capability, lesser driving voltage and very low cost. As there are no moving parts in PIN diode switches they also exhibits very high reliability. However these diodes requires a dc bias current in their ON state which consumes a more amount of dc power. On the other hand the RF MEMS switches are biased by high dc voltage and are actuated by built up static charges. For the same reason no current is drawn and they consume negligible power. Moreover RF MEMS offer low insertion loss, relatively high linearity, high isolation and very wide bandwidth [2].

3.4.3 Various Switching Conditions

The greatest benefit of RF MEMS switches are observed in terms of Isolation and Insertion loss. Here the switches are represented by including or omitting Copper strips of size 400 µm × 200 µm. In this case the OFF state is represented by taking out this rectangular strip of the antenna. The ON state is activated by mounting it in the same location. The five switches are considered as S1-S5 in left section of the slot and S6-S10 in the right section of the slot. The complete simulation is carried out using HFSS (Figs. 9, 10, 11, 12, 13 and Table 2).

The switch positions of the case 3 results in a wide band ranging from 2.7 to 5.8 GHz as shown in Fig. 14 and at the resonance good return loss of −40 dB is observed which is beneficial for wide band antennas.

Figure 15 shows Gain pattern in both Y-Z and X-Z plane. Omni-Directional pattern is observed in X-Z plane and '8' shaped pattern in Y-Z in all three cases results of Antenna Design.

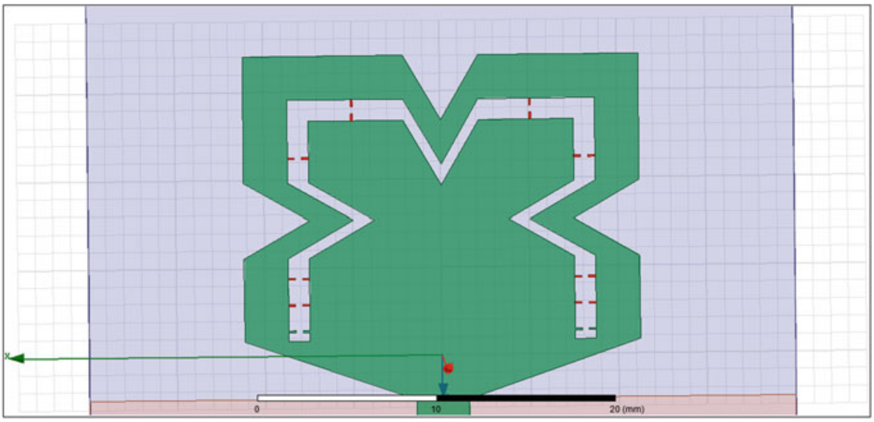


Fig. 9 Case-1 U-Koch antenna structure with switches S1 and S6 ON

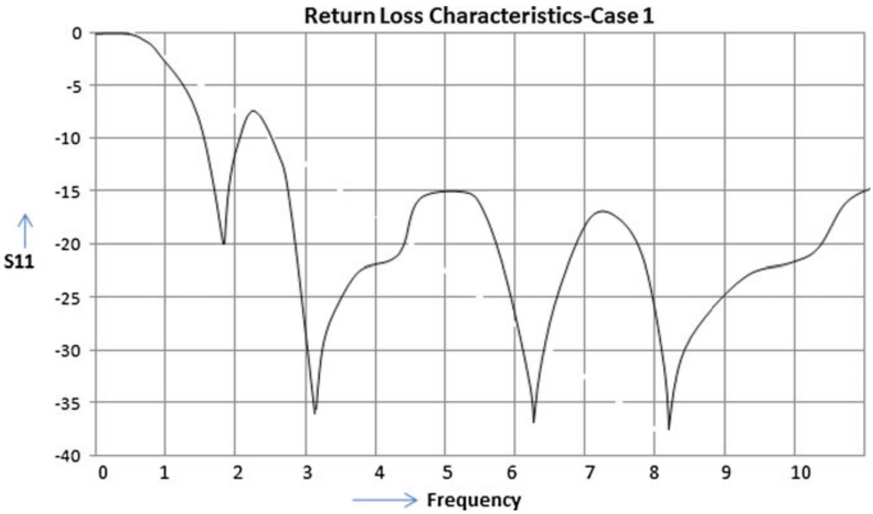


Fig. 10 Return loss characteristics case-1

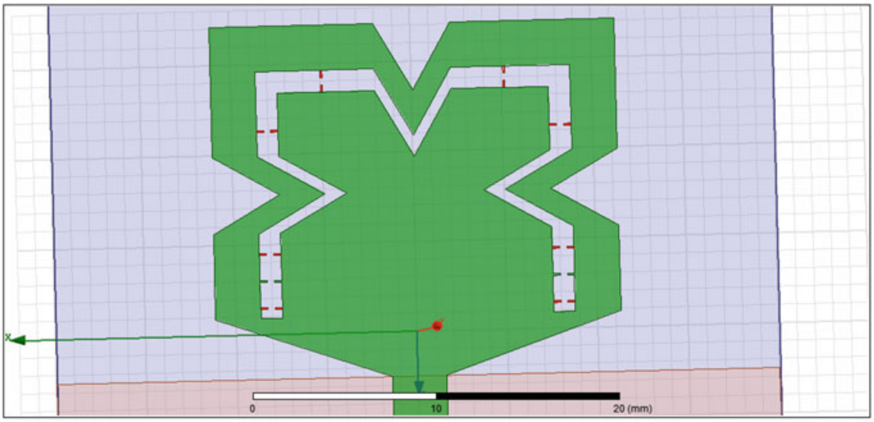


Fig. 11 Case-2 U-Koch antenna structure with switches S2 and S7 ON

4 Results and Discussions

The U-Koch Antenna shows the Multiband and Wideband operation in its first and second iterations as observed in Figs. 6 and 8. Also in case of Koch curve the resonance frequency is lowered as compared to normal structure without increasing the overall size of the Antenna. The Simulation is carried out by HFSS.

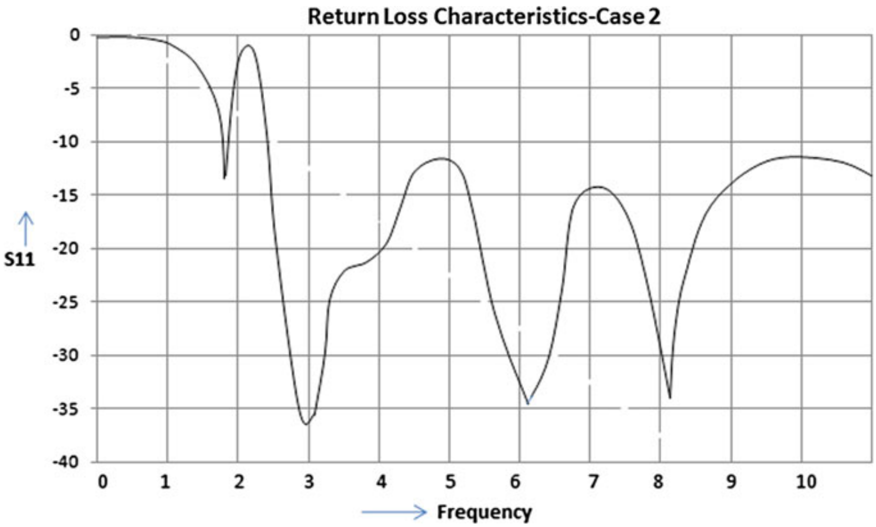


Fig. 12 Return loss characteristics case-2

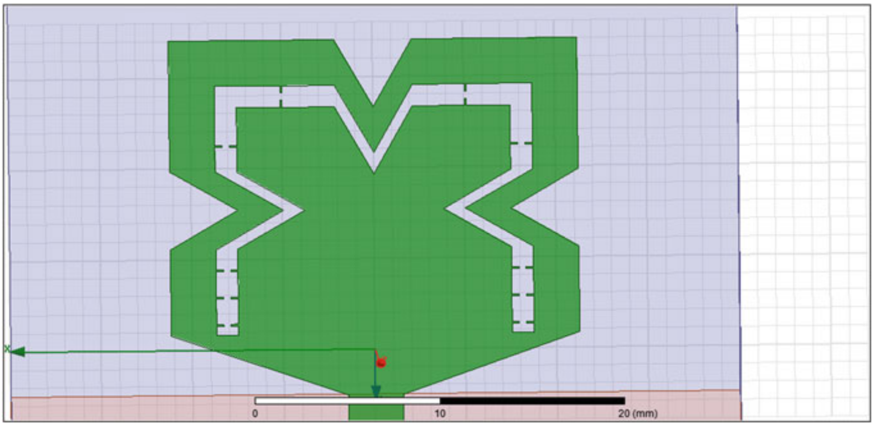


Fig. 13 Case-3 U-Koch antenna structure with all switches ON

Table 2 Status of switches for 3 cases under consideration

Case	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
I	ON	OFF	OFF	OFF	OFF	ON	OFF	OFF	OFF	OFF
II	OFF	ON	OFF	OFF	OFF	OFF	ON	OFF	OFF	OFF
III	ON	ON	ON	ON	ON	ON	ON	ON	ON	ON

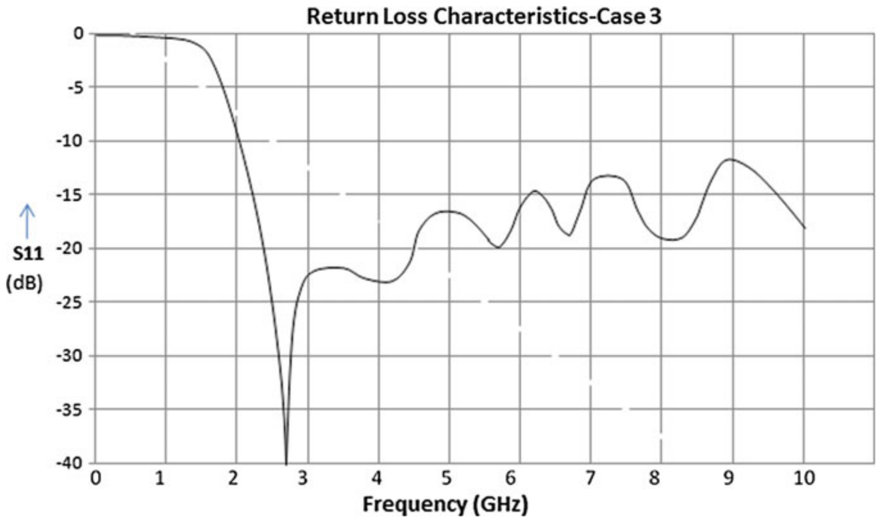
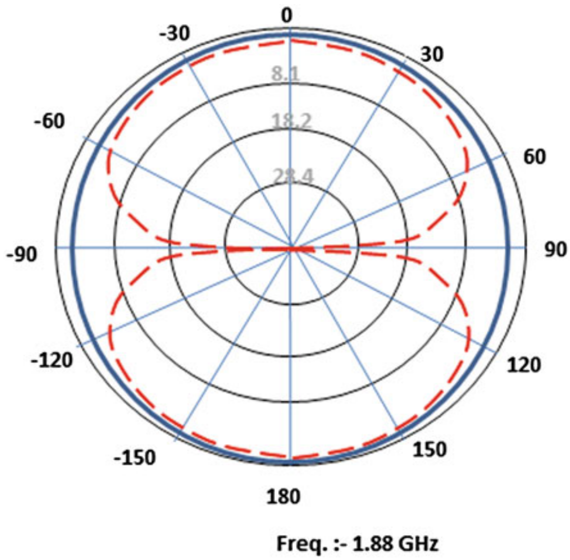


Fig. 14 Return loss characteristics case-3

Fig. 15 Radiation pattern of an antenna in X-Z (*solid line*), Y-Z plane (*dotted line*)



Even though the Fractal Koch curve Antenna shows the Multiband and Wideband operations, the Reconfiguration concept is enhancing the performance of the U-Koch Antenna Design by obtaining the more resonant frequencies as shown in Figs. 10 and 12. The results of the case I shows the first frequency of operation at 1.88 GHz and wideband at 3.1–6.2 GHz. Similar observation is there for case-II.

Table 3 Results for various switching conditions

Case No.	Switches status	Resonant frequency (GHz)	Return loss (dB)
Case-I	S1 and S6 ON Rest in OFF position	1.88	−20.0
		3.15	−35.5
		6.35	−37.0
		8.25	−38.0
Case-II	S2 and S7 ON	1.88	−13.0
		3.05	−36.0
		6.13	−34.3
		8.20	−34.0
Case-III	All switches ON	2.70	−40.0

For case-III first resonance occurs at 2.7 GHz and wideband up to 5.8 GHz. The return loss values for various switching conditions are given in Table 3. The radiation patterns of **‘8’** shaped in Y-Z plane and omni directional in X-Z plane are found in all three cases and are good for required frequency of operation.

5 Conclusion

A Koch fractal geometry is incorporated in U-slotted rectangular patch antenna which is simulated by HFSS software and the placement of the RF MEMS switches and its optimization has been done. The ON-OFF state of switches are considered by connecting or disconnecting of the metal strip. The result reveals several resonant frequency with good return loss i.e. also Multiband and wideband operation as a result of reconfiguration. The three different switching Conditions are analysed and the radiation pattern, return loss characteristics are observed by simulation. The reconfigurability is obtained by mounting RF Switches at selected locations across the U slot. The radiation pattern resembles 8-shaped and is satisfactory. This wideband and multiband design can be used in Satellite communication, Microwave Imaging and Wireless industry applications.

References

1. Anagnostou, D.E., Chryssomaillis, M.T., Papapolymerou, J., Guizhen Zheng, D.E.: Design, fabrication, and measurements of an RF-MEMS-based self-similar reconfigurable antenna. IEEE Trans. Antennas Propag. **54**(2), 2006
2. Yang, S., Zhang, C., Pan, H.K., Fathy, A.E.: Frequency reconfigurable antennas for multiradio wireless platforms. IEEE Microw. Mag., **1**, 66–74 (2009)
3. Krishna, D.D., Gopikrishna, M., Anandan, C.K., Mohanan, P., Vasudevan, K.: CPW fed Koch fractal slot antenna for WLAN/WiMAX applications. IEEE Antennas Wirel. Propag. Lett. **7**, 389–392 (2008)

4. Ramadan, A.H., Kabalan, K.Y., El-Hajj, A., Khoury, S., Al-Husseini, M.: A reconfigurable U-Koch microstrip antenna for wireless applications. Prog. Electromagnet. Res. PIER **93**, 355–367 (2009)
5. Spasos, M., Charalampidis, N., Mallios, N.: On the design of an Ohmic RF MEMS switch for reconfigurable microstrip antenna applications. WSEAS, Trans. Commun. **8**, 153–161 (2009)
6. Anagnostou, D.E., Papapolymerou, J., Tentzeris, M.M., Christodoulou, C.G.: A printed Log-periodic Koch dipole array(LPKDA). IEEE Antennas Wirel. Propag. Lett. **7**, 456–460 (2008)
7. Liu, W.C., Chen, W.R., Wu, C.M.: Printed double S-shaped monopole antenna for wideband and multiband operation of wireless communications. IEEE Proc. Microwave Antenna Propag. **151**(6), 473–476 (2004)
8. Peroulis, D., Sarabandi, K., Katehi, L.P.: Design of reconfigurable slot antennas. IEEE Trans. Antennas Propag. **53**(2), 645–654 (2005)
9. De Los Santos, H.J.: RF MEMS Circuit Design for Wireless Communications. Boston, London (2002)
10. Balanis, C.A.: Antenna Theory, Analysis and Design. Wiley, Hoboken (2005)
11. Varadan, V.K., Vinoy, K.J.: RF MEMS and their Application. Wiley, Hoboken (2003)
12. Yeom, I., Choi, J., Kwoun, S.: Analysis of RF front end performance of reconfigurable antenna with RF switches in the far field. Int. J. Antennas Propag. **2014**, 14 (2014)

Comparison of Meta-heuristic Algorithms for Vehicle Routing Problem with Time Windows

Teerapun Saeheaw and Nivit Charoenchai

Abstract This paper proposes three meta-heuristic algorithms, namely cuckoo search, central force optimization, and chemical reaction optimization for solving vehicle routing problem with time windows (VRPTW). A comparison study between different meta-heuristic algorithms aims to identify their respective strengths and weaknesses . The objective of VRPTW is to serve all customers, at different geographic locations, with varying demands and within specific time windows. The performance evaluation is tested on Solomon's 56 benchmark instances of 100 customer problems, and yielded 24 solutions better than or equal to the best known solution provided by published papers. This paper is also among the first to document the implementation of all the three meta-heuristic algorithms for VRPTW together with their comprehensive results.

1 Introduction

Solving the vehicle routing problem (VRP) is a key to efficient transport management.

The vehicle routing problem with time windows (VRPTW) is an extension of VRP occurring within many distribution systems. It combines both the vehicle routing and scheduling problem with time window constraints which often arises in many real world applications. The objective of VRPTW is to serve a number of customers, at different geographic locations, with various demands and within specific time windows.

It has been extensively studied by many optimization techniques such as heuristic methods, exact approaches and meta-heuristic algorithms to increase efficiency to find optimal solutions. Due to the high complexity level of VRPTW

T. Saeheaw (✉) · N. Charoenchai
Department of Industrial Engineering, Chiang Mai University, Chiang Mai, Thailand
e-mail: murinee@hotmail.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_108

and its wide applicability to real-world scenarios, meta-heuristic algorithm is capable of solving high-quality solutions in short computational time.

In this paper, three meta-heuristic algorithms are applied to solve VRPTW, namely cuckoo search, central force optimization, and chemical reaction optimization. The local search is added to the proposed algorithms to obtain high-quality feasible solutions.

The rest of this paper is organized as follows. Section 2 starts with a description of the VRPTW. In Sect. 3, the proposed meta-heuristic algorithms are presented in detail. The computational results are discussed in Sect. 4. Finally, conclusions and future research directions are provided in Sect. 5.

2 Problem Description

The vehicle routing problem with time windows is a fleet of vehicles based at a depot which is required to serve a number of customers, at different geographic locations, with various demand and within specific time windows for minimizing the total travel distance.

The VRPTW constraints consist of a set of K vehicles, a set of N customer nodes, and a depot node, denoted by customer 0, and a network connection between the depot and customers x_{ijk} where $i, j \in \{0, 1, 2, \ldots, N\}$; $k \in \{0, 1, 2, \ldots, N\}$; $i \neq j$ is equal to 1 if vehicle k drives from node i to node j, and 0 otherwise. All route must start from the depot, visit customer nodes, and then return to depot. The travel cost c_{ij}, the travel time t_{ij} and the Euclidean distance between the customer nodes equal to each other. The travel time t_{ij} includes a service time at customer i. Each vehicle has a same capacity q_k. Each customer has a varying demand m_i. At each customer, the start of the service time s_i must be within a given time interval, called the time window, $[e_i, l_i]$. If vehicles have the arrival time a_i before the time window, the wait time w_i occurs until a service becomes possible.

$$\text{minimize} \sum_{i=0}^{N} \sum_{j=0}^{N} \sum_{k=1}^{K} c_{ij} x_{ijk} \tag{1}$$

$$\text{subject to:} \quad \sum_{k=1}^{K} \sum_{j=1}^{N} x_{ijk} \leq K \quad for\ i = 0 \tag{2}$$

$$\sum_{j=1}^{N} x_{ijk} = \sum_{j=1}^{N} x_{jik} \leq 1 \quad for\ i = 0,\ \forall k \in K \tag{3}$$

$$\sum_{k=1}^{K} \sum_{j=0}^{N} x_{ijk} = 1 \quad \forall i \in N \tag{4}$$

$$\sum_{k=1}^{K}\sum_{i=0}^{N} x_{ijk} = 1 \quad \forall j \in N \tag{5}$$

$$\sum_{i=1}^{N} m_i \sum_{j=0}^{N} x_{ijk} \leq q_k \quad \forall k \in K \tag{6}$$

$$\sum_{i=0}^{N}\sum_{j=0}^{N} x_{ijk}(a_{ij} + s_i + w_i) \leq r_k \quad \forall k \in K \tag{7}$$

$$a_0 = w_0 = s_0 = 0 \tag{8}$$

$$\sum_{k=1}^{K}\sum_{i=0}^{N} x_{ijk}(a_i + a_{ij} + s_i + w_i) \leq a_j \quad \forall j \in N \tag{9}$$

$$e_i \leq (a_i + w_i) \leq l_i \qquad \forall i \in N \tag{10}$$

The objective function is equation (1). Equation (2) specifies that the maximum routes are going out of the depot. Equation (3) ensures that every route starts and ends at the depot. Equations (4) and (5) define that each customer must be assigned to exactly one vehicle. Equation (6) is the capacity constraint. Equation (7) guarantees that a travel time is not exceeded a maximum route time r_k. Equation (8) through equation (10) ensure that all time windows are respected.

3 Meta-heuristic Algorithms

Meta-heuristic algorithms are mathematical optimization techniques which simulate the physical world, the natural phenomenon, the evolution of living organisms, and the swarm behaviors with relatively few modifications to solve optimization problems. They perform the generation process combined with a variety of intelligently different concepts to explore and exploit the search space. Each algorithm has its strengths and weaknesses. Meta-heuristics not only compute high quality solutions in short computation time, but also escape from local minima than traditional methods. Meta-heuristics have become powerful techniques to solve a wide variety of problems. The proposed algorithms are explained in detail.

3.1 Cuckoo Search

Cuckoo Search [1] is a population based stochastic global search inspired by breeding behavior of some cuckoo species. Cuckoos are fascinating birds because of their aggressive reproduction strategy. Evolutionary advantages of cuckoo breeding behavior is laying their eggs in the nests of other host birds, and let these host birds raise their chicks. In CS idealized, an egg is defined as a solution. Each cuckoo lays an egg at once, and dumps its egg in a randomly chosen nest. The best nests with high quality eggs will carry out to the next generation. When generating new solutions, a Levy flight distribution is performed to find a new nest. The number of available host nests is fixed, and the eggs laid by cuckoos can be discovered by the host bird with a probability $P_a \in [0, 1]$. The host birds can either throw the eggs away or abandon the nests, and built a completely new nest. CS used

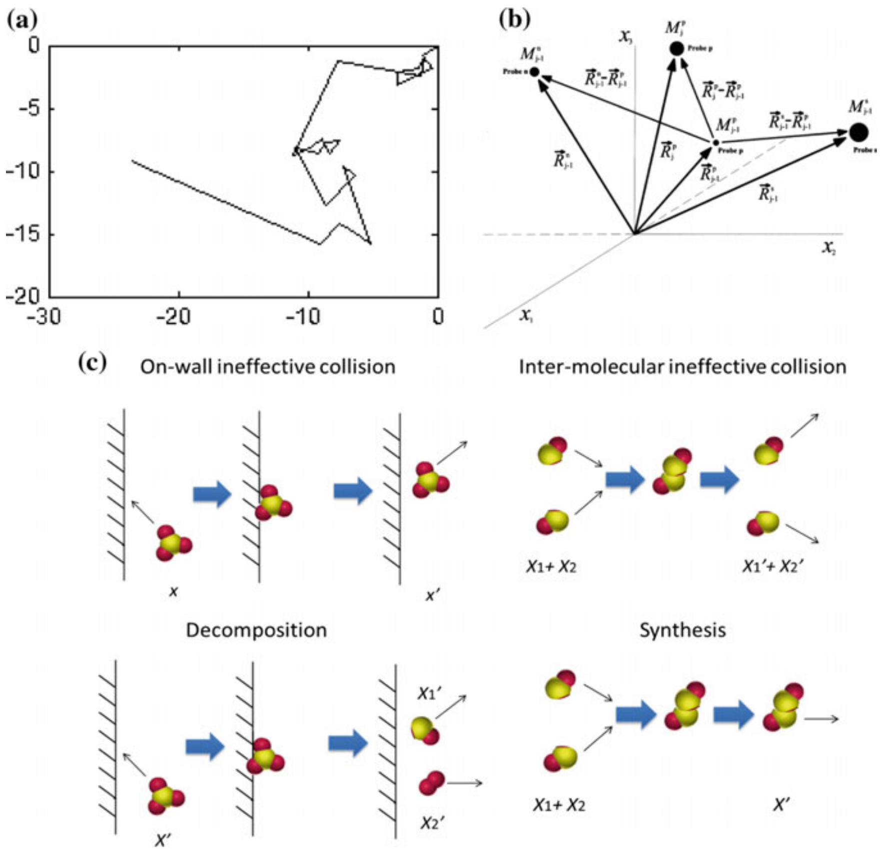


Fig. 1 Comparison between different optimization techniques. (**a**) Levy flight distribution in CS (**b**) Multi-dimensional search in CFO (**c**) Four types of chemical reactions in CRO.

Levy flights in Fig. 1a for making non-Brownian random walks with long jumps around the best solution obtained so far. The pseudocode is summarized below.

```
begin
   Objective function f(x)
   Generate an initial population of n host nest Evaluate
   fitness and rank eggs
     while (t > MaxGeneration) or Stop criterion t=t+1
     Get a cuckoo generate a new solution by Lévy flights
            Evaluate quality/fitness, Fi
            Choose a random nest j
            if(Fi>Fj)
                  Replace j by the new solution
            end if
            Worst nest is abandoned with probability pa
            and new nest is built
            Evaluate fitness and Rank the solutions and
            find current best
      end while
            Post process results and visualization

end
```

3.2 Central Force Optimization

Central force optimization [2] is a deterministic multi-dimensional search evolutionary algorithm based on gravitational kinematics from Newton's law of universal gravity and Newton's second law of motion. In the physical universe, objects traveling through three dimensional spaces have become trapped in close orbits around highly gravitating masses. In the CFO metaphor, mass is the value of the objective function to be maximized. CFO's search uses probes as its basic populations as shown in Fig. 1b. As time progresses, these probes are slowly scattered through an adaptive decision space move towards the probes, according to two simple equations of motion derived from its gravitational metaphor. Equations are presented for the position and velocity over decision making by comparing the influence of gravity in the universe that is consistent with the objective function. The deterministic search on the law of gravity and mass interactions lead to be fast and well perform reproducible. CFO has its self-tuning to reactive implementations of accelerating fitness convergence towards the optimal solutions. The procedure can be summarized as the below pseudocode.

```
begin
   Objective function f(x)
   Generate initial parameters: the objective function's
   dimension ND, boundaries xmin, xmax, total number of
   probes NP, gravitational constant G, and acceleration
   parameters α and β
   Initialize position vectors R and acceleration vectors
   A of each probe to be zero
   Compute the initial probe distribution X
   Calculate initial fitness matrix M
   Repeat update fitness
      if probe flies out of the boundary
         Update probe positions
         Retrieve errant probes
         Calculate fitness matrices of each probe
         Compute new acceleration vectors A
         Increase time step
      end
   Until stopping criteria is reached
   Stop and put out the best solution

end
```

3.3 Chemical Reaction Optimization

Chemical reaction optimization [3] is a population-based meta-heuristic inspired by the first principles simulations of chemical reactions, as well as microscopic phenomena in chemical reactions. CRO provides different characteristics of search techniques to find optimal solutions. The chemical reaction system is presumed to emulate a function. The molecular structure is a solution to the problem, the potential energy (*PE*) is the objective function, the kinetic energy (*KE*) is measurement tolerances for accepting an inferior solution, the number of hits are total number of moves, the minimum structure is the optimal solution, the minimum value is the optimal function value, and the minimum hit number is the current optimal solution found by the number of moves. The collision changes the operating mechanism to find new solutions. Four kinds of elementary steps take place during an iteration, are based upon following synopsis as shown in Fig. 1c. In the on-wall ineffective collision, there is only one molecule colliding with the wall of container. In the decomposition, one molecule is divided into several. The inter-molecular ineffective collision involves more than one molecule interacting with each other. The synthesis combines many molecules into one. The fundamental assumption of CRO ensures the conservation of energy when new solutions are generated with the operators. The number of molecules remains the same and only the neighborhoods of original solution are searched. The intensification is mainly contributed by on-wall ineffective collision and inter-molecular ineffective collision, while decomposition and synthesis perform diversification, synchronously. The main concern of CRO is the inter-changing of energy from *PE* to *KE* between different molecules. The CRO algorithm has two advantages, both searching neighborhoods of the original solution and escaping local extrema entrapment, automatically operated by itself. The basic steps of CRO can be summarized as the pseudocode shown below.

```
begin
   Objective function f(x)
   Generate initial parameters: PopSize,KELossRate,
MoleColl, buffer, InitialKE, α, and β
      Create PopSize number of molecules
   while the stopping criteria not met do
      Generate b∈[0,1]
      if b>MoleColl then
            Randomly select one molecule Mω
            if Decomposition criterion met then
                  Trigger Decomposition
            else
                  Trigger OnwallIneffectiveCollision
            end if
      else
            Randomly select two molecules Mω1 and Mω2
            if Synthesis criterion met then
                  Trigger Synthesis
            else
                  Trigger IntermolecularIneffective
                  Collision
            end if
      end if
      Check for any new minimum solution
   end while
   Show the best solution found and its objective function
   value
end
```

4 Computational Result

The proposed algorithms described in Sect. 3 were coded in Java, and all experiments were performed on an Intel Core i7 CPU at 3.40 GHz with 8 GB of RAM on a Windows XP7 platform. The parameters set as follows were found to be robust for most of the test problems according to our pilot tests. CS parameters are: pa = 0.25, and n = 100. CRO parameters are: MoleColl = 0.2, buffer = 0, InitialKE = 1000, PopSize = 10, KELossRate = 0.1, α = 1, and β = 10. CFO parameters are: NP = 2, ND = 8, Nt = 100, G = 2, α = 2, and β = 2.

The Solomon's 56 benchmark instances of 100 customer problems [4] were selected to evaluate the effectiveness of our proposed algorithms for VRPTW, and can be downloaded from the website http://www.sintef.no/Projectweb/TOP/VRPTW/Solomon-benchmark/100-customers/. The problems are categorized into six problem types, namely C1, C2, R1, R2, RC1 and RC2. For C1 and C2 problems, the customer locations are grouped in clusters, while on the contrary, the customers are randomly distributed in R1 and R2 problems. The RC1 and RC2 problems consist of both random and clustered customer locations. Type 1 problems also differ from type 2 problems in terms of the scheduling horizon. The C1, R1 and RC1 problems have narrow time windows and small vehicle capacity, while C2, R2 and RC2 problems have longer time windows and large vehicle capacity. Therefore, the solutions to type 2 problems have significantly fewer routes and considerably more customers per route.

Table 1 lists a snap comparison of three tested algorithms with the best known solutions provided by published papers. The quality of solutions are derived from the data listed in Appendix A. The first row of each problem describes the average number of vehicles used, while the second represents the average total distance in all six categories.

Table 1 Average performance of the proposed algorithms concerning the quality of solutions

Problem	BKS	CS	CFO	CRO
C1	10.00	10.00	10.00	10.00
	828.38	828.38	828.38	828.38
C2	3.00	3.00	3.00	3.00
	589.86	589.86	589.86	589.86
R1	11.92	13.08	13.08	13.25
	1209.76	1185.95	1192.61	1192.74
R2	2.73	4.55	4.36	4.82
	941.09	904.90	906.33	888.65
RC1	11.50	12.75	12.88	13.50
	1384.17	1350.54	1361.31	1375.45
RC2	3.25	5.00	5.13	5.88
	1119.24	1072.52	1116.01	1018.94

Table 1 shows that all algorithms are able to obtain the same optimal solutions for C1 and C2 problems. CS seems to be most efficient approach in R1, RC1 and RC2 problems. Both CS and CFO perform equally well for R2 problems.

In general, CS outperforms the other algorithms in most of the problems. For those in which CS is not the best, CRO can achieve many good results in RC1 and RC2 with random-clustered distribution. However, CFO still remains a strong competitor to CRO as it is a good compromise of quality solutions in R1 and R2 with random distributions.

In Appendix A, we compare the performance of our proposed algorithms with the list of non-dominated solutions [5] for each instance of the problem. The bolded numbers indicate that the corresponding solution is exactly equal to ever published results. In addition, the bolded numbers with underline indicate an improvement solution. Our numerical results demonstrate that the proposed algorithms are able to generate reasonable good solutions for most of VRPTW in terms of solution quality. Twenty-four out of fifty-six test problems can be solved successfully by the proposed algorithms.

5 Conclusions and Further Study

Although the parameter setting may simply not be the best one, the results can converge to optimal solutions.

Computational results show that the performance of proposed algorithms are competitive for solving VRPTW in terms of solution quality when compared with the best solutions published so far. The effectiveness comes from the combination of the following reasons. First, CS maintains its stochastic behavior with Levy flights that can efficiently achieve optimal solutions. Second, CFO basically uses a deterministic search strategy which can be quickly coverage solutions, and well perform reproducible. Third, CRO has its intensification and diversification between molecules which adapt to the problems automatically making its flexibility for searching optimal solutions. Fourth, the solution quality is improved by 2-opt in the route construction.

For future research, we will develop hybrid meta-heuristics among the proposed algorithms to extend their application to VRPTW.

Acknowledgments The author would like to thank the financial support provided by the Thailand Research Fund through the Royal Golden Jubilee Ph.D. Program (Grant No. PHD/0113/2552) and by the Graduate School, Chiang Mai University, Thailand.

Appendix A

A comparison of the solutions obtained by proposed algorithms for Solomon's 56 benchmark instances of 100 customer problems is given below.

Problem	BKS		CS		CFO		CRO	
	NV	TD	NV	TD	NV	TD	NV	TD
C101	10	828.94	10	**828.94**	10	**828.94**	10	**828.94**
C102	10	828.94	10	**828.94**	10	**828.94**	10	**828.94**
C103	10	828.06	10	**828.07**	10	**828.06**	10	**828.06**
C104	10	824.78	10	**824.78**	10	**824.78**	10	**824.78**
C105	10	828.94	10	**828.94**	10	**828.94**	10	**828.94**
C106	10	828.94	10	**828.94**	10	**828.94**	10	**828.94**
C107	10	828.94	10	**828.94**	10	**828.94**	10	**828.94**
C108	10	828.94	10	**828.94**	10	**828.94**	10	**828.94**
C109	10	828.94	10	**828.94**	10	**828.94**	10	**828.94**
C201	3	591.56	3	**591.56**	3	**591.56**	3	**591.56**
C202	3	591.56	3	**591.56**	3	**591.56**	3	**591.56**
C203	3	591.17	3	**591.17**	3	**591.17**	3	**591.17**
C204	3	590.60	3	**590.60**	3	**590.60**	3	**590.60**
C205	3	588.88	3	**588.88**	3	**588.88**	3	**588.88**
C206	3	588.49	3	**588.49**	3	**588.49**	3	**588.49**
C207	3	588.29	3	**588.29**	3	**588.29**	3	**588.29**
C208	3	588.32	3	**588.32**	3	**588.32**	3	**588.32**
R101	19	1650.80	20	1643.79	20	1648.34	19	1651.36
R102	17	1486.86	18	1474.28	18	1476.61	18	1474.13
R103	13	1292.67	14	1215.59	14	**1213.62**	14	1215.59
R104	9	1007.31	11	981.57	11	976.61	11	981.56
R105	14	1377.11	15	**1360.78**	15	1366.69	15	**1360.78**
R106	12	1252.03	13	1240.53	13	1254.16	13	1253.56
R107	10	1097.01	11	1083.89	11	1079.78	11	1082.68
R108	9	960.88	10	946.74	10	951.10	10	972.16
R109	11	1194.73	12	1154.73	13	1166.32	12	1169.19
R110	10	1118.84	12	1105.98	12	1087.95	12	1111.71
R111	10	1096.73	11	**1054.23**	12	1084.56	11	1059.86
R112	9	982.14	10	969.34	10	1007.14	11	978.70
R201	4	1252.37	6	1183.87	7	1158.82	7	1177.47
R202	3	1191.70	5	1074.57	6	1049.50	5	1066.33
R203	3	939.50	4	921.22	6	881.12	4	930.48
R204	2	825.52	4	**739.76**	4	**738.41**	3	766.95
R205	3	994.43	5	957.58	5	964.72	5	958.46
R206	3	906.14	5	884.89	4	892.13	4	902.81
R207	2	890.61	4	818.56	3	817.88	4	818.56

(continued)

(continued)

Problem	BKS		CS		CFO		CRO	
	NV	TD	NV	TD	NV	TD	NV	TD
R208	2	726.82	4	733.86	4	705.33	4	720.30
R209	3	909.16	5	886.62	5	869.91	5	866.40
R210	3	939.37	5	932.73	5	**922.26**	4	941.60
R211	2	776.32	3	820.29	4	775.11	3	820.29
RC101	14	1696.95	15	1627.71	16	1642.03	16	1644.45
RC102	12	1554.75	14	1466.44	15	1501.02	14	1466.44
RC103	11	1261.67	12	1285.70	12	1329.89	12	1282.66
RC104	10	1135.48	10	1143.06	11	1170.77	10	1206.99
RC105	13	1629.44	15	1547.44	16	1552.89	15	1533.29
RC106	11	1424.73	13	1379.71	14	1403.96	13	1394.88
RC107	11	1230.48	12	1218.37	13	1254.77	12	**1211.11**
RC108	10	1139.82	11	1135.87	11	1148.27	11	1150.64
RC201	4	1406.94	7	1280.19	7	1284.29	7	1441.92
RC202	3	1365.64	6	1276.48	8	**1095.64**	6	1439.79
RC203	3	1049.62	4	952.80	5	932.70	4	952.80
RC204	3	798.46	3	809.42	4	806.57	4	795.48
RC205	4	1297.65	7	1220.22	8	1181.28	6	1258.06
RC206	3	1146.32	5	1108.06	5	1073.38	6	1118.09
RC207	3	1061.14	5	1056.18	6	977.34	5	1003.94
RC208	3	828.14	3	876.83	4	800.31	3	917.99

Notes: BKS is the best known solution provided by published papers. *NV* is the number of vehicles. *TD* is the total distance travelled.

References

1. Yang, X.-S., Deb, S.: Cuckoo search via Lévy flights. In: World Congress on Nature and Biologically Inspired Computing, pp. 210–214 (2009)
2. Formato, R.A.: Central force optimization: a new metaheuristic with applications in applied electromagnetics. Prog. Electromag. Res. **77**, 425–491 (2007)
3. Lam, A.Y.S., Li, V.O.K.: Chemical-reaction-inspired metaheuristic for optimization. IEEE Trans. Evol. Comput. **14**(3), 381–399 (2010)
4. Solomon, Marius M.: Algorithms for the vehicle routing and scheduling problems with time window constraints. Oper. Res. **35**, 254–265 (1987)
5. Chiang, T.-C., Hsu, W.-H.: A knowledge-based evolutionary algorithm for the multiobjective vehicle routing problem with time windows. Comput. Oper. Res. **45**, 25–37 (2014)

Reconfigurable Hardware Acceleration of RGB to HSL Converter

Voon Siew Soon, Chee Yuen Lam and Ehkan Phaklen

Abstract The RGB colour model is an additive colour model in which red, green, and blue light are added together in numerous ways to reproduce a broad array of colours. But there is limitation for RGB colour model to provide sufficient information for image analysis. Hue, saturation, luminance (HSL) is a colour model that different from RGB. HSL provides the information such as colour in degree, the saturation of the colour and the brightness of the colour. This information is much more suitable for image analysis. In this paper the transformation between RGB to HSL is implemented in FPGA. Parallelism capability of FPGA has simplified the processing steps and only consumes one clock cycles to produce the HSL value for each pixel without data latency. The proposed architecture achieved 98.44 % accuracy with hardware verification.

1 Introduction

Digital image becomes very important and useful in daily life. The colour in digital image can be represented in various colour model such as RGB, HSV, HSL, YCbCr and more. Each of this colour model has its specific application. The RGB colour model is commonly used in digital system and imaging due to its simplicity. However, HSL colour model is able to further enhance in image processing and colour. For the work of Pérez et al. [1] and Wang et al. [2] using hue and saturation histogram for image object tracking. The work from Cho et al. [3] has used hue

V.S. Soon (✉) · C.Y. Lam · E. Phaklen
School of Computer and Communication Engineering, University Malaysia Perlis, Pauh Putra Campus, 02060 Arau, Perlis, Malaysia
e-mail: svsiew@hotmail.com

C.Y. Lam
e-mail: cheeyuen_lam@hotmail.com

E. Phaklen
e-mail: phaklen@unimap.edu.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_109

component to perform image or video segmentation. Hue and saturation color model is closer to human eye perception and stable when illumination change. Since HSL and HSV are widely apply in machine vision, there is a need to develop a fast and efficient transformation method from RGB to HSL color model.

This paper proposes digital hardware architecture to implement the conversion algorithm to convert the RGB colour value into HSL value in FPGA. The advantages of the FPGA is parallelism and scalable. All the arithmetic operation in the conversion algorithm are implemented using hardware such as adder, subtractor, multiplier and divider. By using FPGA, parts of the conversion algorithm are implemented in parallel manner in order to simplify the conversion steps and speed up process.

Section two of this paper is about the HSL colour model and the RGB to HSL conversion algorithm. Section three discusses on the implementation hardware architecture of the conversion algorithm while section four is the analysis and discussion of the result for this converter.

2 Hue Saturation Luminance (HSL) Colour Model

As similar to the RGB colour model, three components are hue, saturation and luminance. According to Agoston et al. [4], hue component is represented in angle of degree value from 0 to 360° and the saturation and luminance component are represented in percentage or range between 0 and 1. Fortner and Meyer [5] says hue is referred as the colour of the rainbow. In other words, it is the monochromatic light span from 400 to 700 nm in wavelength in the visible light spectrum. Saturation is defined as the purity of the colour. For 100 % of saturation means there is a pure colour without mixing with other light frequency. A 0 % means the colour is all presented and becomes a white colour. Figure 1 shows the HSL colour

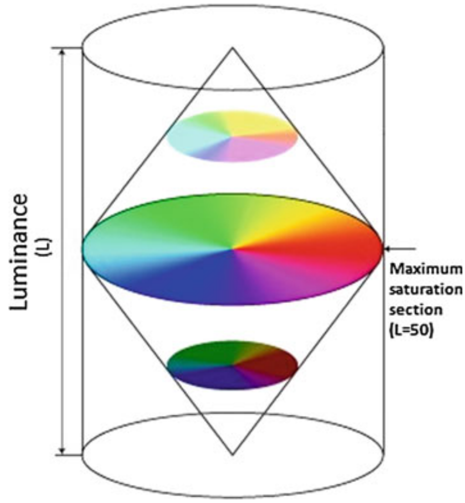


Fig. 1 HSL dual cone in cylinder model [6]

model represented by cylindrical model. The ‘lightness’ in Fig. 1 is same as the luminance which is used in this paper.

Figures 2 and 3 describes the generation of hue from primary colour. Figure 2 shows how the changes of the primary colour red, green, blue generate the hue. The hue is constructed by a series of variation of primary colour which are red, green and blue from 0 to 360°and producing a rainbow spectrum. Figure 3 shows the relation of hue and saturation under the condition of 100 % luminance. When the saturation of a colour increases, that particular colour is slowly faded until become white colour when the saturation come to 100 % regardless the degree of hue.

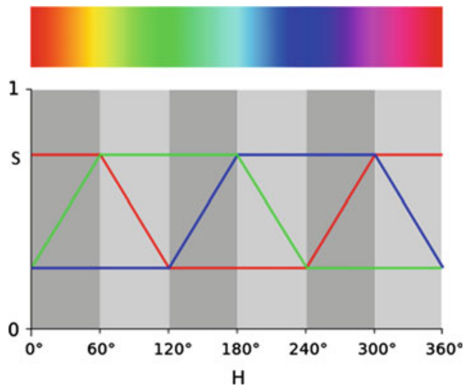


Fig. 2 Generation of hue rainbow colour

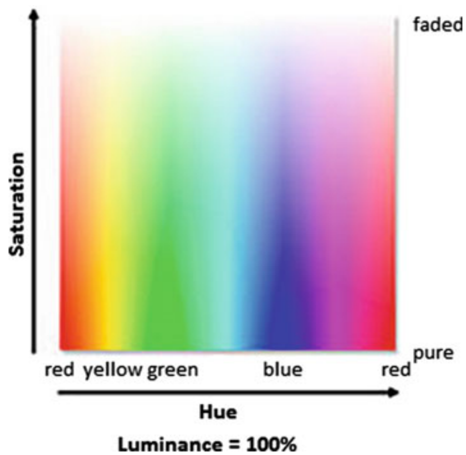


Fig. 3 Relation of saturation with hue

2.1 RGB to HSL Conversion Algorithm

This section describes the RGB to HSL conversion algorithm. This algorithm was adopted from Aguston's book and it requires red, green and blue value as the input sources. The algorithm from the book is translated into mathematical Eq. (1).

$$\begin{bmatrix} r \\ g \\ b \end{bmatrix} = \frac{1}{255} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{1}$$

From (1), the 'R', 'G' and 'B' is the binary input of RGB value respectively. The 'r', 'g' and 'b' are the normalize RGB value in the range of 0–1. Since this conversion algorithm is applied in digital hardware, the input of RGB values will be in the digital binary format. Therefore a format change is required. The required input value of RGB is divided by 255 in decimal order to satisfy the conversion algorithm. The normalized rgb value is then used for the rest of the conversion algorithm. The maximum and minimum values of the rgb are selected. A difference, d, of the maximum and minimum value is obtained by using formula (2). This difference value is an important variable and it is used to calculate the saturation and hue in the coming steps.

$$d = \max(r, g, b) - \min(r, g, b) \tag{2}$$

Luminance, L is obtained by averaging value of the maximum and minimum rgb value as define in (3).

$$L = \frac{\max(r, g, b) + \min(r, g, b)}{2} \tag{3}$$

The saturation, S depends on the value of luminance with either two conditions are valid or not. The first condition is less than 0.5 and the second condition of equal or greater than 0.5 as shown in (4). On the other hand, hue, H depends on the maximum value of rgb. For any of r, g, b value is the maximum value, it will decided the equation to calculate hue which is shown in (5). Any of the condition to calculate the hue is required to multiply with 60° in order to obtain the exact hue angle.

$$S = \begin{cases} \dfrac{d}{\max(r, g, b) + \min(r, g, b)}, & \text{when L} < 0.5 \\ \dfrac{d}{2 - \max(r, g, b) - \min(r, g, b)}, & \text{when L} \geq 0.5 \end{cases} \tag{4}$$

$$H = \begin{cases} 60\left(\frac{g-b}{d}\right), & \text{when } r_{max} \\ 60\left(2.0 + \frac{b-r}{d}\right), & \text{when } g_{max} \\ 60\left(4.0 + \frac{r-g}{d}\right), & \text{when } b_{max} \end{cases} \quad (5)$$

When the value of the hue angle is less than zero or a negative number, it needs to add with 360° to obtain a positive hue angle since hue is represented by a round colour disk. The value of the HSL output requires a round integer number. This is to simplify the implementation of the conversion algorithm in digital hardware. Next section of this paper will be discussed on the detail implementation of this algorithm in digital hardware.

3 Proposed Conversion Digital Hardware Architecture

In the implementation process of this conversion algorithm in digital hardware, primitive digital device such as adder, subtractor, comparator, multiplexer and others are preferred. This is because primitive device has low complexity and fast propagation time. There are total of four functional block developed in this converter. These functional blocks are called *devide255*, *max_min_selector, hue, saturation and luminance.* The function of these functional block are described in Table 1.

Figure 4 shows the top level entity of this reconfigurable RGB to HSL converter from Altera Quartus II software. The RGB input for this converter in 8 bits wide, the output of hue, saturation and luminance is 10, 8 and 8 bits wide corresponding respectively. The Hue, saturation and luminance block are arranged in parallel form. Therefore the HSL output value can be calculated in the same time and yet to save a propagation time. This converter design is truly a static device which did not required any clocking to drive the circuit.

Table 1 Top level entity functional block description

Functional block	Description
divide255	To perform independent division of input RGB value with 255 as shown in (1)
max_min_selector	To select maximum and minimum value from the input set of rgb To sum maximum and minimum value To subtract minimum value from maximum value
hue	To calculate a hue value
saturation	To calculate a saturation value
luminance	To calculate a luminance value

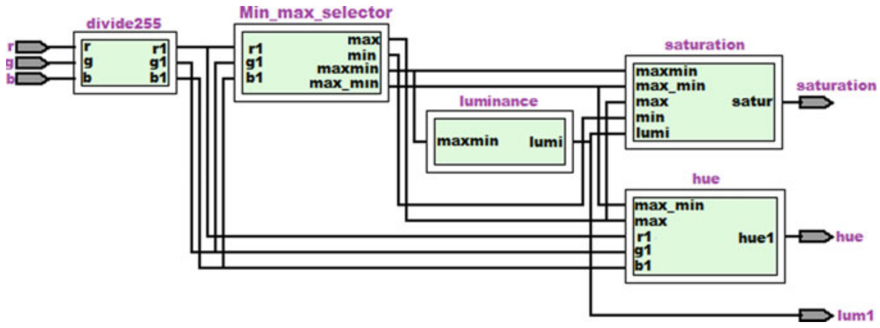


Fig. 4 RGB to HSL converter top level architecture

3.1 *Internal Architecture of Saturation and Luminance Module*

Figure 5 is the internal architecture of the saturation functional block. As observed from Fig. 5 and equation in (4) of the saturation, the input luminance value is going through a comparator to compare with the value of 0.5. The comparator result is then controlling a multiplexer to select the output one of both equations in (4). Besides, the input of *max_min* is the difference of the rgb maximum and minimum values. This input value is then adjusted by a multiplier so that it is compatible to the divider and producing desire an output value. The two divider in this saturation module is used to calculate the saturation value for both condition in parallel mode and let a luminance input to decide the output.

Hue and luminance modules also constructed by the hardware arithmetic device. Luminance is easier to be implemented due to its simple equation in (3). The maximum and minimum value is sum at the max_min_selector module.

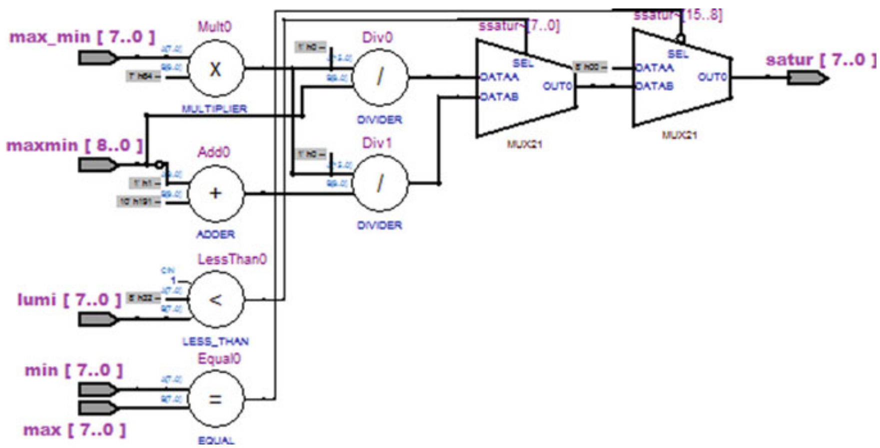
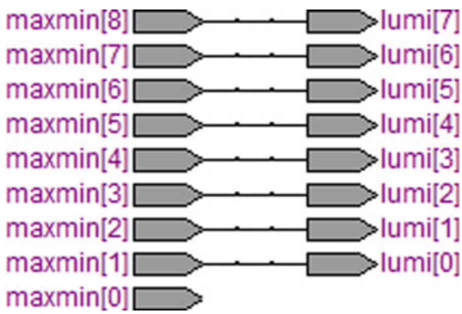


Fig. 5 Internal architecture of saturation module

Fig. 6 Architecture of luminance module



The division of 2 is implemented by using bit shift-away method as shown in Fig. 6. This is because shifting away the least significant bit (LSB) in little endian format to the right, the remaining value is same as divided by 2.

4 Results

4.1 Simulation Result

The simulation test has been conducted by using Altera Quartus II software simulation tool. The simulation is to test the functional capability of the designed module based on the top level entity of this RGB to HSL converter and also based on the HSL output value. Figure 7 shows part of the simulation result of the RGB to HSL converter.

Since the simulation test shows correct result, the accuracy of this simulation result is then analyzed. There are total of 50 random RGB sample sets used as the simulation input values and also used as the input values for a C program RGB to HSL converter based on the same algorithm to generate the ideal HSL reference output. Both output results are then compared and analyzed. The accuracy of the simulation result is calculated based on (6) [7]. H_C is the output result of the C program RGB to HSL conversion while H_{VHDL} is the output result from the simulation of Altera Quartus II. The error of this converter is caused by the decimal point number deviation. The output accuracy of this converter able to achieved

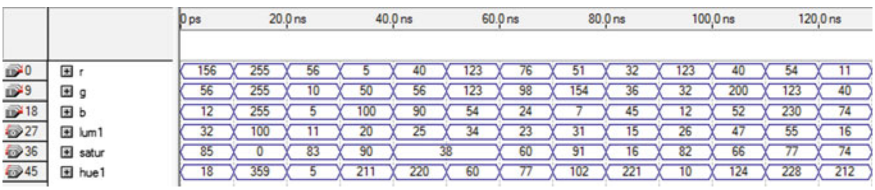


Fig. 7 Simulation result of RGB to HSL converter

98.44 %. This converter has the advantage of direct produce the HSL result without using any pipelining with one clock one result and no clock latency.

$$A = 1 - \left| \frac{H_C - H_{VHDL}}{H_C} \right| \times 100\,\% \tag{6}$$

5 Conclusion

The simulation results has been thoroughly verified. A reconfigurable converter of RGB–HSL was successfully developed using Altera Quartus II software. The percentage of accuracy achievement is 98.44 %. This conclude the HSL output value from this converter is accurate enough for future application on embedded machine vision and embedded image processing system. A small number of logic elements were occupied by this converter and leaving large expansion space for future development.

References

1. Pérez, P., Hue, C., Vermaak, J., Gangnet, M.: Color based probabilistic tracking. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2350, pp. 661–675. Springer, Heidelberg (2002)
2. Wang, M.J., Lee, G.G., Lin, H.Y.: Extraction of perceptual hue feature set for color image/video segmentation. In: Huang, Y.M., Xu, C., Cheng, K.S., Yang, J.F., Swamy, M.N.S., Li, S., Ding, J.W. (eds.) Advances in Multimedia Information Processing 2008. LNCS, vol. 5353, pp. 875–878. Springer, Heidelberg (2008)
3. Cho, J.U., Jin, S.H., Dai Pham, X., Kim, D., Jeon, J.W.: FPGA based real time visual tracking system using adaptive color histograms. In: IEEE International Conference on Robotics and Biomimetics (ROBIO 2007), pp. 172–177 (2007)
4. Agoston, M.K.: Computer Graphic and Geometric—Implementation and Algorithms. Springer, London (2005)
5. Fortner, B., Meyer, T.E.: Number by Colors. Springer, New York (1997)
6. Canon: Picture Style. http://www.canon.co.jp/imaging/picturestyle/editor/matters05.html. Accessed 26 Jan 2015
7. Kalsi, H.S.: Electronic Instrumentation. Tata McGraw-Hills Pub, Noida (2005)

Design Implications for Quality User eXperience in Mobile Augmented Reality Applications

Shafaq Irshad and Dayang Rohaya Awang Rambli

Abstract Mobile augmented reality (MAR) technology is a new trend that provide users with the augmented view of digital information in real world. This paper contains design implications for improving the User eXperience (UX) of MAR applications. The scope is limited to MAR and its application in advertisement industry. Results from an experimental study are used to draw the guidelines that can be followed by the designers and practitioners to ensure a quality experience to their users.

Keywords Augmented reality · Mobile augmented reality · User experience · Design guidelines · Advertising · Human computer interaction

1 Introduction

The way user interacts with computers and hand-held devices is changing drastically with the emergence of new technologies. Augmented Reality (AR) is one of those leading-edge and innovative technology that allows the user to observe real world objects supplemented with computer-generated content [1].

When AR technology enables interaction with the users physical environment through smart phones, hand held devices or tablets and superimpose virtual content on top of the real world it is called Mobile Augmented Reality (MAR) [2]. MAR is a revolutionary technology that augments the real world by overlaying digital data on top of it.

MAR is an exciting and emerging technology with a potential to offer an innovative interface to contextual data in the frequently varying mobile situations, the User Experience (UX) of MAR services is very challenging to forecast or envision.

S. Irshad (✉) · D.R.A. Rambli
Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Bandar Seri Iskandar, Malaysia
e-mail: shafaqirshad223@gmail.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_110

Hence, it is insightful to apprehend what users anticipate of the experience, therefore attaining latest insight about the type of experimental design targets, limitations and requirements. Besides discovering the possible UX, it is essential to define a set of design implications or guidelines for enhancing the quality UX of MAR technology.

Keeping in mind the need for UX design guidelines a pilot study was conducted based on the latest MAR advertising applications. The study highlighted several shortcomings in the overall design of MAR applications thus providing a solid background to propose certain implications that can be applied in the designing process. However, the scope of our research is limited only to advertising MAR applications. Thus, this paper contributes to the research society by giving a set of design guidelines based on the results obtained from an experimental study.

The remaining paper shows the related work in Sect. 2 highlighting the concepts such as UX, AR and MAR in more detail. Section 3 contains methodology followed by results and discussion. Design implications for a quality UX of MAR applications based on the study are discussed in Sect. 4. The main contribution of the paper lies in this section. The guidelines are perceived from the previous study [3] which is highlighted briefly in methodology section as well. Finally, it is concluded that if the UX designers strive for an optimal solution, success of the product is unavoidable and it invokes a positive and quality UX for the users.

2 Background and Related Work

2.1 User Experience

Although usability is essential in making any technology apparent, it is not sufficient in todays competitive situation. User centered design still holds the core importance in designing for the humans and is modified with a concept of User Experience (UX) [4]. UX involves a person's attitudes, behaviors, and emotions towards a specific artifact. ISO 9241-210, international standard on ergonomics of human system interaction, outlines UX as "a person's perceptions and responses resulting from the use or anticipated use of a product, system or service" [5].

The concept of UX awakens a broader image changing the old acronyms like (what you see is what you get) WYSIWYG to WYXIWYG, What You eXperience is What You Get [6]. According to Marc Hassenzahl [7] UX is not just good business design, functionality or pretty interfaces. It is about crafting a lifelong experience that go beyond the materials and resources. Hassenzahl and Tractinsky [8] describe UX as a aftermath of users internal state, the features and characteristics of the designed system, and the context within which the interaction occurs.

UX was first defines in 1996 by Lauralee Alben. According to the author UX is an all-encompassing term that depicts how an interactive product look, feels, serve and contributes to the context and quality of ones life [9].

Forlizzi and Fords [10, 11] also defined UX as an interaction with product model, demonstrating components that affect UX. They state that user, product, social and cultural factors and context of use are to be considered when looking at the UX of a product. In Kankainens [12] opinion, UX is the outcome of an inspired and motivation driven action in a per-definite context (i.e. people, place and things surrounding the user in the interaction).

Overall, UX is considered a subjective and universal concept and in addition to the customary usability aspects, UX necessitates value-sensitive design, social and cultural interaction, and emotional impact including interaction experiences such as joy of use, excitement, pleasure, and aesthetics [13].

Furthermore, UX is fundamentally subjective, as it occurs only in users mind. Because of the aforementioned features, designing UX is not possible, but rather designing for UX is, thus the designers should design for user experience to enhance the quality of their products.

2.2 Mobile Augmented Reality

Augmented reality (AR) concept is the combination of digital information and the real environment generated by existing computer in the users view of somatic reality, so that they appear as one environment [14].

AR is adopting efficiently for state-of-the-art mobile domain. It has been made possible due to the viable solutions for the interaction between mobile handling, sensor technology, mobile based GPS and wireless technology and image recognition techniques. Not only the Mobile AR technical enablers and respective prototypes have been studied extensively for decades, but also the first widely available mobile applications has recently been introduced. When the AR experience is delivered on mobile devices it is termed as Mobile Augmented Reality (MAR) [15].

MAR resourcefully integrates objects as components of the user interface, thus enabling virtual interaction with the content displayed on digital devices. The digital information aligned to the physical world can be interacted with in real time hence, it serves as a versatile user interface for services of ambient intelligence and smart environments [16]. In fact, MAR can be considered a fundamental technology in the paradigm shift from desktop based interaction towards ubiquitous computing, enabling smart services anytime and anywhere [17, 18].

2.3 Related Work: UX for 3D Applications

Three-dimensional (3D) applications help users perform tasks in various innovative and interactive systems such as virtual environments, ubiquitous computing and augmented reality (AR). User Interfaces (UI) involving physical interaction in 3D space is prominent universally and go by many names such as freehand gestures, motion controls or natural user interfaces [19].

Instead of playing with the joysticks and pushing buttons on the keyboard, gamers now swing their hands in the air, move their arms, jump up and down or bend their whole bodies to play 3D virtual games. People make mid-air gestures instead of using a remote control the electronic devices in their homes. Instead of looking for places on a 2D map, people just point their mobile devices to get an augmented view of the real world [20].

Various studies have been reported that address the design part of UX in 3D applications and services. One prominent contribution in our field is by Jumisko-Pyykk et al. [21]. It is about designing for UX and expectations of people regarding the 3D technology such as tv and video. Guidelines highlighting the characteristics of system, user and specific context of use were proposed but they cannot be generalized to MAR technology.

Another prominent contribution is the study regarding effects of UX while navigating in 3D virtual environments (VEs). Different desktop navigation aids are compared and the effects of UX are evaluated with 3D desktop VEs [22]. UI design guidelines for public information kiosk systems are reviewed by [23].

In augmented reality technology, the prominent work is done by T. Hollerer and colleagues. His major contribution is the study of UI management techniques for collaborative MAR applications [24]. He states that maximum benefits of MAR technology can be achieved if the UI is properly managed e.g. by maximizing the relevance of content and minimizing the confusions of augmented materials. The article addressed some important steps involved in this process, centering on design and the layout of virtual environment overlaid on the mobile devices.

From the prominent literature review we can clearly see the lack of design implications for UX of MAR. This paper aims to bridge this knowledge gap by providing a set of design guidelines based on a survey study. In this research study the focus is on how to design for the quality UX of mobile augmented reality advertising applications. Many studies has been performed on UX of various products, services and systems. However, methods and techniques to design for 3D interaction to ensure a quality UX and design implications to follow while designing specifically for mobile AR are still scarce.

3 Methodology

A study was designed to fulfill the objectives i.e. proposing a set of design implications for improving the quality of UX for MAR applications. The fundamental purpose of this study was to evaluate the experience of people towards latest MAR applications and assess the shortcomings they have towards this latest technology. We conducted the survey using easily available MAR advertising applications. The applications used in the study are demonstrated in Figs. 1 and 2. Experimental study was conducted in controlled lab environments. The respondents

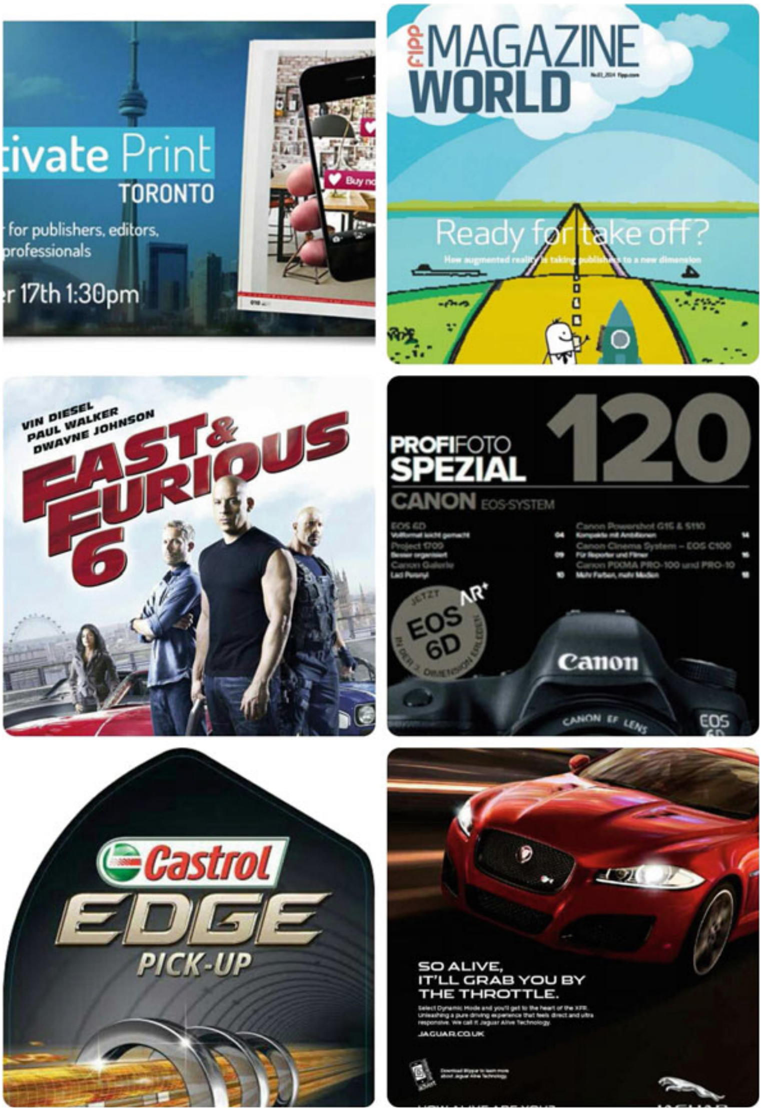


Fig. 1 Mobile Augmented Reality active prints used in experimental study

were shown various MAR applications and their utility in a short demonstration and were allowed to use the technology by themselves later on in the session. The respondents were asked to fill a questionnaire latter on adopted from the already existing questionnaire measuring various product centered features. A combination of quantitative and qualitative research scheme was adopted.

Fig. 2 3D Model of Canon pops up when the mobile device is pointed on the active print

3.1 Study Design and Structure

A brief overview of the survey is presented in this section which was performed to evaluate the latest MAR marketing applications. Questionnaires were used to collect data from the selected respondents. In initial design phase the already existing survey questionnaire was adopted and improved. Finally, a pilot survey study was performed with the selected respondents to ensure the validity and reliability of the research. A personal survey technique was used to ensure the accuracy of the results. The survey was administered closely to ensure reliable outcomes and results.

3.2 Respondents

There were fifteen respondents who took part in the study. Several sessions were conducted separately to complete the survey thus ensuring satisfying results. The details of respondents who participated in the study are demonstrated in Fig. 3. From the results it was observed that although all participants except one had smart phones yet they were not all familiar with the term AR or MAR. Furthermore only 1 participant used the MAR technology before.

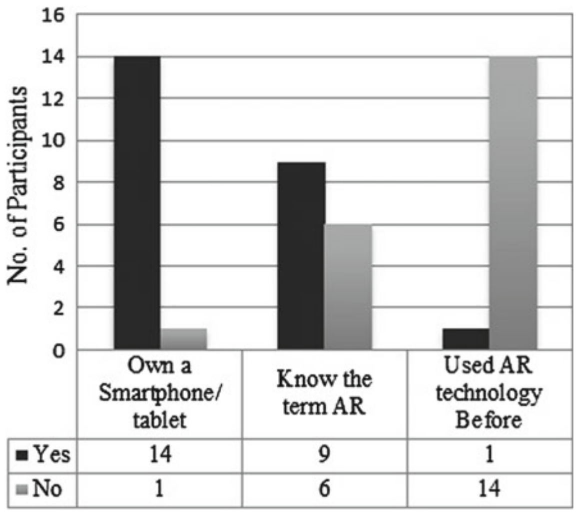


Fig. 3 Table shows the number of respondents who possess a mobile phone, know the term AR and have used AR before

3.3 Survey Results and Statistical Analysis

The results captured through the survey are demonstrated in graphical form for the ease of understanding. Various factors affecting the aesthetic experience such as captivity, inspiration, playfulness and liveliness were measured as part of the product UX (shown in Fig. 4).

Similarly Figs. 5 and 6 shows the feedback of participants based on their experience with the user interface of mobile AR applications. Keeping in mind the diverse UX that MAR invokes, various factors like Consistency, cleanliness and Ease of use were monitored. Furthermore control over the content, interaction with product, its color texture and quality were also evaluated.

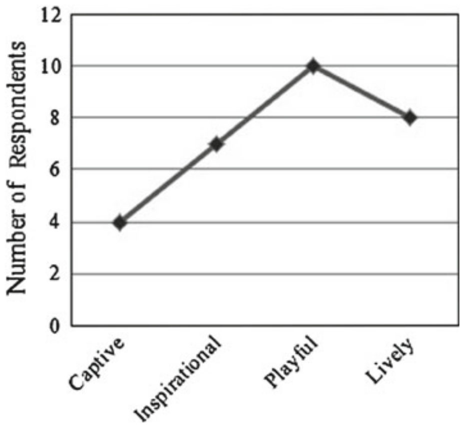


Fig. 4 Graph depicting the aesthetic experience of participants

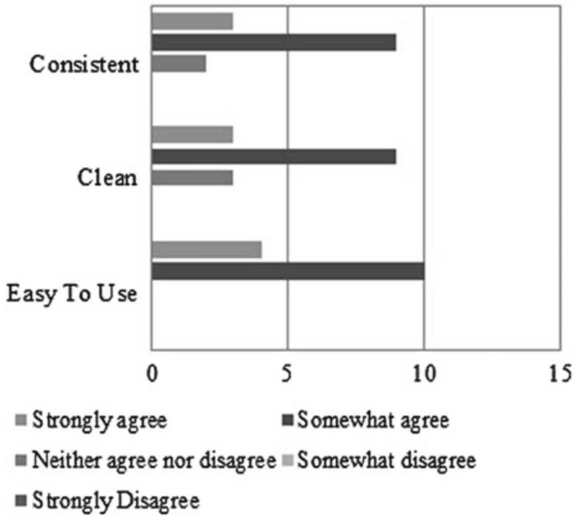


Fig. 5 Participants feedback on the UI design elements

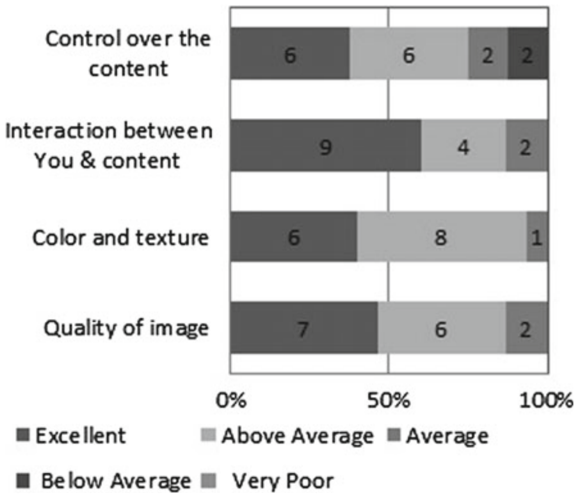


Fig. 6 Graph depicting results from various UI design dimensions that make up the experience of participants

The results obtained from this study were further quantified and the standard deviation and median were calculated to know how these results can be applicable to a larger population.

Standard deviation (SD) of the main product centered design elements such as its ease to use, clean design, consistency, quality, color and texture, interaction of user with the 3D content and control over the 3D content clearly shaped our design implications.

Further more, while examining the overall UX and design elements that directly affects the UX e.g. control, 3d content, efficiency startup time of MAR, flexibility of

Table 1 Statistical analysis of UI elements

UI of MAR (n = 15)	Mean	SD
Easy to use	4.1	0.718
Clean	4.0	0.632
Consistent	3.93	0.77
Quality of image	4.3	0.699
Color and texture	4.6	0.59
Interaction with content	4.46	0.718
Control over the content	4.46	0.88
Control in your hands	4.13	0.718
Examine content in 3D	4.13	0.618
Efficient experience	4.33	0.86
Startup time	4.13	0.805
Flexibility	3.86	0.718
Productivity	3.93	0.771
Response time	3.73	0.853

the system, its productivity and response time of MAR application were also analyzed (see Table 1). The detailed mean and SD values of these elements are given in the table.

From the acquired data, elements with the SD of less that 0.80 were taken in consideration for further improvement. The design implication in Sect. 4 understandably couples with the above mentioned results.

4 Design Implications for Quality UX of MAR Applications

The survey results along with the literature review helped in designing the guidelines. Keeping in mind the characteristics of UX and the dynamic nature of MAR applications, design guidelines from several studies and our empirical observations [3] are presented below:

- Design for natural 3D interaction on the mobile devices. MAR applications should promote natural or high fidelity user interfaces (UI) [25]. For example when users point their devices on an advertisement in a moving car, they should encounter a natural, on-the-go experience that requires no special expertise. The interface design should be fun, engaging and realistic at the same time.
- Design for quick and effective interfaces. The UI should be designed in such a way that the user can perform their tasks quickly and effectively. User should not be constrained to things they can do in the real world. They should have the ability to perform actions that can enhance their physical abilities, perceptual abilities and even cognitive abilities (e.g., navigation through a 3D world, x-ray vision through the advertising products etc.)

- Design for the ease and efficacy of user. Although users ability to interact in the 3D MAR applications should not be constrained but it should provide the constraints that help the user to easily and efficiently interact with the system. For example in interior design AR applications, uses should be allowed to place the furniture anywhere in the physical space but the position should be upright. Users should be allowed to control certain parameters like 2D position, rotation in vertical axis. Helpful boundaries can be added to the application design thus promoting the ease of user. Absence of appropriate constraints can cause the user to become fatigue decreasing their efficiency.
- Maintain the functional integrity of the design. The design should always work as intended. The applications should not freeze or crash when certain operations are invoked.
- Always design for user comfort. In 2D designs physical comfort is not considered a major issue but its not the case with 3D UIs as they involve large scale physical movement. Unnatural rotations of the 3D content and demand for rapid movements to navigate through the 3D content should be avoided.
- Design for target audience to enhance the utility. A quality applications will never overlook the utility, usefulness, importance and interest of their audience. Although utility is a continuous quality as its relative to the users of any particular application, it is important to study the target audience before designing the UI.
- Design for an engaging, productive and enjoyable user experience. The basic human computer interaction principles should be applied to the UIs. When the MAR applications provide the viewers with an enjoyable and productive experience they will be forced to use the application again and again. Further more a uniform and engaging experience ensures a quality UX.
- Design for usable experience. MAR application design should be ease to learn, efficient and easy to use. If an application design has high fidelity but it is difficult to use, chances are that the users will quit using it.
- Design for persuasiveness of users. Persuasiveness refers to the encouragement and promotion of a specific user behaviors. This trait is specially important for advertising MAR applications because the company using the application might want to promote their sales. For this scenario the UX should be designed in such a way that it provides complete information regarding the product and vendors. The quality and location of information on a particular UI is also very important and directly affects the persuasiveness of users.
- Provide Graphic design that enhances the branding and appeal of an applications interface. The look and feel of Augmented reality applications is equally important. The graphic design of 3D content in MAR applications have a large impact on how users perceive experience. Colors, images, modeling other media used can invoke positive emotional reactions in users. It directly effects the excitement of users thus producing a quality user experience.

5 Conclusion

It is important to design for enhancing the UX of MAR services. This paper states some important concepts like UX, AR and MAR highlighting a few prominent contributions and pointing out the need for design implications for MAR services. The particular focus of this study is on enhancing the UX design of MAR advertising applications. From the experimental results and statistical analysis UX design guidelines are proposed that when applied to the MAR services specifically advertising applications can enhance the UX. These guidelines can ensure a positive UX for MAR applications. These implications are an initial step for the future research. Further studies will be done to validate our guidelines and improve them through development of prototype and a general framework.

References

1. Azuma, R., Baillot, Y., Behringer, R., Feiner, S., Julier, S., MacIntyre, Blair: Recent ad-vances in augmented reality. Comput. Graph. Appl. IEEE **21**(6), 34–47 (2001)
2. Hllerer, T., Feiner S.: Mobile augmented reality. In: Telegeoinformatics: Location-Based Computing and Services. Taylor and Francis Books Ltd., London, 21 (2004)
3. Irshad, S., Rambli, D.R.A.: User experience evaluation of mobile AR services. In: Proceedings of the 12th International Conference on Advances in Mobile Computing and Multimedia (MoMM '14), pp. 119–126. ACM, New York (2014)
4. Karapanos, Evangelos: User Experience Over Time. Modeling Users' Experiences with Interactive Systems, pp. 57–83. Springer, Berlin Heidelberg (2013)
5. International Organization for Standardization: Ergonomics of Human System Interaction—Part 210: Human-centered design for interactive systems (formerly known as 13407). ISO FDIS 9241-210:2009 (2009)
6. Lee, G.A., Kim, G.J., Billinghurst, M.: Immersive authoring: what you experience is what you get. Commun. ACM **48**(7), 76–81 (2005)
7. Hassenzahl, M.: The thing and I: understanding the relationship between user and product. In: Blythe, M.A.,Overbeeke, K., Monk, A.F., Wright, P.C.: (eds.) Funology. From Usability to Enjoyment, pp. 31–42. Kluwer Academic Publishers, Dordrecht (2003)
8. Hassenzahl, M., Tractinsky, N.: User experience a research agenda. Behav. Inf. Technol. **25**(2), 91–97 (2006)
9. Alben, L.: Quality of experience: defining the criteria for effective interaction design. Interactions **3**(3), 11–15 (1996)
10. Forlizzi, J., Battarbee, K.: Understanding experience in interactive systems. In: Proceedings of the 2004 Conference on Designing Interactive Systems (DIS2004): Processes, Practices, Methods, and Techniques, pp. 261–268. ACM, New York (2005)
11. Forlizzi, J., Ford, S.: The building blocks of experience: an early framework for interaction designers. In: Proceedings of the 3rd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques (DIS 00), pp. 419–423. ACM, New York (2000)
12. Kankainen, A.: Thinking model and tools for understanding user experience related to information appliance product concepts. Doctoral dissertation, Helsinki University of Technology, Finland (2002)
13. Kuniavsky, M.: Observing the User Experience: A Practitioner's Guide to User Research. Morgan kaufmann, Burlington (2003)

14. Hllerer, T., Feiner, S.: Mobile augmented reality. In: Karimi, H., Hammad, A. (eds.) Telegeoinformatics: Location-Based Computing and Services, Taylor & Francis Books Ltd., UK (2004)
15. Irshad, S., Bt Awang Rambli, D.R.: User experience of mobile augmented reality: a review of studies. In: 3rd International Conference on User Science and Engineering (i-USEr), pp. 125–130, 2–5 Sept 2014
16. Remagnino, P., Foresti, G.L.: Ambient intelligence: a new multidiciplinary paradigm. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **35**(1) (2005)
17. Satyanarayanan M.: Pervasive computing: vision and challenges. IEEE Pers. Commun. 10–17 (2001)
18. Weiser, M.: The computer for the 21st century. Sci. Am. **265**(3), 94–104 (1991)
19. Bowman, D.A., Kruijff, E., LaViola Jr, J.J., Poupyrev, I.: 3D user Interfaces: Theory and Practice. Addison-Wesley, Bpston (2004)
20. Burdea, G., Coiffet, P.: Virtual reality technology. Presence Teleoperators Virtual Environ. **12** (6), 663–664 (2003)
21. Jumisko-Pyykk, S., Weitzel, M., Strohmeier, D.: Designing for user experience: what to expect from mobile 3d tv and video?. In: Proceedings of the 1st International Conference on Designing Interactive User Experiences for TV and VIDEO (UXTV '08). ACM, New York (2008)
22. Burigat, S., Chittaro, L.: Navigation in 3D virtual environments: effects of user experience and location-pointing navigation aids. Int. J. Hum. Comput. Stud. **65**(11), 945–958 (Nov 2007). ISSN: 1071-5819
23. Maguire, M.C.: A review of user-interface design guidelines for public information kiosk systems. Int. J. Hum. Comput Stud. **50**(3), 263–286 (1999)
24. Hllerer, T.,Feiner, S., Hallaway, D., Bell, B., Lanzagorta, M., Brown, D., Julier, S., Baillot, J., Rosenblum, L.: User interface management techniques for collaborative mobile augmented reality. Comput. Graph. **25**(5), 799–810 (Oct 2001)
25. Bowman, D.A., Gabbard, J.L., Hix, D.: A survey of usability evaluation in virtual environments: classification and comparison of methods. Presence Teleoperators Virtual Environ. **11**(4), 404–424 (2002)

Microstrip Patch Antennas for Terahertz Sensing Applications

S. Sreenath Kashyap and Vedvyas Dwivedi

Abstract The Paper presents an investigation on electromagnetically coupled patch antennas on two different substrates namely Arlon and Polymide at terahertz frequency. The analysis is carried out with two parameters namely by varying the distance of separation of patches and varying the substrates. The electrical performance of the proposed models like return loss, gain, and bandwidth are compared at same resonant frequency. This proposed model is useful In applications like terahertz sensing and communications. The proposed model with dielectric substrate having low permittivity value shows the significance improvement in terms of gain and reflection. The analysis is carried out using finite integral technique by way of CST microwave studio a proven electromagnetic tool.

Keywords Terahertz antennas · Terahertz sensing · Terahertz communications return loss · Gain · Bandwidth · Microstrip patch antennas

1 Introduction

In the field of science and technology there is a rapid growth in the area of wireless communication devices [1]. The advancements in Antenna technology is play a vital role in the wireless communication system [2]. The significant progress in the design of Microstrip patch antennas with high bandwidth and gain enhanced operations has been reported over the past years [2]. MSTPA offers advantages like

S. Sreenath Kashyap (✉) · V. Dwivedi
C.U.Shah University, Wadhwan, Gujarat, India
e-mail: kashyap.foru3@gmail.com

V. Dwivedi
e-mail: provc@cushahuniversity.org

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_111

low profile, small volume, resistant to shock and vibrations, conformability, low cost. Due to these advantages MSTPA has gained attention of various researchers for potential applications in Terahertz regime [3]. There is a necessity of high gain antennas at terahertz frequency in order to avoid path loss and successful implementation of terahertz wireless system [3, 4]. Gain and bandwidth can be increased by different methods one of the method is by varying height of substrate, reducing the ε_r of substrate etc. [4, 5]. The variation in height of the substrate leads to the shock waves between the interface of air and dielectric substrate [4, 6]. Generally for multiple resonance and enhancement of gain reflector arrays are used at terahertz frequencies [7, 8]. One of the best method is by placing multiple patches [9, 10]. A second patch is placed at a height in-front of the first one so called stacked patch antenna [11].

In this present contribution, the analysis of the effect of stacked patch antenna on two different substrates is carried out by varying the distance between the patches with respect to the thickness of dielectric substrate. In Sect. 2 the geometry of the conventional and staked patch model along with the design parameters of the proposed models on two different substrates namely Arlon and Polymide are explained. In Sect. 3 the simulation results of the proposed model along with the comparison of performances of the antennas in terms of return loss. Gain and bandwidth are tabulated.

2 Antenna Configuration

The basic geometry of conventional Microstrip Patch Antenna is as shown in Fig. 1 [12]. The patch is printed on a ground plane with Arlon as a substrate of thickness ‘t’. Further, another patch is placed at a distance ‘d’ above the lower patch as shown in Fig. 2 [13].

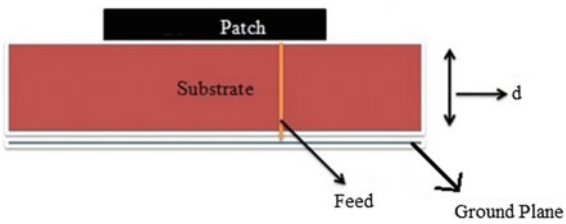


Fig. 1 Geometry of conventional patch model

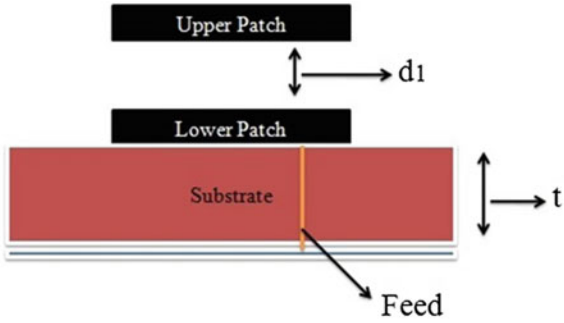


Fig. 2 Geometry of staked patch model

The investigation on the proposed antenna is carried out by varying the distance between the patches by considering the thickness of substrate as reference and also by varying the material of dielectric substrate. Instead of direct coupling, EM coupling is used for electrically thick antennas as the lower patch is directly connected to the feed line.

The width of the patch can be calculated using the Eq. (1) [13] where W is width of patch, c is speed of light, f_0 is the resonant frequency and ε_r is the permittivity of substrate

$$W = \frac{C}{2f_0}\sqrt{\frac{\varepsilon_r + 1}{2}} \tag{1}$$

Due to the fringing effect, the effective dielectric permittivity (ε_{eff}) is to be considered. The effective dielectric permittivity can be calculated using Eq. (2) [13, 14] where h is height of dielectric substrate, w is width of the patch.

$$\varepsilon_{eff} = \frac{\varepsilon_r + 1}{2} + \frac{\varepsilon_r - 1}{2}\left[\frac{1}{\sqrt{1 + 12(h/W)}}\right] \tag{2}$$

The length (L) of the patch can be calculated using the Eq. (3).

$$L = \frac{C}{2f_0\sqrt{\varepsilon_{eff}}} - 0.824h\left(\frac{(\varepsilon_{eff} + 0.3)(\frac{W}{h} + 0.264)}{(\varepsilon_{eff} - 0.258)(\frac{W}{h} + 0.8)}\right) \tag{3}$$

2.1 *Proposed Model 1*

In this proposed model 1 the analysis is carried out on Arlon substrate for conventional patch, stacked patch model. The Fig. 3a, b shows the proposed model of the conventional MSTPA and stacked patch model.

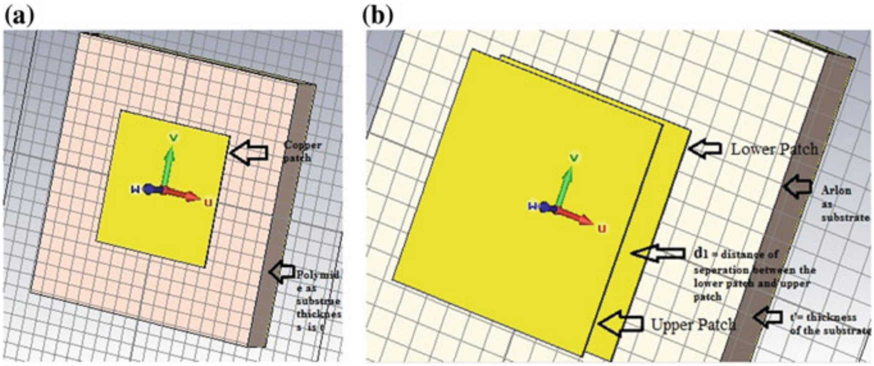


Fig. 3 **a** Proposed model of conventional patch model without slot using arlon as substrate. **b** Proposed model of staked patch model without slot using arlon as substrate

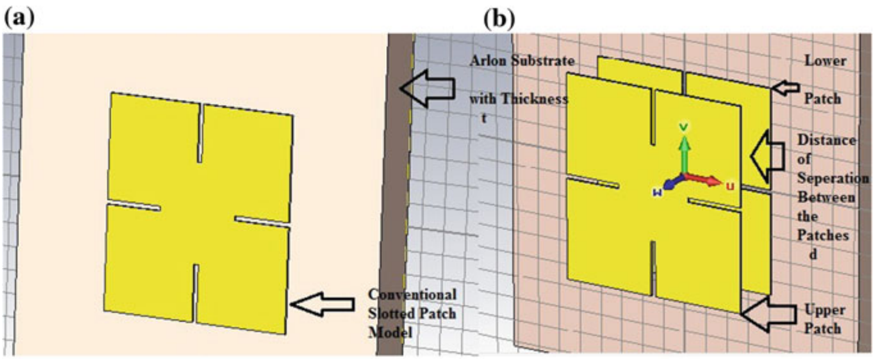


Fig. 4 **a** Proposed model of conventional patch model with slot using arlon as substrate. **b** Proposed model of staked patch model with slot using arlon as substrate

Further another model is designed by introducing a slot on the patch of the antenna. The Fig. 4a, b shows the proposed model of the conventional patch and stacked patch model with slot.

2.2 Proposed Model 2

In this proposed model 2 the analysis is carried out on polymide substrate. The Figs. 5a, b and 6a, b shows the proposed model of the antenna.

The Fig. 6a, b shows the proposed model of the conventional patch and stacked patch model with slot on polymide substrate.

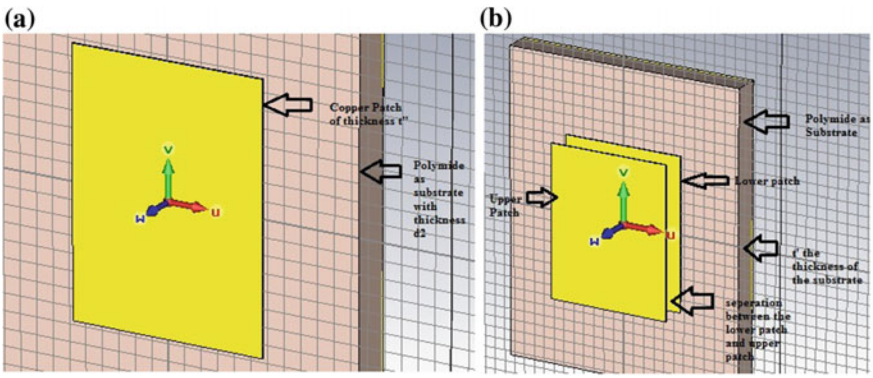


Fig. 5 **a** Proposed model of conventional patch model without slot using polymide as substrate. **b** Proposed model of staked patch model without slot using polymide as substrate

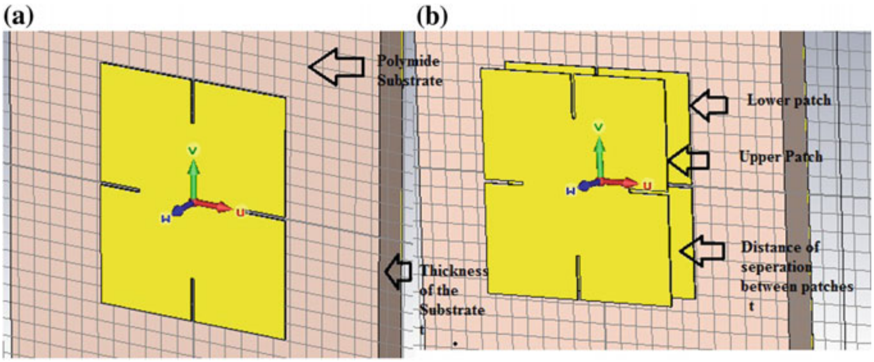


Fig. 6 **a** Proposed model of conventional patch model with slot using polymide as substrate. **b** Proposed model of staked patch model with slot using polymide as substrate

3 Simulation Results

The design and analysis is carried out in CST Microwave studio which uses the finite integral technique.

3.1 Proposed Model 1—Arlon as Substrate

Case 1 Conventional patch model without slot (Fig. 7):
Case 2 Stacked patch model without slot d = t (Fig. 8):

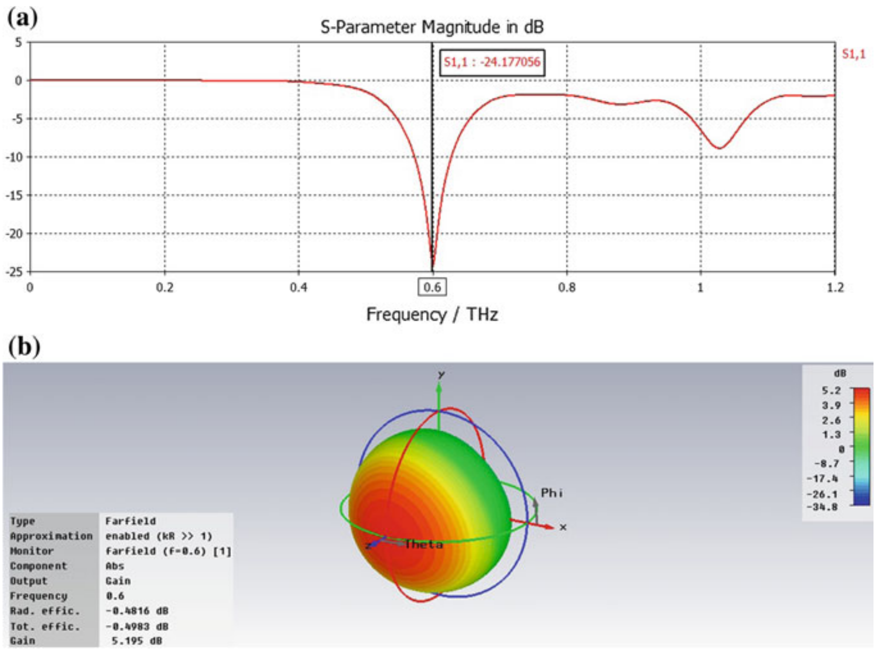


Fig. 7 **a** Return loss plot of case 1 conventional patch model without slot using arlon as substrate. **b** Gain plot of case 1 conventional patch model without slot using arlon as substrate

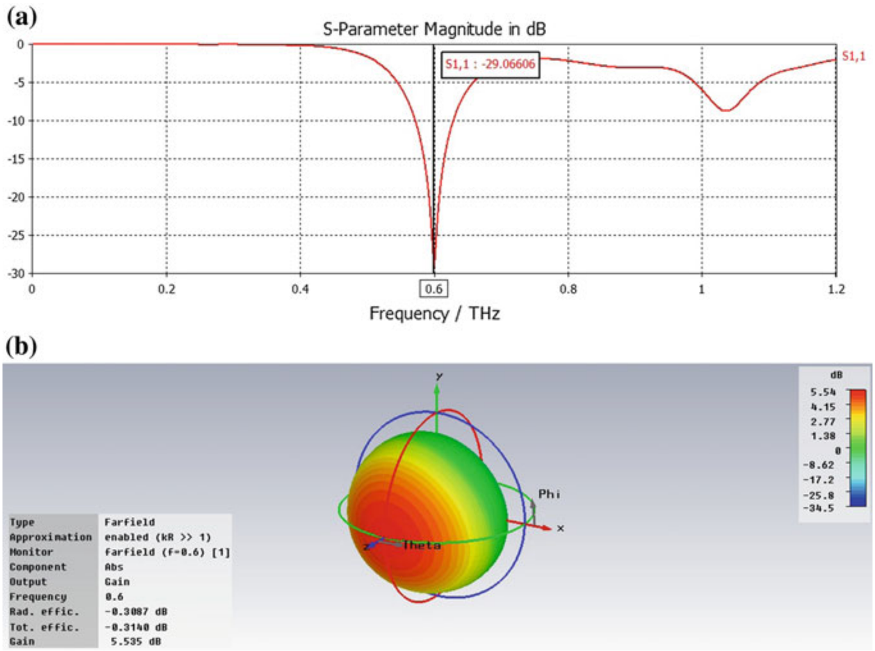


Fig. 8 **a** Return loss plot of case 2 stacked patch model without slot d = t using arlon as substrate. **b** Gain plot of case 2 stacked patch model without slot d = t using arlon as substrate

Case 3 Stacked patch model without slot d = t/2 (Fig. 9):
Case 4 Conventional patch model with slot (Fig. 10):
Case 5 Stacked patch model with slot d = t (Fig. 11):
Case 6 Stacked patch model with slot d = t/2: (Fig. 12):

3.2 Proposed Model 2—Polymide as Substrate

Case 1 Conventional patch model without slot (Fig. 13):
Case 2 Stacked patch model without slot d = t (Fig. 14):
Case 3 Stacked patch model without slot d = t/2 (Fig. 15):
Case 4 Conventional patch with slot (Fig. 16):
Case 5 Stacked patch model d = t with slot (Fig. 17):
Case 6 Stacked patch model d = t/2 with slot (Fig. 18; Table 1):

The electrical parameters like return loss, gain and bandwidth of the proposed models of the antenna are shown in the simulation results. The electrical performance parameters are compared for proposed models of the antenna and tabulated in Table 2. In case of proposed model in which Arlon is used as substrate only one

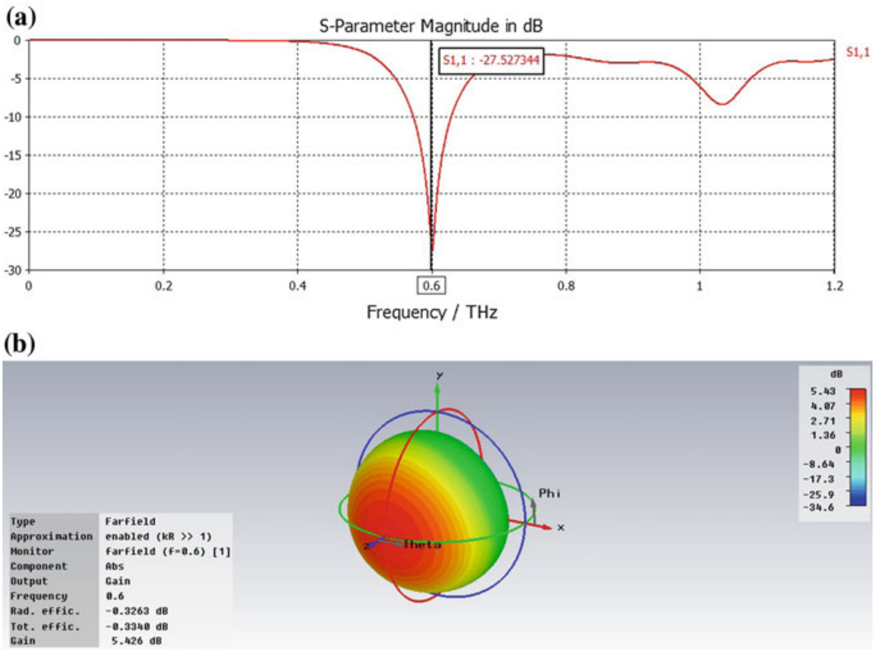


Fig. 9 **a** Return loss plot of case 3 stacked patch model without slot d = t/2 using arlon as substrate. **b** Gain plot of case 3 stacked patch model without slot d = t/2 using arlon as substrate

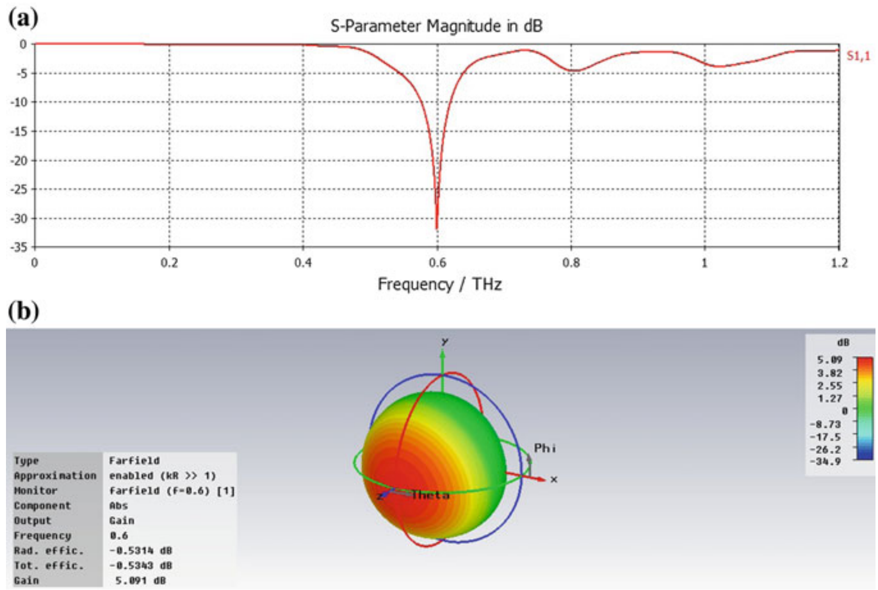


Fig. 10 **a** Return loss plot of case 4 conventional patch model with slot using arlon as substrate. **b** Gain plot of case 4 conventional patch model with slot using arlon as substrate

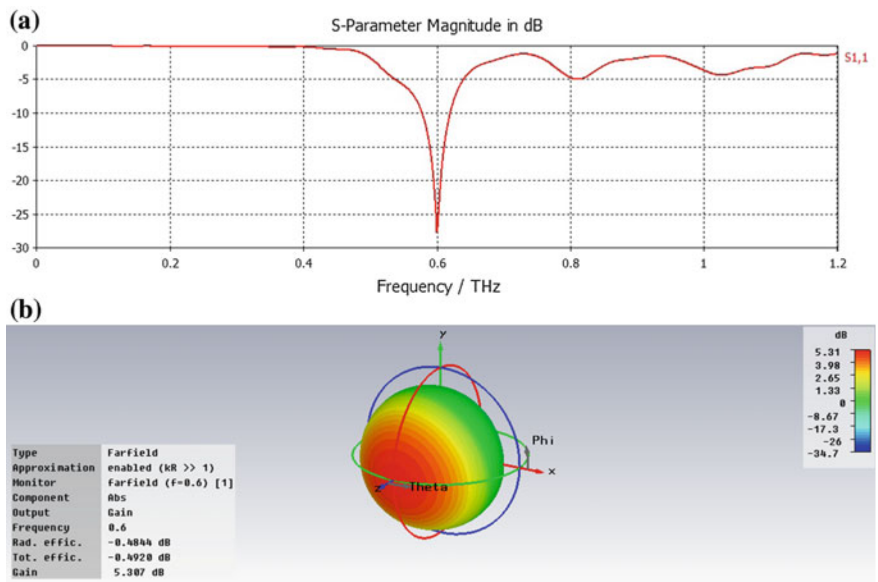


Fig. 11 **a** Return loss plot of case 5 stacked patch model with slot d = t using arlon as substrate. **b** Gain plot of case 5 stacked patch model with slot d = t using arlon as substrate

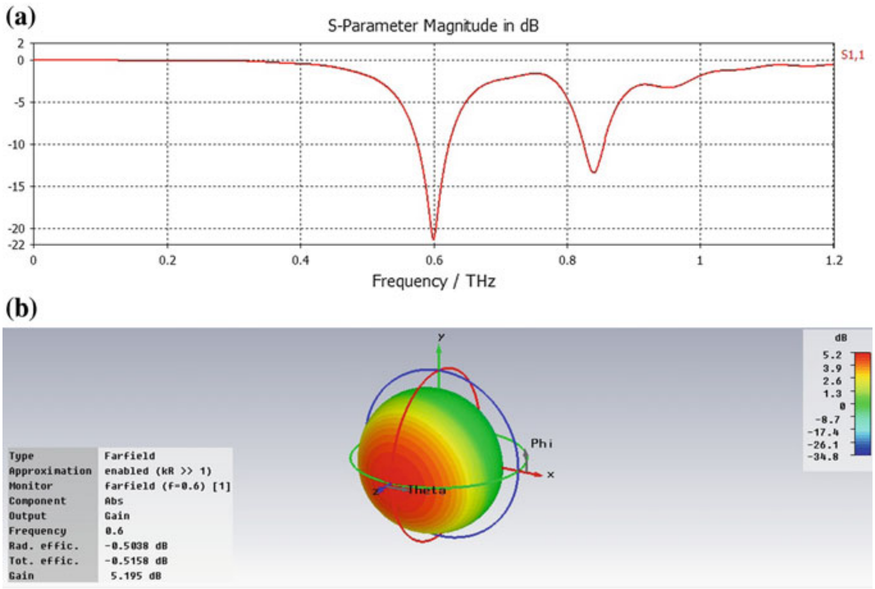


Fig. 12 **a** Return loss plot of case 6 stacked patch model with slot d = t/2 using arlon as substrate. **b** Gain plot of case 6 stacked patch model with slot d = t/2 using arlon as substrate

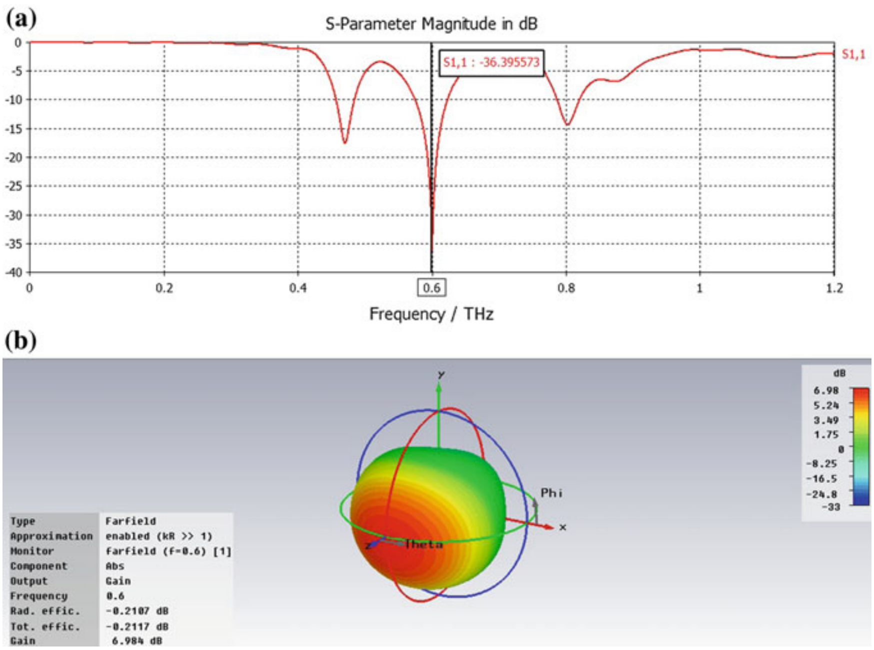


Fig. 13 **a** Return loss plot of case 1 conventional patch model without slot using polymide as substrate. **b** Gain plot of case 1 conventional patch model without slot using polymide as substrate

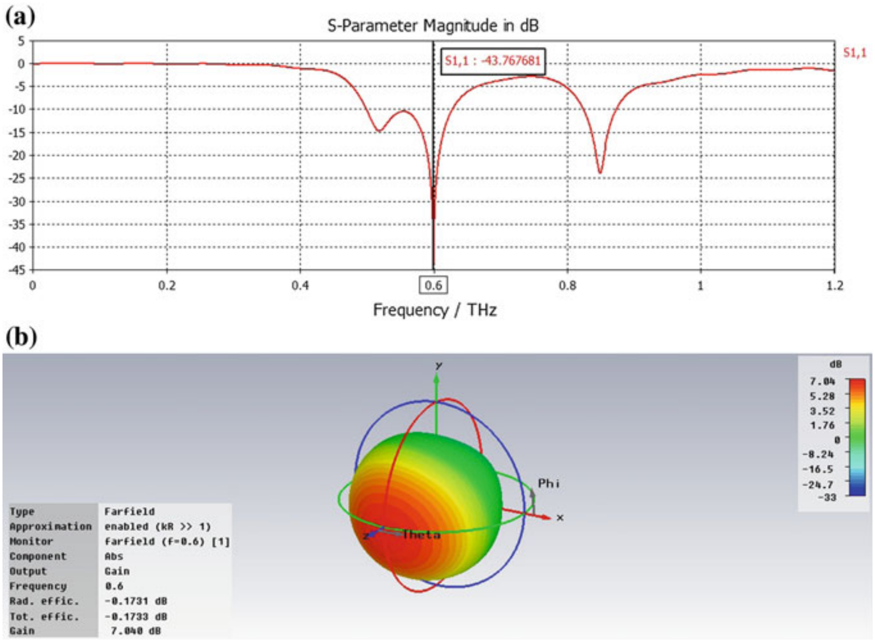


Fig. 14 **a** Return loss plot of case 2 stacked patch without slot d = t using polymide as substrate. **b** Gain plot of case 2 stacked patch model without slot d = t using polymide as substrate

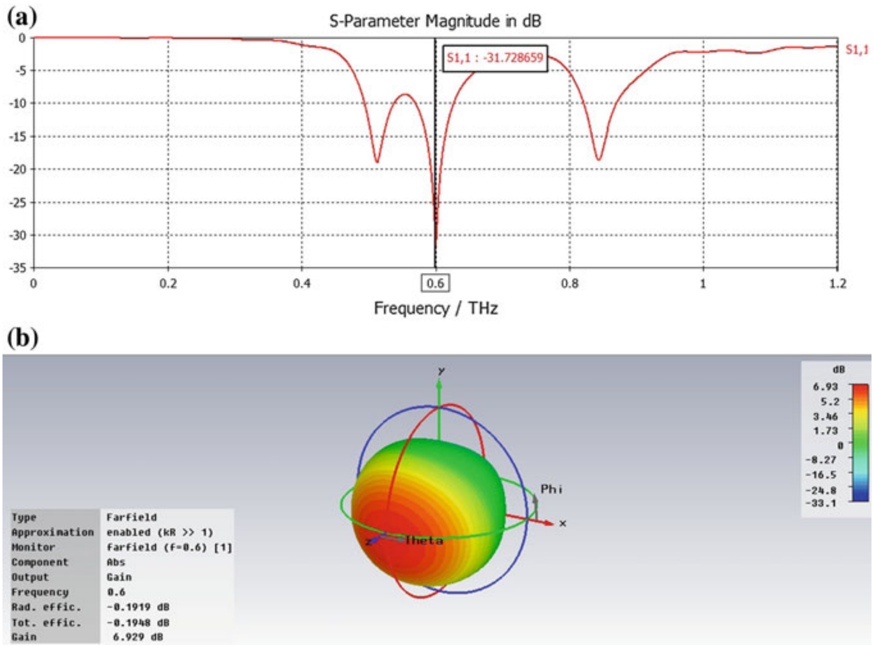


Fig. 15 **a** Return loss plot of case 3 stacked patch without slot d = t/2 using polymide as substrate. **b** Gain plot of case 3 stacked patch model without slot d = t/2 using polymide as substrate

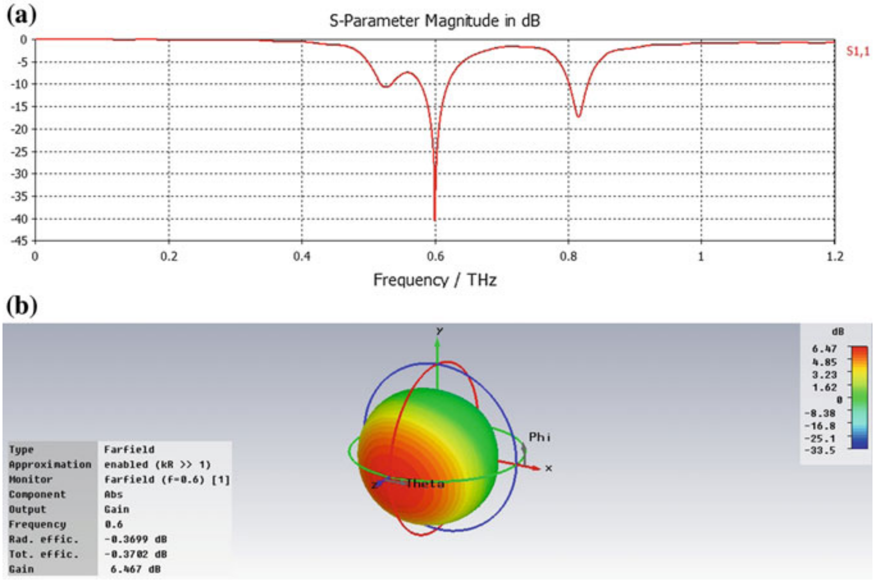


Fig. 16 **a** Return loss plot of case 4 conventional patch model with slot using polymide as substrate. **b** Gain plot of case 4 conventional patch model with slot using polymide as substrate

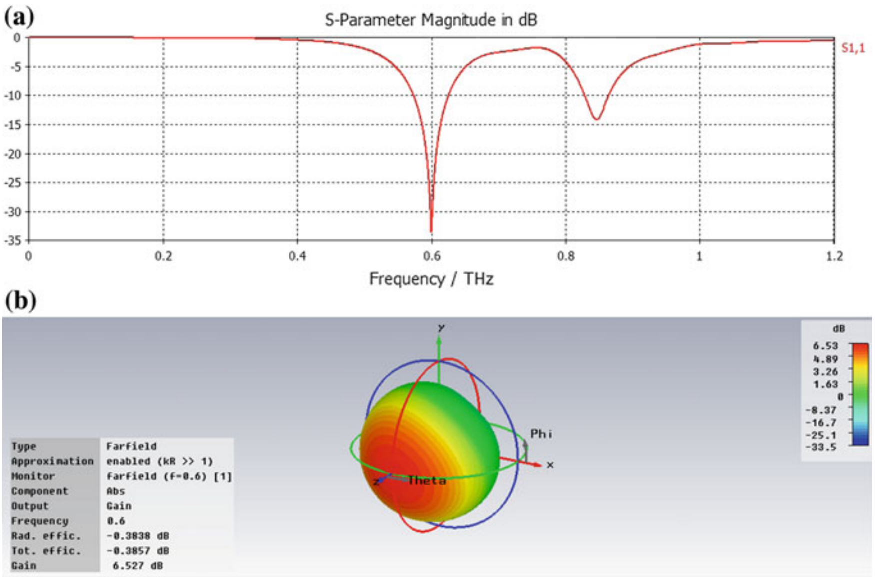


Fig. 17 **a** Return loss plot of case 5 stacked patch model with slot d = t using polymide as substrate. **b** Gain plot of case 5 stacked patch model with slot d = t using polymide as substrate

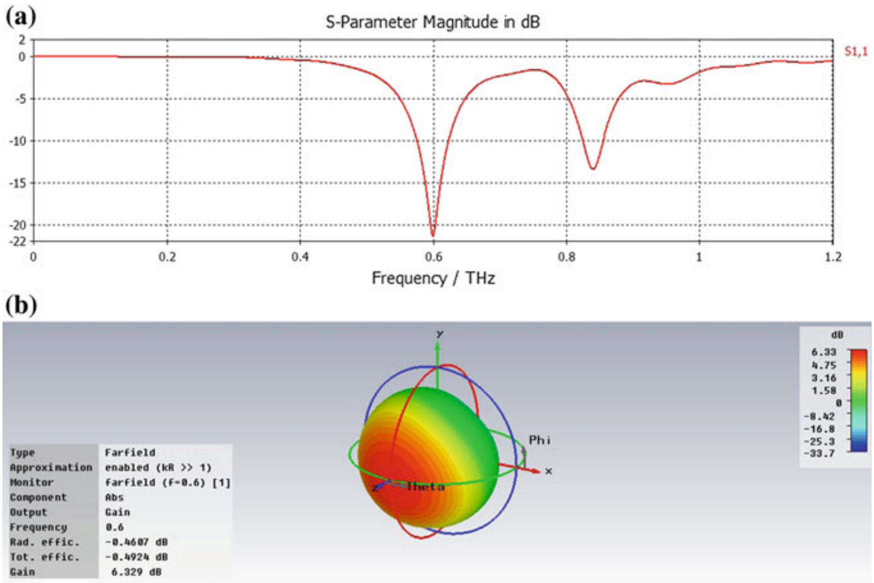


Fig. 18 **a** Return loss plot of case 6 stacked patch model with slot d = t/2 using polymide as substrate. **b** Gain plot of case 6 stacked patch model with slot d = t/2 using polymide as substrate

Table 1 Design parameters of the proposed models of the Antenna

Sr. No	Parameter	Proposed model 1	Proposed model 2
1.	Dielectric permittivity	6.0 (Arlon)	3.5 (polymide)
2.	Thickness of lower patch	1 µm	1 µm
3.	Thickness of upper patch	1 µm	1 µm
4.	Distance of separation between the patches	*d* µm	*d* µm
5.	Length of the patch	91.94 µm	121.3 µm
6.	Width of the patch	133.5 µm	166.6 µm
7.	Resonant frequency	0.6 THz	0.6 THz

transmission valley with magnitude of 0.6 THz in all conditions where as the high value of return loss of −31.82 dB occurs for the condition of conventional patch model with slot. This is due to the high loss tangent of Arlon, at terahertz frequencies the dielectric loss becomes quite high. It is observed that the model designed using Polymide as substrate is effective in enhancing the gain and bandwidth and realizing the triple band resonance of 0.49, 0.6 and 0.8 THz. The highest value of reflection −43.79 dB, gain of 7.04 dB and Bandwidth of 10.1 % occurs for condition stacked patch model without slot d = t where as the gain is 6.929 dB and bandwidth is 9.7 % is observed for a conventional patch model with slot. It is to be noted that the gain of 6.825 dB, return loss of −36.80 dB, and

Table 2 Comparison of performance parameters of the proposed models of the antenna

		Arlon				Polymide			
Sr. No	Condition	Freq (THz)	S_{11} (dB)	Gain (dB)	BW (%)	Freq (THz)	S_{11} (dB)	Gain (dB)	BW (%)
1.	Conventional patch model without slot	0.6	−24.17	5.196	9.5	0.4 0.6 0.8	−25.10 −36.80 −21.05	6.825	8.9
2.	Stacked patch model d = t without slot	0.6	−29.06	5.535	9.5	0.6 0.8	−43.79 −23.86	7.04	10.1
3.	Stacked patch model d = t/2 without slot	0.6	−27.52	5.426	9.7	0.5 0.6 0.8	−18.95 −32.72 −18.58	6.929	9.73
4.	Conventional patch model with slot	0.6	−31.82	5.091	9.5	0.6 0.8	−40.44 −17.34	6.467	6.6
5.	Stacked patch model d = t with slot	0.6	−21.6	5.307	7	0.6 0.8	−33.55 −14.14	6.527	8.9
6.	Stacked patch model d = t/2 with slot	0.6	−25.61	5.195	7.3	0.6 0.8	−21.38 −13.39	6.329	7.5

bandwidth of 8.9 % at 0.6 THz is obtained for the condition of conventional patch model without slot which is also appreciable. Therefore such responses with good resonances open avenues for Terahertz frequency sensing and communication applications.

4 Conclusion

In this paper, a conventional Microstrip patch antenna, stacked patch antenna has been analyzed with and without slot and simulated at terahertz frequency on two substrates having different dielectric permittivity values. It is proved in the simulated results of the proposed model of antenna that with the implementation of Electromagnetic coupling technique the enhancement in the value of reflection, gain and bandwidth occurs. The triple band resonance can be obtained by carefully selecting the substrate. The multi band, multi frequency, high gain and bandwidth antennas can be had by implementing the technique of Electromagnetic coupling. This proposed model of antennas can be used for Terahertz sensing and communication applications. The substrate with low dielectric permittivity value shows significance improvement in terms of gain, bandwidth, return loss and the analysis can be extended further on materials of different dielectric permittivity values.

Acknowledgments The authors are thankful to Marwadi Education Foundation Group and C.U. Shah University for encouraging us to carry out the research work.

References

1. Jha, K.R., Singh, G.: Terahertz planar antennas for future wireless communications: a technical review. Infrared Phys. Technol. **60**, 71 (2013)
2. Young, M.: The Technical Writer's Handbook. University Science, Mill Valley (1989)
3. I.J.Bhal, P.Bhartia: Microstrip Antennas, Artech house incorporation, 1980
4. Dwivedi, V., Kosta, Y.P., Jyoti, R.: An investigation on design and application issues of miniaturised compact MSTPA for RF communication system using metamaterials: a study. In: IEEE International RF and microwave conference proceedings, Malysia, 2–4 Dec 2008
5. Chen, T., li, S.: Metamaterials applications in sensing. Sensors **12**, 2742–2765 (2012)
6. Dwivedi, V., Kosta, Y.P.: Introduction to Frequency Electronics Microwave Engineering. Roopal and Nirav Publication, India (2004)
7. Ohkubo, S., Onuma, Y.: THz wave propagation on strip devices, properties and applications. Radio Eng. **17**(2), 48–55 (2008)
8. Saito, Y., Sheng, P.: Terahetz Electric response of fractal metamaterial structures, American Physics society, Phys. Rev. Lett. B **77**, 045124 (2008)
9. Limaj, O., Lupi, S.: Investigations on terahertz and mid infrared metamaterials, Universita di roma, October (2011)
10. Christian, D,: Frequency selective surfaces for high frequency sensitivity terahertz sensing. Appl. Phys. Lett. **93**, 083507 (2008)
11. Hussian, S., Lee, J.Y.: A dual planar metamaterial based hybrid structures in terahertz regime. In: Progress in electromagnetic symposium, Russia, Aug 19–23 (2012)
12. K.R.Jha, G.Singh, Analysis and design of Terahertz MSTPA on PBG material, J comput Electron, 2012
13. Sreenath kashyap, S., Dwivedi, V.: Swastika shaped microstrip patch antenna for Terahertz applications. In: IEEE International Conference Proceedings of ET2EN 2014
14. Chen, Z.N., et.al.: A stacked suspended plate antenna. Microwave Opt Lett. **37**, 337–339 (2003)

Game Development for the Visually Impaired

Wacheerapan Kaewprapan

Abstract Electronic games can be a powerful educational and entertainment platform. However, blind people cannot typically access these advantages because most games are primarily visually oriented and designed for persons with normal vision. This research is focused on understanding how the visually impaired relate to and interact with electronic games as well as their preferences for gameplay. This research also involved some preliminary, basic development towards a game environment and scenario for the visually impaired to participate in electronic/computer game play. This research is a step towards the goal of the visually impaired having access to this mode of entertaining, mentally stimulating and engaging education that is available widely to persons with normal visual ability. The results of the data collected during this research with 40 participants at the school for the blind, Bangkok, led us to the conclusion that the visually impaired sample could derive enjoyment and benefit from electronic games to a similar degree that normal sighted people do when the games are adapted or created with their special needs in mind or have features that can compensate for visual impairment. We also determined the types of games that the visually impaired sample had the most interest in: action, adventure and racing games.

1 Introduction

Throughout the history of the development of electronic and computer technology, computer and electronic gaming based on the new technology has expanded in use and improved in quality and capability jointly with the advancement of the technologies. Many of these games have documented cognitive, social, health and

W. Kaewprapan (✉)
Department of Electrical Technology Education, King Mongkut's University of Technology Thonburi, Thung Khru, Bangkok, Thailand
e-mail: wacheerapan.kae@kmutt.ac.th

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_112

educational benefits and are a ubiquitous part of almost all children's and adolescents' lives, with 97 % playing for at least one hour per day in the United States [1]. Unfortunately, the accessibility of computer and electronic games for blind and visually impaired and other disabled persons did not improve very much during this time. A study in 2013 stated "individuals with disabilities are unable to fully, if at all, engage with many commercial games" [2]. According to an estimate by Microsoft, the number of disabled individuals that could benefit by improved accessibility changes using existing technology and programming methods is in the 10s of millions in the U.S. alone [3]. June 06 MSDN.Microsoft.com—Making Video Games Accessible: Business Justifications and Design Considerations. But even now accessibility is very limited for the visually impaired user compared to those users with normal vision. And although the game technology has become an important part of various facets of society and culture, because of "widespread and often critical accessibility issues, many gamers with disabilities are finding themselves cut off from full participation" [4]. The denial of access to such a valuable and widely used resource is so severe that some academics have proposed a legal mandate to ensure equal access [5].

The computer game and games on electronic devices such as a gamepad, computer, smart phone and tablet can be are classified into many genres: Point and Click; Narrative; memory games; arcade style; logic; puzzle; strategy; racing; point and shoot/first person shooter; adventure, action, social, card, word, sports, music, dance and many other types of games. There exists substantial evidence of the playability, functionality, cost effective development of many possible programming and technical methods for incorporating accessibility into games current games. Almost every category of games mentioned above has been adapted or created to be accessible to visually impaired users over the past decade using available and affordable programming and interface devices [6, 7]. The application of recent advances in user interfaces such as Haptics (tactile feedback and touch sensation) and full body digital tracking (gesture interface) offer further opportunity and incentive to meet the challenge of creating more accessible games as has been demonstrated by several innovative examples in the past few years. An instance of the application of computer haptics, auditory cues and high contrast graphics to increase comprehension of the visually impaired in these subjects is described by Darrah et al. [8] and Mathias [9].

The positive human impacts, beneficial interactions, and other advantages of different types of games have been studied by psychologists [10] and these could be reasonably inferred to apply to the visually impaired as well. Similar research into the application of gaming technologies for impaired or disabled person could have positive implications, especially for educational success in difficult, tedious and sometimes abstract areas. For example, a team of researchers used a computer game enhanced with an Audio-based virtual reality environment simulator called AbES to teach navigation skills to the blind and enable them to perceive the spatial layout of a previously unfamiliar building [11].

In this research we wanted to develop and test a virtual reality (VR) game designed for visually impaired users. As a first step toward creating a playable and enjoyable gaming environment and scenario for our target audience we did a survey conducted by means of interviews to determine their preferences for game types and interfaces. Based on that data we designed and programmed an audio cue based VR test environment with touch screen interaction and accelerometer sensors. This game can be played without visual display or graphical feedback and instead uses audio based techniques such as audio cues or narrative speech. The results were based on the interviews and data collected from our experiences with the 40 visually impaired students at the Bangkok School for the Blind. We found that they shared the common frustrating access experience of visually impaired people attempting to play computer and electronic games designed for players with normal vision. They shared the sense of disparity and deprivation of being "digital outcasts" regarding the use of commonly available entertainment and educational computer gaming technology. But they did reveal that, despite their frustrations, they enjoyed computer game play, had definite preferences about game genres they enjoyed playing and desperately desired to be able to participate in this almost universal experience that is a part of our culture. Our goal is to contribute to the development of an effective design framework and paradigm for games that are accessible by the visually impaired.

1.1 The Proposes of this Study

To acquire data on the preferences and experiences of visually impaired user related to computer games and develop a game environment model that would meet their needs for successful environment perception and interaction.

1.2 The Scope of Study

The data were collected using interview techniques from a sample population possessing visual acuity less than 20/70 or completely blind. A preliminary prototype VR game environment would be designed and constructed and used in usability testing with feedback from subsequent user surveys driving further design and development.

2 Data Collection

2.1 The Samples

40 students in the School for the Blind Bangkok.

2.2 The Research Tool

The data were collect using interview technique. A questionnaire was designed in two parts: the first part was for general information regarding the sample; the second part was related to the preferred format of game design for the users.

2.3 The Research Methodology

The research was conducted as follows.

1. Researcher did field study to collect the general information and data gathered in the research.
2. The results were summarized and analyzed. The criteria for analysis are as follows; Gender ages, visual acuity level, Game devices, Types of games, The way to interact with games and graphic needs.

3 The Results

From the survey interviews of the 40 members of the sample, the games that the visually impaired preferred and their gameplay experiences and preferences can be summarized as follows:

The interviews revealed that the most of the majority (56.76 %) had serious vision problem, which are Legally and includes Low vision with the Visual acuity in the early 20/200—20/1000 and another group is completely blind (42.50 %).

The frequency of game playing show that the samples up to 50 % play games 1–3 times/week and 22.5 % play games almost every day. The few of samples are not play games much just only ones a month (12.5 %) and some (15 %) never play game because they don't have their own game play devices and most of games are for normal visions.

Sample members with limited vision. Most of their leisure activities are listening to the drama on cell phones and play music on music players.

For devices that they use to play the game. The samples use deference devices for gameplay. The device has been the top three ranked are the tablet (32.65 %), computer (30.61 %) and mobile phones (28.57 %), and other devices are only 8.16 % respectively, consistent with research on "Games and Children's Brains: What Is the Latest Research?" [10] Found that at present. Most gamers will play through these game devices. Because these devices have evolved to the player easily accessible, convenient and easily portable (tablet and mobile) and can be connected to the Internet. The Internet is considered as one factor in access to certain types of games as well. And normally the visually impaired can play through these devices has had normal vision and assistance in how to play the game.

Wearing headphones showed that the most majority (48.72 %) had been wearing headphones and like wearing headphones. The reason is making it privacy easy to hear and not disturb others. For those who did not like to wear headphones (23.08 %) are feel tinnitus or too loud. Only 28.20 % never worn headphones as they don't have them.

The response of the game found that the samples need the game response. Using sound and touching sense then vibration because listening is the primary sensory perception for visual impaired. The research on "A New Way of Increasing Access and Understanding of Math and Science for Students Who are Blind and Visually Impaired" [12] found that the response of the game using Haptic is a good way to increate immersion and attract players better in harmony with the results of the interview above.

The demand for graphics Showed that the majority (69.44 %) Graphic. Because some of the samples have a minimal level vision, they are not completely blind but can see a little. However, the users who were completely blind wanted to have graphics output so others could see their gameplay. The graphics for the visually impaired have to be in High contrast [13] which means that there is a high difference between the darkest and brightest for visual impaired to be able to see and identify objects clearly.

For the narrative sound 75.76 % of the samples want Thai (their mother tongue) as it is the national language, easy to understand and communicate as the samples are relatively low English proficiency. For some argued that English is the international language and like the sound of English rather than Thai (Table 1).

In terms of game modes that the game style that has been popular with most 3 is a Action, Racing and Adventure, respectively, consistent with research on Games and Children's Brains: What Is the Latest Research? [10] Found that genre Action is a game format that has been most popular.

For the game, and Adventure Racing is a game that is similar in style, but the Action Racing will race on the speed but Adventure more focuses on the adventures in the game. Because the Action genre is a game that relies on player's speed in responding. A thriller make players pay more attention and love to play. Also in the interview, showed that most of samples want the Action game as it is so excited and the battle scenes are the main part of the game (Table 2).

Table 1 Game type

Game type	Total (%)
Action	20.62
Adventure	15.46
Strategy	10.32
RPG	11.34
Simulation	13.40
Sports	11.34
Racing	16.49
Other	1.03

Table 2 Game type in favor between low vision and blind

Game type	Low vision (%)	Totally blind (%)
Action	22.64	22.22
Adventure	15.09	19.44
Strategy	7.55	13.89
RPG	9.43	13.89
Simulation	13.21	13.89
Sports	13.21	11.11
Racing	18.87	16.60
others	0.00	2.78

In addition, the favorite game of the low vision and blind people are in the same direction. The most 3 game types that low vision like are Action, Racing and Adventure and the blind favorite Action, Adventure and Racing respectively.

4 Conclusion

From the Interview the game format that should fit people who are visually impaired would consider that can be played on devices such as smart phones and tablets. The game can be used with headphones. Players can interact using both audio and touch sense. The graphics should be created with high contrast. The narration should be in Thai as it player native language Action is most favorite game type and also can put some adventure and racing type mix in as well.

5 Limitations

Since the survey interviews were conducted by the research team and with a small sample from one school in one location we acknowledge that the responses may not be representative of the wider population of visually impaired students. Also the game itself is a prototype and very incomplete and constrained in functionality to only provide the users with a basic experience to obtain preliminary usability data.

6 Future Research

The present study did not examine the visually impaired users experience with the test game environment. We intend to conduct another survey and use their feedback to further develop the VR environment to correct any negative usability experiences. We also would like to add a haptic interface and voice interaction to the game. We would also like to make the application portable to send an easily deployable package to other schools and institutions where a larger population of visually impaired students can experience the game and also provide a larger survey population.

References

1. Granic, I., Lobel, A., Engels, R.: The benefits of playing video games. Am. Psychol. **69**, 66–78 (2014)
2. Porter, J., Kientz, J.: An empirical study of issues and barriers to mainstream video game accessibility. In: ASSETS'13 Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility. Bellevue, Washington, USA (2013)
3. Zahand, B.: Making video games accessible: business justifications and design considerations. In: MSDN Library Technical Articles. Accessed 6 June 2007. Available at Hyperlink "https://msdn.microsoft.com/enus/library/windows/ desktop/ee415219(v = vs.85).aspx" https://msdn.microsoft.com/enus/library/windows/desktop/ee415219(v=vs.85).aspx
4. Porter, J.: Understanding and addressing realworld accessibility issues in mainstream video games. ACM SIGACCESS Accessibil. Comput. Newsl. **108**, 42–45 (2014)
5. Powers, G., Nguyen, V., Frieden, L.: Video game accessibility: a legal approach. Disabil. Stud. Q. **35**(1) (2015)
6. Folmer, E., Yuan, B.: Blind hero: enabling guitar hero for the visually impaired. In: ASSETS'08 Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility, pp. 169–176. Halifax, Nova Scotia, Canada (2008)
7. Austin Community College: Accessible games list from disabled world library. In: Library Research Guides at Austin Community College. Accessed 3 Mar 2015. Available at Hyperlink "http://researchguides.austincc.edu/ content.php?pid = 238827&sid = 3110874" http://researchguides.austincc.edu/content.php?pid=238827&sid=3110874
8. Darrah, M., Aviles, W., Murphy, K., Speransky, K.: Computer haptic effects in math and science lessons support student learning. In: Kouroupetroglou, G. (ed.) Proceedings of ICEAPVI. Athens, Greece, 12–14 Feb 2015
9. Mathias, N.: The sightlence game: designing a haptic computer game interface. In: DiGRA'13—Proceedings of the 2013 DiGRA International Conference: DeFragging Game Studies, pp. 473–487. Atlanta, GA, USA (2014)
10. Gill, R.: Games and Children's Brains: What Is the Latest Research?. Cornell University, Ithaca (2013)
11. Merabet, L., Connors, E., Sanchez, J., Chrastil, E.: Action video game play and transfer of navigation and spatial cognition skills in adolescents who are blind. Front. Hum. Neurosci. **8** (2014)
12. Darrah, M.: Computer haptics: a new way of increasing access and understanding of math and science for students who are blind and visually impaired. J. Blindness Innov. Res. **3**(2) (2013)
13. Anne Lesley Corn, J.: Foundation of Low vision. American Foundation for the Blind, New York (2010)
14. Blumberg, F.: Developmental differences at play: children's selective attention and performance in video games. J. Appl. Dev. Psychol., 615–624 (1998)

Design of Planar Dipole Array Antenna Above Square Reflector for Energy Harvesting at 2.4 GHz Band

Khanet Pookkapund and Chuwong Phongcharoenpanich

Abstract This paper presents a rectifying antenna (rectenna) which can harvest the RF wireless power at 2.45 GHz band. The proposed rectenna is designed to convert the wireless RF signal into DC power. The antenna structure consists of four printed dipoles located perpendicularly to one another to combine the pattern and increase the gain. The compact antenna radiates unidirectional pattern with the high gain. The rectifying circuit can convert the efficiency of 40 % (200 mV) with resistor of 1 MΩ when the input power is 2 mW. The rectifying circuit part is designed based on the voltage diode with stub matching circuit.

1 Introduction

Rectifying antenna (rectenna) is the an important component in a wireless power transmission (WPT) system which can convert RF energy to DC power in free space without cable for the battery supplies or Un-interruptible Power Supply (UPS) and charging electrical devices [1–5]. Historically, the first WPT technology was started in the 1985 by Heinrich Hertz. In the 1890, Nikola Tesla converted the electromagnetic waves into DC power [6, 7]. Then, researchers have been developed the WPT technology in many applications such as microwave-powered helicopter prototype, space solar power transmission (SSPT), and many others. The rectenna is used to obtain the DC power converted from the received RF power [8]. The rectenna consists of an antenna and a rectifier [9, 10]. The designed rectifier comprises of four parts namely; a schottky diode, a matching circuit, an output bypass capacitor, and a load resistor [11]. In this paper, there are four printed dipole

K. Pookkapund · C. Phongcharoenpanich (✉)
Faculty of Engineering, King Mongkut's Institute of Technology, Ladkrabang, Bangkok 10520, Thailand
e-mail: kpchuwon@kmitl.ac.th; pchuwong@gmail.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_113

located perpendicularly to one another to combine the pattern and increase the gain. The compact antenna radiates unidirectional pattern with the high gain. The rectifying circuit can convert the efficiency of 40 % (200 mV) with resistor of 1 MΩ when the input power is 2 mW.

2 Antenna Design and Results

The diagram of the rectenna system consists of four parts. First, the antenna receives the RF signal. Next, the received signal will be converted to DC power by the rectifier circuit. Finally, the DC power will be transferred to the load of the actual applications as shown in Fig. 1.

The antenna structure is composed of a reflector and four printed dipoles located perpendicularly on each side of a square array. The reflector is applied to achieve the unidirectional beam as illustrated in Fig. 2.

In Fig. 3, the rectifying circuit of the rectenna consists of the series ship coupling capacitor connected to the input open stub matching. The schottky diode part

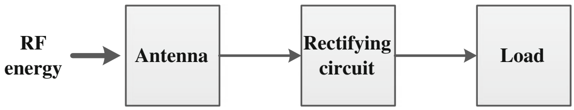


Fig. 1 Diagram of rectenna system

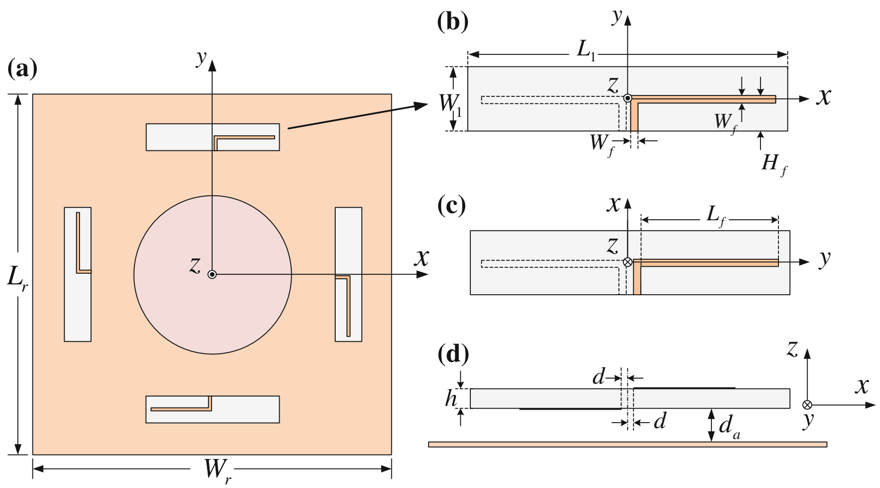


Fig. 2 Geometry and associated dimension of the proposed antenna

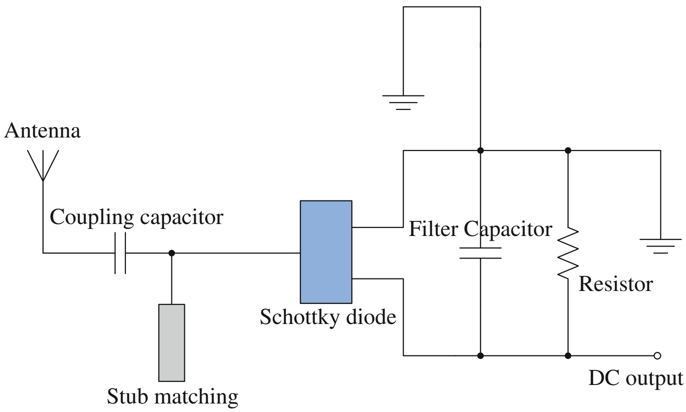


Fig. 3 Schematic of the rectifier circuit

Table 1 Antenna designed parameters

Parameter	Description	Dimension (mm)
W_r	Width of reflector	300.0
L_r	Length of reflector	300.0
W_1	Width of printed structure	10.2
L_1	Length of printed structure	13.0
W_f	Width of strip line	1.5
L_f	Length of strip line	18.0
H_f	Height of strip line	2.2
d	Distance from the center of the antenna to the strip line	0.63
d_a	Distance from the reflector to the printed dipole	16.0
h	Thickness of substrate	1.5

number is SMS-7650-006 located between the open stub matching and the ship filter capacitor. The shunt ship resistor is used as the load. The antenna parameters are given in Table 1.

The photograph of the prototype antenna is depicted in Fig. 4. The antenna is fabricated on FR-4 substrate with the dielectric constant of 4.3 and the thickness of 1.6 mm. The SMA connector is used to connect between the antenna and the transmission line.

The $|S_{11}|$ of each antenna are revealed in Figs. 5 and 6. The simulated and measured results show $|S_{ii}|$ of less than −10 dB covering cover the frequency range of 2.30–2.70 GHz and 2.28–2.60 GHz, respectively. The percentage bandwidth is 13.5 %.

The proposed antenna radiates the unidirectional beam as illustrated in Fig. 7. The simulated and measured results of HPBW in xz and yz-planes are 32° and 35° respectively. The front-to-back ratio (F/B ratio) is 24.5 dB.

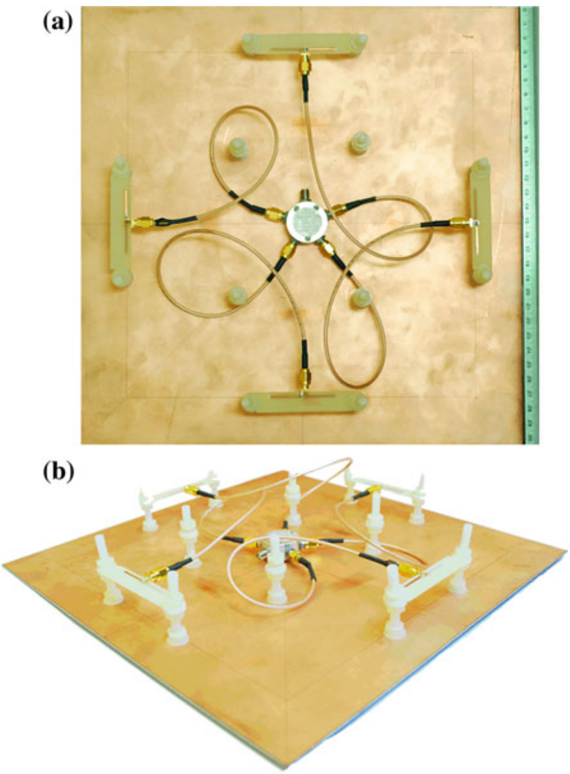


Fig. 4 The prototype antenna. **a** Top view. **b** Perspective view

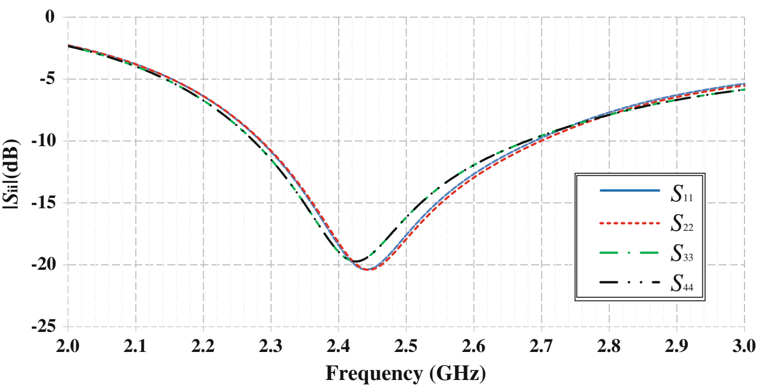


Fig. 5 Simulated $|S_{ii}|$ versus frequency

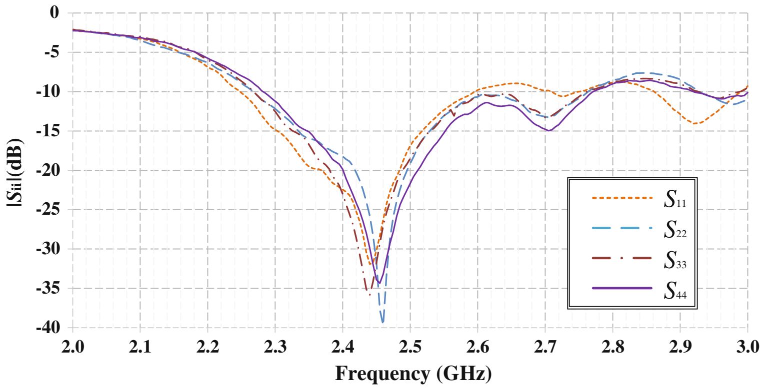


Fig. 6 Measured $|S_{ii}|$ versus frequency

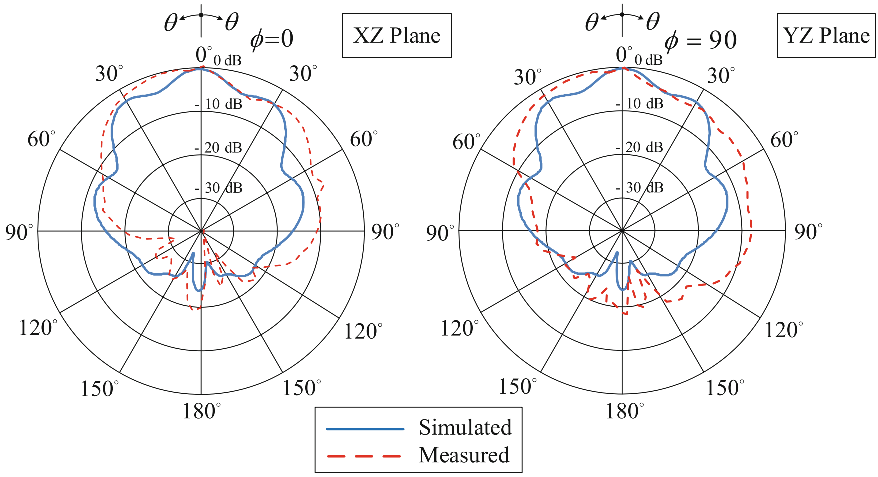


Fig. 7 Radiation pattern in the xz and yz planes at 2.4 GHz

The simulated and measured antenna gains are about 10 dBi, considering at the center frequency as shown in Fig. 8.

3 Rectifier Circuit Design

In Figs. 9 and 10 the proposed rectifying circuit is shown simulated by Advance Design System (ADS) software. The resistor and capacitor components are used the chip case with the width and the length of 0.3 and 0.6 mm, respectively. The rectifying circuit is composed of the series chip capacitor coupling for block the DC

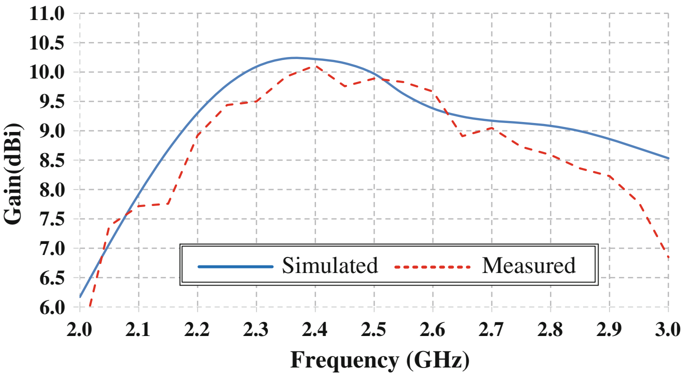


Fig. 8 Simulated and measured gain versus frequency

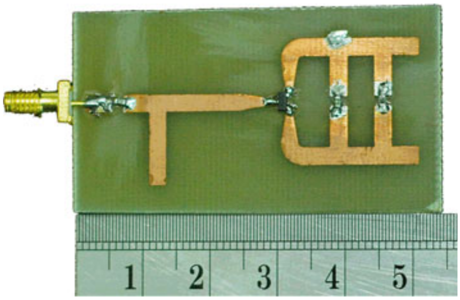


Fig. 9 The photograph of the proposed rectifier

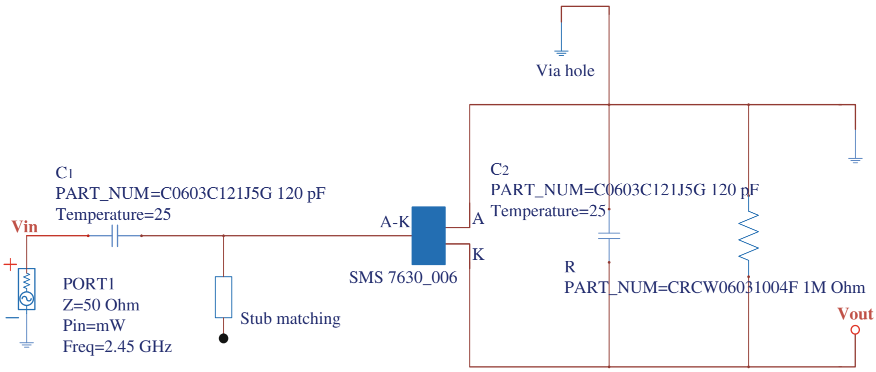


Fig. 10 Layout of the rectifying circuit

power and it is connected to the open stub matching. The rectifying doubler voltage diode is located between the open stub matching and the chip capacitor filter. The capacitor filter is reduced the ripple of the DC power Moreover, it can reject the harmonics generated by the diode. The resistor is the shunt acting as the peak voltage and the conversion efficiency (Table 2; Figs. 11 and 12).

From Fig. 13, the measured output voltage is varied as the frequency and distance. The distance between the transmitting and receiving antennas are varied as 0.5, 1.0 and 1.5 m. The RF input power is varied from −20 to 14 dBm with the

Table 2 Rectifier designed components

Components	Value
C_1, C_2	120 pF
Diode (SMS7630)	–
Resistor	1 MΩ
Stub matching	10.86 mm

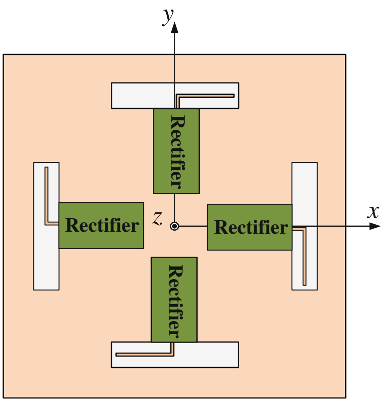


Fig. 11 Geometry of the proposed rectenna

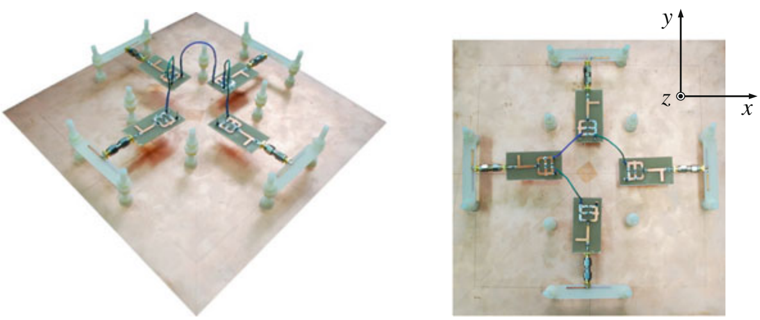


Fig. 12 The photograph of the proposed rectenna

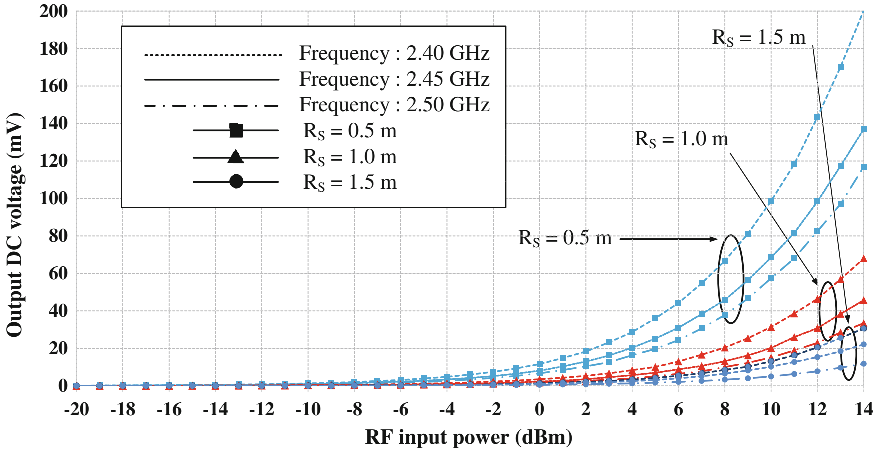


Fig. 13 The output DC voltage of the prototype rectenna

different frequencies of 2.40, 2.45 and 2.50 GHz, respectively. Obviously, the output DC voltage is 200 mV when the RF input power is 14 dBm at 2.40 GHz and the distance between the transmitting and receiving antennas is 0.5 m.

Figure 14 illustrates the conversion efficiency versus the RF input power of the prototype rectenna. The maximum conversion efficiency is 40 % when the transmitting antenna is the standard dipole considering at the frequency of 2.4 GHz. The

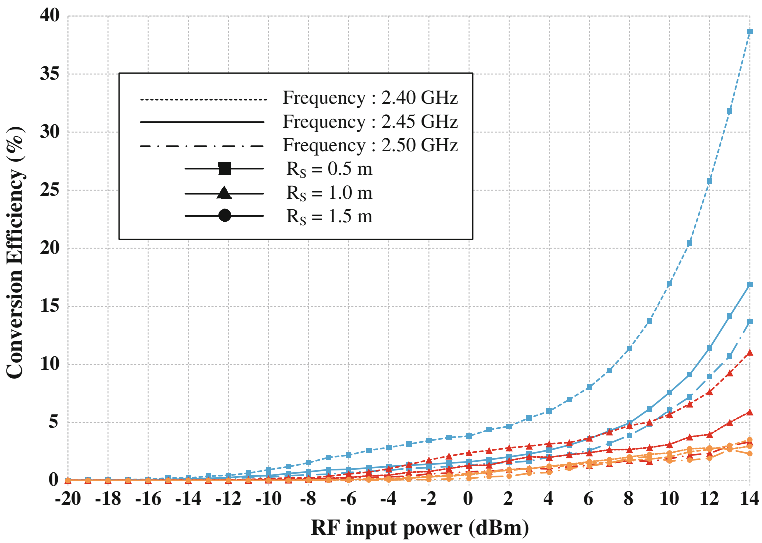


Fig. 14 Conversion efficiency of the prototype rectenna

distance between the transmitting and receiving antennas is 0.5 m and the RF input power is 14 dBm.

4 Conclusion

The proposed antenna is designed by using four printed dipole located perpendicularly to form square array. The measured $|S_{ii}|$ is covered from 2.28 to 2.60 GHz. The radiation pattern of the proposed antenna is the unidirectional beam. The antenna gain is around 10 dBi. The rectifying circuits is used to convert the RF power to the DC power. The rectifying circuit can convert the efficiency of 40 % (200 mV) with resistor of 1 MΩ when the input power is 2 mW.

References

1. Shinohara, N.: Development of rectenna with wireless communication system. In: Proceedings of the 5th European Conference on Antennas and Propagation (EuCAP), pp. 3970–3973 (2011)
2. Cataldo, S.V., Karmakar, N.C.: Wireless battery design for clock and DC motor. In: Proceedings of the Asia-Pacific Microwave Conference (APMC 2008), pp. 1–4 (2014)
3. Popvic, Z., Korhummel, S., Dunbar, S., Scheeler, R., Dolgov, A., Zane, R., Falkenstein, E., Hagerty, J.: Scalable RF energy harvesting. In: IEEE Transactions on Microwave Theory and Techniques, vol. 62, pp. 1046–1056 (2014)
4. Tsukamoto, S., Iizasa, N., Yoshitomi, K., Pokharel, R., Yoshida, K., Hattori, R., Kanaya, H., Kanemoto, D.: Development of a rectenna for batteryless electronic paper. In: Proceedings of the 2013 IEEE Region 10 Conference (TENCON 2013), pp. 1–4 (2013)
5. Tudise, D.S., Voinescu, A.: Rectifier antenna design for wireless sensor network. In: Proceedings of the 19th International Conference on Control Systems and Computer Science (CSCS 2013), pp. 184–188 (2013)
6. Chin, C.K., Xue, Q., Chan, C.H.: Design of a 5.8-GHz rectenna incorporating a new patch antenna. IEEE Antennas Wirel. Propag. Lett. **4**, 175–178 (2005)
7. Marincic, A.S.: Nikola tesla and the wireless transmission of energy. IEEE Trans. Power Apparatus Syst. **PAS 101**(10), 4064–4068 (1982)
8. Chin, C.K., Xue, Q., Chan, C.H.: Design of a 5.8-GHz rectenna incorporating a new patch antenna. IEEE Antennas Wirel. Propag. Lett. **4**, 175–178 (2005)
9. Sun, H., Guo, Y., He, M., Zhong, Z.: Design of a high-efficiency 2.45-GHz rectenna for low-input-power energy harvesting. IEEE Antennas Wirel. Propag. Lett. **11**, 929–932 (2012)
10. Van Hoang, T.Q., Séguenot, E., Ferrero, F., Dubard, J.-L., Brachat, P., Desvilles, J.-L.: 3D voltage pattern measurement of a 2.45 GHz rectenna. IEEE Trans. Antennas Propag. **61**(6), 3354–3356 (2013)
11. Zbitou, J., Latrach, M.: Hybrid rectenna and monolithic integrated zero-bias microwave rectifier. IEEE Trans. Microw. Theory Tech. **54**(1), 147–152 (2006)

Erratum to: The Design Trends of Keystream Generator for Stream Cipher for High Immunity Attacks

Saif Uldun Mostfa Kamal, Hayder Saad, Mustafa Musa Jaber, Mohammed Hasan Ali and Karam Dhafer

Erratum to:
Chapter 74 in: H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362, DOI 10.1007/978-3-319-24584-3_74

The subjected book was inadvertently published with an incorrect author's name as Saifuldun Mostafa. The correct author name is "Saif Uldun Mostfa Kamal". The erratum chapter is updated with the correct name.

The updated original online version for this chapter can be found at 10.1007/978-3-319-24584-3_74

S.U. Mostfa Kamal (✉) · H. Saad · K. Dhafer
Faculty of Computer Science, UKM, Selangor, Malaysia
e-mail: saifaldeenalsamir@gmail.com

H. Saad
e-mail: haydersaad@gmail.com

K. Dhafer
e-mail: karamdhafer@gmail.com

M.M. Jaber
Biomedical Computing and Engineering Technologies (BIOCORE) Applied Research Group, Universiti Teknikal Malaysia Melaka (UTeM), Melaka, Malaysia
e-mail: mustafamusajaber@yahoo.com

M.H. Ali
Faculty of Computer Science and Software Engineering, Universiti Malaysia Pahang, Pahang, Malaysia
e-mail: eng.mohammed12@yahoo.com

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_114