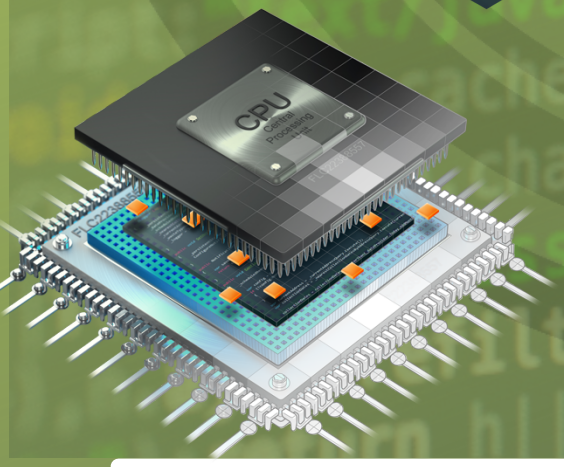
ASTES

Advances in Science, Technology & Engineering Systems Journal



VOLUME 2-ISSUE 1|JAN-FEB 2017

www.astesj.com
ISSN: 2415-6698

Editorial

Advances in Science, Technology and Engineering Systems Journal (ASTESJ) is an online-only journal dedicated to publishing significant advances covering all aspects of technology relevant to the physical science and engineering communities. The journal regularly publishes articles covering specific topics of interest.

Current Issue features key papers related to multidisciplinary domains involving complex system stemming from numerous disciplines; this is exactly how this journal differs from other interdisciplinary and multidisciplinary engineering journals. This issue contains 34 accepted papers in Computer Science & Electronics domains.

Editor-in-chief
Prof. Passerini Kazmersk

ADVANCES IN SCIENCE, TECHNOLOGY AND ENGINEERING SYSTEMS JOURNAL

Volume 2 Issue 1 **January-February 2017**

CONTENTS

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 1-5 (2017)

www.astesj.com

ASTESJ

ISSN: 2415-6698

Adaptive Intelligent Systems applied to two-wheeled robot and the effect of different terrains on performance

Sender Rocha dos Santos*, Jorge L. M. Amaral, José Franco M. Amaral

Department of Electronics and Telecommunication Engineering, Rio de Janeiro State University, UERJ, Rio de Janeiro, 20550-900, Brazil

ARTICLE INFO

Article history:
Received : 26 November, 2016
Accepted : 22 December, 2016
Online : 28 January, 2017

Keywords :
Two wheeled robot
Neuro-fuzzy control
Artificial neural net

ABSTRACT

This work discuss two different intelligent controllers: Online Neuro Fuzzy Controller (ONFC) and Proportional-Integral-Derivative Neural Network (PID-NN). They were applied to maintain the equilibrium and to control the position of a two-wheeled robot prototype. Experiments were carried out to investigate the equilibrium control and movement of the two-wheeled robot first on flat terrain, then in other situations, where terrain may not be flat, horizontal surface. The effectiveness of each controller was verified by experimental results, and the performance was compared with conventional PID control scheme applied for the prototype.

1. Introduction

Two-wheeled robots have some advantages over other types of mobile robots. The wheel configuration makes them highly maneuverable and still easier to control than legged robots. Having no more than two wheels means more room for larger wheels, potentially allowing them to traverse rougher terrain. In the last decade, auto-balance two-wheeled robots have been intensively discussed [1], [2], [3].

The dynamics of two-wheeled robots have some particular features which complicate their control. Specifically, they are non-minimum phase, under-actuated and also unstable in open loop. Therefore, robust mechanisms of auto-balance are very important to these dynamics systems, requiring proper control based techniques based on suitable sensing.

Many controller designs have been investigated for the auto-balance two-wheeled robot [4]. Proportional-Integral-Derivative (PID) controllers are widely used in many areas in industry, because no model of the plant is required, with tuning of just three gains [5]. However the purpose of an intelligent control method is to minimize the stress of control applied for complex plants.

The hybrid systems based on fuzzy logic have demonstrated their capacity of resolve several types of problems in many applications, as in robotic. Using one algorithm based on the gradient descent method, with a cost function using only the error signal, is possible to obtain one controller with good performance, [6], [7]. An important example is the ONFC controller (Online Neuro Fuzzy Controller) [8], which represents one structure with three synaptic weights.

Also, artificial neural networks (ANN) are commonly used as a strategy of adaptive control for auto-balance two-wheeled robots. The algorithm executes the learning process of neural network, where the synaptic weights are adjusted, during the execution of the network as the goal to reach the desired result [9]. The equilibrium will be improved during the neural network learning process.

Moreover, for outdoor applications the robot should also be able to stabilize on inclined or uneven terrain. This invites researchers to design controllers for the control of the stability of two-wheeled robots on inclined terrain at the desired speeds and implemented in real time.

The proposal of this work is to implement intelligent algorithms to control an auto-balance two-wheeled robotic system that can be able to move in terrains with some irregularity.

2. System Formulation

*Corresponding Author: Sender Rocha dos Santos. Email: senderrocha@yahoo.com.br

https://dx.doi.org/10.25046/aj020101

2.1. Two-Wheeled Robot Dynamic Model

Figure 1 shows the structure of an two-wheeled robot where m_1 [kg] is the mass of each wheel; m_2 [kg] is the mass of the body; I_1 [kgm^2] is the inertial moment of the wheel in relation to the gravity center; I_2 [kgm^2] is the inertial moment of the body in relation to the gravity center; r [m] is the wheel radius; L [m] is the distance of the mass center of body to axis rotation of wheel; τ_1 [Nm] is the wheel torque and τ_2 [Nm] is the body torque; θ_1 is the rotation angle of each wheel in degrees; θ_2 is the inclination angle of the body in degrees.

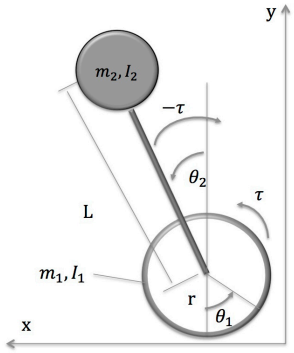


Figure 1: Autobalance diagram of a two-wheeled robot [10]

Based on Newton's equations of motion, the dynamic modeling of the robot is described by

$$\begin{cases} \tau_2 = H_1\ddot{\theta}_1 + H_3\ddot{\theta}_2 - \left(\frac{1}{2}m_2 rLsin\theta_2\right)\dot{\theta}_2{}^2 \\ \tau_1 = \frac{1}{2}H_3\ddot{\theta}_1 + H_2\ddot{\theta}_2 - m_2 gLsin\theta_2 \end{cases} \quad (1)$$

where H_1, H_2 and H_3 is defined as

$$\begin{bmatrix} H_1 & H_2 \\ H_2 & H_3 \end{bmatrix} = \begin{bmatrix} I_1 + r^2\left(m_1 + \frac{1}{2}m_2\right) & m_1 rLcos\theta_2 \\ m_2 rLcos\theta_2 & m_2 L^2 + I_2 \end{bmatrix} \quad (2)$$

and

$$\tau_1 = -\tau_2 \quad (3)$$

3. Controller Design

The primary objective in the control of two-wheeled robots is always to remain balanced and avoid toppling. Secondary objectives may include tracking a certain speed or trajectory.

In the prototype, there were implemented the balance and position control. For the balance control all variables are subtracted of the desired reference value. Therefore, null vector is the condition of auto-balance:

$$[\theta, \dot{\theta}, y, \dot{y}] = [0,0,0,0] \quad (4)$$

For the position control, the variable of the encoder, *y,* is feedback and subtracted of the reference value. The position control uses the proportional controller gain, K_{pos}. The complete strategy of control is showed on the diagram of Figure 2.

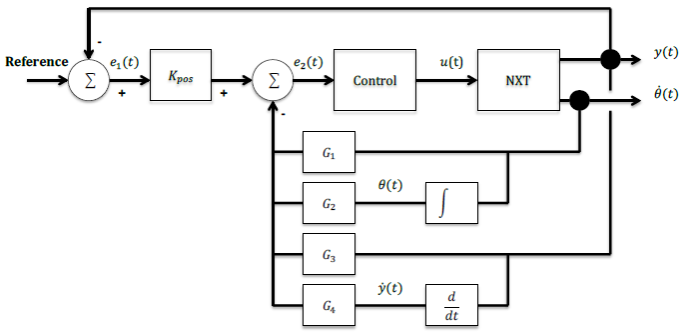


Figure 2: Autobalance control diagram

3.1. Online Neuro Fuzzy Control

Reference [3] presents the Online Neuro Fuzzy Control (ONFC) to the implementation of the online algorithm. Figure 3 shows one block diagram of the generic system using the ONFC control and its internal structure. The ONFC control uses only three membership functions (triangular and complementary) as it can be seen in Figure 4. For all values in the universe of discourse, the sum of the three membership functions μ_p, μ_z and μ_n should be equal to (5).

$$\mu_p + \mu_z + \mu_n = 1 \quad (5)$$

The control output, u, is defined based on Sugeno model of zero order. Since the membership functions are complements, the output can be written as:

$$u(n) = \mu_p(n)w_p(n) + \mu_z(n)w_z(n) + \mu_n(n)w_n(n) \quad (6)$$

If the error is limited and zero average, the weight w_z will be limited through of cost function. The same is not valid to the weights w_n and w_p, that could be unlimited for any disturbance signal in the error or for intermittent variations of the reference.

The update of weights is realized by gradient descent defined by the following cost function, J, described by (7)

$$J = \frac{e(n)^2 + \gamma w_p(n)^2 - \gamma w_n(n)^2}{2} \quad (7)$$

where n is the discrete time, e is the error between the desired and obtained signal of the system and γ is one factor between 0 and 1 and will determine the decay rate of the weights adjusted according to the plant used.

Through this cost function, it´s possible to obtain the equations (8) to perform the update of the weights in order to minimize

$$\begin{aligned} w_p(n) &= w_p(n-1) - \alpha\beta e(n)\mu_p(n) + \gamma w_p(n-1) \\ w_n(n) &= w_n(n-1) - \alpha\beta e(n)\mu_n(n) + \gamma w_n(n-1) \end{aligned} \quad (8)$$

where β is one parameter dependent of the plant and unknown.

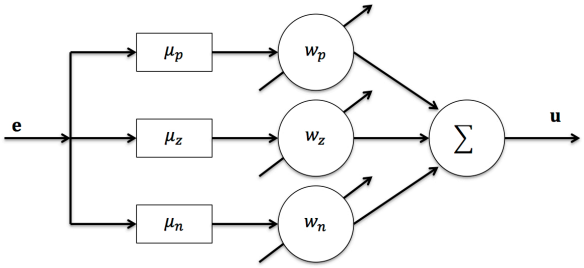


Figure 3: Structure of ONFC controller with three weights

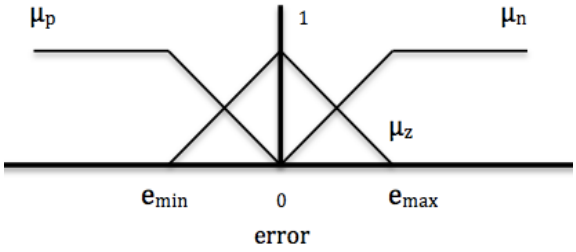


Figure 4: Complementary membership functions used in the ONFC control.

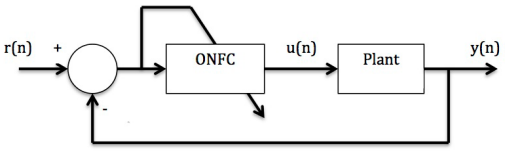


Figure 5: ONFC Control block diagram.

3.2. Artificial Neural-Networks Control

The Proportional-Integral-Derivative Neural-Networks (PID-NN) controller [4] is based on the discrete equation model of the transfer function of the PID control.

First, the PID-NN is trained off-line and the data are generated by one simulated PID system and used for training the neural network. The learning is supervised and the desired goal is to minimize the sum of the squared errors, E_{sq}, of the trained data.

Combining the equations of the discrete time PID controller, we can obtain one recurrent neural network (Figure 6). The input of network is the error signal *e(n)* and the output of the control signal is *u(n)*. Activation functions of the neurons are linear. The Figure 6 also shows the gains K_p, K_i and K_d of the neural network control structure. They are the weights, which can be adjusted. The rest of the weights present constant values. Considering $O_1(n)$, $O_2(n)$ and $O_3(n)$ the outputs of neurons of hidden layer, the output of the network could be computed by the equations:

$$e(n) = O_1(n) \tag{9}$$

$$v(n) = O_2(n) = O_2(n-1) + \Delta t * e(n) \tag{10}$$

$$w(n) = O_3(n) = \frac{1}{\Delta t}[e(n) - e(n-1)] \tag{11}$$

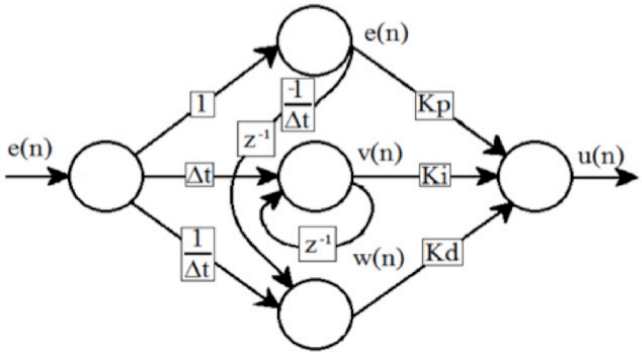


Figure 6: Structure of the PID-NN control.

The equation of control in discrete time can be seen in (12)

$$u(n) = K'_p e(n) + K'_i v(n) + K'_d w(n) \tag{12}$$

and:

$$v(n) = v(n-1) + ne(n) \tag{13}$$

$$w(n) = \frac{1}{\Delta n}[e(n) - e(n-1)] \tag{14}$$

After the off-line training, the PID-NN is combined in series with the system for online tuning, as observed in Figure 7. In this form of learning, the update of the weights is done before the presentation of each example of training. The difference between the response of the output of the plant and the reference signal generates the error e_1 that will be the input signal of the controller. The error e_2 is generated by difference between the response of the output of the plant and the response of the reference of the model. This signal is directly related with the update of the online weights.

According to the method of gradient descent, the weights update follows the equations:

$$K'_p(n+1) = K'_p(n) + \alpha e(n) O_1(n) \tag{15}$$

$$K'_i(n+1) = K'_i(n) + \alpha e(n) O_2(n) \tag{16}$$

$$K'_d(n+1) = K'_d(n) + \alpha e(n) O_3(n) \tag{17}$$

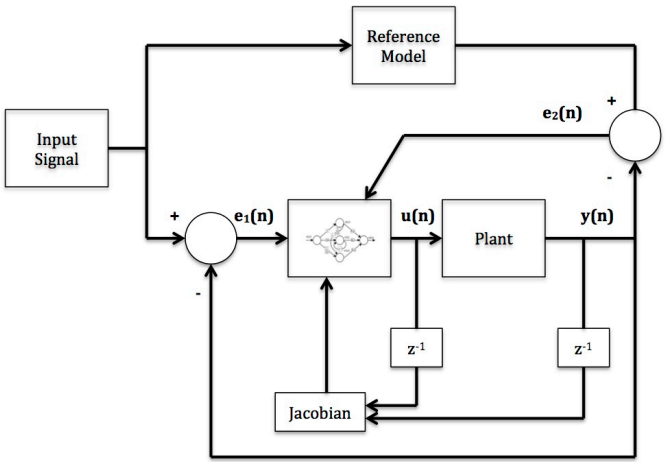


Figure 7: PID-NN Control diagram of online tuning [9].

4. Experimental Study

4.1. Hardware Implementation

Figure 8 shows the robot system used in the experiments. The hardware of whole system includes a microcontroller (AT91SAM7S256), gyro sensor (NGY1044), ultrasonic sensor, servomotors and structural support parts. Although the system has ultrasonic sensor it was not used in these experiments at this moment.

4.2. PID Control

Table 1 shows the gains of the Proportional-Integral-Derivative (PID) control of the two-wheeled robot based on the Ziegler-Nichols method. These results will be necessary to compare the performance of the other controllers.

The experimental tests have shown that K_d gain needs to be reduced to 0.000504 to ensure system stability. The proportional gain K_{pos} to position control was obtained by trial and error. The best value found was $K_{pos} = 0.495$.

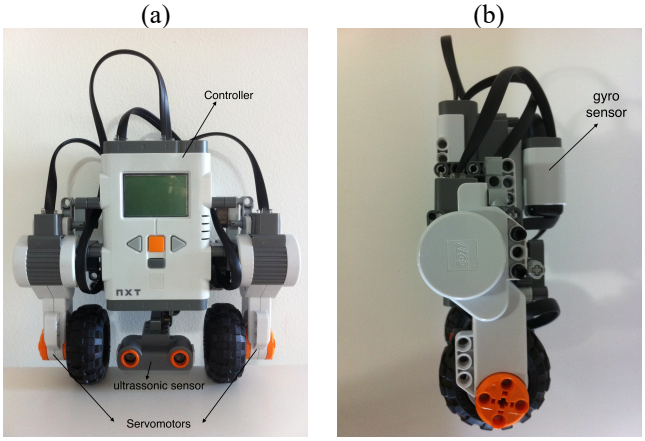
4.3. Computational Model

The simulation platform of the two-wheeled robot was developed in the MatLab-Simulink software. The computational model was validated to reproduce the same response of the auto-balance two-wheeled robot during its operation in the real world. The correlations between the simulated and experimental results were satisfactory considering the nonlinearity of the system.

Table 1: Initial PID gain based on Ziegler-Nichols method

PID Gain	Control
	PID
K_p	0.03360
K_i	0.26888
K_d	0.00105

Figure 8: Autobalance two-wheeled robot NXT (a) front view (b) right view.



4.4. ONFC Control Implementation

To evaluate the effectiveness of the proposed architecture, the proposed ONFC controller is implemented and equipped in the two-wheeled robot. The proposed control laws are represented in equations (5-8). To tune the ONFC control, several real experiments in the two-wheeled robot NXT was realized to evaluate the effect of the following parameters: learning rate α, regularization rate γ, and the universe of discourse of the membership functions. The best parameter values found were: α=0.05 and γ=0.005. The best universe of discourse of the membership functions was -100 to +100. In Figure 9 it is possible to observe the update of the weights through time for these parameters. The addition of regularization ensures a satisfactory performance.

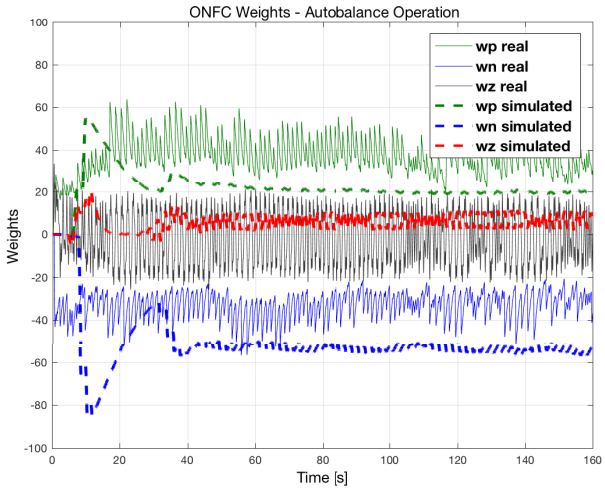


Figure 9: Real time weight update of the ONFC control and comparison with simulated data.

4.5. PID-NN Control Implementation

In order to evaluate the proposed algorithm, it was simulated on the Simulink. In this simulation platform several experiments were realized to obtain the best parameter values. The proposed control laws are represented in equations (9-17). Considering the learning rate value $\alpha_1=1.0*10^{-7}$, the average weight of the gains obtained to the online and off-line training could be defined in Table 2. The implementation of the online training was developed in Simulink software.

Table 2: Weights of PID-NN off-line training

PID-NN Gain	Average weight	
	Off-line training	Online training
K'_p	0.0364	0.0360
K'_i	0.1998	0.1998
K'_d	0.0150	0.00036

4.6. Controllers Performance

The performance of each controlled system was observed for its

1. Stationary balancing on a horizontal flat surface,
2. Balancing during motion at a desired speed on a horizontal flat surface, and

3. Balancing during motion at a desired speed on inclined surface, 10 degrees.

The initial valures of inclination angle were assumed zero and the desired speed were set to be 10% of the maximum velocity. The results indicate that the increase in speed and terrain inclination unfavourably affects the performance of the speed control. Figure 10 shows the range of the oscillation of the inclination angle of figure 10 (a) is smaller than plot 10 (b) and 10 (c). Then the error stays close to null value (θ_2=0º). To quantify the analysis of controllers some comparisons among the integral of plots in Figure 10 is showed in Table 3.

Through this calculus get the error between the function of curve and the zero value (equilibrium point). If the value is smaller so the error is smaller. In the case of ONFC, the mean square error in 1200 seconds of the test was of the 0.0172 while for the same interval the PID controller means square error was of the 0.276. Therefore the ONFC control is more stable and it has a better performance for the two-wheeled robot.

Table 3: Comparative performance of tilt control in different scenarios

Controller	Error	Flat Terrain		Inclined Terrain
		Stationary	Run	Run
ONFC	MSE	0.0241	0.0302	0.0172
	RMS	0.1562	0.1718	0.0887
PID	MSE	0.0279	0.0276	0.0306
	RMS	0.1663	0.1644	0.1502
PIDNN	MSE	0.0351	0.0266	0.0613
	RMS	0.1883	0.1610	0.2663

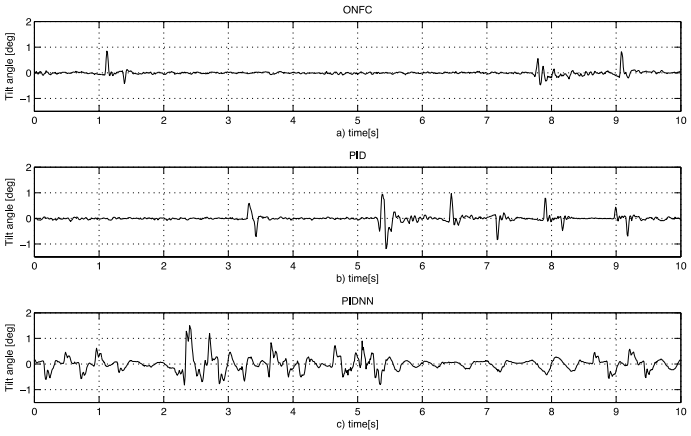


Figure 10: Balance control results of tilt control of the two-wheeled robot prototype by employing ONFC, PID, and PID-NN control strategies in inclined terrain.

5. Conclusion

The ONFC and PID-NN controllers are developed for a two-wheeled robot prototype and they are successfully applied to control the equilibrium of the two-wheeled robot. Compared with conventional PID control schemes, the intelligent controls present better performance, which was verified by experiments. It was verified the effects of terrain inclination in real experiment. The design of such controllers shows versatility, simplicity, improved performance and robustness.

Conflict of Interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] P. Deegan, B. Thibodeau, and R. Grupen, “Designing a self-stabilizing robot for dynamic mobile manipulation”, In Proceedings of the Robotics: Science and Systems Workshop on Manipulation for Human Environments, Philadelphia, Pennsylvania, August 2006.

[2] G. H. Lee, and S. Jung, “Line Tracking Control of a Two-Wheeled Mobile Robot Using Visual Feedback”, International Journal of Advanced Robotic Systems. 2013.

[3] K. Su and Y. Chen, “Balance Control for Two-Wheeled Robot via Neural-Fuzzy Technique”, SICE Annual Conference 2010.

[4] R. P. M. Chan, K. A. Stol, C. R. Halkyard “Review of modeling and control of two-wheeled robots”, Annual reviews in control 37, pp. 89-103, 2013.

[5] D. Jin, “Development of a Stable Control System for a Segway,” Royal Institute of Technology, 2013.

[6] M. R. Gouvêa, “Controle Neurofuzzy de Motor de Indução Com Estimação de Parâmetros e Fluxo de Estator”, Tese de Doutorado, Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Minas Gerais, Belo Horizonte, MG, 2005.

[7] A. V. Pires, “Controladores baseados em técnicas de inteligência computacional: Análise, projeto e aplicação,” Dissertação de Mestrado, Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Minas Gerais, Belo Horizonte, MG, 2007.

[8] L. M. Spadini, P. E. Silva, L. F. F. Campos, C. J. F. Araújo, and A, Nied, “Desenvolvimento de Controle Neurofuzzy para Plantas Não Lineares: Aplicação em Tanques Acoplados”, Simpósio Brasileiro Automação Inteligente, SBAI 2013, 2013.

[9] A. C. K. Ferrari, “Controlador PID sintonizado por redes neurais artificiais”, Monografia, Curso de Engenharia Elétrica, Universidade Federal do Paraná, Curitiba, 2011.

[10] M. R. Bageant, “Balancing a Two-Wheeled Segway Robot”, Thesis (S.B.)-Massachusetts Institute of Technology, Dept. of Mechanical Engineering, 2011.

[11] B. Costa, “A practical implementation of self-evolving cloud-based control of a pilot plant”, in Proceedings of 2013 IEEE International Conference on Cybernetics (CYBCONF 2013), Lausanne, Switzerland, 2013, pp. 7-12.

[12] R. Precup, M. Radac, M. L. Tomescu, E. M. Petriu, S. Preiti, “Stable and convergent iterative feedback tuning of fuzzy controllers for discrete-time SISO systems”, Expert Systems with Applications, vol. 40, no. 1, pp. 188-199, 2013.

[13] A. Castro, R. Carballo, G. Iglesias, J.R. Rabunal, “Performance of artificial neural networks in nearshore wave power prediction”, Applied Soft Computing, vol. 23, pp. 194-201, 2014.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 6-15 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

System Testing Evaluation for Enterprise Resource Planning to Reduce Failure Rate

Samantha Mathara Arachchi*, Siong Choy Chong, Alik Kathabi

Management and Science University (MSU), University Drive, Off Persiaran Olahraga, Section 13, 40100 Shah Alam, Selangor Darul Ehsan, Malaysia

Email: saman17lk@yahoo.com, eddyscchong@yahoo.com, alik@msu.edu.my

ARTICLE INFO

Article history:
Received: 26 November, 2016
Accepted: 23 December, 2016
Online: 28 January, 2017

Keywords:
ERP
Risk factors
Security risks
Failures and Evaluations

ABSTRACT

Enterprise Resource Planning (ERP) systems are widely used applications to manage resources, communication and data exchange between different departments and modules with the purpose of managing the overall business process of the organization using one integrated software system. Due to the large scale and the complexity nature of these systems, many ERP implementation projects have become failure. It is necessary to have a better test project management and test performance assessing system. To build a successful ERP system these processes are important. The purpose of the Test project management is verification and validation of the system. There was a separate stage to test the quality of software in the software development lifecycle and there is a separate independent Quality Assurance and testing team for a successful ERP development team. According to best practice testing principles it is necessary to, understand the requirements, test planning, test execution, identify and improve processes. Identify the necessary infrastructure; hardware and software are the major areas when developing test procedures. The aim of this survey is to identify ERP failures associated with the ERP projects, general and security within the Asian region, so that the parties responsible for the project can take necessary precautions to deal with those failures for a successful ERP implementation and bring down the ERP failure rate.

1. ERP Testing Overview

System testing checklist is an excellent way to ensure the success of a software testing process during the software development life cycle as it helps to make proper planning for any particular software [1], based on its requirements and expected quality [2,3]. Software testing plays a major role in any software development life cycle or methodology. Therefore, it has to be done accurately by professional software testers. Although professional testers have a good knowledge and experience in software testing, there are situations where they also create a big mess with the test planning and process testing activities. So the best way to solve this problem is to prepare a system testing checklist in order to make an accurate and suitable plan to test each software [1]. When a QA team has checklists, they can easily process all the required test activities in proper order without missing anything [2, 3]. A system testing checklist is a document or may be a software tool that systematically plans and prepares software testing and helps to define a framework for test environment, testing approach, staffing, work plan, issues, test sets and test results summary. As mentioned earlier the software testing phase is vital for any implementation team because, if the developed system doesn't show the expected quality it would be a total waste of time and money since clients may not prefer to use the software in the future [2,3,4].

2. Literature Review

ERP software quality evaluation system developed on the bases of the testing [1]. Finally, the project manager must assess the system and also the factors used for the new system Therefore, it is important to maintain documents for each phase. In order to do this the allocated and used resources must be compared and it is also important to assess the test performances [5].

2.1. Requirements for Test Procedure

The comprehensive evaluation and the comparison of different testing strategies for scenarios is a critical part in the

*Corresponding Author: Samantha Mathara Arachchi, Management and Science University (MSU), University Drive, Off Persiaran Olahraga, Section 13, 40100 Shah Alam, Selangor Darul Ehsan, Malaysia, Email: saman17lk@yahoo.com

https://dx.doi.org/10.25046/aj020102

current research on software testing [6]. Mutation analysis helps to determine the effectiveness of a test strategy and compare the different test strategies based on their effectiveness measures [7]. Test strategy which is reliable for all programs cannot be established [6]. Boolean and Relational Operator (BRO) and Boolean and Relational Expression (BRE) testing are two condition based testing strategies [8]. These are unlike the existing condition based upon the detection of errors such as Boolean errors and relational expression error in a condition [8]. In accordance with the empirical studies of the algorithm SBEMIN and SBEMINSEN along with theoretical properties of BRO and BRE it is confirmed that BRO and BRE testing are practical and effective for testing programmes with complicated conditioning [8].

Software based self-test strategy is well suited for a low cost embedded system, which does not require immediate detection of error [9].

Three software state of the practice testing strategies are: code reading by stepwise abstraction, functional testing using equivalence portioning and boundary value analysis, and finally structural testing using 100% statement coverage criteria [10]. Adaptive testing is an online testing strategy, adapted to reduce the variance in the results of the software reliability assessment [11]. Complex system test strategies are derived through the performance of the software, calculated based on the computer line of code. Test policies for the complex system are chosen based on the derived test strategy [12].

Optimally refined proportional sampling is a strategy which is simple and low in cost [13]. Empirical study was made through a sample programme with seeded error, and found that this strategy is better than random testing [13].

Test data generation strategy is used to generate the test data [7]. Test strategy can only be considered reliable for a particular programme, if and only when it produces a reliable bunch of test data for that programme [7].

Path analysis testing strategy is a method used to analyze the reliability of the path testing. In this strategy data are generated which enables the different paths of the system to be executed [14] and [13].

Debugging strategy based on the requirement of testing is focused on the situation where the selected testing requirement does not indicate the fault site but apparently provides helpful information for fault localization [15].

Path pre fix testing strategy is a user interactive. Adaptive testing strategy [16] and Test path which are used previously will be used to select the subsequent paths for testing [16].

2.2. Instruction for Test Procedures

It is necessary to have a better test project management and test performance assessing system. To build a successful ERP system these processes are important. Test project management is different from the other project management roles [3]. The purpose of the Test project management is verification and validation of the system.

There was a separate stage to test the quality of software in the software development lifecycle and there is a separate independent Quality Assurance and testing team for a successful ERP development team. According to best practice testing principles it is necessary to, understand the requirements, test planning, test execution, identify and improve processes (Process Improvement, Defect Analysis, requirements review and risk mitigation) [17].

A Development Model's Implications for testing is based on: Review the user interface early, Start writing the test plan as early as possible, Start testing when the work is on the critical path, Plan to staff the project very early, Plan waves of usability tests as the project grows more complex, Plan to write the test plan and Plan to do the most powerful testing as early as possible [18,19]. Identify the necessary infrastructure; hardware and software are the major areas when developing test procedures.

2.3. Get Ready for Test Preparation

Emphasis is on the typical test description or the schedule and milestones for each of the tasks. Therefore, the project manager has complete control over the project. The development timeline will be shown as; Product Design, Fragments coded: first functionality, Almost alpha, Alpha software, Pre-beta, Beta, User Interface Freeze, Pre-final and Final Integrity Test Release [18-21,3].

2.4. Create Test Plan

As soon as the test plan is created incorporating risk management, communication management, test people management and test cases, it has to be reviewed by an independent group and approved by an authority that is not influenced by the project manager responsible for the testing [4].

Test plans help to achieve the test goal in an organized and planned manner. It says, build test plan, define metric objective and finally review and approve the plan as a step involved in the development of test plan [18]. Test plan is an ongoing document as it changes when the system changes, and it is very true in the spiral environment.

Test plans are the main factors which lead to the success of the system testing [4]. Testing is done at different levels and the master test plan is considered to be the first level in order to follow the methodology of testing.

In addition to the master level test plan, level specific test plan according to the phases such as acceptance test, system test, integration test and unit test plans will be created. The main goal of test plan is to address issues such as test strategy and resource utilization.

2.5. Create Test Cases

Software failures in a variety of domain have major implications for testing and further emphasize on exhaustive testing of computer software where all errors in the system are identified by a combination of fewer parameters and testing the n-tuples of parameters, which enable effective testing [22,23].

"Proportional Sampling Strategy" which facilitate partition testing, i.e. testing in each partition has higher probability in identifying faults compared to random testing. The amount of test cases used in each partition will be proportional to the size of the partition [24]. Therefore, identifying and naming each test case is significant.

A good set of test cases would have to develop and most of the developers have similar experiences in developing software products rather than responding to a good test set [25]. Since software programmes have many branches, paths or sub-domain, testing every part is not practical. Test cases for special functions and cases are important [26].

2.6. Testing Methods

There are several testing methods that incorporate to test the ERP application. These methods can be used in a different stage as well as for different purposes.

Structural Testing: the software entity is viewed as a "white box". The selection of test cases is based on the implementation of the software entity. The expected results are evaluated on a set of coverage criteria. Structural testing emphasizes the internal structure of the software entity; [27,28]. Grey-box testing, is defined as testing software while already having some knowledge of its underlying code or logic [28]. The testers are only aware of what the software is supposed to do, not how it does it [29, 28]. Black-box testing methods include: equivalence partitioning, boundary value analysis, all-pairs testing, state transition tables, decision table testing, fuzz testing, model-based testing, use case testing, exploratory testing and specification-based testing [30]. Validation is basically done by the testers during the testing. While validating the product if some deviation is found in the actual result as against the expected result then a bug is reported or an incident is raised. Hence, validation helps in unfolding the exact functionality of the features and helps the testers to understand the product in a much better way. It helps in making the product more user friendly [18, 28].

2.7. Testing Levels

Testing Levels have been used to test the application in a different level. There are different testing levels also to make sure of the accuracy of the Applications.

Unit testing is the basic level of testing that focuses separately on the smaller building blocks of a program or system. It is a process of executing each module to confirm that each performs its assigned function. It permits the testing and debugging of small units. Therefore, it provides a better way of integrating the units into larger units [3]. Integration testing is testing for whole functionality and also for the acceptance of them. It also verifies some nonfunctional characteristics. Some examples of system testing include usability testing, stress testing, performance testing, compatibility testing, conversion testing, and document testing and so on, as described by Lewis [3]. System testing is tested as a whole for functionality and fitness of use based on the system test plan. Systems are fully tested in the computer operating environment before acceptance testing carried out. The source of the system tests are the quality attributes specified in the software quality assurance plan. System testing is a set of tests to verify quality attributes [3].

2.8. Testing Types

ERP type of testing is a continuous task performed before implementation of the system and even after being implemented in order to ensure the quality of the ERP system. ERP systems are very critical to a company's operation as all its functions are integrated into one holistic system. According to William E. Lewis in 2005 and 1983, apart from the traditional testing techniques, various new techniques necessitated by the complicated business and development logic were realized to make software testing more meaningful and purposeful [3]. Software testing is an important means of assessing the software to determine its quality [17]. Since testing typically consumes 40 ~ 50% of development efforts and consumes more effort for systems requiring higher reliability it is an essential part of software engineering [27]. Various techniques reveal different quality aspects of a software system, and there are two main categories of techniques, functional and structural [27].

Alpha testing of an application is performed when development is nearing completion; minor design changes may still be made as a result of such testing. Typically, the alpha testing is done by end-user or others, not by programmers or testers [3]. Deciding on the entry criteria of a product for beta testing and deciding the timing of a beta test poses several conflicting choices to be made. Therefore, the success of a beta programme depends heavily on the willingness of the beta customers to exercise the precut in various ways, being well aware that there may be defects [3, 4].

The load testing is an application tested under heavy loads. Such testing of ERP applications run through the internet under a range of loads to determine at what point the system's response time degrades or fails [3].

Volume test checks are applied to test whether there are any problems when running the system under test with realistic amounts of data, or even a maximum. Typical problems are full or nearly full disks, databases, files, buffers, counters that may lead to overflow. Maximal data amounts in communications may also be a concern [31]. Acceptance Testing is applied to test whether the software products meet the requirements of the users or contract. [32]. This ensures the overall acceptability of the system.

Automated acceptance testing has been recently added to testing in agile Software development process [33]. The concern is on good communication and greater collaboration which enable testing in each stage to decide whether all functions meet the requirements and push the process to arrive at an acceptable software product [34].

The Compatibility testing is done to check how well or not the software performs in a particular hardware, software, operating system, network environment and other infrastructure compatible [3]. Conformance testing, also known as compliance testing, is a methodology used in engineering to ensure that a product, process, computer programme or system meets a defined set of standards.

Manual trial and error testing approach has made the ERP system complicated. There is a growing demand for test automation due to a shortfall in manual testing such as, high cost and resource utilization, problem in detecting errors, and timeline issues in the manual testing [12]. Not all the testing should be automated, only up to 40% -60% should be automated [31,35].

Software testing is an important, but time consuming and costly process in the software development life cycle [36]. Testing process will be efficient only if the testing tool would enable to organize and present it in an easily understandable manner in addition to executing the test [37].

Factors that support the use of test automation are, low human involvement in testing and steady underlying technology reusability [38]. In Industry, companies are trying to achieve efficiency in testing through automating the software testing for both hardware and software components [36]. Currently, testing tools have enhanced to a level where one can automate the interaction with software systems at the GUI level [39]. Even though there is such advancement, the industry currently faces a lack of knowledge on the usability and applicability of testing tools [40].

Preparing application for the automated testing, organizing the automation team members, designing automation test plan, defining criteria to measure the success of automation, developing appropriate test cases, appropriate selection of test tools and finally, implementation of the automated testing are considered test automation best practices to be followed [1,41]. Software test automation framework (STAF) is a multi-platform, multi-language approach based on the underlying principle of reusable service [36]. STAF is used to automate major activities of the testing process and automate resource intensive test suit [36].

The use of test automation started in the mid-1980s after the invention of the automated capture or replay tool [17]. Now we have very sophisticated tools used throughout the testing process such as Regression testing tool, Test design tools, Load/performance tools Test management tools Unit testing tools, Test implementation tools, Test Evaluation tools, Static test analyzer Defect Management tool, Application performance monitoring /tuning tool and Run time analysis testing tool [17,27].

System testing or software testing is a major area in the software industry. In traditional waterfall development life cycle the system testing / software testing phase plays a big role. There are several definitions available and used for system testing, and according to the IEEE definition, system testing is “the process of analyzing a software item to detect the differences between existing and required conditions (that is, bugs) and to evaluate the features of the software item”. Basically, this means that the main intention of system testing is to identify whether there is a gap between the developed functions or system with the expected outcomes which are defined according to the previous and the current customer requirements. “Validation” and “Verification” are very important processes in system testing. There are various types of techniques and methodologies used in the software development industry to test and debug the software system and these techniques are different from one another according to the type of system or software or the organization perspective. System testing is an overall activity which started from the requirement analysis phase and goes through each and every phase in the software/ system development life cycle.

2.9. Verification Tests

Verification is intended to check that a product, service, or system meets a set of design specifications. In the development phase, verification procedures involve performing special tests to model or simulate a portion, or the entirety of a product, service or system. The verification procedures involve regularly repeating tests devised specifically to ensure that the product, service, or system continues to meet the initial design requirements, specifications, and regulations as time progresses [42].

2.10. Vulnerability Testing

Vulnerability analysis is a process that defines, identifies, and classifies the security holes (vulnerabilities) in a computer or in an application such as ERP, network, or communications infrastructure. In addition, vulnerability analysis can forecast the effectiveness of proposed countermeasures and evaluate their actual effectiveness after they are put into use [3].

3. Research Method

This research employs the descriptive method. This method is used in any fact-finding study that involves adequate and accurate interpretation of findings. Relatively, the method is appropriate to the study since it aims to describe the present condition of ERP failure analysis [43].This method describes the nature of a condition as it takes place during the time of the study and explore the system or systems of a particular condition at each and every phase of the SDLC.

Specifically, direct-data survey using questionnaire was used in the study. This is a reliable source of first-hand information because, since the researcher directly interacts with the respondents, rational and sound conclusions and recommendations can be derived at. The respondents were given ample time to assess the failures in the ERP testing faced by the software development companies in Sri Lanka. Their own experiences at the testing phase of software development are necessary in identifying their strengths and limitations.

The primary data are derived at from the responses provided by the respondents through the self-administered questionnaires prepared by the researcher. The constructs are based on recent literatures related to reducing the failure rate of ERP in the software development industry in Sri Lanka as well as the challenges and the concepts cited by respondents during the pre-survey. In terms of approach, the study employed the quantitative approach which focused on obtaining numerical findings.

4. Variables and Hypotheses

Variables are anything that can take on differing of varying values [44]. Based on the literature, the following variables have been identified as shown in the research framework. Independents variables will usually affect the expected output,

that is, the dependent variable. The System Testing is the independent variables in this study. A dependent variable is a measure based on the independent variable. In other words, the dependent variable responds to the independent variable. ERP failure has been considered as the dependent variable in this study. A mediating variable can influence the relationship between an independent variable and the dependent variable. In other words, it determines whether the indirect effect of the independent variable on the dependent variable is significant [45]. Furthermore a variable may be considered a mediator to the extent to which it carries the influence of a given independent variable to a given dependent variable. In this research, the testing method mediators which is tested through the following hypotheses:

H1 There is a negative relationship between System Testing and ERP Failure

H2 Many Testing Methods do mediates the positive relationship between System Testing and ERP Failure

5. Study Setting

The researcher has critically examined the ERP development companies under the Companies Registration Act and found that such development work is undertaken by software development companies registered under the Sri Lanka BOI and the Public Limited Companies (PLC) Act of Sri Lanka. The software development companies registered under the PLC Act have not been considered in this study due to their lack of investment capacity and involvement of ERP application development [46]. In addition, the PLCs do not have enough capital and resources as well as bank guarantee when signing vendor agreements [47]. Hence, they are not engaged in developing total solutions for ERP applications. Further, there are some companies which registered under the BOI but having less than twenty employees, making little investments and having less capital. Such PLCs have not been considered as well. The same justification has been applied for the BOI companies to select the appropriate ERP development companies for data collection.

The direct-data survey aims at collecting pertinent data to achieve the research objectives. Accordingly, direct-data survey is used to reveal the status of some phenomenon amongst the people identified who are engaged in developing ERP applications at software testing phase of the SDLC in the software development industry in Sri Lanka.

6. Unite of Analysis

The unit of analysis consists of employees engaged in testing ERP applications at the phase of the SDLC in software development industries in Sri Lanka. All of the respondents were selected using stratified sampling method by considering staff members with similar educational qualifications and working experience. Under this sampling method, each member of a population has an equal opportunity to become part of the sample based on the software testing phase of SDLC prescribed. As all the members have an equal chance of becoming research participants, this sampling method is said to be the most efficient sampling procedure [44]. Using this sampling strategy, the researcher first defined the population, that is, 48 companies registered under the BOI, and then listed all the members and selected members to determine the sample based on the testing phase of the SDLC.

6.1. The Survey Questionnaire

There were 400 participants for the questionnaire survey in soft testing stage of the SDLC. Questionnaire was given to each participant individually or the head of divisions in each and every software development company.

6.2. Content Analysis

The survey items in the study were developed as a result of analyses of previous studies, discussions with practitioners in the field as well as a review of relevant literature which also takes into consideration the research framework and hypotheses developed.

Reliability and validity are important aspects of questionnaire design. According to Suskie (1996), a perfectly reliable questionnaire elicits consistent responses [48]. Robson (1993) indicates that a highly reliable response is obtainable by providing all respondents with the same set of questions. Validity is inherently difficult to establish using a single statistical method. If a questionnaire is perfectly valid, these inferences drawn from it will also be accurate. In addition, Suskie (1996) reports that reliability and validity are enhanced when the researcher takes certain precautionary steps. Accordingly, it is important to have people with diverse backgrounds and viewpoints to view the survey before it is administered to determine if: (1) each item is clear and easily understood; (2) they interpret each item in the intended way; (3) the items have an intuitive relationship to the topic and objectives of the study; and (4) the intent behind each item is clear to colleagues knowledgeable about the subject [48].

The testing stage of the SDLC was used to test the overall ERP application and to detect whether there were any bugs. The questionnaire developed included48 key points as items clustered under seven factors. In addition, there is another set of questionnaire which was used to test the quality or failure of ERP applications.

6.3. The Sample Design

According to the statistics of BOI, there are 66 companies registered for BOI software development projects [49]. After applying the justifications mentioned under the study setting, only 48 companies fulfilled the conditions and are capable of developing ERP applications for local and foreign industries. The company name and sensitive details such as job hierarchy were not disclosed due to taxation and policies as per government rules and regulations.

In the study, the representative samples were selected using stratified sampling technique to select the employees (groups/strata), followed by applying the random sampling approach to distribute the questionnaires. The stratified random sampling is an appropriate methodology to make the samples proportionate so that meaningful comparisons between the sub-groups in the population are possible.

Accordingly, everyone in the sampling frame is divided into ‘strata’ (groups or categories). There are four strata as follows:

1. Employees engaged in Business Requirement Analysis (BA)

2. Employees engaged in System Design (SD)
3. Employees engaged in System Implementation (SI)
4. Employees engaged in Quality Assurance for Software Testing(ST)

Within above stratum, a simple random sample is selected to the Quality Assurance for Software Testing (ST) group. Since the employees engaged in the same job role, possessing the same academic qualifications and experience in the same field within the strata (group), the random sampling method was used to distribute the questionnaires to within the above employee group.

To ensure that a sample of 400 from a group of 7745 employees in the four strata (groups) that Software Testing employees in same proportions as in the population (i.e. the group of 7745), the researcher has identified the following sample size who engaged in Quality Assurance for Software Testing(ST) for this research.

No. of ST in sample = (400 / 7745) x 1413 = 73

On the basis of this calculation, the researcher has determined the number of respondents for software testers and for this research 73 sample size has been identified as software testers.

7. Methods of Analysis

Data analysis generally begins with the calculation of a number of descriptive statistics such as the mean, median, standard deviation scores and the like. The aim at this stage is to describe the general distributional properties of the data, to identify any unusual observations (outliers) or any unusual patterns of observations that may cause problems for later analyses carried out on the data [50]. Taking the cue from Kristopher (2010), severalstatistical methods were used to formulate the output of data analysis using Statistical Package for the Social Sciences (SPSS) and Sobel test as described in the following sub-sections.

8. Hypotheses Testing for System Testing

The researcher has tested the hypotheses of H4: There is a negative relationship between System Testing and ERP Failure. It signifies that if the System Testing does not do a proper job it could be a case of failure for the ERP application. If they do a better job according to the factors that the researcher has identified as per the literature they might be able to achieve success of developing ERP application by doing a proper System Testing.

8.1. Overall Descriptive Analysis for Mean in each Factor in System Testing

The overall Descriptive Analysis for mean in each factor is given as follows as shown below table 1;

Table -1: Overall Descriptive Analysis for mean in each factor in System Testing

Descriptive Statistics			
	Mean	Std. Deviation	N
TOTALST	3.2566	.26071	73
RD	3.4188	.56960	73
RTP	3.3973	.71720	73
ITP	3.6438	.60369	73
RDTP	3.6545	.35744	73
TM	3.4452	.57168	73
TT	3.5845	.43706	73

Based on the 73 samples of respondents Required Documents (RD), Requirements for Test Procedure (RTP), Instruction for Test Procedures (ITP), Require Document for Test Preparation (RDTP), Testing Methods (TM), Testing Types (TT) are also close to the very important scale. The Requirements for Test Procedure (RTP) has the biggest variance of points (0.71720) and Require Document for Test Preparation (RDTP) has the lowest variance of points (0.35744) as shown in the above table 1.

8.2. Descriptive Statistics of System Testing (ST)

The descriptive statistics of system testing is given bellow table in 2;

Table 2 : Descriptive statistics of system testing

Descriptives				
			Statistic	Std. Error
TOTALST	Mean		3.2566	.03051
	95% Confidence Interval for Mean	Lower Bound	3.1957	
		Upper Bound	3.3174	
	5% Trimmed Mean		3.2555	
	Median		3.1875	
	Variance		.068	
	Std. Deviation		.26071	
	Minimum		2.73	
	Maximum		3.85	
	Range		1.13	
	Interquartile Range		.40	
	Skewness		.267	.281
	Kurtosis		-.611	.555

The mean system Implementation for ERP applications for the 72 samples is 3.2 with the standard deviation of 0.26071. The maximum and minimum ranks are 3.85 and 2.73. The median is 3.2, indicating at least 50% of the developers ranked more than 3.2. The mode value, obtained using the frequency procedure, is 3.2 Thus the most frequent rank among the developers was 3. According to this, the Factors in the system implementation questionnaire for ERP system testing is moderately important.

Since the mean and median values are very close to each other, perhaps the data could be symmetrical. The skewness value is 0.267. If it shows long tailed data to the right it is said to be positive skewed or skewed to the right The Kurtosis value is -611, which is within ±1. Hence, the data can be assumed to be platykurtik.

8.3. Correlation Analysis for Overall System Testing for ERP Failure

The correlation analysis for overall system testing for ERP is as follows according to the table 3;

According to the correlation analysis in the above table the Factors Required Documents (RD), Requirements for Test Procedure (RTP), Require Document for Test Preparation (RDTP), Testing Methods (TM), Testing Types (TT) have correlation of 0.632, 0.383, 0.411, 0.439, and 0.886 respectively.

They have almost a linear association with the dependent variable ERP Failure (FA) in respect to system testing.

Table -3: correlation statistics of system testing

Correlations		TOTALST	RD	RTP	ITP	RDTP	TM	TL	TT
Pearson Correlation	TOTALST	1.000	.632	.383	.217	.411	.439	.268	.886
	RD	.632	1.000	.081	.117	.094	.062	.116	.443
	RTP	.383	.081	1.000	-.078	.240	.404	.455	.129
	ITP	.217	.117	-.078	1.000	.028	-.286	-.282	.236
	RDTP	.411	.094	.240	.028	1.000	-.117	-.086	.299
	TM	.439	.062	.404	-.286	-.117	1.000	.599	.274
	TT	.886	.443	.129	.236	.299	.274	.010	1.000
Sig. (1-tailed)	TOTALST	.	.000	.000	.032	.000	.000	.011	.000
	RD	.000	.	.248	.163	.215	.300	.164	.000
	RTP	.000	.248	.	.257	.020	.000	.000	.138
	ITP	.032	.163	.257	.	.407	.007	.008	.022
	RDTP	.000	.215	.020	.407	.	.163	.236	.005
	TM	.000	.300	.000	.007	.163	.	.000	.009
	TT	.000	.000	.138	.022	.005	.009	.465	.

The correlation coefficient for Required Documents (RD) and Testing Types (TT), r= 0.443, which is more than 0.3 Thus, there is an association between Required Documents (RD) and Testing Types (TT), in System Testing.

The correlation coefficient for Requirements for Test Procedure (RTP) and Testing Methods (TM), r= 0.404,which is more than 0.3 Thus, there is an association between Test Procedure (RTP) and Testing Methods (TM).

8.4. Regression Analysis on Factors of System Testing for ERP Failure

The Regression analysis on Factors of System Testing for ERP Failure is show according to the given table 4 below;

Table -4 : Regression analysis on Factors of System Testing for ERP Failure

Model Summary[b]					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.995[a]	.990	.989	.02780	1.688

a. Predictors: (Constant), TT, RDTP, ITP, RD, RTP, TM
b. Dependent Variable: TOTALST

According to the above table 4.42, R^2 = 99%of the variation in ERP failure is explained by System Testing to reduce failure rate.

Table -5: ANOVA analysis on Factors of System Testing for ERP Failure

ANOVA[a]						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.844	7	.692	895.143	.000[b]
	Residual	.050	65	.001		
	Total	4.894	72			

a. Dependent Variable: TOTALST
b. Predictors: (Constant), TT, RDTP, ITP, RD, RTP, TM

The p-value from the ANOVA table 5, is less than 0.001 which is 0.000. It means System Testing can be used to predict ERP failure. In the above ANOVA table, the large F-value (895.143), indicated by a small p-value (<0.05) which is 0.000, implies good fit. It means that the least one of the independent variables can explain the outcome.

Table -6: Coefficient analysis on Factors of System Testing for ERP Failure

Coefficients[a]									
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
	B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1 (Constant)	.081	.054		1.497	.139	-.027	.190		
RD	.139	.007	.304	21.107	.000	.126	.153	.759	1.318
RTP	.038	.006	.105	6.863	.000	.027	.049	.669	1.495
ITP	.061	.006	.140	9.863	.000	.048	.073	.780	1.282
RDTP	.157	.011	.216	14.823	.000	.136	.179	.746	1.340
TM	.101	.008	.222	11.898	.000	.084	.118	.455	2.196
TT	.345	.010	.578	33.952	.000	.325	.365	.544	1.838

a. Dependent Variable: TOTALST

The correlation coefficient for items in each factors are as shown in the above table and Required Documents (RD), Requirements for Test Procedure (RTP), Instruction for Test Procedures (ITP), Require Document for Test Preparation (RDTP), Testing Methods (TM), Testing Types (TT) represent p-values as 0.000 respectively, which are less than 0.05 Thus, there are significant predictors of ERP failures among these items in each factor when considering system testing to reduce failure rate as shown in the table 6.

The Equation: FAST = 0.081 +0.139 (RD) +0.038 (RTP) +0.061 (ITP) +0.157 (RDTP) + 0.101 (TM) + 0.345 (TT)

The Equation: FAST = 0.081 +0.139 (Required Documents) +0.038 (Requirements for Test Procedure) +0.061 (Instruction for Test Procedures) +0.157 (Require Document for Test Preparation) + 0.101 (Testing Methods) + 0.345 (Testing Types)

Thus for every unit increase in Required Documents (RD), System Testing for ERP Failure is expected to drop by 0.139. Every unit increase in Requirements for Test Procedure (RTP), System Testing for ERP Failure is expected to drop by 0.038 and for every unit increase in Instruction for Test Procedures (ITP), System Testing for ERP Failure is expected to drop by 0.061 and also for every unit increase in Require Document for Test Preparation (RDTP), System Testing for ERP Failure is expected to drop by 0.157. Furthermore, every unit increase in Testing Methods (TM), System Testing for ERP Failure is expected to drop by 0.055 and also for every unit increase in Testing Types (TT), System Testing for ERP Failure is expected to drop by 0.345.

The 95% Confidence Interval (CI) for Required Documents (RD) is [0.126, 0.153], Where the value of 0 does not fall within the interval, again indicating Required Documents (RD) is a significant predictor.

The 95% Confidence Interval (CI) for Requirements for Test Procedure (RTP) is [0.027, 0.049]. Where the value of 0 does not fall within the interval, again indicating Requirements for Test Procedure (RTP) is a significant predictor.

The 95% Confidence Interval (CI) for Instruction for Test Procedures (ITP) is [0.048, 0.073]. Where the value of 0 does not fall within the interval, again indicating Instruction for Test Procedures (ITP) is a significant predictor.

The 95% Confidence Interval (CI) for Require Document for Test Preparation (RDTP) is [0.136, 0.179]. Where the value of 0 does not fall within the interval, again indicating Require Document for Test Preparation (RDTP) is a significant predictor.

The 95% Confidence Interval (CI) for Testing Methods (TM) is [0.084, 0.118]. Where the value of 0 does not fall within the interval, again indicating Study of Testing Methods (TM) is a significant predictor.

The 95% Confidence Interval (CI) for Testing Types (TT) is [0.325, 0.365]. Where the value of 0 does not fall within the interval, again indicating Testing Types (TT) is a significant predictor.

8.5. Mulicollinearity for System Testing

Required Documents (RD), Requirements for Test Procedure (RTP), Instruction for Test Procedures (ITP), Require Document for Test Preparation (RDTP), Testing Methods (TM), Testing Types (TT) are respectively 1.318, 1.495, 1.282, 1.340, 2.196, 1.890 and 1.838 and as they are below 5 the multicollinearity is not serious. Hence there is no problem of multicollinearity.

9. The Mediating effect of Testing Methods between System Testing and ERP Failure (H1)

The illustration of mediating effect of testing methods between system testing and ERP failure is given bellow (figure 1);

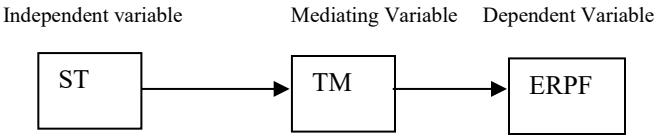


Figure -1: Mediating effect of Testing Methods between System Testing and ERP Failure

The Sobel test is the most appropriate test to measure the mediating effect of a variable on the relationship between the independent variable and dependent variable. According to the result given by the Sobel statistical, results of the Sobel test is 7.55930821, significant p-value is 0.000 which is less than 0.05. It implies that the Testing Methods (TM) is a significant mediator between System Testing (ST) and ERP Failure (ERPF)as shown in the below figure 2.

– a=.962sa =.234

– b=.007sb =.046

Standard error of ab is approximately square root of $b^2s_a^2 + a^2s_b^2$

(Sobel test)–ab =$\sqrt{(.962^2 * .046^2 + .007^2 * 0.234^2)}$= .04421

P<0.001

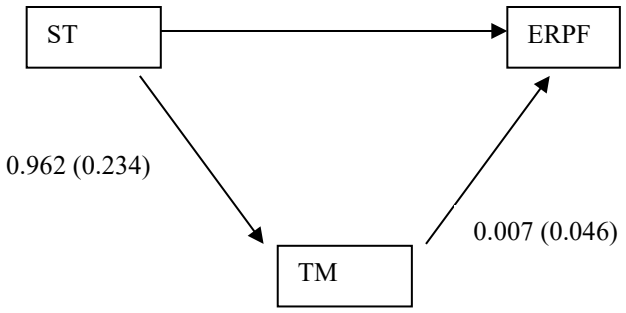


Figure -2: Mediating coefficient values for Testing Methods between System Testing and ERP Failure

Thus, Testing Methods (TM) mediate the relationship. Hence H1 can be accepted.

10. Full Path Analysis

The full path analysis is as follows for the conceptual framework.

FA=FABA+FASD+FASI+FAST+FATM+FABT+FAPT+FAPL +FASDM

FA= -0.121 +0.152 (CSI)+0.179(SES) -0.168(INBP)+0.144(SSS)-0.082(DNC) +0.142 (CI)+0.159 (IC)+0.000 +0.097 (ADA) +0.111(SUD) + 0.147(BUD) + 0.129(DD) + 0.101(DDBS) + 0.208 (UID) + 0.081 (CA)+0.000 +0.146 (ALC) +0.208 (BESI) +0.146 (POPE) +0.083 (VD) + 0.083 (MO) + 0.063 (SA) + 0.083 (CF)+ 0.188 (CS)+0.081 +0.139 (RD) +0.038 (RTP) +0.061 (ITP) +0.157 (RDTP) + 0.101 (TM) + 0.345 (TT) + 0.442(FATM) + 0.126(FABT) + 0.089(FAPT) + 0.000(FAPL) + 0.067(FASDM)

11. Conclusion

The p-value from the ANOVA table is less than 0.001 which is 0.000. It means System Testing can be used to predict ERP failure. In the above ANOVA table, the large F-value (895.143), indicated by a small p-value (<0.05) which is 0.000, implies good fit. It means that the least one of the independent variables can explain the outcome.

The correlation coefficient for items in each factors are as shown in the above table and Required Documents (RD), Requirements for Test Procedure (RTP), Instruction for Test Procedures (ITP), Require Document for Test Preparation (RDTP), Testing Methods (TM), Testing Types (TT) represent p-values as 0.000 respectively, which are less than 0.05 Thus, there are significant predictors of ERP failures among these items in each factor when considering system testing to reduce failure rate.

Thus for every unit increase in Required Documents (RD), System Testing for ERP Failure is expected to drop by 0.139. Every unit increase in Requirements for Test Procedure (RTP), System Testing for ERP Failure is expected to drop by 0.038 and for every unit increase in Instruction for Test Procedures (ITP), System Testing for ERP Failure is expected to drop by 0.061 and also for every unit increase in Require Document for Test

Preparation (RDTP), System Testing for ERP Failure is expected to drop by 0.157. Furthermore, every unit increase in Testing Methods (TM), System Testing for ERP Failure is expected to drop by 0.055 and also for every unit increase in Testing Types (TT), and System Testing for ERP Failure is expected to drop by 0.345.

11.1. 95% Confidence Interval (CI) for System Testing

According to the regression Analysis the coefficient table says that 95% CI where the value 0 does not fall within the interval indicates it as a significant predictor. The following are the significant predictors that support the H1 hypotheses as shown in the below table 7.

Table -1 : 95% Confidence Interval (CI) for System Testing

	Cluster name	95% CI Values	Significant or Not
01	Required Documents (RD)	[0.126, 0.153]	Significant predictor
02	Requirements for Test Procedure (RTP)	[0.027, 0.049]	Significant predictor
03	Instruction for Test Procedures (ITP)	[0.048, 0.073]	Significant predictor
04	Require Document for Test Preparation (RDTP)	[0.136, 0.179]	Significant predictor.
05	Testing Methods (TM)	[0.084, 0.118]	Significant predictor
06	Testing Types (TT)	[0.325, 0.365]	Significant predictor

All seven clusters are significant and therefore, there is a negative relationship between System Implementation and ERP Failure.

H1 There is a negative relationship between System Testing and ERP Failure -Accepted

H2 Many Testing Methods do mediates the positive relationship between System Testing and ERP Failure -Accepted

In accordance with the above discussion the above summary gives a clear picture. Based on the hypotheses the researcher has tested in order to reduce the ERP failure rate under the software testing phase. The findings in this study considered the H1 and H2 acceptable and aligned with software testing process in order to reduce the ERP failure rate.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

I would like to express my sincere thanks and heartfelt gratitude to my supervisor Prof. (Dr.) Siong Choy Chong for his invaluable advice and guidance throughout the course of my work. It has been a great privilege to have worked with such an outstanding person, and his ever willing help and cooperation contributed towards the smooth progress of my work. I also extend my gratitude to Prof. (Dr.) Ali Katibi, Dean, Faculty of Graduate Studies, University of Management and Science, Malaysia for his advice and guidance throughout the research, as without his invaluable support and assistance given to me in continuing this study, his cooperation and tolerance, my endeavour would not have become a reality.

References

[1] AL-Hossan A., AL-Mudimigh S., & Abdullah, "Practical Guidelines for Sucessful ERP Testing, **27:** 11 – 18 (2011).
[2] Galin, D.Software Quality Assurance. Harlow: Pearson Education Limited (2004).
[3] Lewis, W. E. Software Testing and Continuous Quality Improvement. New York: Auerbeach Publications, (2005).
[4] Srinivasan Desikan. Software Testing. Chennai: Dorling Kindersley Pvt. Ltd (2011).
[5] Chris P. Learn to Program, (2nd ed.). Pragmatic Bookshelf (2009).
[6] Jae H. Y. ,Jeong S.K.& Hong S.P. Test Data Combination Strategy for Effective Test Suite Generation. IT Convergence and Security (ICITCS), 2013 International Conference, 1-4 (2013).
[7] Andrews J., Briand L., Labiche Y. Using Mutation Analysis for Assessing and Comparing Testing Coverage Criteria. Software Engineering, IEEE Transactions, **32**: 608 – 624 (2006).
[8] Tai K. Condition-based software testing strategies. Computer Software and Applications Conference, 1990. COMPSAC 90, 564 – 569 (1990).
[9] Gizopoulos D., Paschalis A. Effective software-based self-test strategies for on-line periodic testing of embedded processors. Computer-Aided Design of Integrated Circuits and Systems, **24**: 88 – 99 (2004).
[10] Basili V., Selby R., Comparing the Effectiveness of Software Testing Strategies. Software Engineering, **13**: 1278 – 1296 (2006).
[11] Junpeng L, Beihang U., Kai-yuan C., On the Asymptotic Behavior of Adaptive Testing Strategy for Software Reliability Assessment. Software Engineering, **40**: 396 – 412 (2014).
[12] Yachin D. SAP ERP Life–CycleManagement Overcoming the Downside of Upgrading. IDC Information and Data, (2009).
[13] Fraser G., Arcuri A. The Seed is Strong: Seeding Strategies in Search-Based Software Testing. Software Testing, **5**: 121 – 130 (2012).
[14] Howden W. Reliability of the Path Analysis Testing Strategy. Software Engineering, **2**(3): 208 – 215 (2006).
[15] Chaim M., Maldonado J., Jino M. A debugging strategy based on requirements of testing. Software Maintenance and Reengineering, **17**, 160 – 169 (2003).
[16] Prather, Ronald E., Myers J. The Path Prefix Software Testing Strategy. Software Engineering, **13**: 761 – 766 (2006).
[17] William E., Lewis E., Gunasekaran W. Software Testing and Continuous Quality Improvement, **2**: (2005).
[18] Cem Kaner C., Jack falk J. , Hung H. Q.N. Testing Computer Software. New Yor: John Wiley & Son,2nd ed., k, USA. (1999).
[19] Robert, R., Futrell, T., Donald, F., Shafer, F., & Linda L. Quality Software Project Management (2002).
[20] Andrei A., Daniel I.V. Automating Abstract Syntax Tree construction for. 14th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (2012)
[21] Demilli R., Offutt A., Constraint -based automatic test data generation. Software Engineering, **17:** 900-910 (2002).
[22] Kuhn D., Wallace D, Gallo A. Software fault interactions and implications for software testing. Software Engineering, **30**(6): 418 – 421 (2013).
[23] Wallace D, Kuhn D. & Gallo A. Software fault interactions and implications for software testing. Software Engineering, **30**(6): 418 - 421. (2013).
[24] Chan F., Chen T., Mak I. Proportional sampling strategy: guidelines for software testing practitioners. Information and Software Technology, **28**: 775–782 (1996).
[25] Satyanarayana, C., Niranjan, V., and Vidyadara, C. Method, process and technique for testing erp solutions. U.S. Patent Application, (2012).
[26] Dennis, D., Shahin, S., & Zadeh, J. Software Testing and Quality Assurance, **4**: 4 (2008).
[27] Software Testing Techniques, Hariprasa, P.,Hariprasa, P, (2015).
[28] Laurie Williams, Testing Overview (2006).
[29] Jovanovi. Software Testing Methods and Techniques. The IPSI BgD Transactions on Internet Research, **30** (2014).

[30] Wikipedia. Software Testing. Retrieved from wikipedia.org: http://www.wikipedia.org/wiki/software_testing (2014).
[31] Anthony A., John J. Acceptance Testing. Encyclopedia of Software Engineering, 1st edition (2002).
[32] Elfriede D., Jeff R., John P. Automated Software testing, (2008).
[33] Ahimbisibwe, A., Cavana, R. Y., & Daellenbach, U. A contingency fit model of critical success factors for software development projects: A comparison of agile and traditional plan-based methodologies. Journal of Enterprise Information Management, **28**(1): 7-33 (2015).
[34] Thummalapenta S., Sinha S., Mukherjee D., Chandra S. Automating Test Automation, (2011).
[35] Haugset B., Hanssen G. Automated Acceptance Testing. A Literature Review and an Industrial Case Study, 27-38 (2008).
[36] Rankin C. The Software Testing Automation Framework. IBM Systems Journal, **41**: 129-139 (2002).
[37] Karhu K., Repo T., Taipale O., Smolander K. Empirical Observations on Software Testing Automation. Software Testing Verification and Validation, **9**: 201-209 (2009).
[38] Bansal A., Muli M., Patil K. Taming complexity while gaining efficiency: Requirements for the next generation of test automation tools. AUTOTESTCON, 1 – 6 (2013).
[39] Borjesson E., Feldt R. Automated System Testing Using Visual GUI Testing Tools: A Comparative Study in Industry. Software Testing, **5**: 350 – 359 (2012).
[40] Guide, G. C. Best practices for developing and managing capital Program Costs **GAO-09-22**: (2009).
[41] Jain, A., Jain, M., & Dhankar, S. A Comparison of RANOREX and QTP Automated Testing Tools and their impact on Software Testing. IJEMS, **1**(1): 8-12 (2014).
[42] Creswell, J. "Research Design: Qualitative and Quantitative Approaches. Sage Publications, (1994)
[43] Sekaran & Bougie. Research Methods for Business A Skill Building Approach. (5th, Ed.) John Wiley & Sons Ltd, (2009)
[44] Kristopher J. P. Calculation for the Sobel Test. Retrieved June 10, 2015, from http://quantpsy.org/sobel/sobel.html (2010).
[45] Gerrard, P. Test Methods and Tools for ERP Implementations. PO Box 347, Maidenhead, Berkshire, SL6 2GU, UK: Gerrard Consulting, (2015).
[46] Central Bank, O. S. Annual Report 2014. Central Bank of Sri Lanka, (2014).
[47] Suskie,. A. "Questionnaire Survey Research: What Works". Assn for Institutional Research, (1996).
[48] Board of Investment. BOI. Retrieved from http://www.investsrilanka.com/: http://www.investsrilanka.com, (2012).
[49] Colombo Stock Exchange. CSE. Retrieved from /www.cse.lk: https://www.cse.lk (2012).
[50] Karuthan. Statistical Analysis using SPSS. Lecture Notes. Management and Science University. (2014).

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 16-22 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

The Class Imbalance Problem in the Machine Learning Based Detection of Vandalism in Wikipedia across Languages

Arsim Susuri*, Mentor Hamiti, Agni Dika

Faculty of Contemporary Sciences and Technologies, South East European University, Tetovo, Macedonia

A R T I C L E I N F O

Article history:
Received: 30 November, 2016
Accepted: 14 January, 2017
Online: 28 January, 2017

Keywords:
Machine learning
Wikipedia
Vandalism
Class imbalance

A B S T R A C T

This paper analyses the impact of current trend in applying machine learning in detection of vandalism, with the specific aim of analyzing the impact of the class imbalance in Wikipedia articles. The class imbalance problem has the effect that almost all the examples are labelled as one class (legitimate editing); while far fewer examples are labelled as the other class, usually the more important class (vandalism). The obtained results show that resampling strategies: Random Under Sampling (RUS) and Synthetic Minority Oversampling Technique (SMOTE) have a partial effect on the improvement of the classification performance of all tested classifiers, excluding Random Forest, on both tested languages (simple English and Albanian) of the Wikipedia. The results from experimentation extended on two different languages show that they are comparable to the existing work.

1. Introduction

Ever since its inception, in 2001, Wikipedia has continuously grown to become one the largest information source on the Internet. One of its unique features is that it offers the ability to anyone to edit the articles. This popularity, in itself, means that, a number of articles can be read, edited, and enhanced by different editors and, inevitably, be subject to acts of vandalisms through illegitimate editing.

This paper is an extension of work originally presented in [1], by addressing the issue of class imbalance in the detection of vandalism in Wikipedia articles across languages.

Vandalism means any type of editing which damages the reputation of an article or a user in Wikipedia. A list of typical vandalisms along with their chances of appearance was created as a result of empirical studies done by Priedhorsky et al. [2]. Typical examples include massive deletions, spam, partial deletions, offences and misinformation. In order to deal with vandalism, Wikipedia relies on the following users:

- Wikipedia users' ability and willingness to find (accidentally or deliberately) damaged articles
- Wikipedia administrators and
- Wikipedia users with additional privileges

These users use special tools (e.g. Vandal Fighters) to monitor recent changes and modifications that enable retrieval of bad expressions or which are implemented by blacklisted users.

Wikipedia was subject to different statistical analysis from various authors. Viégas et al. [3] uses visualization tools to analyze the history of Wikipedia articles. When it comes to vandalism, authors were able to identify (manually) massive deletions as a jump in the history flow of a particular article page. Since late 2006, some bots (computer programs designed to detect and revert vandalism), have appeared on Wikipedia. These tools are built on the primitive included in the Vandal Fighters. These use lists of common phrases, and consult databases containing blocked users or IP addresses in order to separate legitimate editing from vandalism.

One drawback of these approaches is emphasized that these world use static list of obscenities and grammatical rules which are difficult to maintain and easily "fooled". These detect only 30% of vandalisms committed. Consequently, there is a need to improve the detection of this kind. One of the possible improvements is the application of machine learning.

The prior success implemented in interference detection, spam filtering for email, etc., is a good indicator for the opportunity that the machine learning shows in improvements in detecting vandalism in Wikipedia [4].

*Corresponding Author: Arsim Susuri, Faculty of Contemporary Sciences and Technologies, South East European University, Tetovo, Macedonia
Email: arsimsusuri@gmail.com

2. Wikipedia Vandalism Detection

To define the vandalism detection task, we have to define some key concepts of MediaWiki (the wiki engine used by Wikipedia).

An article is composed of a sequence of revisions, commonly referred to as the article history. A revision is the state of an article at a given time in its history and is composed of the textual content and metadata describing the transition from the previous revision [5]. Revision metadata contains, among others, the user who performed the edit, a comment explaining the changes, a timestamp, etc. An edit is a tuple of two consecutive revisions and should be interpreted as the transition from a given revision to the next one.

Evaluating a vandalism detection system requires a corpus of pre-classified edits. Our focus is on four different corpora:

- PAN-WVC-10 - The PAN Wikipedia Vandalism Corpus 2010 (PAN-WVC-10), compiled in 2010 via Amazon's Mechanical Turk comprises 32439 edits from 28468 English Wikipedia articles of which 2394 have been annotated as vandalism. The dataset was created by 753 human annotators by casting 193022 votes, so that each edit has been annotated at least three times, whereas edits that were difficult to be annotated received more than three votes [6]. The PAN-WVC-10 was first used in the 1st International Competition on Wikipedia Vandalism Detection [7].
- PAN-WVC-11 - The PAN Wikipedia Vandalism Corpus 2011 (PAN-WVC-11) from 2011 is an extension of the PAN-WVC-10. It was used in the 2nd International Competition on Wikipedia Vandalism Detection [8]) and is the first multilingual vandalism detection corpus. The corpus comprises 29949 Wikipedia edits in total (9985 English edits with 1144 vandalisms, 9990 German edits with 589 vandalisms, and 9974 Spanish edits with 1081 vandalism annotations).
- Wikipedia History Dump - Wikipedia records all revisions of all articles and all other Wikipedia pages and releases them as XML or SQL dump files. For this research dumps 29.10.2015 have been used for:
    - simplewiki and
    - sqwiki.

2.1. Approaches based on Machine Learning

Since 2008 Wikipedia vandalism detection based on machine learning approaches has become a field of increasing research interest. In table 1 existing vandalism detection approaches from the literature are shown, depicting Precision, Recall and PR-AUC (Precision Recall – Area Under Curve) values.

Potthast et al. [9] contributed the first machine learning vandalism detection approach using textual features as well as basic metadata features with a logistic regression classifier. Smets et al. [10] used a Naive Bayes classifier on a bag of words edit representation and were the first to use compression models to detect Wikipedia vandalism. Itakura and Clarke [11] used Dynamic Markov Compression to detect vandalism edits on Wikipedia.

Mola Velasco [12] extended the approach of Potthast et al. [9] by adding some additional textual features and multiple wordlist-based features. He was the winner of the 1st International Competition on Wikipedia Vandalism Detection [7]. West et al. [13] were among the first to present a vandalism detection approach solely based on spatial and temporal metadata, without the need to inspect article or revision texts.

Similarly, a vandalism detection system on top of their WikiTrust reputation system was built by Adler et al. [14 and 15]. Adler et al. [16] combined natural language, spatial, temporal and reputation features used in their aforementioned works [12, 13 and 14].

Besides Adler et al. [16], West and Lee [17] were the first to introduce ex post facto data as features, for whose calculation also future revisions have to be considered.

Supporting the current trend of creating cross language vandalism classifiers, Tran and Christen [18] evaluated multiple classifiers based on a set of language independent features that were compiled from the hourly article view counts and Wikipedia's complete edit history.

3. Objectives

The objectives of this research work were to experimentally compare four classifiers on unbalanced data with and without resampling on four different corpora: PAN-WVC-10, PANWVC-11, Simple English Wikipedia (simplewiki) and Albanian Wikipedia (sqwiki) history dumps, respectively.

We compare four different classifiers Logistic Regression, RealAdaBoost, BayesNet, and Random Forest regarding their performances using RUS and SMOTE.

Based on this experiment we try to build a model that would be able to represent the impact of class imbalance on the detection of vandalism across languages, for small scaled datasets.

3.1. The Class Imbalance Problem

The problem of used vandalism corpora (Webis-WVC-07, PAN-WVC-10 and PAN-WVC-11) is that they offer data that are highly skewed. What this means is that the ratio between the number of vandalism and regular edits is highly imbalanced (5-7% of all samples are annotated as vandalism edits).

Learning traditional classifiers with such datasets can cause lower detection performance. Based on surveys of classifying imbalanced data by He and Garcia [20] and Ganganwar [21] we can list three reasons for performance decline:

- If classifier learning is based on minimizing the overall error, then the minority class instances contribute little to the error. This results in an increase of bias of the classifier towards the majority class.
- Many classifiers assume a balanced class distribution of the minority and the majority class, which is not often the case when working with realistic scenarios.
- Often classifiers implicitly assume equal costs for misclassification for both classes, which is often not sensible: for example, the cost for classifying cancer as non-cancer is way higher than the other way round.

Table 1: Vandalism detection classification obtained from various authors

Authors	Data	Classifier	Precision	Recall	PR-AUC	Corpora
Smets et al. [10]	Unbalanced	Probabilistic Sequence Modeling	0.3209	0.9171	-	Simplewiki
Smets et al. [10]	Unbalanced	Naive Bayes	0.4181	0.5667	-	Simplewiki
Tran and Christen [18]	Balanced	Gradient Tree Boosting	0.870	0.870	-	Historical Dump
Potthast et al. [9]	Unbalanced	Logistic Regression	0.830	0.870	-	Webis-WVC-07
Velasco [5]	Unbalanced	Random Forest	0.860	0.570	0.660	PAN-WVC-10
Adler et al. [14]	Unbalanced	ADTree	0.370	0.770	0.490	PAN-WVC-10
Adler et al. [16]	Unbalanced	Random Forest	-	-	0.820	PAN-WVC-10
West and Lee [17]	Unbalanced	ADTree	0.370	0.770	0.490	PAN-WVC-10
Harpalani et al. [19]	Unbalanced	LogitBoost	0.606	0.608	0.671	PAN-WVC-10
West and Lee [17]	Unbalanced	ADTRee	-	-	0.820	PAN-WVC-11

In general, there are two approaches to overcome the class imbalance problem:

- The data level that involves several training data resampling techniques.
- The algorithmic level that involves adjusting the misclassification costs or probabilistic estimates, e.g., at the tree leaves of decision tree classifiers, as well as learning classifier models solely based on minority class samples (so-called one-class classification).

During the examination of the impact of training dataset resampling on vandalism detection performance we find that, in most cases, resampling reduces the performance of the tested classifiers. Logistic Regression, Real Ada Boost and Bayesian Network classifiers benefit from certain resampling strategies, whereas a Random Forest classifier turns out to be relatively unaffected by resampling approaches.

3.2. Evaluating Resampling Techniques

One approach to overcome performance issues of classifiers is resampling the training dataset in order to balance the classes. There are several common approaches to do so, namely random under sampling, random oversampling, directed over and under sampling and hybrid methods which combine the aforementioned [20].

3.2.1. Resampling Strategies

Random under sampling (RUS) removes a certain amount of randomly picked majority class instances from the training dataset. RUS leads to class balancing or, in an extreme case, even to majority class removal. However, a disadvantage of RUS is the loss of possibly decisive instances. Since important information for the class separation is likely to be removed, this technique might induce a lower classification performance.

Random over sampling (ROS) reproduces a certain amount of randomly chosen minority class samples. Thus, the class distribution can be adjusted towards a uniform distribution. Since classifiers, after oversampling, are trained by using some minority class values multiple times, the learned model is likely to over fit.

The Synthetic Minority Oversampling TEchnique (SMOTE) by Chawla et al. [22] over samples the minority class by computing artificial instances. The feature values of these samples are calculated by random interpolation of the K-nearest neighbors' feature values (typically K = 5). The method aims at avoiding over fitting while oversampling minority class instances. Han et al. [23] extend SMOTE to use only the minority class samples at the class borderline (borderlineSMOTE) in order to generate artificial data which is more important for classification.

3.2.2. The Classifiers

Our focus is on Logistic Regression and Random Forest, since they are used by Potthast et al. [9], Mola Velasco [12], and Adler et al. [16]. Additionally, we consider RealAdaBoost as a state-of-the-art Boosting algorithm and a Bayesian Network classifier as a Bayes approach that is reported to outperform the Naive Bayes classifier used by Smets et al. [10].

Logistic Regression analysis estimates the relationship between a dependent variable and one or more independent variables.

RealAdaBoost (Friedman et al. [24]) is a boosting algorithm based on Adaptive Boosting (AdaBoost) by Freund and Schapire [25]. Boosting is a method to enhance classification performance by combining many weak base classifiers (weak hypotheses) in order to create a more powerful classifier.

A Bayesian Network (Pearl and Russell [26]) is a directed acyclic graph. The nodes in the graph represent random variables; the arcs signify direct correlations between these variables. Tree Augmented Naive Bayes (TAN) described by Friedman et al. [27] has been used in the experiments. In our experiment, each attribute in the graph has only the class value and at most one other attribute as parents.

Random Forest Random Forest (Breiman [28]) is an ensemble learning technique, constructing a defined number of decision tree predictors and combining them to a predictor set (forest). The individual trees are learned from randomly chosen feature subsets and represent independent and identically distributed random vectors. Each tree is grown to full depth.

To classify a new data sample, the final class is determined by the mode of classes that are predicted by the individual trees.

3.2.3. Datasets

We use two datasets: the complete edit history of Wikipedia in Simple English and Albanian[1], and the hourly article view count[2].

The edit history data dump is that of 1 December 2015 for both, the Simple English Wikipedia and for Albanian Wikipedia [1]. In figure 1 the number of articles and edits revisions (per month and per year) are shown.

3.2.3.1. Labeling Vandalized Revisions

From the raw revision data, every revision is reduced to a vector of features described in table 2. The reasons for selecting these features are independence from language and simplicity.

Every revision's comment is scrutinized for keywords of "vandal" and "rvv" (revert due to vandalism), which would signal a vandalism in the previous revision. Afterwards, these revisions are marked as vandalism.

In order to correctly arrange the timestamp of revisions with the corresponding article view dataset, we round up the revision time to the next hour. This ensures that the hourly article views reference the correct revision when combining the two datasets.

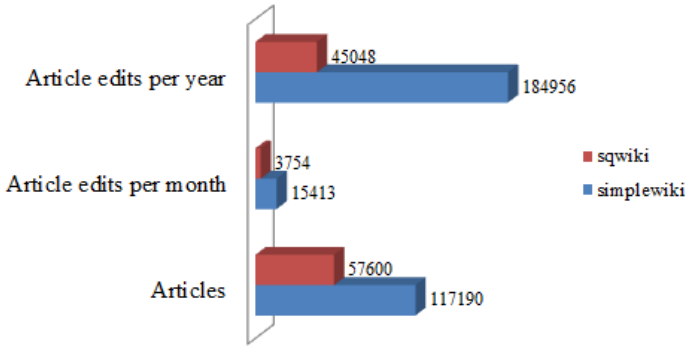


Figure 1: Statistical data of editing history for simplewiki and sqwiki

The arrangement is implemented on all revisions and should not affect classification.

1.1.1.1. Article Views

The raw article view dataset is structured by views of article aggregated by hour. We use the process of applying the transformation and filtering of articles viewed in the revision dataset above (table 3), as used in [1].

Extraction of the redirect articles from the revision dataset is applied and then all access to redirect articles is changed to the canonical article. These extra view counts are aggregated accordingly. These article views are important to seeing the impact of vandalism on Wikipedia [2]. Having in mind that the average time vandalism is active is 2.1 days [5], a lot hours are for unsuspecting readers to face vandalized content.

Table 2: Description of editing history dataset

Attribute	Description
Title of Article	Unique identifier of a Wikipedia article.
Hour Timestamp	The timestamp of the revision.
Anonymous Edit	The editor of this revision is considered to be anonymous if an IP address is given. 0 for an edit by a registered user, and 1 for an edit by an anonymous user.
Vandalism	This revision is marked as vandalism by analyzing the comment of the following revision. 0 for regular edit, and 1 for vandalism.

However, the behavior of vandals may also be seen in a change in access patterns, which may be from vandals checking on their work, or that article drawing attention from readers and their peers [29].

Table 3: Description of article view dataset

Attribute	Description
Name of Project	MediaWiki project, (Simple English - "simplewiki" and Albanian -"sqwiki").
Hour Timestamp	In the format of DDMMYYYY-HH0000.
Title of Article	The title of the Wikipedia article.
Number of Requests	The number of requests during that hour.

The edit history dataset is scanned in order to be sure that these article views occurred when articles are in a vandalized state. Then we apply labelling of all article views of observed vandalized or non-vandalized revisions.

The unknown views from revisions made before 2015, or articles without revisions in this 4-month period under study, are discarded. Thus, we have an article view dataset labelled with whether the views are of vandalized revisions.

The resulting size of the data is identical to the resulting dataset in the following subsection. This labeled article view dataset allows us to determine whether view patterns can be used to predict vandalism.

From this resulting dataset, we split the "Hour Timestamp" attribute into an "hour" attribute. This allows the machine learning algorithm to learn daily access patterns.

1.1.1.2. Resulting Dataset

The resulting dataset is created by merging two-time series datasets for each language. The dataset is constructed by adding

[1] http://dumps.wikimedia.org/backup-index.html

[2] http://dumps.wikimedia.org/other/pagecounts-raw/

features from the labeled revision dataset to the labeled article view dataset by repeating features of the revisions. Thus, for every article view, we have information on what the properties are and whether this revision was viewed.

We use the "hour" attribute split from the timestamp in the article views dataset. Thus, we have the following 6 features in our resulting dataset: hour, size of the comment, size of article, anonymous edit, number of requests, and vandalism.

These features are language independent and capture the metadata of revisions commonly used, and access patterns. Article name is not included in the resulting dataset because access patterns of vandalized articles may be similar to other vandalized articles, regardless of the name of the article.

To apply the classification algorithms, we split the resulting dataset by date into a training set (September to November) and a testing set (December), as shown in figure 2.

2. Experiments and Results

For SMOTE oversampling we use an amount of 50% and 100% of the original vandalism class instances ($SMOTE_{50}$ and $SMOTE_{100}$).

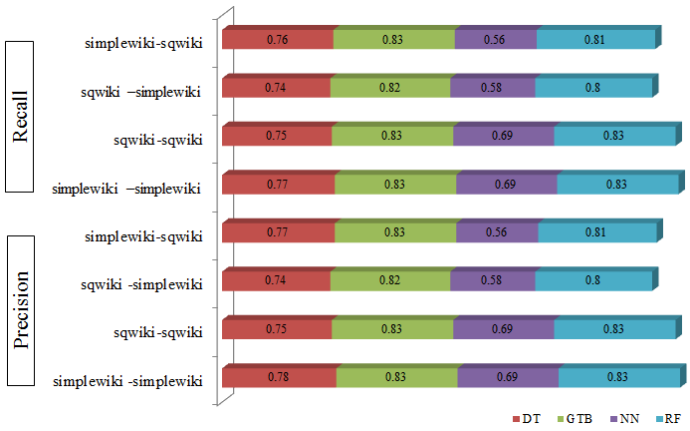


Figure 2: Revisions Dataset – Classifications results

Additionally, we chose an oversampling amount of 1300% which leads to an almost balanced dataset without loss of any majority class instances - 13613 (48.81%) vandalism, 14281 (51.19%) regular.

On PAN-WVC-11 we use a SMOTE oversampling of 1100% instead of 1300%, due to the different class distribution in that corpus (1100% oversampling leads to 28728 (48.88%) vandalism and 30045 (51.12%) regular samples). Table 5 provides the corresponding PR-AUC values.

Using RUS on the training data, all classifiers but RealAdaBoost on PAN-WVC-11 show a performance drop on all four corpora.

For Logistic Regression (on both corpora) and Random Forest (on PAN-WVC-10) RUS leads to the lowest overall performance.

If a classifier already handles class imbalance internally, RUS only removes majority class data that is needed to train the model without benefiting from a balanced data set. For Real Ada-Boost the loss of regular samples seems not to be as influential as the training on a balanced dataset.

SMOTE Logistic Regression benefits from $SMOTE_{50}$ (on PAN-WVC-10) and from $SMOTE_{100}$ (on PAN-WVC-11). Both oversampling strategies result in the best overall performances on the respective corpora. Similar results have also been obtained on simplewiki and sqwiki.

SMOTE oversampling leads to a high performance drop for the RealAdaBoost classifier. Over-sampling the target class with SMOTE causes a slight decrease of performance if the Random Forest classifier is used on both corpora. The performance for BayesNet increases for lower oversampling proportions (50% and 100%) on PAN-WVC-11, $SMOTE_{50}$ even leads to the highest overall performance. On PAN-WVC-10 all SMOTE proportions result in a performance drop for the BayesNet classifier.

On PAN-WVC-10 for Logistic Regression and Random Forest, a higher oversampling proportion leads to lower performance. This is also the case for all classifiers but Logistic Regression on PAN-WVC-11. For all classifiers on both corpora $SMOTE_{1100/1300}$ lead to the lowest classification performance using SMOTE approaches. An exception is the RealAdaBoost classifier (on PAN-WVC-11), for which $SMOTE_{1100}$ outperforms the other proportions.

A reason for the observed lower performance using SMOTE might be the absence of significant data in the training and test corpora. If the vandalism samples given in the test dataset represent other vandalism types than those given in the training set, some kinds of vandalism will never be found.

Wikipedia vandalism has been found to be a heterogeneous problem [30]. Hence, an underrepresentation of vandalism edits from certain categories in the training corpora would not be surprising, since the samples have been chosen randomly ([6], and [8]). In the case of missing decisive vandalism samples, oversampling would not produce a more accurately defined vandalism class region, but would only insert further weak samples.

3. Conclusions and Future Work

Summarizing our experiments, we can conclude that RealAdaBoost is most affected by the imbalance of the training data. Random Forest shows only little sensitivity to resampling approaches. However, it turns out to be the best performing classifier of all evaluated approaches, without applying resampling strategies, as shown in table 4.

We compared different resampling strategies applied on four classifiers, Logistic Regression, RealAdaBoost, BayesNet and Random Forest. We observed that examined resampling strategies (RUS and SMOTE) had a partial increase of the classification performance for all tested classifiers, except for Random Forest. However, regarding the total classification performance, Random Forest, trained with the original data set, outperforms all other approaches.

The reasons for the poor improvement by resampling techniques can be found in the class overlapping or due to class imbalance of the four corpora training datasets, given our chosen feature set.

With these experiments we have shown that the class imbalance has a similar impact on various datasets across languages in terms of the detection of vandalism rates.

Table 4: Values of PR-AUC for the resampling strategies

Classifier	Resampling strategies				
	None	RUS	$SMOTE_{50}$	$SMOTE_{100}$	$SMOTE_{1300/1100}$
PAN-WVC-10					
Logistic Regression	.584	.478	.586	.578	.545
RealAdaBoost	.612	.561	.476	.485	.443
BayesNet	.627	.596	.615	.616	.537
RandomForest	.675	.621	.667	.664	.634
PAN-WVC-11					
Logistic Regression	.633	.615	.636	.646	.616
RealAdaBoost	.534	.536	.495	.453	.499
BayesNet	.651	.643	.663	.656	.562
RandomForest	.744	.726	.738	.732	.706
Simple English					
Logistic Regression	.636	.615	.632	.623	.608
RealAdaBoost	.498	.536	.497	.443	.487
BayesNet	.623	.643	.652	.642	.560
RandomForest	.736	.734	.726	.722	.716
Albanian					
Logistic Regression	.621	.615	.636	.626	.610
RealAdaBoost	.487	.536	.495	.439	.479
BayesNet	.611	.632	.628	.640	.552
RandomForest	.726	.728	.718	.717	.709

For future work, more investigation is needed to point out the within-class imbalance properties in the PAN-WVC corpora and in the Wikipedia history dumps regarding certain feature sets.

References

[1] Susuri, M. Hamiti and A. Dika, Machine Learning Based Detection of Vandalism in Wikipedia across Languages. In proceedings of the 5th Mediterranean Conference on Embedded Computing (MECO), Bar, Montenegro, (2016).

[2] R. Priedhorsky, J. Chen, S. T. K. Lam, K. Panciera, L. Terveen, and J. Riedl. Creating, destroying, and restoring value in Wikipedia. In proceedings of the international ACM conference on supporting GroupWork (GROUP), Sanibel Island, FL, 259-268 (2007).

[3] F. B. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In proceedings of the ACM Conference on human factors in computing systems (CHI), Vienna, Austria, 575-582 (2004).

[4] M. Hamiti, A. Susuri and A. Dika, Machine Learning and the Detection of Anomalies in Wikipedia, Recent Advances in Communications, Proceedings of the 19th International Conference on Communications, Zakynthos Island, Greece, (2015).

[5] Santiago M. Mola-Velasco. Wikipedia Vandalism Detection. - WWW 2011 - Ph.D. Symposium, Hyderabad, India. (2011).

[6] Martin Potthast. Crowdsourcing a wikipedia vandalism corpus. Proceeding of the 33rd international ACM SIGIR conference on research and development in information retrieval - SIGIR '10, 789, (2010).

[7] Martin Potthast, Benno Stein, and Teresa Holfeld. Overview of the 1st International Competition on Wikipedia Vandalism Detection. In Martin Braschler, Donna Harman, and Emanuele Pianta, editors, Working Notes Papers of the CLEF 2010 Evaluation Labs, September (2010).

[8] Martin Potthast and Teresa Holfeld. Overview of the 2nd International Competition on Wikipedia Vandalism Detection. In Vivien Petras, Pamela Forner, and Paul D. Clough, editors, Notebook Papers of CLEF 11 Labs and Workshops, September (2011).

[9] Martin Potthast, Benno Stein, and Robert Gerling. Automatic vandalism detection in wikipedia. In advances in information retrieval, 663-668. Springer Berlin Heidelberg, (2008).

[10] Koen Smets, Bart Goethals, and Brigitte Verdonk. Automatic vandalism detection in wikipedia: Towards a machine learning approach. In WikiAI '08: Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence, (2008).

[11] Kelly Y. Itakura and Charles L. a. Clarke. Using dynamic markov compression to detect vandalism in the Wikipedia. Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09, 822, (2009).

[12] Mola Velasco Santiago Moisés Mola Velasco. Wikipedia vandalism detection through machine learning: Feature review and new proposals - lab report for pan at clef 2010. In CLEF (Notebook Papers/LABs/Workshops), (2010).

[13] Andrew G. West, Sampath Kannan, and Insup Lee. Detecting wikipedia vandalism via spatio-temporal analysis of revision metadata. In Proceedings of the Third European Workshop on System Security, EUROSEC '10, 22-28, New York, NY, USA, (2010).

[14] B. Thomas Adler, Luca De Alfaro, and Ian Pye. Detecting wikipedia vandalism using wikitrust. Notebook papers of CLEF, (2010).

[15] B. Thomas Adler and Luca De Alfaro. A content-driven reputation system for the wikipedia. Proceedings of the 16th international conference on World Wide Web WWW 07, 7(Generic):261, (2007).

[16] B. Thomas Adler, Luca De Alfaro, Santiago M. Mola-Velasco, Paolo Rosso, and Andrew G. West. Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part II, 277-288, Berlin, Heidelberg, (2011).

[17] Andrew G. West and Insup Lee. Multilingual vandalism detection using language-independent & ex post facto evidence - notebook for pan at clef 2011. In CLEF (Notebook Papers/Labs/Workshop), (2011).

[18] Khoi-Nguyen Tran and Peter Christen. Cross Language Prediction of Vandalism on Wikipedia Using Article Views and Revisions. Advances in Knowledge Discovery and Data Mining, 268-279 (2013).

[19] Manoj Harpalani, Michael Hart, S Signh, Rob Johnson, and Yejin Choi. Language of Vandalism: Improving Wikipedia Vandalism Detection via Stylometric Analysis. ACL (Short Papers), 83-88 (2011).

[20] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, **21**(9): 1263-1284 (2009).

[21] Vaishali Ganganwar. An overview of classification algorithms for imbalanced datasets. International Journal of Emerging Technology and Advanced Engineering, **2**(4): 42-47 (2012).

[22] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research," **16**(1): 321-357 (2002).
[23] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: A new over-sampling method in imbalanced data sets learning. In Proceedings of the 2005 International Conference on Advances in Intelligent Computing - Volume Part I, ICIC'05, 878-887, Berlin, Heidelberg, (2005).
[24] J. Friedman, T. Hastie, and R. Tibshirani. Additive Logistic Regression: a Statistical View of Boosting. The Annals of Statistics, **38**(2): (2000).
[25] Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In Lorenza Saitta, editor, Machine Learning, Proceedings of the Thirteenth International Conference (ICML '96), Bari, Italy, July 3-6, 1996, 148-156. Morgan Kaufmann, (1996).
[26] Judea Pearl and Stuart Russell. Bayesian networks. In M. Arbib, editor, Handbook of Brain Theory and Neural Networks, number R-277. MIT Press, (2001).
[27] Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. Machine Learning, **29**(2-3): 131-163 (1997).
[28] Leo Breiman. Random forests. Machine Learning, **45**(1): 5-32 (2001).
[29] K. Tran, P. Christen, "Cross-language prediction of vandalism on wikipedia using article views and revisions". Proceedings of the 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) (2013).
[30] Si-Chi Chin and W. Nick Street. Enriching wikipedia vandalism taxonomy via subclass discovery. In LDH, 19-24, (2011).

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 23-41 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

Operational Efficiencies and Simulated Performance of Big Data Analytics Platform over Billions of Patient Records of a Hospital System

Dillon Chrimes[1*], Belaid Moa[2], Mu-Hsing (Alex) Kuo[3], Andre Kushniruk[3]

[1]*Database Integration and Management, Quality Systems, Vancouver Island Health Authority, V8R1J8, Canada*

[2]*Compute Canada/WestGrid/University Systems, Advanced Research Computing, University of Victoria, V8P5C2, Canada*

[3]*School of Health Information Science, Faculty of Human and Social Development, University of Victoria, V8P5C2, Canada*

ARTICLE INFO

Article history:
Received: 30 November, 2016
Accepted: 12 January, 2017
Online: 28 January, 2017

Keywords:
Big Data
Big Data Analytics
Data Mining
Data Visualizations
Healthcare
Hospital Systems
Interactive Big Data
Patient Data
Simulation
Usability

ABSTRACT

Big Data Analytics (BDA) is important to utilize data from hospital systems to reduce healthcare costs. BDA enable queries of large volumes of patient data in an interactively dynamic way for healthcare. The study objective was high performance establishment of interactive BDA platform of hospital system. A Hadoop/MapReduce framework was established at University of Victoria (UVic) with Compute Canada/Westgrid to form a Healthcare BDA (HBDA) platform with HBase (NoSQL database) using hospital-specific metadata and file ingestion. Patient data profiles and clinical workflow derived from Vancouver Island Health Authority (VIHA), Victoria, BC, Canada. The proof-of-concept implementation tested patient data representative of the entire Provincial hospital systems. We cross-referenced all data profiles and metadata with real patient data used in clinical reporting. Query performance tested Apache tools in Hadoop's ecosystem. At optimized iteration, Hadoop Distributed File System (HDFS) ingestion required three seconds but HBase required four to twelve hours to complete the Reducer of MapReduce. HBase bulkloads took a week for one billion (10TB) and over two months for three billion (30TB). Simple and complex query results showed about two seconds for one and three billion, respectively. Apache Drill outperformed Apache Spark. However, it was restricted to running more simplified queries with poor usability for healthcare. Jupyter on Spark offered high performance and customization to run all queries simultaneously with high usability. BDA platform of HBase distributed over Hadoop successfully; however, some inconsistencies of MapReduce limited operational efficiencies. Importance of Hadoop/MapReduce on representation of platform performance discussed.

1. Introduction

Gantz and Reinsel [1] predicted in their 'The Digital Universe' study that the digital data created and consumed per year would reach 40,000 Exabyte by 2020, from which a third will promise value to organizations if processed using big data technologies. In fact, global digital patient data expected to reach 25 Exabytes (1018 bytes) in 2020 [2]. Furthermore, A McKinsey Global Institute stated US healthcare that uses Big Data effectively could create more than $300 billion in value from cost savings annually [3]. At the same time in 2013, Canada Health Infoway [4] produced a white paper to solidify the meaningful use of patient data to cut healthcare costs for each of the Provinces. However, the increase in digital data and fluid nature of information- processing methods and innovative big data technologies has not caused an increase of implementations in healthcare in Canada and abroad. Furthermore, there are very few if any of the 125 countries surveyed by the World Health Organization with any Big Data strategy for universal healthcare coverage of their eHealth profiles [5]. Conventional systems in healthcare are very expensive to implement and establish that further reduces the uptake of open source software like Hadoop/MapReduce frameworks.

Health Big Data is a large complex-distributed highly diversified patient data requiring high performance analytical tools to utilize large volume of data for healthcare application [6-8]. Big Data Analytics (BDA) is a platform with analytical technologies

*Corresponding Author: Dillon Chrimes, Quality Systems, IMIT, Vancouver Island Health Authority, 1952 Bay Street, V8R1J8, Victoria BC Canada
Phone: 1 (250)-370-8111 ext.13975 | Email: dillon.chrimes@viha.ca

frame worked to extract knowledge in real-time for evidence-based medicine, medical services and transport of inpatients in hospital wards, onset of in-hospital acquired infections, and treatments linked to health outcomes including scientific and clinical discoveries [3, 4, 9-11]. However, a BDA platform is of little value if decision-makers do not understand the patterns it discovers and cannot use the trends to reduce costs or improve processes. Subsequently, BDA research is important and highly innovative to effectively utilize data quickly for the ongoing improvement of health outcomes in bioinformatics, genome sequencing, and tertiary healthcare [12-16].

After extensive literature review, many Big Data technologies with Hadoop/MapReduce are available but few applied in healthcare (Table 1 in Appendix). BDA platform called Constellation deployed at the Children's Mercy Hospital in Kansas City, US successfully matched patient data of children to their whole-genome sequencing for treatment of potentially incurable diseases [43]. Their big data analytics study showed that, in emergency cases, the differential diagnosis for a genetic disease in newborn patients was identifiable in 50 hours [42]. Further improvement using Hadoop reduced the analysis time for the whole genome sequencing from 50 to 26 hours [44]. Unfortunately, even with these success stories of fully functional BDA platforms in bioinformatics, there are few studies and published reporting of BDA platforms used by health providers in hospital systems.

There are many alternative solutions for databases in Big Data platforms; choice of the best solution depends on the nature of the data and its intended use, e.g., [37]. In practice, while many systems fall under the umbrella of NoSQL systems and are highly scalable (e.g., [45], these storage types are quite varied). Big Data is characterized in several ways: as unstructured [19], NoSQL [31, 36], key-value (KV) indexed, text, information-based [46], and so on. In view of this complexity, BDA requires a more comprehensive approach than traditional data mining; it calls for a unified methodology that can accommodate the velocity, veracity, and volume capacities needed to facilitate the discovery of information across all data types [4]. Furthermore the KV data stores represent the simplest model of NoSQL systems: they pair keys to values in a very similar fashion to show a map (or hashtable) works in any standard programming language. Various open-source projects provide key-valued NoSQL database systems; such projects include Memcached, Voldemort, Redis, and Basho Riak, e.g., [22]. HBase was chosen because it simplified the emulation of the columns using the metadata in each column rather than the data types and the actual relationships among the data. HBase also has a dynamic schema that can be uploaded via other Apache applications; therefore, the schema can be changed and tested on the fly. Another benefit of using HBase is that further configurations can be accomplished for multi-row transactions using a comma-separated value (.CSV) flat file [47].

The KV class of store files in databases is the heart of data storage in HBase that provides inherent encryption. Privacy mandates are a major barrier for any BDA implementation and utilization. The Health Insurance Portability and Accountability Act (HIPAA), as well as Freedom of Information and Protection of Privacy (FIPPA) Act requires the removal of 18 types of identifiers, including any residual information that could identify individual patients, e.g., [48]. Therefore, privacy concerns can be addressed using new database technologies, such as key-value (KV) storage services. For example, Pattuk, Kantarcioglu, Khadilkar, Ulusoy, and Mehrotra [49] proposed a framework for securing Big Data management involving an HBase database – called Big Secret – securely outsources and processes encrypted data over public KV stores.

In a hospital system, such as for the Vancouver Island Health Authority (VIHA), the capacity to record patient data efficiently during the processes of admission, discharge, and transfer (ADT) is crucial to timely patient care and the quality of patient-care deliverables. The ADT system is the source of truth for reporting of the operations of the hospital from inpatient to outpatient and discharged patients. Proprietary hospital systems for ADT also have certain data standards that are partly determined by the physical movement of patients through the hospital rather than the recording of diagnoses and interventions. Among the deliverables are reports of clinical events, diagnoses, and patient encounters linked to diagnoses and treatments. Additionally, in most Canadian hospitals, discharge records are subject to data standards set by Canadian Institute of Health Information (CIHI) and entered into Canada's national Discharge Abstract Database (DAD). Moreover, ADT reporting is generally conducted through manual data entry to a patient's chart and then it is combined with Electronic Health Record (EHR), which could also comprise auto-populate data, to conglomerate with other hospital data from provincial and federal health departments [50]. These two reporting systems, i.e., ADT and DAD, account for the majority of patient data in hospitals, but they are seldom aggregated and integrated as a whole because of their complexity and large volume. A suitable BDA platform for a hospital should allow for the integration of ADT and DAD records to query the data to find unknown trends at extreme volumes of the entire system. However, there are restrictions that limit the data that gets recorded, especially on discharging a patient a physician is legally required only to record health outcomes rather than the details of interventions. For these and other reasons, health informatics has tended to focus on the structure of databases rather than the performance of analytics at extreme volumes.

Currently, the data warehouse at VIHA has over 1000 (relational) tables that include alias pools for data integrity of patient encounters in ADT of the hospital system. Its total size estimated at one billion records or 14 Terabytes (TB), and it is one of the largest continuous patient records in Canada [51]. Huge volumes of highly diversified patient data are continuously added into this collection; this equates to annually 500 million records or five TB. Currently, at VIHA, numerous data manipulations and abstracting processes put into place via non-enterprise platforms to combine patient data from the relational databases of the hospital system customized to apply clinical and operational queries. Neither business intelligence (BI) tools nor data warehouse techniques are currently applied to both data sets of ADT and DAD at the same time and over its entire volume of data warehouse. Therefore, we propose an enterprise BDA platform with applications to query patient data of a database representing a hospital system comprised of ADT and DAD databases.

2. Interactive Healthcare Big Data Analytics (HBDA) Framework

2.1. Conceptual Framework

The conceptual framework for a BDA project in healthcare is similar to that of a traditional health informatics analytics. That is, its essence and functionality is not very different from that of

conventional systems. The key difference lies in data-processing methodology. In terms of the mining metaphor, data represent the gold over the rainbow while analytics systems represent the leprechaun that found the treasure or the actually mechanical minting of the metals to access it. Moreover, healthcare analytics is defined as a set of computer-based methods, processes, and workflows for transforming raw health data into meaningful insights, new discoveries, and knowledge that can inform more effective decision-making [23]. Data mining in healthcare has traditionally been linked to knowledge management, reflecting a managerial approach to the discovery, collection, analysis, sharing, and use of knowledge [52-53]. Thus, the DAD and ADT are designed to enable hospitals and health authorities to apply knowledge derived from data recording patient numbers, health outcomes, length of stay (LOS), and so forth, to the evaluation and improvement of hospital and healthcare system performance. Furthermore, because the relational databases of hospitals are becoming more robust, it is possible to add columns and replicate data in a distributed filing system with many (potentially cloud-based) nodes and with parallel computing capabilities. The utility of this approach is that columns are combined (i.e., columns from the DAD and ADT database). In addition, such a combination can mimic data in the hospital system in conjunction with other clinical applications. Through replication, generation and ingestion, the columns can form one enormous file then queried (while columns added and removed or updated).

BDA platform(s) should offer the necessary tools currently performed by hospitals and its managed relational databases to query the patient data for healthcare. Furthermore, the end user experience should include analytical tools with visualizations using web-based applications. To achieve this, a dynamic interactive BDA platform, and following our preliminary results from Chrimes, Moa, Zamani, and Kuo [54], established for healthcare application at the University of Victoria (UVic), in association with WestGrid, and Vancouver Island Health Authority (VIHA), Victoria, BC, Canada. Thus, our Healthcare BDA (HBDA) platform provided a proof-of-concept implementation and simulation of high performance using emulated patient data generated on WestGrid's Linux database clusters located at UVic. Emulation consisted of reaching nine billion health records that represented the main hospital system of VIHA and its clinical reporting via its data warehouse (Figure 1). The emulated data utilized in simulation could capture the appropriate configurations and end-user workflows of the applications, while ultimately displaying health trends at the hospital and patient levels.

Our team of collaborators existed between UVic, Compute Canada\WestGrid and VIHA that thru requirements gathering, usability testing and software installations established the framework of our HBDA platform. It comprised innovative technologies of Hadoop Distributed File System (HDFS) with MapReduce's Job/Task Scheduler, and noSQL database called HBase. The database construct was complex and had many iterations of development over the past three-four years. There were many configurations of components included such as Apache Phoenix, Apache Spark and Apache Drill, as well as Zeppelin and Jupyter Notebook web-client interfaces. Furthermore, we required a proof-of-concept to implement in simulation before applying it to real patient data after rigorously approved by research ethics with guaranteed highly secured patient data. Our aim was effectively to query billions records of patient data stored in the VIHA with health professionals and providers to reveal its fast and reliable queries, as well as unveil unknown health trends, patterns, and relevant associations of medical services with health outcomes.

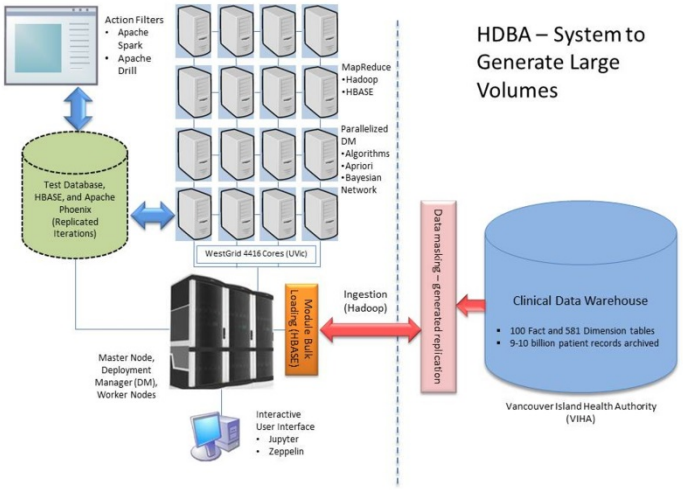


Figure 1. The proposed framework of Health Big Data Analytics (HBDA) platform with Vancouver Island health authority (VIHA) under replicated Hadoop Distributed File System (HDFS) to form noSQL database via MapReduce loads with big data analytic tools interacting under parallelized deployment manager (DB) at Westgrid, UVic, Victoria.

We constructed 50 million to nine billion patient records to form different levels of testing of our HBDA platform based on known data profiles and metadata of the patient data in the hospital system. Henceforward, our simulation should show significant improvements of query times; usefulness of the interfaced applications; and the applicability and usability of the platform to healthcare. To deal with the implementation challenges, we viewed HBDA as a pipelined data processing framework, e.g., [11], and worked in conjunction (interviewed) with VIHA experts (and potential stakeholders) to identify the metadata of important inpatient profiles (Figure 2). Additionally, the stakeholder provided the required patient data and workflow processes for both generating the reports and the application used for querying results to achieve visualizations.

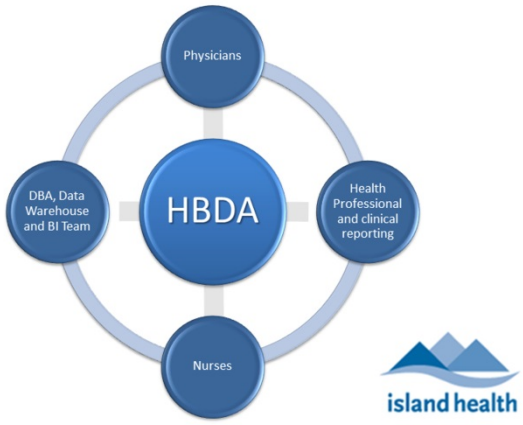


Figure 2. Main stakeholder groups of Physicians, Nurses, Health Professionals (i.e. epidemiologists and clinical reporting specialists), and Data Warehouse and Business Intelligence (BI) team (includes application platform and database integration) involved in clinical reporting at VIHA.

2.2 Overview

The objective of this platform implementation and simulated performance was to establish an interactive framework with large representative patient data aligned with front-end and interfaced applications linked to HDFS stack and noSQL database with visualization capabilities that allows users to query the data. For the implementation, advanced technologies created a dynamic BDA platform while generating emulated patient data over a distributed computing system, which is currently not in use at VIHA and many other health institutions in Canada. The implemented HBDA was big data centric and designed to make big data capabilities, including analytics using mainly SQL-like compatibilities with data warehouse team and front-end architecture for correct visualization, accessible to different stakeholders, especially physicians and healthcare practitioners (Figure 3).

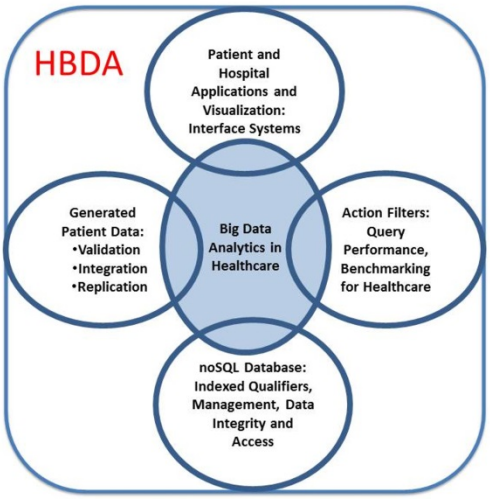


Figure 3. The main components of our Healthcare Big Data Analytics (HBDA) platform envisioned by stakeholders and derived from our team. It is important to note that the center of this diagram is truly the SQL-like code used to query the data because without it the other interconnected components would not provide any linkage to user and no meaningful use.

The established BDA platform allowed for testing the queries' performance that included actions and filters that corresponded to current reporting of VIHA's clinical data warehouse. To test performance, the first step was to emulate the metadata and data profiles of VIHA's ADT model, which is Cerner's (proprietary) hospital system. The metadata derived for the patient encounter(s) combined with VIHA's hospital reporting to CIHI of its DAD formed our test database. The data aggregation represented both the source system of patient encounters and its construct, which represented patient data collected during the encounter (for different encounter types) before patient discharge. At VIHA, and its Cerner system, there are hundreds of tables that comprise the ADT and all are part of a relational database with primary and foreign keys. These keys are important for data integrity that link clinical events for that patient only. Therefore, it was necessary to use constraints in emulating data for the noSQL database.

Hadoop/MapReduce framework proposed to implement the BDA platform and analyze emulated patient data over a distributed computing system is net new to acute patient care settings at VIHA and other health authorities in Canada. Innovative technologies from Hadoop's ecosystem with MapReduce programming, and a NoSQL database, called HBase were utilized to form a complex database construct. HBase is an open-source, distributed key-value (KV) store based on Google's BigTable [55] — persistent and strictly consistent NoSQL system using HDFS for data storage. Furthermore, with all these technical components to construct the platform, the build also took into account the workflow by clinical reporting workgroups with the same metadata from hospital datasets.

2.3 Data Generation

To generate accurate representations of patient records, we originally constructed the emulated database in Oracle Express 11g using SQL to establish the column names and metadata using constrained data profiles for each randomized row of dummy patient data. For example, data for the diagnostic column was ICD-10-CA codes and set that data to those standardized characters. The data was populated for each column based on a list of metadata set in the script to generate for each of the columns. Furthermore, important data profiles and dependencies established through primary keys over selected columns. Ninety columns from DAD and ADT were eventually established, generated, and ingested into the HBDA platform. Since data warehouse team working with health professionals for clinical reporting relies on comma-separated value (CSV) formats when importing and exporting their data, we opted to use the ingested CSV files directly for analytics instead of HBase. This was carried out previously on our platform using Apache Phoenix and its SQL-like code on HBase [51, 56]. Two data sizes of 50 million and one billion records established a benchmark check on how different packages (Apache Spark and Drill) scaled with data size with an additional three billion tested.

2.4 Infrastructure

Using existing computational resources and architecture was an essential requirement. We relied on the existing WestGrid clusters at UVic to run the platform. Among ~500 nodes available on the WestGrid clusters hosted at UVic, we planned to use as many nodes as possible while benchmarking and testing scalability. Currently, we only used six dedicated nodes due to configuration issues and ongoing maintenance. Local disks had 6TB per node that then could reach ten billion records at this 36TB storage to illustrate operational platform. The PBS/Torque resource manager managed the clusters that allowed access to launch our interactive Hadoop/HBase jobs and modules for Apache Spark/Drill while loading Jupyter Notebook or Zeppelin on those nodes. The nodes had 12 cores, 24GB and three 2TB-disks each, and were similar in technical specification to the existing servers at VIHA. Thus, comparisons could be made with VIHA based on Westgrid's high performance supercomputing (HPC), referred to architectural and technical specifications for more details [51].

Given that the WestGrid clusters are traditional HPC clusters, we had to customize our setup scripts to launch a dynamic HBDA platform that runs when the job starts and terminates when the job finishes. This allows our HBDA platform to exist within an HPC cluster and thus eliminate special treatment to the environment with extending the job wall-time when necessary.

2.5 Big Data Analytics (BDA) and Visualization

After interviews with different VIHA stakeholders, identified requirements applied to our proof-of-concept and visualization simulation: interactive, simple, SQL-like, and visualization-enabled interface. Moreover, the platform should be able to offer the necessary tools to implement new analyses and act as an expert

system when needed; this was required because clinicians were interested in answering specific scenarios. An example of a scenario was:

"Clinicians suspect that frequent movement of patients within the hospital can worsen outcomes. This is especially true of those who are prone to confusion (or mental relapse) due to physical and environmental changes in their current state that can exacerbate the confusion. If this occurs, a physician may attribute their confusion to the possible onset of an infection, a disease or a mental illness, resulting in unnecessary laboratory tests and medications. Moreover, by moving to a new location, new healthcare workers may detect subtle changes in a patient's mental baseline and frequent patient movements can cause sepsis onset."

For the HBDA platform to be successful it should process 1) a large number of patient records (data space) over a wide range of computationally intensive algorithms (algorithmic space); as well as 2) to easily generate visualizations (visualization space); 3) offer libraries; and, 4) application tools (tool space) to support these four spaces. Deploying our platform covered all the four of these spaces while being interactive was challenging. In previous study [51, 54], we reported that Hadoop, HBase, and Apache Phoenix provided an excellent platform to perform SQL-like queries with high performance and accurate patient results. However, the ingestion of one billion records from Hadoop HFiles bulkloaded to HBase (indexed files) took weeks, while three billion took over two months on the infrastructure before finalizing the configurations. Moreover, SQL-like queries and tools were only a small fraction of the interest and importance mentioned by stakeholders, who included other factors of their workflow and professional consideration like algorithmic, tool, visualization spaces (e.g., epidemiology versus hospitalization reporting). In our simulation, therefore, we expanded on those spaces that our first article of Hadoop and HBase did not investigate with three-nine billion records while using Zeppelin and Jupyter with Spark and Drill's separate interface at various testing. Moreover, we investigated SQL-like capabilities of Spark and SQL-ANSI of Drill on CSV files indexed by HBase compared to Phoenix on HBase.

The data construction framework used by this study extracted the appropriate data profiles from the data dictionaries and data standard definitions for the ADT system and DAD abstract manual [57]. The data used to test performance was circumvented with the interviews with different VIHA stakeholders, several requirements and scenarios, and 17 clinical cases were outlined that were identified for a realistic proof-of-concept BDA and data visualization: the interface should be interactive, simple, ANSI-SQL or SQL-like, and visualization enabled/embedded in the browser. Moreover, the platform should be able to offer the necessary tools to implement newly advanced analytics to act as a recommendation of an expert system; this is required, as clinicians are interested in answering specific scenarios of inpatient encounters with accurate data.

2.6 Testing & Evaluation

The functional platform tested performance of data migrations or ingestions of HFiles via Hadoop (HDFS), bulkloads of HBase, and ingestions of HFiles to Apache Spark and Apache Drill. Test speed of performance to complete ingestion or queries were proof-of-concept testing using simulated data with the same replicated metadata and very large volume, but this did not consist of comparing performance of SQL (relational database) with NoSQL or different data models with real patient data. The SQL can be very similar with real data but cannot be compared at this time. Furthermore, the bulk of the methodology is generating the emulated data and queries with Hadoop configurations and other software, i.e., HBase, Spark and Drill. Most of the configurations established after installing the components, which were defaulted for the distributed filing and MapReduce in Java, Python and Scala to perform as expected; therefore, the platform established by Hadoop-MapReduce configurations to run and integrate with all other big data tools.

For SQL-like analytics, to test and evaluate Zeppelin versus Jupyter with Spark and SQL-ANSI with Drill, we selected 22 SQL query tests based on the interviews conducted and same as published in Chrimes et al., [51]. Eventhough the queries had different level of complexity; our performance results showed that all had similar times to generate the patient results or reports. An example of SQL statements used to test system was:

```
SELECT Count(Episode_Duration) as EDCount,
Count(Anesthetistic_Technique) as ATCount,
Count(Interven_Location) as ILCount,
Count(Medical_Services) as MSCount,
Count(Unit_Transfer_Occurrence) as UTOCount FROM
DADS1 where EncounterID<1000 Group By age;
```

For other kinds of analytics, especially the machine learning algorithms, we performed simple tasks such as computing the correlations between different pairs of columns, such 'age' and the 'length of stay at the hospital' or LOS, as well as simple clustering. The data was synthetic or emulated; therefore, using machine learning to answer intelligent scenarios or find interesting patterns were only applied narrowly to find randomized patterns of the health outcome parameters that was already known. Nevertheless, its configurations and performance was ultimately the information attained towards using real patient data over our HBDA.

For visualization, we utilized the common set of graphs that healthcare providers would use to generate such reporting as table, pie chart, scatter plot, and histogram visualizations. In producing each graph and SQL-like query, we recorded and documented all processes and connectivity times.

3. Implementation of HBDA Platform

3.1. Overview of High Performance System(s)

The beauty of a BDA platform with open-source software in Hadoop's ecosystem is that there are a wide range of tools and technologies for bulkloading and accessing large datasets from different sources. Sqoop, for example, is useful to ease the transfer between existing relational databases and a BDA platform [36]. For collecting and gathering unstructured data, such as logs, one can use Flume. Since the performance tests of queries on the platform relied on data emulation, it was used, as a proof-of-concept, the usual high-speed file transfer technologies (such as SCP and GridFTP) to transfer data to the HPC parallel file system (GPFS). It was then used the Hadoop and HBase as NoSQL database bulkload utilities to ingest the data.

The established BDA platform will be used to benchmark the performance of end users' querying of current and future reporting of VIHA's clinical data warehouse (i.e., in production at VIHA, spans more than 50 years of circa 14 TB). To accomplish this, Hadoop environment (including the Hadoop HDFS) from a source installed and configured on the WestGrid cluster, and a dynamic Hadoop job was launched. Hadoop (version 2.6.0) and its HDFS was configured by hdfs-site.xml and a MapReduce frame [26],

configured via mapred-site.xml, that was run under the Hadoop resource manager Yarn (with configuration file yarn-site.xml). The number of replicas was set to three. To interact with HDFS, command scripts were run to automate the ingestion step (generate data replication in the exact format specified by SQL script to the nodes).

The BDA platform was built on top of the available open-source database software (HBase). HBase (NoSQL version 0.98.11) is a NoSQL database composed of the main deployment master (DM) and fail-over master, the RegionServers holding HBASE data, and ZooKeeper, which contained services to allocate data locality [28], of three nodes, that orchestrated that ensemble. The xml configuration files were HBase-site.xml and the HBase-env.sh adjusted to improve the performance of HBase. HBase was chosen due to its NoSQL services and many other features, especially linear and modular scalability. In addition, it allows for SQL-like layers via Apache Phoenix configured on top of HBase big-tables.

The construction and build of the framework with HBase (NoSQL) and Hadoop (HDFS) coincided with and enforced by the existing architecture of the WestGrid clusters at UVic (secure login via LDAP directory service accounts to deployment database nodes and restricted accounts to dedicated nodes). The data were distributed in parallel on the nodes via a balanced allocation to each local disk with running part of the batch jobs in a serial computing process. Deployment of the Hadoop environment on the nodes was carried out via a sequence of setup scripts that the user calls after loading the necessary modules. These setup scripts create an initial configuration depending on the number of nodes chosen when launching the job. The user can adjust those configurations to match the needs of the job and its performance.

Making the DBA platform InfiniBand-enabled was challenging, as most of the Hadoop environment services rely on the hostname to get the IP address of the machine. Since the hostnames on a cluster are usually assigned to their management network, the setup scripts and the configuration files required adjustment. The InfiniBand was used because it offers low latency (in us) and high bandwidth (~40Gb/s) connectivity between the nodes. Yarn, Hadoop's resource and job manager [33], unfortunately still partly used the Gig-Ethernet interface when orchestrating between the nodes, but the data transfer was carried out on the InfiniBand.

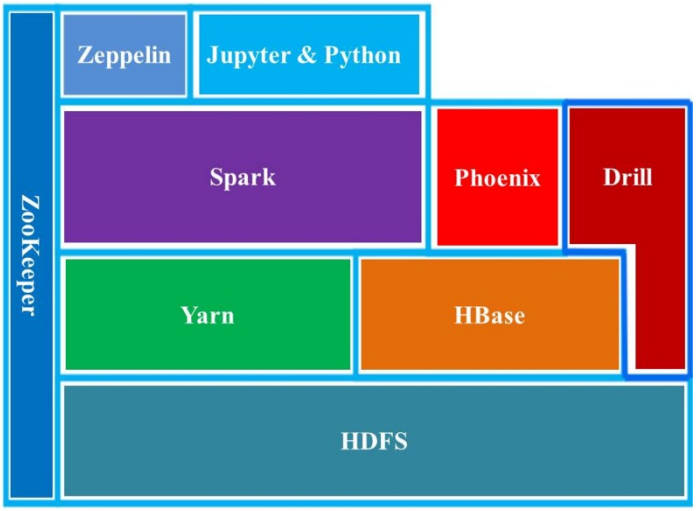
The bulk of the framework was comprised of open-source packages of Hadoop's ecosystem. Even though several configurations were done to the Hadoop ecosystem to optimize running on the WestGrid dedicated cluster, no hardware modification was needed; possible future changes could be made to meet minimum recommended RAM, disk space, etc. requirements per node, e.g., refer to guidelines via Cloudera [58]. Hadoop provides the robust, fault-tolerant HDFS inspired by Google's file system [55], as well as a Java-based API that allows parallel processing across the nodes of the cluster using the MapReduce paradigm. The platform was used in Python with Java to run jobs and ingestions. Hadoop comes with Job and Task Trackers that keep track of the programs' execution across the nodes of the cluster. These Job and Task Trackers are important for Hadoop to work on a platform in unison with MapReduce and other ingestion steps involved with HBase, ZooKeeper, Spark, and Drill. There have been many contributors, both academic and commercial (Yahoo being the largest), to using Hadoop over a BDA platform, and a broad and rapidly growing user community [59].

The software stack used in the platform has at its bottom is HDFS (Figure 4). HDFS is the distributed file system of Hadoop and known to scale and perform well in the data space. Yarn was the resource manager of Hadoop and services of scheduling incongruent to running the Hadoop jobs. In addition to MapReduce component, Yarn, and HDFS constitute the main components [59]. In our platform, we built, configured and tested different versions of Hadoop, and managed the version via its respective module environment packages. The following module command lists all the available Hadoop versions:

```
$module avail Hadoop

Hadoop/1.2.1(default)Hadoop/2.2.0Hadoop/2.6.0
Hadoop/2.6.2
```

For our study, we used Hadoop2.6.2 with replication set to three in all our runs. Instead of using the MapReduce as our main service for the algorithmic space, we chose Apache Spark, as it covers larger algorithmic and tool spaces and outperforms Hadoop's MapReduce to HBase as bulkload.



eeper is a resource allocator; Yarn is resource manager; HDFS is Hadoop Distributed File System; HBase is noSQL database; Phoenix is Apache Phoenix (query tool on HBase); Spark is Apache Spark (query tool with specialized Yarn); Drill is Apache Drill (query tool with specialized configuration to Hadoop via ZooKeeper); and Zeppelin and Jupyter are interfaces on local web hosts (clients) using Hadoop ingestion and Spark module.

The queries via Apache Phoenix (version 4.3.0) resided as a thin SQL-like layer on HBase. This allowed ingested data to form structured schema-based data in the NoSQL database. Phoenix can run SQL-like queries against the HBase data. Similar to the HBase shell, Phoenix is equipped with a python interface to run SQL statements and it utilizes a CSV file bulkloader tool to ingest a large flat file using MapReduce. The load balancing between the RegionServers (e.g., "salt bucket") was set to the number of slaves or worker nodes that allowed ingested data to be balanced and distributed evenly. The pathway to running ingestions and queries from the build of the BDA platform on the existing HPC was as follows: CSV flat files generated → Module Load → HDFS ingestion(s) → Bulkloads/HBase → Apache Phoenix Queries. Under this sequence, the Apache Phoenix module loaded after Hadoop and HBase SQL code was directed and then iteratively run to ingest 50 million rows to the existing NoSQL HBase database

(see Results for details). This pathway was tested in iteration up to three billion records (once generated) for comparison of the combination of HBase-Phoenix versus Phoenix-Spark or an Apache Spark Plugin [60].

After the metadata, used in the emulated database, was verified from the existing relational database, it was established as a patient-centric SQL-like queries using Phoenix on top of the non-relational NoSQL HBase store. To establish data structure, the EncounterID was set as a big data integer (so that it can reach billions of integers without limitation) and indexed based on that integer via HBase for each unique at every iteration that followed. This indexed-value column, unique for every row, causes MapReduce to sort the KV stores for every one of the iterations that can increase the integrity of the data and increase its secured access once distributed. Once distributed, this allowed for SQL-like scenarios to be used in the evaluation of the performance of the BDA platform and accuracy of the emulated patient database.

3.2. Implementation of Apache Spark Technology System

Apache Spark (version 1.3.0) was also built from source and installed to use on HBase and the Hadoop cluster. The intent was to compare different query tools like Apache Spark over the BDA platform with Zeppelin and Jupyter then cross-compared with Drill's ANSI-SQL and Apache Phoenix using similar SQL-like queries. Currently, Apache Spark, its action filters and transformations and resilient distributed dataset (RDD) format, is at the heart of the processing queries over the platform aligned with Hadoop [33]. It covered a substantial data space because it could read data from HDFS and HBase. Its speed can be fast because of its judicious use of memory to cache the data, because Spark's main operators of transformations to form actions/filters were applied to an immutable RDD [32-33]. Each transformation produces an RDD that needs to be cached in memory and/or persisted to disk, depending on the user's choice [32, 61]. In Spark, transformation was a lazy operator; instead, direct acyclic graphs (DAG) of the RDD transformations build, optimize, and only executed when action applied [32]. Spark relied on a master to drive the execution of DAG on a set of executors (running on the Hadoop DataNodes). On top of the Spark transformation engine there was a suite of tool space: Spark SQL and data frames, Machine Learning Library (MLLib), Graph library (GraphX), and many third-party packages. These tools run interactively or via a batch job. Similar to Hadoop, all of this software is available through the module environments package [54, 61]. Since the investigation was for user interactivity and usability the following was elaborated on:

i) Spark terminal is available for users to interact with Spark via its terminal. Similar to the usual command line terminals, everything displayed as text. For Scala language (Java and Python), the user can call spark-shell to get access to the Scala (i.e., Python with Java libraries) terminal. For Python, the user should run Pyspark (module). For this study's platform, the scale mode of use was only recommended for advanced users. For all stakeholders, selecting a terminal is not user-friendly and the platform will be viewed as too cumbersome to use; therefore, it ran the module in simulation as an administrator with the end user.

ii) Jupyter notebook is a successful interactive and development tool for data science and scientific computing. It accommodates over 50 programming languages via its notebook web application, and comprises all necessary infrastructures to create and share documents, as well as collaborate, report, and publish. The notebook was user-friendly and a rich document able to contain/embed code, equations, text (especially markup text), and visualizations. For utilization, the simultaneous running of all queries over the fast ingestion of the entire database was a significant achievement. Furthermore, corroboration of the results, after selecting a port and local host to run over a browser, started from 50 million and extending to three billion records, and validated the queries over the time of application to display. The existing software installations of Jupyter 4.0.6 on python 2.7.9 were implemented without customization.

iii) Zeppelin was another interface used to interact with Spark. Zeppelin is a web- based notebook; it is not limited to Spark. It supports a large number of back-end processes through its interpreter mechanism [32]. In Zeppelin, any other back-end process can be supported once the proper interpreter is implemented for it. SQL queries were supported via the SparkSQL interpreter and executed by preceding the query with a %sql tag. What makes Zeppelin interesting, from the HBDA platform perspective, was the fact that it has built- in visualizations and the user can, after running an SQL query, click on the icons to generate the graph or chart. This kind of at-the-fingertip visualization was essential for the zero-day or rapid analytics. Spark 1.5.2 and Zeppelin 0.6.0 were built from source, configured with a port and local host had to run over a browser, and used for the testing and benchmarking of the platform.

3.3. Implementation of Apache Drill Technology System

Inspired by Google's big query engine Dremel, Drill offers a distributed execution environment for large-scale ANSI-SQL:2003 queries. It supports a wide range of data sources, including .csv, JSON, HBase, etc... [62]. By (re)compiling and optimizing each of the queries while it interacting with the distributed data sets via the so-called Drillbit service, Drill showed capacity of the query with performance at a low latency SQL query. Unlike the master/slave architecture of Spark, in which a driver handles the execution of the DAG on a given set of executors, the Drillbits were loosely coupled and each could accept a query from the client [63]. The receiving Drillbit becomes the driver for the query, parsing, and optimization over a generated efficient, distributed, and multiphase execution plan; it also gathers the results back when the scheduled execution is done [19, 63]. To run Drill over a distributed mode, the user will need a ZooKeeper cluster continuously running. Drill 1.3.0 and ZooKeeper 3.4.6 were installed and configured on the framework of the platform over a port with a local host.

4. Findings and Significant Benchmarking Results

4.1. Technical Challenges

From a technical perspective, very few issues were faced during software installation of Spark and Drill. Besides the lack of documentation, a major issue with Hadoop's ecosystem of packages was that many of them failed to include the dependencies that are local to the system, like with ZooKeeper for Spark and the version of Hadoop required with version of Python (Table 2 in Appendix).

To establish the BDA framework with different big data technologies, therefore, additional configurations had to be made in order to customize the platform without failed services and ongoing running of Hadoop at high performance (Figure 5), which was required to run ingestions and query. Additionally, for

improved efficiency, the platform used native libraries because Hadoop needs to be recompiled from source [19, 22, 23, 30, 61, 63-64]. Compression of HBase store files reduced the performance by ~100 minutes from 399 to 299 minutes on average. And it was found that many packages do not offer a profile that allowed them to be used, as local Hadoop even avoids downloading components. Moreover, given that the end users were allowed to run their Hadoop module on WestGrid's clusters. Therefore, it was important that the configuration directories had most of the Hadoop's ecosystem of packages built in to minimize reliance on them. Unfortunately, the configuration of Hadoop was not propagated properly to all the remote services running on the nodes, which caused the default configuration on those nodes to be used.

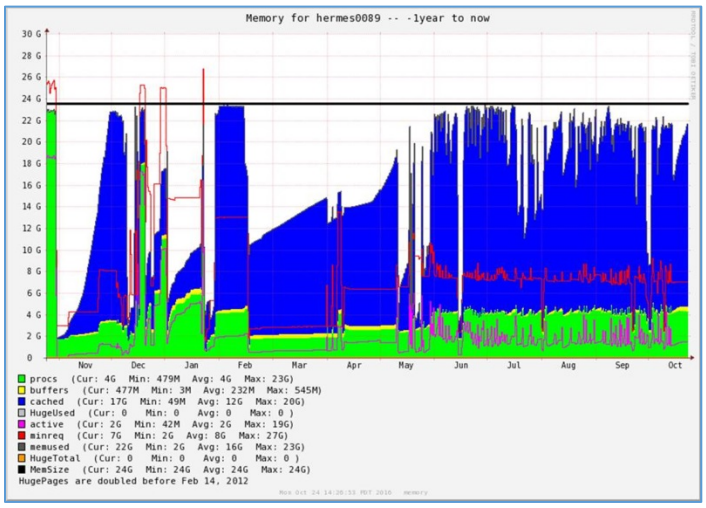


Figure 5. A month-to-month varied iteration lengths with 24GB ram consumption on Hermes89 node reported from WestGrid showing variation in the duration of the ingestion of 50 Million records over each of the iterations over an entire year (November 2015 to October 2016), with more activity mid-May to October. The graph shows the following cached memory (in blue), active passes (in green), buffers (in yellow), as well as active (in pink), minimum required (in red) and available memory (black line).

Four CSV files were ingested to HDFS: 1) one with 10 records was used for quick testing; 2) one with 50 million records was used for the actual benchmarking; and, 3) another two with one and three billion records. The results for 3 billion are shown in Table 3 (in Appendix) while the other results for 50 million and 1 billion were previously published [51]. Furthermore, the results also revealed that unbalanced nodes had slower performance compared to more balanced nodes (after manually running compression was run for 24 hours on HBase store files).

There is a tool called Hannibal available to check Regions over Hadoop/HBase and showed that the five Regions averaged ~1TB and this coincided with 176-183 Store files for each (comprised of Hadoop's HFiles) for the entire distributed table (ingested into HBase) for three billion records. Total equals 900 Regions and 900 Store files. The slight imbalance and an initial peak in the sizes because compaction was disabled to run automatically (due to performance issues) and manually running bot minor (combines configurable number of smaller HFiles into one larger HFile) and major reads the Store files for a region and writes to a single Store file) compaction types. Furthermore, the durations of the iterations continued to fluctuate with no distinctive trend for one billion. For three billion, at more operational level at high core CPU over May-October 2016, had similar fluctuations of 299-1600 minutes with average of ~300-375 minutes for ingestions and roughly the same amount of time for compaction (major and minor) manually run before the next ingestion iteration, which accounted for 2 months to ingest files to reach three billion.

Another major challenge faced was that Apache Drill did not have any option to force it to bind to a specific network interface. As a result, its Drillbits started used the network management instead of the InfiniBand network. Zeppelin used databricks from Spark to ingest the file that was run over the interface. Similarly, it also intermittently occurred when initializing Zeppelin and this added an additional 30 minutes to the ingestion and query of 1-3 billion rows. In addition to having to move from the 0.5.0 version (as it did not support Spark 1.5.2) to the latest snapshot (0.6.0), the Zeppelin's Pyspark timed out the first time %pyspark was run. In fact, it was found that the user has to wait for a couple of seconds and execute the cell one more time to get Pyspark working. Another challenge with Zeppelin, even though it offers a pull-down list of notebooks, was that the notebook could only be exported as a JSON file. Saving the notebook as an html page produced empty webpages and that exhibited very poor usability. Therefore, the Zeppelin notebooks can only be open under Zeppelin. There were, however, some off-the-shelf buttons to plot data as a data visualization tool for Zeppelin.

4.2. Data Size and Ingestion Time

Three CSV files ingested to HDFS: 1) 10 records and was used for quick testing, and 2) others used for the actual benchmarking: 50 million records and the other with one and three billion records. Table 4 shows results of ingestion times.

Table 4. Hadoop ingestion time (minutes) for 50 million to 3 billion records.

Data Size	50 Million Records (23 GB)	1 Billion Records (451 GB)	3 Billion Records (10 TB)
Ingestion Time	~6 min	~2 h 5 min	~5 h 2 min

If we set equation (1) Ti (N) to be the time to ingest N records, then the data ingestion efficiency (IE) for 50 million (50M), one billion (1B) and three billion (3B) was therefore:

$$IE = \frac{1(3)B \times T_i(50M)}{50M \times T_i(1(3)B)} = 0.93 \quad (1)$$

4.3. SQL Query Results

The queries were run on Zeppelin, spark-shell, and Pyspark; all took approximately the same amount of time. This was not surprising, as they all rely on Spark SQL. Therefore, it was only reported as a single time for all three. Spark was configured to run on a Yarn-client with 10 executors, four cores with 8 GB of RAM each; therefore, each node had two executors with a total of eight cores and 16 GB memory (tests using one executor with 16 GB and eight cores on each node was slightly less efficient).

The Zeppelin code snippet used was:

```
%pyspark
from time import time t0 = time()
df=sqlContext.read.format('com.databricks.spark.
csv').option('header','true').option('inferS
chema','true').load('hdfs://h
ermes0090:54310/dad/big_dad_all.csv')
df.registerTempTable('DADS1') df.cache()
```

```
%sql SELECT Count(Episode_Duration) as
EDCount, Count(Anesthetistic_Technique) as
ATCount, Count(Interven_Location) as ILCount,
Count(Medical_Services) as MSCount,
Count(Unit_Transfer_Occurrence) as UTOCount from
DADS1 where EncounterID<1000 Group By age
```

Drill was configured to run the Drillbits on the Hadoop DataNodes (with a single Drillbit on each node). Each Drillbit was configured with a VM heap of 4 GB and a maximum direct memory of 16 GB (standard configuration). The queries were run from both the web interface of one of the Drillbits and the sqlline command line, using the code snippet below as example. Both gave the same timing.

```
SELECT Count(Episode_Duration) as EDCount,
Count(Anesthetistic_Technique) as ATCount,
Count(Interven_Location) as ILCount,
Count(Medical_Services) as MSCount,
Count(Unit_Transfer_Occurrence) as UTOCount from
dfs.`/dad/big_dad_all.csv` where
EncounterID<1000  Group By age
```

The results clearly showed that Drill was five to 7.6 faster than Spark, which indeed justified its use as a low latency query-engine tool. The Phoenix and HBase benchmarking reported earlier [51] showed that HBase outperformed all of them and the HBase queries were all within a second or two. This was an astonishing performance and HBase was used, when possible, as the underlying data source for the HBDA platform.

It was surprising that the query efficiency (QE) of Drill was only 76%. It was believed at the time of running both engines that this was due to a lack of binding to the InfiniBand interface. The Drill Developers were contacted about this and there was work with them to debug for an eventual break fix. Drill's query processes, however, were still not as efficient as that of Spark with increased database size to ingest.

We defined Tq (N) to be the query time and its efficiency (QE) in equation (2) for 50 million (50M), one billion (1B) and three billion (3B) was defined as:

$$QE = \frac{1(3)B \times T_q\,(50M)}{50M \times T_q\,(1(3)B)} \quad (2)$$

The results of the benchmarking in Table 5 clearly shows that Drill was 5 to 7.6 was faster than Spark and indeed justified its use as a low latency query engine tool.

Our Phoenix and HBase benchmarking reported earlier showed that HBase outperformed all of them and the HBase queries were all within a second or two [51]. These results were astonishing and high performance using HBase when possible, as the underlying data source for our HBDA platform. Furthermore, Spark did show improved query efficiencies from three to six billion patient records compared to less relative improvement by Drill (Figure 6).

Table 5. SQL query time (seconds) for 50 million to 3 billion records.

SQL Engine	Spark SQL (seconds)	Drill SQL (seconds)
50 Million Records	194.4	25.5
1 Billion Records	3409.8	674.3
3 Billion Records	5213.3	1543.2
Query Efficiency	1.14-1.52	0.76-0.84

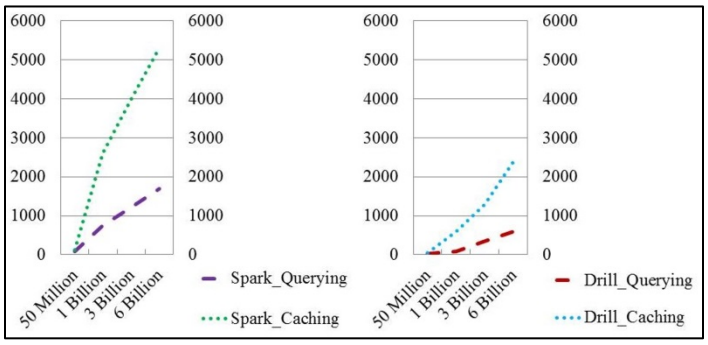


Figure 6. Projected ingestion and query results for 50 million to 6 Billion patient records using Apache Spark or Apache Drill Systems.

4.4. Usability, Simple Analytics and Visualizations

The results showed that the ingestion time of one billion records took circa two hours via Apache Spark. Apache Drill outperformed Spark/Zeppelin and Spark/Jupyter. However, Drill was restricted to running more simplified queries, and was very limited in its visualizations that exhibited poor usability for healthcare. Zeppelin, running on Spark, showed ease-of-use interactions for health applications, but it lacked the flexibility of its interface tools and required extra setup time and 30-minute delay before running queries. Jupyter on Spark offered high performance stacks not only over the BDA platform but also in unison, running all queries simultaneously with high usability for a variety of reporting requirements by providers and health professionals.

Being able to perform low latency SQL queries on a data source is not enough for healthcare providers, clinicians, and practitioners. Interacting with and exploring the data through different analytics algorithms and visualizations is usually required to get the data's full value. A variety of functionalities and tools for expressing data was an essential quality to test over the platform.

Drill did perform well compared to Spark, but it did not offer any tools or libraries for taking the query results further. That is, Drill proved to have higher performance than Spark but its interface had less functionalities. Moreover, algorithms (as simple as correlations between different columns) were time-demanding if not impossible to express as SQL statements. Zeppelin, on the other hand, offered the ability to develop the code, generate the mark-down text, and produced excellent canned graphs to plot the patient data.

Combined with the richness of Spark and Pyspark, Zeppelin provided a canned visualization platform with graphing icons. The plots under Zeppelin, however, are restricted to the results/tables obtained from the SQL statements. Moreover, algorithms (as simple as correlation between different columns) were time demanding if not impossible to express as SQL statements. Zeppelin, on the other hand, offered the ability to develop the code, generate the mark-downed text, and produced excellent canned graphs to plot the patient data (Figure 7).

Combining with the richness of Spark and Pyspark, Zeppelin provided a canned visualization platform with graphing icons (Figure 8). The plots, however, under Zeppelin were restricted to the results/tables obtained from SQL statement. Furthermore, the results that we directly produced from Spark SQL context did not have any visualization options in Zeppelin. Generating results

from queries via Zeppelin took much longer (over 30 minutes) to establish the platform to run queries on the interface compared to Jupyter.

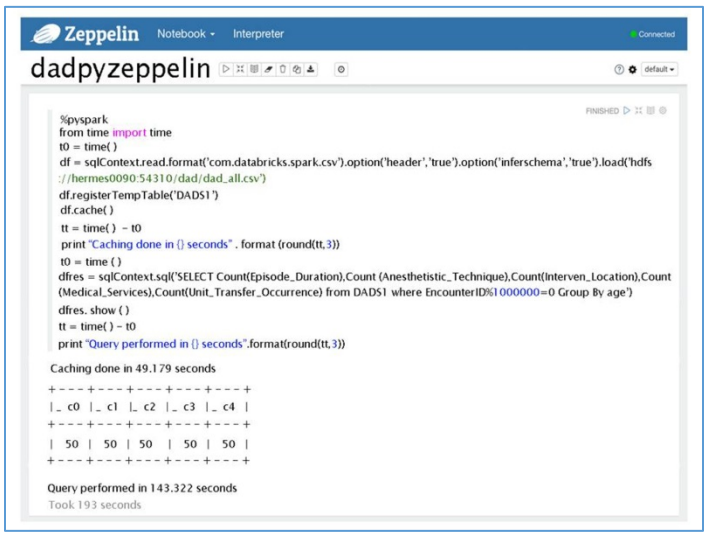


Figure 7. A web-based interface for Zeppelin showing embedded queries and the results.

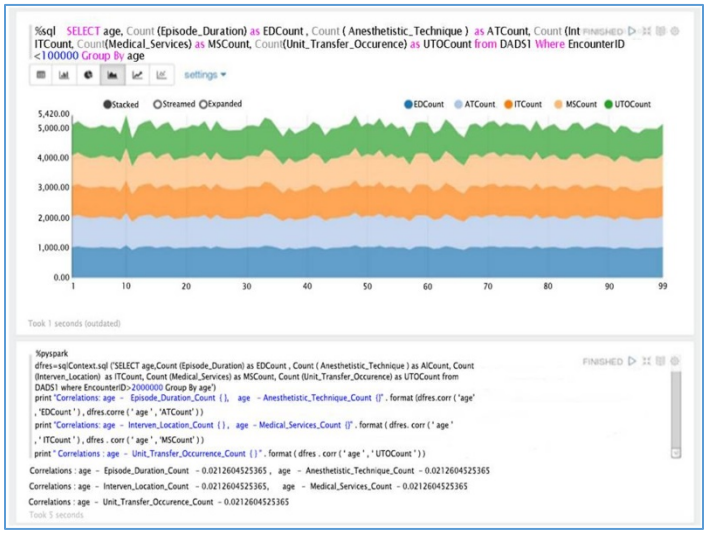


Figure 8. An example of visualization and simple correlation analytics results within Zeppelin using Pyspark.

Furthermore, the results that were produced directly from the Spark SQL context did not have any visualization options in Zeppelin. Generating results from queries via Zeppelin took much longer (over 30 minutes). Establishing the platform to run queries on the interface and generate results via Zeppelin took longer than Jupyter.

With Jupyter, more configurations with the data queries were tested. It exhibited similar code to ingest the file, same Spark databricks initialized in the interface and its SQL to query as Zeppelin (Figure 9).

At the expense of writing the visualization code, using the *matlplotlib* Python package in addition to other powerful tools, such as Pandas, i.e., a powerful Python data analysis toolkit. The local host was added to Hermes node to access Jupyter via the BDA platform to compensate for the lack of visualization options via the Zeppelin interface. Figure 10 shows a small snippet from the output of a Jupyter/Spark interaction that uses both *matlplotlib* and Java's Pandas.

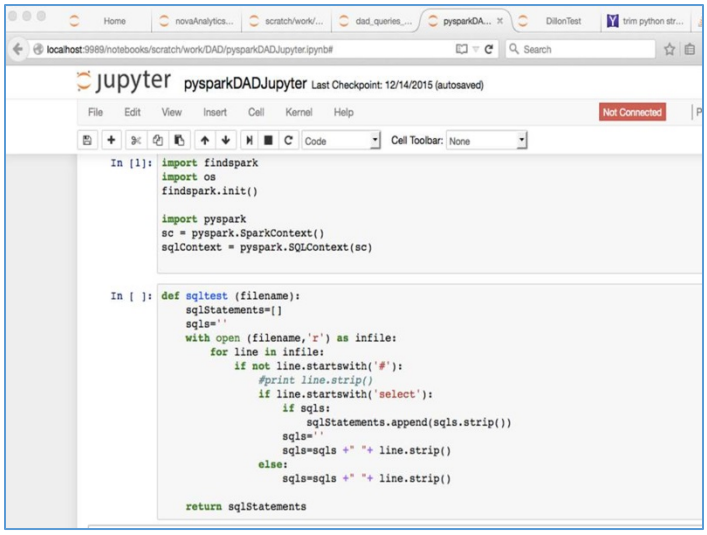


Figure 9. Spark with Jupyter and loading file before SQL is placed.

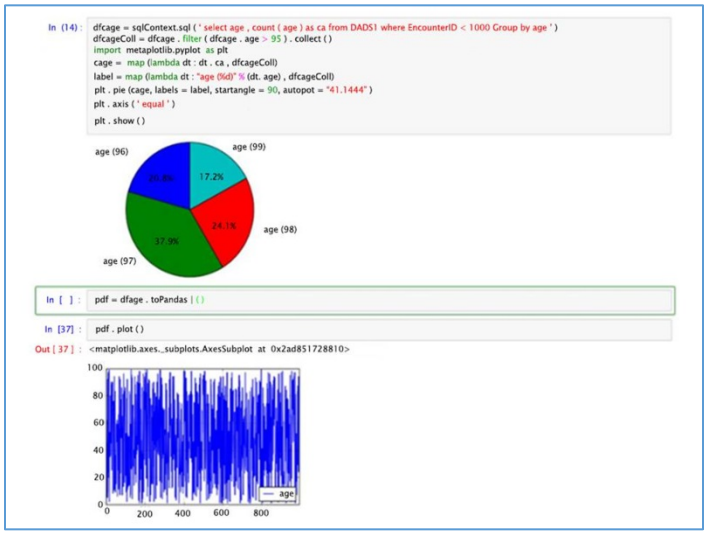


Figure 10. A simple Jupyter/Spark interaction on its web-based interface with data visualizations using Pandas and graphed results.

Usability of the platform did validate the proof-of-concept of querying patient data with high performance in generating results and visualization over the interface. Performance to generate results had the same number of sequence steps for the end users as HBase, Spark and Drill.

Running Apache Spark took the same amount of time for queries on Zeppelin and Jupyter; however, the initial setup (to start the queries) took much longer via Zeppelin than Jupyter. Drill seemed to have simplified steps to the setup interface compared to Spark and took significantly less time; therefore, it appeared to have better usability. Nevertheless, Jupyter supplied more visualization defaults and customization than Drill for its distributed mode and its interface to run the query where severely lacking any visualization.

Usability testing of our HBDA by health professionals is limited in this study. Running modules in sequence from Hadoop to Spark or Drill with web clients is too technical for most end

users and it would require additional refinements to the interface for producing clinical reports. We placed code in both Zeppelin and Jupyter Notebooks to run all queries at once over the database and Hadoop running. However, the only stakeholders that would benefit from this code change or utilization would be the data warehouse team.

5. Discussion

A Hadoop/MapReduce framework successfully formed our HBDA platform for health applications. Very few studies have applied big data technologies to patient data of hospital system for healthcare applications. Moreover, no studies have tested a variety of big data tools in Hadoop's ecosystem of packages. The platform successfully implemented a BDA platform and tested it for healthcare applications with moderate resources and able to run realistic ANSI-SQL (Drill) SQL-like (Spark) queries on three billion records and perform interactive analytics and data visualizations. An integrated solution eliminates the need to move data into and out of the storage system while parallelizing the computation, a problem that is becoming more important due to increasing numbers of sensors and resulting data in healthcare. Furthermore, usability goals based end user computing of the Big Data technologies and leveraging the existing tools from data warehouse at a health authority is important to make a stand on to use the best Apache tools. In this case, Apache Drill and Spark with Zeppelin or Jupyter proved to be an important test over the one-three billion records because they had not only different performances but also much greater differences in their usability during operational simulation. Therefore, our study accomplished a working Hadoop ecosystem that is applicable to large volumes of patient data.

The sheer volume of 3 billion indexed and ~9-10 billion generated shows that one platform could not only be operational and productive for one hospital but many and even at a provincial scale in Canada or statewide in other countries. The volume achieved at a productive and operational level in our study can also further lead to simulations that are more rigorous. Part of the established simulations and its representation of health informatics metadata were formulated in the formation of NoSQL HBase database, as well as the ANSI-SQL (Drill) SQL-like (Spark) data queries displaying results. Few studies have produced a platform of the NoSQL database that tested ANSI-SQL and SQL-like data queries of patient data of a hospital system and this study adds to big data technologies, data simulation and healthcare applications over large volumes. Hence, this study achieved the three V's that define Big Data [4]: high performance (velocity) over its generator of detailed data (variety) that formed extremely large volumes (volume) significantly contributed to ongoing development of Information Management and Information Technology (IMIT) in healthcare.

While ingestions in this study were extremely fast, the bulkloads of 50 million rows in iteration to one and three billion were slow and took, collectively, one week and up to two months, respectively. Ingestions beyond three billion were even slower, but these times are as fast as or faster than current data migrations (of a similar size) estimated at VIHA. The ingestion times achieved required several reconfigurations of HBase and Hadoop to increase the time to distribute the data; these involve changes in the site, yarn, and RegionServer XMLs. MapReduce was to blame for the time needed to ingest 50 million rows varied widely (from 2 to 12 hours). The corresponding Java coding, Java Virtual Machines (JVMs), and Java services were a performance bottleneck, which is common on most platforms [22, 39], especially memory loss due to Reducer process of MapReduce [65]. This demonstrated operationally that while Hadoop/MapReduce did have high performance efficiencies its clusters did break that required ongoing maintenance and this is common across Hadoop clusters (cf. [66]). However, significance of using the MapReduce programming model on top of the Hadoop cluster proved a process of large volumes of clinical data can be accomplished. More importantly, query times were less than two seconds for all queries, which is significantly faster than current estimated query times. Since there were no differences in query durations observed, and since HBase is linearly scalable, it is expected that query durations would decrease with an increase in the number of nodes and be within a few milliseconds as nodes approached 100, even at ten billion rows.

Clearly, this study showed that Drill, a software addition to Hadoop developed recently in 2015-2016 [62-63], significantly outperformed the other Apache tools. Drill outperformed Spark and Phoenix in HBase processes before queries generated. However, its interface lacked any functionality to customize and mine the data, which is what health professionals require (because of the complexity of the data and what its clinical reporting should reveal). Furthermore, running modules Spark or Drill with web clients over Hadoop cluster is far too technical for most end users to generate clinical reports. This poor usability contradicts the recommendations by Scott [62] in that in most cases Drill should be used instead of Spark in this showdown of using SQL over big data platforms. Drill had on Drill's interface to test usability but Spark had other interfaces (like Jupyter and Spark), which no other studies have produced results on. Furthermore, this study provided new insight into more customized Java coding with the combination of Jupyter with Spark that enhanced the platform's dynamic interactivity. It is therefore recommended that Spark with Jupyter be used with scripted coding to run ingestion and queries simultaneously. The code can be placed in both Zeppelin and Jupyter Notebooks to run all queries at once over the database. However, the only stakeholders that would truly benefit from this code change and its utilization would be the data warehouse team. Besides, Scott [62] indicated that the battlefield for the best Big Data software solutions is between Spark and Drill and that Drill can emulate complex data much more efficiently than Spark because Spark requires elaborate Java, Python and Scala coding to do so. Nonetheless, both Spark and Drill were significantly faster than HBase in ingesting files directly into Hadoop via Drillbits (Drill) with ZooKeeper and MapReduce, and RRD transformations with MapReduce (Spark). In fact, the ingestion and queries for both Spark and Drill could be run in sequence instead of having to run compaction as well. However, it is

difficult to compare since neither Spark nor Drill indexed the files. Absence of indexing increases the risk of inaccuracies (even though the framework was more fault-tolerant when running Spark and Drill). Therefore, the big data tools and inherent technologies highly influence the health informatics of the data established and resulting data from queries.

The most impactful technology of the Big Data technical components in this study was MapReduce (and Java code therein). MapReduce methodology is inherently complex as it has separate Map and Reduce task and steps in its default-programming framework as this study discovered. This study's platform was highly dependent on the efficiency of MapReduce in ingesting files over the six nodes, using this workflow: Input → Map → Copy/Sort → Reduce → Output similar to a study by Chen, Alspaugh, and Katz [67]. The Map part of the platform showed high performance but the Reduce took more than tenfold longer to complete its schedule; however, once configurations in Yarn, ZooKeeper, and others the Reducer optimized at iterations of 50 million rows. According to blogs and technical resolutions involved enabling or disabling services or xml settings over the platform as expected to be carried because the system relied heavily on InfiniBand (IB) bandwidth at low latency over WestGrid nodes. Furthermore, there are known issue with the combination of MapReduce to HBase, although studies have shown that additional indexing and reduction processes can be added and/or modified at the reducer with an advanced programming method [22, 23, 39, 68-69]. However, a customized reduction at the Reducer level of this platform proved to be difficult to overcome and maintain at less than 3 hours for each of the iterations at only 50 million rows at file size 258GB.

The complex nature of HBase means that it is difficult to test the robustness of the data in emulations based on real data. This complexity somewhat rejects our hypothesis that noSQL database accurately simulates patient data. Nevertheless, several steps are standardized by hospitals to prepare the DAD database for statistical rendering to CIHI. Moreover, the actual columns used in this study are similar ones used by VIHA. Additionally, the DAD data also makes calculations by add columns in the data warehouse. Adding columns to a NoSQL database is much easier than adding columns to a SQL relational database, and von der Weth and Datta [70] showed good performance of multi-term keyword searches over noSQL. Therefore, it is an advantage to have a large database with row keys and column families already set up; Xu et al., [36] support this, as their middleware ZQL could easily convert relational to non-relational data.

Essentially this study is proposing a row-column key-value (KV) model to the data distributed over a customized BDA platform for healthcare application. Wang, Goh, Wong, and Montana [71] support this study's claim in their statement that NoSQL provided high performance solutions for healthcare, being better suited for high-dimensional data storage and querying, optimized for database scalability and performance. A KV pair data model supports faster queries of large-scale microarray data and is implemented using HBase (an implementation of Google's BigTable storage system). The new KV data model implemented on HBase exhibited an average 5.24-fold increase in high-dimensional biological data query performance compared to the relational model implemented on MySQL Cluster and an average 6.47-fold increase on query performance on MongoDB [22]. Freire et al., [40] showed highest performance of CouchDB (similar to MongoDB and document store model) but required much more disk space and longer indexing time compared to other KV stores. The performance evaluation found that KV data model, in particular its implementation in HBase, outperforms and, therefore, supports this studies use of NoSQL technology for large-scale BDA platform for a hospital system. HBase schema is very flexible, in that new columns can be added to families at any time; it is therefore able to adapt to changing application requirements [72-73]. HBase clusters can also be expanded by adding RegionServers hosted on commodity class servers, for example, when a cluster expands from 10 to 20 RegionServers, it doubles both in terms of storage and processing capacity. Sun [74] lists the following notable features of HBase: strongly consistent reads/writes; "Automatic sharding" (in that HBase tables distributed on the cluster via regions can be automatically split and re-distributed as data grows); automatic RegionServer failover; block cache and "blooming" filters for high-volume query optimization; and built-in web-applications for operational insight along with JMX (i.e., Java memory) metrics. However, HBase settings had to be purged and cleaned after each of the ingestions due to unknown tracers or remnants of transactions that then later caused failure, and compaction was run manually to improve performance; therefore, robustness of HBase needs further investigation.

The present study showed that performing maintenance and operational activities over the platform were essential for high availability. Unbalanced ingestions required removing files and starting again. Some studies have shown that Hadoop can detect task failure and restart programs on healthy nodes, but if the RegionServers for HBase failed, this process had to be started manually and other studies confirm this problem [37, 39, 75-77]. Our study showed that compaction improved the number of successful runs of ingestion; however, it did not prevent failure of the nodes, a finding that is supported by other studies [23, 24, 31, 68, 78]. If a node failed, the partly ingested file had to be cleaned up, re-run, and re-indexed. The platform, therefore, showed a single point of failure.

Data privacy in healthcare involves restricted access to patient data but there are often challenging situations when using hospital systems and attempting to find new trends in the data. For instance, on one hand there are workarounds to access patient data in critical situation like sharing passwords that goes against HIPAA and FIPPA Acts [79]. There are strict rules and governance on hospital systems with advanced protection of privacy of patient data based on HIPAA [80-81] that must take into consideration when implementing a BDA platform. It's

processing and storage methods must adhere to data privacy at a high level and also the accessibility of the data for public disclosure [82-84]. One method of ensuring that patient data privacy/security is to use indexes generated from HBase, which can securely encrypt KV stores [36, 85-87], and HBase can further encrypt with integration with Hive [35]. Scott [62] also stated that Drill is already setup for encryption for HIPPA but we did not find this out-of-the-box and attempting to encrypt was time consuming.

There are a large number of developers working on additional libraries for Hadoop like Lucene and Blur [38]. For example, Hadoop R, in particular, provides a rich set of built-in as well as extended functions for statistical, machine learning, and visualization tasks such as: data extraction, data cleaning, data loading, data transformation, statistical analysis, predictive modeling, and data visualization. Also, SQL-like queries can be run via Hive as a data warehouse framework for ad hoc querying that can be used with HBase, although no real-time complex analyses can be performed [35]. More investigation of this study's different libraries of a variety of packages offered in Hadoop's ecosystem (many of have not been used in healthcare applications) is crucial to ascertaining the best possible BDA platform.

Conclusion

Our HBDA platform showed high performance tested for healthcare applications. With moderate resources, we were able to run realistic SQL queries on three billion records and perform interactive analytics and data visualization using Drill, Spark with Zeppelin or Jupyter. The performance times proved to improve over time with repeated sessions of the same query via the Zeppelin and Jupyter interfaces. An ingesting and using CSV file on Hadoop also had its advantages (i.e. simplicity, CSV exports and imports commonly carried out in healthcare applications, fast ingestion compared to HBase) but was expensive when running Spark. Drill offers better low latency SQL engine but its application tool and visualization were very limited to customization, and, therefore, had lower usability. Useful knowledge gained from this study included the following challenges and specific objectives:

(1) data aggregation – actual storage doubled compared to what was expected due to HBase key store qualifiers, Spark and Drill had the same procedure to ingest Hadoop file before running SQL queries;

(2) data maintenance – ingestions to the database required continual monitoring and updating versions of Hadoop-MapReduce and HBase with limitations persisting for MapReduce (ultimately Java performance in the Reducer) from one to three billion;

(3) data integration –

i. combination of ADT and DAD possible with simulated patient data and followed current clinical reporting but required a data model of the row keys and column families and this needs to be further tested;

ii. study's three billion indexed data at 30TB equalled six times more rows than current production and archived at most health authorities, which is said to be 500 million rows on average for a health authority with up to three billion for the entire Province;

iii. large volumes at different scales, i.e. hospital, health authority, Province, and multiple Provinces, can be achieved if ADT and DAD can be formed to flat file of CSV format

iv. data model was only verified via simplified analytical queries of simulated data as a benchmark test, but not fully integrated to a defined patient data model and health informatics.

(4) data analysis – high performance of 3.5 seconds for three billion rows and 90 columns (30TB of distributed files) achieved with increasing complexity of queries with high performance of Drill to run queries and high usability with customized Spark with Jupyter; and

(5) pattern interpretation of application – randomized patterns found via Spark with Jupyter interface; however, health trends cannot be found via the application and further investigation required using Hadoop's Machine Learning Libraries (MLLib).

Conflict of Interest

Authors have no conflict of interest.

Acknowledgment

The authors would like to acknowledge WestGrid for their support. This project was partially supported by 2015-2016 Research Seed Grant Competition, VIHA. Special thanks go to Mr. Hamid Zamani for his exceptional work as research assistant.

References

[1] J. Gantz, D. Reinsel, "The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East," Study report, IDC,URL [www.emc.com/leadership/digital-universe/index.htm. (2012).

[2] J. Sun, C.K. Reddy, "Big Data Analytics for Healthcare," Tutorial presentation at SIAM Inter. Conf. Data Mining, Austin, TX., (2013).

[3] J. Manyika, M. Chui, J. Bughin, B. Brown, R. Dobbs, C. Roxburgh, B. Hung, "Big Data: The next frontier for innovation, competition, and productivity," URL:http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation, (2014).

[4] Canada Health Infoway, Big Data Analytics: Emerging Technology Series, White Paper (Full Report), Ottawa, ON, (2013).

[5] World Health Organization (WHO), Atlas of eHealth Country Profiles – The use of eHealth in support of universal health coverage, WHO Press, Geneva, Switzerland, ISBN 978-92-4-156521-9, (2016).

[6] J. Alder-Milstein, A.K. Jha, "Healthcare's "Big Data" challenge," Am. J. Manag. C., **19**(7): 537-560 (2013).

[7] E. Baro, S. Degoul, R. Beuscart, E. Chazard. "Toward a literature-drive definition of big data in healthcare," BioMed Res. Intern., ID 639021, 9 pages, (2015).

[8] M.-H. Kuo, T. Sahama, A.W. Kushniruk, E.M. Borycki, D. Grunwell, "Health Big Data Analytics: Current Perspectives, Challenges and Potential Solutions," Int. J. Big Data Intel, **1**(12): 114–126 (2014).

[9] M.M. Hansen, T. Miron-Shatz, A.Y.S. Lau, C. Paton, "Big Data in Science and Healthcare: A Review of Recent Literature and Perspectives," Yearb. Med. Inform., **9**(1): 21-26 (2014).

[10] R. Nelson, N. Staggers, Health Informatics: an interprofessional approach, Mosby, imprint of Elsevier Inc.; 2014. Saint Louis, MO, (2014).

[11] B. Wang, R. Li, W. Perrizo, Big Data Analytics in Bioinformatics and Healthcare, 1st edition, Med. Info. Sci. Ref. - IGI Global, Kansas City, MO, (2014).

[12] D. Agrawal, P. Bernstein, E. Bertino, S. Davidson, U. Dayal, M. Franklin, J. Gehrke, L. Hass...etc, "Challenges and Opportunities with Big Data," Big Data - White Paper, Computing Research Association, (2012).

[13] H. Chen, H.L. Chiang, C. Storey, "Business intelligence and analytics: from Big Data to big impact," MIS Q **36**(4): 1-24 (2012).

[14] H. Demirkan, D. Delen, "Leveraging the capabilities of service-oriented decision support systems: Putting analytics and Big Data in cloud," Decis Support Syst., **55**(1): 412-421 (2013).

[15] L.P. Jr. Garrison, "Universal Health Coverage-Big Thinking versus Big Data," Value Health **16**(1): S1-S3 (2013).

[16] N.H. Shah, D. Tenenbaum, "The coming age of data-driven medicine: translational bioinformatics' next frontier," J. Am. Med. Inform. Assoc., 19: e2-e4 (2012).

[17] T. White, Hadoop – The Definitive Guide: Storage and analysis at internet scale, O'Reilly Media, San Francisco, CA, ISBN 978-1-491-90163-2, (2015).

[18] Hadoop, "Apache Hadoop," [http://Hadoop.apache.org.], (2016).

[19] R. Jurney, Agile data science: building data analytics applications with Hadoop, O'Reilly Media, San Francisco, CA, (2013).

[20] R. Karim, C.F. Ahmed, B.-S. Jeong, H.J. Choi, "An Efficient Distributed Programming Model for Mining Useful Patterns in Big Datasets," IETE Tech. Rev., **30**(1): 53-63 (2013).

[21] P. Langkafel, Big Data in Medical Science and Healthcare Management.: Diagnosis, Therapy, Side Effects, Walter de Gruyter Verlag GmbH, Berlin/Boston, (2016).

[22] S. Sakr, A. Elgammal, "A. Towards a comprehensive data analytics framework for smart healthcare services," Big Data Research, **4**: 44-58 (2016).

[23] R.C. Taylor, "An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics," BMC Bioinformatics, **11**(12): S1 (2010).

[24] J. Dean, S. Ghemawat, "MapReduce: A Flexible Data Processing Tool," Comm. ACM, **53**(1): 72-77 (2010).

[25] W. Raghupathi, V. Raghupathi, "Big data analytics in healthcare: promise and potential," Health Info. Sci. Syst., **2:3**, 10 (2014).

[26] E.A. Mohammed, B.H. Far, C. Naugler, "Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends," BioData Mining, 7(22): 23 (2014).

[27] T. Dunning, E. Friedman, Real-World Hadoop, O'Reilly Media, San Francisco, CA, (2010).

[28] ZooKeeper, "ZooKeeper - Apache Software Foundation project home page," [http://Hadoop.apache.org/ZooKeeper/], (2016).

[29] HBase, "HBase - Apache Software Foundation," [http://Hadoop. apache.org/HBase/], (2016).

[30] R.S. Chang, C.-S. Liao, K.Z. Fan, C.-M. Wu, "Dynamic Deduplication Decision in a Hadoop Distributed File System," Int. J. Distrib. Sens. Networks, Article ID 630380, 14 (2014).

[31] A.B.M Moniruzzaman, S.A. Hossain, "NoSQL Database: new era of databases for Big Data Analytics – Classification, Characteristics and Comparison," Int. J. Database Theo. App., **6**(4): 1-14 (2013).

[32] H. Karau, A. Konwinski, P. Wendell, M. Zatharia, Learning Spark: Lightning-Fast Big Data Analysis, O'Reilly Media, San Francisco, CA, (2015)

[33] K. Sitto, M. Presser, Field Guide to Hadoop - An Introduction to Hadoop, Its Ecosystem, and Aligned Technologies, O'Reilly Media, San Francisco, CA, (2015).

[34] Hive, "Hive - Apache Software Foundation project home page," [http://Hadoop.apache.org/hive/], (2016).

[35] Hive HBase, "Hive-HBase Integration project home page," [http://wiki.apache.org/Hadoop/Hive/HBaseIntegration], (2016).

[36] J. Xu, M. Shi, C. Chen, Z. Zhang, J. Fu, C.H. Liu, "ZQL: A unified middleware bridging both relational and NoSQL databases," IEEE 14th Int. Conf. Dependable, Autonomic and Secure Computing, Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress, 730-737, (2016).

[37] A. Lith, J. Mattson, "Investigating storage solutions for large data: a comparison of well performing and scalable data storage solutions for real time extraction and batch insertion of data," MSc Thesis, Chalmers University of Technology, Göteborg, Sweden, (2010).

[38] W. Seo, N. Kim, S. Choi, "Big Data Framework for Analyzing Patents to Support Strategic R&D Planning." IEEE 14th Int. Conf. Dependable, Autonomic and Secure Computing, Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress, 746-753 (2016).

[39] M. Maier, "Towards a Big Data Reference Architecturem," MSc Thesis. Eindhoven University of Technology, Netherlands, (2013).

[40] S.M. Freire, D. Teodoro, F. Wei-Kleiner, E. Sundsvall, D. Karlsson, P. Lambrix, "Comparing the Performance of NoSQL Approaches for Managing Archetype-Based Electronic Health Record Data," PLoS ONE **11**(3): e0150069 (2016).

[41] H, Nordberg, K. Bhatia, K. Wang, Z. Wang, "BioPig: a Hadoop-based analytic toolkit for large-scale sequence data," Bioinformatics **29**(23): 3014–3019 (2013).

[42] N.A. Miller, E.G. Farrow, M. Gibson, L.K. Willig, G. Twist, B. Yoo, T. Marrs, S. Corder, L. Krivohlavek, A. Walter, J.E. Petrikin, C.J. Saunders, I. Thiffault, S.E. Soden, L.D. Smith, D.L. Dinwiddie, S. Herd, J.A. Cakici, S. Catreux, M. Ruehle, S.F. Kingsmore, "A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases," Genome Med. Sep 30, **7**(1): 100 (2015).

[43] G.P. Twist, A. Gaedigk, N.A. Miller, E.G Farrow, L.K. Willig, D.L. Dinwiddie, J.E. Petrikin, S.E. Soden, S. Herd, M. Gibson, J.A. Cakici, A.K. Riffel, J.S. Leeder, D. Dinakarpandian, S.F. Kingsmore, "Constellation: a tool for rapid, automated phenotype assignment of a highly polymorphic pharmacogene, CYP2D6, from whole-genome sequences," NPJ Genomic Med. **1**, 15007, (2016).

[44] C.J. Saunders, N.A. Miller, S.E. Soden, D.L. Dinwiddie, A. Noll, N.A. Alnadi, et al., "Rapid Whole-Genome Sequencing for Genetic Disease Diagnosis in Neonatal Intensive Care Units," Sci. Trans. Med. **4**(154): 154ra135 (2012).

[45] C.J.M. Tauro, S. Aravindh, A.B. Shreeharsha, "Comparative Study of the New Generation, Agile, Scalable, High Performance NOSQL Databases," Int. J. Comp. App. **48(**20): 1-5 (2012).

[46] J.M Tien, "Big Data: unleashing information," J. Syst. Sci. Syst. Eng., **22**(2): 127-151 (2013).

[47] C. Zhang, "Supporting multi-row distributed transactions with global snapshot isolation using bare-bones HBase," GRID, 11th IEEE/ACM Int. Conf., 25-28, Waterloo, Canada, (2010).

[48] K. Moselle, "Data Management in the Island Health Secure Research Environment," Enterprise Architecture at Vancouver Island Health Authority, Working Draft 5, Victoria, BC, (2015).

[49] E. Pattuk, M. Kantarcioglu, V. Khadilkar, H. Ulusoy, S. Mehrotra, "BigSecret: A secure data management framework for key-value stores," Tech. Rep. [http://www.utdallas.edu/~exp111430/techReport.pdf] [Access December 2015], (2013).

[50] M.K. Ross, W. Wei, L. Ohno-Machado, "Big Data" and the Electronic Health Record," Yearb. Med. Inform., 97-104, (2014).

[51] D. Chrimes, M.-H. Kuo, B. Moa, W. Hu, "Towards a real-time big data analytics platform for health applications", Int. J. Big Data Intell., Vol. 4, No. 2, pp.61–80, (2017).

[52] H. Chen, S.S Fuller, C. Friedman, W. Hersh, "Knowledge management, data mining, and text mining in medical informatics," In: H. Chen, S.S. Fuller, C. Friedman, W. Hersh (Eds.). Medical Informatics: knowledge management and data mining in biomedicine. Springer, pp. 20-40, (2005).

[53] D. Li, H.W. Parl, M.I. Ishag, E. Batbaatar, K.H. Ryu, "Design and Partial Implementation of Health Care System for Disease Detection and Behavior Analysis by Using DM Techniques," IEEE 14th Int. Conf. Dependable, Autonomic and Secure Computing, Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress, 781-786, 2016.

[54] D. Chrimes, B. Moa, H. Zamani, M-H. Kuo, "Interactive Healthcare Big Data Analytics Platform under Simulated Performance," IEEE 14th Int. Conf. Dependable, Autonomic and Secure Computing, Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress, 811-818.

[55] F. Chang, J. Dean, S. Ghemawat, W.C. Hsieh, D.A. Wallach, M. Burrows, T. Chandra, A.E. Fikes, "Bigtable: A distributed storage system for structured data," Seventh Symposium on Operating System Design and Implementation (ODI) Seattle, WA: Usenix Association, 205-18, (2006).

[56] M.-H. Kuo, D. Chrimes, B. Moa, X. Hu, "Design and Construction of a Big Data Analytics Framework for Health Applications," IEEE Proceedings Int. Conf. on Smart City/SocialCom/SustainCom together with DataCom 2015 and SC2 2015, Chengdu, China. 631-636 (2015).

[57] Canadian Institute of Health Information (CIHI), "DAD Abstracting Manual: Province-Specific Information for British Columbia," Ottawa, ON. CIHI Publishing, (2012).

[58] Cloudera, "Integrating Hive and HBase - Cloudera Developer Center," [http://http://www.cloudera.com/blog/2010/06/integrating-hive-and-HBase/], (2016).

[59] D. Henschen, "Emerging Options: MapReduce, Hadoop," Young, But Impressive, Inform. Week, **24** (2010).

[60] Apache Phoenix, "Apache Spark Plugin," https://phoenix.apache.org/phoenix_spark.html, (2016).

[61] B. Dhyani, A. Barthwal, "Big Data Analytics using Hadoop," Int. J. Comp. App., **108**(12): 1-5 (2014).

[62] J. Scott, "Apache Spark vs. Apache Drill. Converge Blog, Powered by MapR.," [https://www.mapr.com/blog/apache-spark-vs-apache-drill] [accessed October 12, 2016], (2015).

[63] T. Dunning, E. Friedman, T. Shiran, J. Nadeau, Apache-Drill, O'Reilly Media, San Francisco, CA, (2016).

[64] W.K. Lai, Y.-C. Chen, T.-Y. Wu, M.S. Obaidat, "Towards a framework for large-scale multimedia data storage and processing on Hadoop platform," J. Supercomp., **68**: 488–507 (2014).

[65] S.M. Nabavinejad, M. Goudarzi, S. Mozaffari, "The Memory Challenge in Reduce Phase of MapReduce Applications," J. Latex Class Files, Transactions on Big Data IEEE, **14**(8) (2016).

[66] A. Rabkin, R.H. Katz, "How Hadoop Clusters Break," IEEE Software, July/August, 88-95 (2013).

[67] Y. Chen, S. Alspaugh, R. Katz, "Interactive Analytical Processing in Big Data Systems: A Cross-Industry Study of MapReduce Workloads," Proceedings of the VLDB Endowment **5**(12): 1802–1813 (2012).

[68] A.L. Greeshma, G. Pradeepini, "Input split frequent pattern tree using MapReduce paradigm in Hadoop," J. Theo. App. Inform. Tech., **84(**2): 260-271 (2016).

[69] S.C. Yu, Q.-L. Kao, C.R. Lee, "Performance Optimization of the SSVD Collaborative Filtering Algorithm on MapReduce Architectures," IEEE 14th Int. Conf. Dependable, Autonomic and Secure Computing, Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress, 612-619 (2016).

[70] C. von der Weth, A. Datta, "Multi-term Keyword Search in NoSQL Systems," IEEE Internet Computing, January/February, 34-43 (2012).

[71] Y. Wang, W. Goh, L. Wong, G. Montana, "Random forests on Hadoop for genome- wide association studies of multivariate neuroimaging phenotypes," BMC Bioinformatics **14**(16): 1–15 (2013).

[72] S. Nishimura, S. Das, D. Agrawal, A.E. Abbadi, "MD-HBase: design and implementation of an elastic data infrastructure for cloud-scale location services," Springer Science+Business Media, LLC, (2012).

[73] A.V. Nguyen, R. Wynden, Y. Sun, "HBase, MapReduce, and Integrated Data Visualization for Processing Clinical Signal Data," AAAI Spring Symposium: Computational Physiology, (2011).

[74] J. Sun, "Scalable RDF store based on HBase and MapReduce," Advanced Computer Theory and Engineering (ICACTE), 3rd Int. Conf., Hangzhou, (2013).

[75] Y.-J. Chang, C.-C. Chen, J.-M. Ho, C.-L. Chen, "De Novo Assembly of High- Throughput Sequencing Data with Cloud Computing and New Operations on String Graphs," Cloud Computing (CLOUD), IEEE 5th International Conference, 155–161 (2012).

[76] W-C. Chung, H.-P. Lin, S.-C. Chen, M.-F. Jiang, Y.-C. Chung, "JackHare: a framework for SQL to NoSQL translation using MapReduce," Autom. Softw. Eng. **21**: 489–508 (2014).

[77] H. Dutta, J. Demme, Distributed Storage of Large Scale Multidimensional EEG Data using Hadoop/HBase, Grid and Cloud Database Management, New York City: Springer, (2011)

[78] J. Dean, S. Ghemawat, MapReduce: Simplified Data Processing on Large Cluster, OSDI, (2004).

[79] R. Koppel, S. Smith, J. Blythe, V. Kothari, "Workarounds to computer access in healthcare organizations: you want my password or a dead patient?" Stud. Health. Tech. Inform., **208**: 215-20 (2015).

[80] Canada Health Infoway and Health Information Privacy Group, Privacy and EHR Information Flows in Canada (Version 2.0), (2012).

[81] S. Kumar, A. Henseler, D. Haukaas, "HIPAA's effects on US healthcare, Int. J. Health C. Q. Assurance," **22**(2): 183–197 (2009).

[82] J Erdmann, "As Personal Genomes Join Big Data Will Privacy and Access Shrink?" Chemistry & Biology, **20**(1): 1-2 (2013).

[83] S. Spiekermann, C.F. Cranor, "Engineering privacy," IEEE Trans. Soft. Eng. **35**(1): 67-82 (2009).

[84] K.T. Win, W. Susilo, Y. Mu, "Personal health record systems and their security protection," J. Med. Syst., **30**(4): 309–15 (2006).

[85] N.V. Chawla, D.A. Davis, "Bringing Big Data to Personalized Healthcare: A Patient- Centered Framework," J. Gen. Intern. Med., **28**(3): S660–5 (2013).

[86] Z. Chen, S. Yang, S. Tan, L. He, H. Yin, G. Zhang, "A new fragment re-allocation strategy for NoSQL database systems," Front. Comp. Sci., **9**(1): 111–127 (2015).

[87] S. Wang, I Pandis, C. Wu, S. He, D. Johnson, I. Emam, F. Guitton, Y. Guo, "High dimensional biological data retrieval optimization with NoSQL technology," BMC Genomics **15**(8): S3 (2014).

Appendix

Table 1. Literature review of big data technologies using Hadoop with possible applications in healthcare (Yes* or No**).

Big Data Technologies	Description	Purpose	Applied in Healthcare*
Hadoop Distributed File System (HDFS)	The Hadoop Distributed File System (HDFS) is the place in a Hadoop cluster where you store data (Apache Hadoop, 2016). Built for data-intensive applications, the HDFS designed to run on clusters. HDFS optimized for high performance, read-intensive operations, and resilient to failures in the cluster. It does not prevent failures, but likely to lose data, because HDFS by default makes multiple copies of each of its data blocks [17-18].	High capacity, fault tolerant, inexpensive storage of very large datasets [19].	Yes* [20-23]
MapReduce	MapReduce was the first and is the primary programming framework for developing applications in Hadoop. Advanced work in Java to use MapReduce in its original and pure form [24].	A programming paradigm for processing big data.	Yes* [23, 25]
Hadoop	Fully integrated, and linkage between two technologies: HDFS and MapReduce [26].	Processing	Yes* [23]
YARN (Yet Another Resource Negotiator)	Hadoop resource allocator. It is a resource- management platform responsible for managing compute resources in clusters and using them for scheduling of users' applications [27]. Works efficiently and easy configure with Apache Spark.	Resource allocator	Yes* [23]
ZooKeeper	Hadoop and HDFS are effective tools for distributing work across many machines. ZooKeeper is not intended to fill the space of HBase or any other big data key-value store. In fact, there are protections built into ZooKeeper to ensure that folks do not attempt to use it as large data store. ZooKeeper is, however, just right when all you want to do is share a little bit of information across your environment [28]. Works efficiently with Apache Drill.	Coordination	Yes* [23]
HBase	HBase is a NoSQL database system included in the standard Hadoop distributions. It is a key-store, logically. This means that rows are defined by a key, and have associated with them a number of bins (or columns) where the associated values are stored [29]. The only data type is the byte string. Physically, groups of similar columns are stored together in columns families. Most often, HBase is accessed via Java code, but APIs exist for using HBase with Pig, Thrift, Jython (Python based), and others. HBase is not normally accessed via MapReduce but is configurable. It does have a shell interface for interactive use.	NoSQL database with random access	Yes* [26, 30-31]
Spark	MapReduce is the primary workhorse at the core of most Hadoop cluster. While highly effective for very large batch-analytic jobs, MapReduce has proven to be suboptimal for applications like graph analysis that require iterative processing and data sharing. Three core areas: 1) resilient distributed dataset (RDD), transformation, and action [32].	Processing\storage	No**
Spark SQL	Spark outperforms Hive [33]. Easier to configure and less dependent on MapReduce and Indexing.	SQL access to Hadoop data	No**
Hive	The goal of Hive is to allow SQL access to data in the HDFS [34]. The Apache Hive data-warehouse software facilities querying and managing large datasets residing in HDFS. Hive defines a simple SQL-like query language, called HQL that enables users familiar with SQL to query the data [35].	Data Interaction	No** [36]
Cassandra	Key-value datastores are a common fixture in any big data system because they are easy to scale, quick, and straightforward to work with. Cassandra is a distributed key-value database designed with simplicity and scalability in mind [37].	Key-value store	No** [22]
Apache Solr	While Solr is able to use the Hadoop Distributed File System to store data, it is not truly compatible with Hadoop and does not use MapReduce or Yarn to build indexes or respond to queries. There is a similar effort named Blur to build a tool on top of the Lucene framework that leverages the entire Hadoop stack [38].	Document Warehouse	No**
Lucene and Blur	Blur is a tool for indexing and searching text with Hadoop. Because it has Lucene (a very popular text-indexing framework) at its core, it has many useful features, including fuzzy matching, wildcard searches, and paged results. It allows you to search through unstructured data in a way that would otherwise be very difficult.	Document Warehouse	No, not in healthcare but development and patents** [38]
MongoDB	MongoDB is a document-oriented database, the document being a JSON object. In relational databases, you have tables and rows. In MongoDB, the equivalent of a row is JSON document, and the analog to a table is a collection, a set of JSON documents [39].	JSON document-oriented database	Yes* [26, 40]
JSON	JSON is becoming common in Hadoop because it implements a key-value view of the world.	Data description and transfer	Yes* [22]
Oozie	Hadoop's workflow scheduler [17]. A workflow scheduler to manage complex multipart Hadoop jobs.	Task scheduler	No**
Pig	Pig is translated or compiled into MapReduce code and well optimized so that a series of Pig statements do not generate mappers and reducers for each statement and then run them sequentially [41].	High-level data flow language for processing data.	No**
Storm	Many technologies in the big data ecosystem, including Hadoop MapReduce, are built with very large tasks in mind. These systems are designed to perform work in batches, bundling groups of smaller tasks into larger tasks and distributing those large tasks [39].	Streaming ingest.	Yes* [42-43]

Table 2. Operational experiences, persistent issues and overall limitations of big data technologies and components that impacted the Big Data Analytics (BDA) platform.

Technology and Component	Brief Experience	Issue and Limitation	Impact to Platform
Hadoop Distributed File System (HDFS)	Each node requires configuration and monitoring, distributed filing system is unbalanced and local disks will not fail over via Hadoop. If the local disks differed in size 500GB versus 1TB versus 2TB, once full capacity hit, HDFS would crash.	Files need to be distributed with relatively the same. The local disks reaches max 90% and should re- distribute via inherent HDFS's processes but did not quickly re-balance. WestGrid's failover to disks doesn't work because Hadoop is moving the files to re-balance the nodes. No failover for Hadoop from full disks to available (also issue with WestGrid).	Did not reconfigure more than 6 nodes because very difficult to maintain and ongoing issues. If one unbalanced then it will either drastically slow the Reducer or it will be in quasi un-processing state with constantly moving files. Had to add additional 2- 4TB of local disks because issue persisted with running to 3 billion. Impact was on database ingestion inoperable and need to cleared and restarted. Had to implement large local disks of all the same size to avoid Hadoop HDFS crash.
MapReduce	Map component was extremely fast at 3-12mins for each of the iterations. Reduce can start at 10-40% during the Map and this variation was not controllable. Reducer hits 99% after 12 hours and crashes. Optimized iterations at 50 million rows as it took long to run for 100 million and more difficult to monitor.	Reducer fails and ingestion needs to be cleared from nodes and module load restarted. Reducer was extremely slow. Reducer was placing the files only on one node.	Totally failed ingestion and system inoperable. Indexed files need to be removed from node and restarted to complete the iteration. Extremely slow performance to form the database and requires constant monitoring. Current major limitation and more advanced algorithms and the java. Coding for MapReduce needs to be further explored, verified and implemented.
HBase	All five RegionServers need to run in unison otherwise unbalanced HDFS and poor performance. InfiniBand not always accessed by RegionServer. HBase qualifiers and key stores influencing the Reduce part at each of the iterations the key store had to be re- indexed.	There is an error message via a customized script when restarting the bulkload to HBase but sometimes the Reducer will place the data on the ones available even though it should stop. RegionServers slow or killed because of lack of connectivity. RegionServers constantly died but resolution was to run compaction after ingestion. HBase cannot re- index data from either HFiles that crashed and didn't complete at any Reducer level, even at 99-100%.	RegionServers were required for the ingestion to form the database, without them it was not operational. Ongoing monitoring and log checking if the RegionServers were down or not connected to InfiniBand. The script to prompt user that 5 RegionServers dead provided better usability as finding the log files was tedious work and time consuming. Run compaction with shell script after each iteration to HBase. It ran only 50 million because indexing could fail at larger amount. It was easier to clean and restart than at 100 million rows. If index failed or space on local disk maximized, had to re-run all HFiles to bulkload via MapReduce to HBase again. Each iteration set to start + 1 from the last and this setting was manually done.
ZooKeeper and Yarn	ZooKeeper services need to be ongoing and configuration done for InfiniBand in its relation to RegionServers.	ZooKeeper not allocating and/or slow.	Extreme slow performance when ZooKeeper services were not running properly but additional configuration minimized this limitation with few issues thereafter.
Phoenix	Length of columns too long and need to be matching in the schema and on the distributed nodes.	Extremely slow performance in ingesting the files of column names are more than 12 characters. Queries will return an error if the column names do not match.	Maintain a database schema with current names in a file on the nodes such that if the files ingested don't match it will show error and to verify while running queries. Zero times this occurred while ingesting files but many times when running queries.
Spark	Relies on Yarn and advanced Java Programming.	Yarn not allocating and slow process, and more java code (Scala) required.	Potential slow performance if not coded correctly. Valid online code. sources.
Spark with Zeppelin	30min delay to run the SQL-like script in its initial additional ingestion, SQL-like code is more complicated than traditional SQL.	30min delay to start testing or running queries.	30minute delay before running queries, which takes the same amount of time as with Jupyter. Currently, no fix to this issue.
Spark with Jupyter	Need to perform some Java coding to produce graphs.	No graphs produced and no buttons on interface available like Zeppelin.	Once the Java was established it has high usability and excellent performance.
Drill	Can only plugin one SQL code at a time, and relies on ZooKeeper	Poor usability, ZooKeeper not allocating and slow.	Extremely fast but poor usability interfaces. It was recently developed as net new version so better interfaces are forthcoming or at least improved changes to integration with other interface engines.

Table 3. Duration (seconds) of queries run by Apache Phoenix over 3 Billion with unbalanced* and balanced* HBase NoSQL datasets across Hadoop nodes.

Description	Type	Apache Phoenix SQL-like Query	Output Efficiency (OE) *unbalanced (seconds)	Output Efficiency (OE) **balanced (seconds)
#1. Basic selection of encounter data	Simple	`select * from DADS1 where EncounterID<10010 and EncounterID>10004;`	3.87	3.05
#2. Basic selection of encounter data based on admitted via ambulance	Simple	`select * from DADS1 where EncounterID<10010 and EncounterID>10004 and ADMIT_BY_AMBULANCE ='C';`	3.65	1.65
#3. Frequency of Diagnosis (Dx) Code with LOS	Simple	`select Diagnosis_Code, COUNT (Diagnosis_Code), AVG(LOS) from DADS1 where EncounterID<1000100 and EncounterID>1000000 GROUP BY Diagnosis_Code;`	3.11	2.11
#4. Frequency of Diagnosis (Dx) Code with Diagnosis and LOS	Simple	`select Diagnosis_Code, COUNT (Diagnosis_Code) as Frequency, LOS from DADS1 where EncounterID<1000100 and EncounterID>1000000 GROUP BY Diagnosis_Code, LOS;`	3.32	2.32
#5. Diagnosis Code with Discharge date and Discharge time	Simple	`select Diagnosis_Code, Discharge_Date, Discharge_Time from DADS1 where EncounterID<1000010 and EncounterID>1000005;`	3.02	1.02
#6. Diagnosis Code with Unit Transfer Occurrence	Simple	`select Diagnosis_Code, COUNT (Diagnosis_Code), AVG(Unit_Transfer_Occurrence) from DADS1 where EncounterID<1000100 and EncounterID>1000050 GROUP BY Diagnosis_Code;`	3.67	1.67
#7. Diagnosis Code with Location building, Location Unit, Location Room, Location Bed, Discharge Disposition	Simple	`select Diagnosis_Code, Location_Building, Location_unit, Location_Room, Location_Bed, Discharge_Disposition from DADS1 where EncounterID<1000010 and EncounterID>1000000;`	3.23	0.98
#8. Diagnosis Code with Encounter Type and LOS	Simple	`select Diagnosis_Code, Encounter_Type, LOS from DADS1 where EncounterID<1000010 and EncounterID>1000000;`	3.01	0.98
#9. Diagnosis Code with Medical Services and LOS	Simple	`select Diagnosis_Code, Medical_Services, LOS from DADS1 where EncounterID<1000010 and EncounterID>1000000;`	3	1.02
#10. Provider Service with Diagnosis codes	Simple	`select Diagnosis_Code, Provider_Service from DADS1 where EncounterID<1000010 and EncounterID>1000000;`	3.52	1.92
#11. Highest LOS for MRNs with Admit date	Simple	`select LOS, MRN, Admission_Date from DADS1 where EncounterID<1000100 and EncounterID>1000050 GROUP BY LOS, MRN, Admission_Date ORDER BY LOS DESC;`	3.62	1.62
#12. Frequency (or number) of Admit_category with Discharge_Date	Simple	`select Admit_Category, COUNT (Admit_Category) as Frequency, Discharge_Date from DADS1 where EncounterID<1000100 and EncounterID>1000050 GROUP BY Admit_Category, Discharge_Date;`	3.54	1.89
#13. Admitted by Ambulance, Interventions, and Medical Services with Diagnosis	Complex	`select Gender, COUNT (Admit_by_Ambulance), COUNT (Discharge_Disposition), COUNT (Interven_Occurrence), COUNT (Medical_Services), COUNT (Diagnosis_Code), MAX(LOS) from DADS1 where EncounterID<1000010 and EncounterID>1000000 GROUP BY Gender;`	3.67	1.89
#14. Intervention and Location with Admit and Location	Complex	`select Interven_Occurrence, Interven_Episode_St_Date, Interven_Location, Interven_Episode_Start_Dat e, Interven_Attribute_Location, Admission_Time, Location_Unit, Location_Bed, Location_Building from DADS1 where EncounterID<1000010 and EncounterID>1000000;`	3.03	1.87
#15. Medical Services with Unit Transfer Occurrence	Complex	`select Count (Episode_Duration), Count (Anesthetistic_Technique), Count(Interven_Location), Count(Medical_Services), Count (Unit_Transfer_Occurrence) from DADS1 where`	3.47	1.92

		`EncounterID<1000010 and EncounterID>1000000 Group BY age;`		
#16. Admit Category and Discharge with Transfer	Complex	`select LOS, Count(Discharge_Disposition), Count(Most_Responsible_Si te), Max(Transfer_In_Date), Min(Transfer_Out_Date), Max(Transfer_Hours), Max(Days_In_Unit), Count(Patient_Service), Max (Patient_Service_Occurrence) from DADS1 where EncounterID<1000010 and EncounterID>1000000 GROUP BY LOS;`	3.61	1.75
#17. Encounter, Discharge and Transfer	Complex	`select Diagnosis_Code, Encounter_Type, LOS, Admit_Category, Discharge_Date, Discharge_Time, Location_Building, Location_Unit, Location_Bed from DADS1 where EncounterID<1000010 and EncounterID>1000000 ORDER BY Diagnosis_Code DESC;`	3.56	1.53
#18. Medical Services and Days in Unit	Complex	`select Patient_Service_Days, Patient_Service_Occurrence, Transfer_In_Date, Transfer_Out_Date, Days_In_Unit, Medical_Services, Location_Unit from DADS1 where EncounterID<1000010 and EncounterID>1000000;`	3.72	2.34
#19. Admission, Transfer with Intervention and Encounter	Complex	`select LOS, Count(MRN), Count(Admission_Date), Count(Admission_Time), Max(Institute_From), Count(Admit_Category), Count(Encounter_Type), Count(Entry_Code), Count(Diagnosis_Code), Max(Interven_Episode_St_ Date), Count(Interven_Attribute_Extent) from DADS1 where EncounterID<1000010 and EncounterID>1000000 and LOS BETWEEN 0 AND 9999 GROUP BY LOS ORDER BY LOS DESC;`	3.82	3.02
#20. Frequency (or number) of Admit_Category with Patient Service	Complex	`select Admit_Category, AVG(Patient_Services_Occurrence), COUNT (Patient_Service_Type), MAX(Transfer_In_Date), MAX(Transfer_Out_Date), Count (Transfer_Nursing_Unit), Count(Service_Nursing_Area), Count(Medical_Services), Count(Encounter_Type), Count(Diagnosis_Type), count(Location_Unit), count(Provider_Service) from DADS1 where EncounterID<1000010 and EncounterID>1000000 GROUP BY Admit_Category, Discharge_Date; Admit_Category, Discharge_Date;`	3.86	2.61
#21. Provider Occurrence with Nursing	Complex	`select Provider_Service, Provider_Type, Diagnosis_Code, Provider_Occurrence, Transfer_Nursing_Unit, Medical_Services from DADS1 where EncounterID<1000010 and EncounterID>1000000;`	3.78	1.7
#22. Provider with Diagnosis and Intervention	Complex	`select Provider_Service, Provider_Type, Provider_Occurrence, Diagnosis_Code, Diagnosis_Type, Medical_Services, Unit_Transfer_Occurrence, Interven_Code, Interven_Occurrence, Interven_Provider_Service, Interven_Episode_St_Date, Interven_Attribute_Extent from DADS1 where EncounterID<1000010 and EncounterID>1000000;`	3.86	1.71

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 42-50 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

A Web-Based Decision Support System for Evaluating Soil Suitability for Cassava Cultivation

Adewale Opeoluwa Ogunde*, Ajibola Rasaq Olanbo

[1]Department of Computer Science, Redeemer's University, Osun State, 234, Nigeria

ARTICLE INFO

Article history:
Received: 05 December, 2016
Accepted: 15 January, 2017
Online: 28 January, 2017

Keywords :
Classification Rule Mining
Cassava Cultivation
Data Mining
Decision Support Systems
Decision Trees
Machine Learning

ABSTRACT

Precision agriculture in recent times had assumed a different dimension in order to improve on the poor standard of agriculture. Similarly, the upsurge in technological advancement, most especially in the aspect of machine learning and artificial intelligence, is a promising trend towards a positive solution to this problem. Therefore, this research work presents a decision support system for analyzing and mining knowledge from soil data with respect to its suitability for cassava cultivation. Past data consisting of some major soil attributes were obtained from relevant literature sources. This data was preprocessed using the ARFF Converter, available in WEKA. 70% of the data were used as training data set while remaining 30% were used for testing. Classification rule mining was carried out using J48 decision tree algorithm for the data training. 'If-then' construct models were then generated from the decision tree, which was used to develop a system for predicting the suitability status of soil for cassava cultivation. The percentage accuracy of the data classification was 76.5% and 23.5% for correctly classified and incorrectly classified instances respectively. Practically, the developed system was esteemed a prospective tool for farmers, soil laboratories and other users in predicting soil suitability for cassava cultivation.

1. Introduction

The process of evaluating a land for its suitability for crop cultivation is highly important for the farmers to have a pre-knowledge about the given piece of land. According to [1], the major constraints inherent in crop production in Nigeria are the imbalance and deficiencies of nutrients. There exists the urgency in the need to assess the suitability of a farmland purposefully for agricultural practice to boost crop yield. Often, the stakeholders – the farmers being negligent of the nutrient composition of a soil, do engage in cultivating the soil for agriculture. Research has shown that due to such poor management and system of agriculture specifically in most developing countries, most soils had been rendered infertile. Some due to adverse climatic condition had been negatively affected leaving them leached of nutrients. Agricultural practices in the country are immensely suffering from lack of technological method to enhance performance and productivity in crops cultivation. The manual existing system employed by major players, that is, the local farmers had proved inefficient. One of the areas where the manual methods had proved inefficient is the stage where one needs to make some cogent decision as to what to do to produce a maximum yield of a particular crop on a given piece of land. The farmers often time had been put in a state of dilemma in the place of decision making. This research work is therefore geared towards proffering a worthwhile system of decision making with a specific focus on one of the most important entity in agriculture – the soil and its suitability for cassava cultivation, as a case study.

Recently, cassava production attained some level of prominence in terms of its positive impact locally, serving as a means for staple food crop, and employment opportunities. Besides, its potential in increasing the nation's Gross Domestic Product (GDP) cannot be over-emphasized. In the recent times, similar research works had been done to cater for some of these inefficiencies, each employing different methodology. A robust and an efficient decision support system (DSS) in this area will not only support the farmer's decision making but also improve marginally the quality and quantity of farm produce. Reviewed literatures on the topic of soil suitability for cassava production had revealed an expanse of various analyzed soils parameters with respect to cassava cultivation. In this light, this paper focused on the development of a web-based decision support system for evaluating agricultural soil suitability for cassava cultivation using a classification mining technique. The paper is aimed at providing an improved method of soil evaluation by the farmers for cassava plantation with respect to soil nutrients. A web-based decision support system for evaluating agricultural soil suitability for cassava plantation would be developed for easy accessibility by

*Corresponding Author: Adewale Opeoluwa Ogunde, Department of Computer Science, Redeemer's University, Osun State, 234, Nigeria
Email: ogundea@run.edu.ng

farmers. The paper encompasses the provision of a system for alleviating challenges inherent in making decision concerning the potential yield of a piece of farmland. Through its significant influence to the agricultural practices, the application of this system will help boost the quality and quantity of cassava crop produce..

2. Literature Review

Nigeria presently has up to 79 million hectares of arable land, which has 32 million hectares under cultivation. The farm holding is quite predominant among the smallholders, which are mostly subsistence producers accounting for 80%. The potential of both crop and livestock production has remained low. The inherent problem which crops are subjected to such as poor access, uptake of reasonable and quality seeds, and the usage of fertilizer including the poor methods of production had led to this unfavorable trend in agriculture. The vast populace is dependent on the importation of foods such as rice, wheat and fish to mention a few [2]. There is an urgent need to ensure that lands used for agricultural practices be based on its inherent capacity for sustainability and optimization for soil productivity. Senjobi [3] further emphasized this need in terms of the present and generally accepted view of the precision farming which is more particular to developing countries where land is being used without knowledge of its inherent capacity. According to Aderonke and Gbadegesin [4], a major dilemma rampant in the Nigeria agriculture is the poor knowledge couple with suitability appraisal of parcels of land for agricultural practice. Consequently, we are faced with unfortunate farm management systems, minimum yield and needless high cost of production.

The importance of the adoption of a scientific method in handling land evaluation cannot be over-emphasized in respect of assessing the inherent ability of a parcel of land for agricultural production [5]. Lin et al. [6] emphasized that having a pre-knowledge on the ability of an evaluated land aids its alleviation either ahead of or during cropping period. Assessing the performance is hinged on matching features of diverse land units in specific area with the requirements of definite or possible land use types that result in lands classification in respect of their suitability for specific crops [7]. The adoption of Soil Suitability Decision Support System in the Nigeria agricultural sector/domain will help the famers to combine data, knowledge and mathematical models gotten from literature on the production of crop to enhance decision making capabilities in their quest to obtain quality and quantity of cultivated crops [8].

The agriculture sector in Nigeria has not been able to produce to expected capacity due to the irrational use of agricultural lands and the implementation of archaic methods. Consequently, the current food security challenges have been on the increase. A promising advancement in this is decision support system to evaluate soil suitability for agricultural practices. Presently, the field lacks such decision support system [4]. A specific objective of a DSS is to assist in decision processes. Rather than automating decision making, it should support and besides adapt promptly to the changing requirements of decision makers. A knowledge-driven DSS dictates or recommend actions to users. A Knowledge-driven DSS aids in handling tasks and decisions which can be taken and respectively performed by a human expert. These tasks include classification, configuration, interpretation, diagnosis, prediction and planning. Liu et al. [9] outlined that integrating the DSS aided with knowledge management function helps the performance of the decision makers by improving their quality of services especially when human experts are not available. In addition, Alvarado et al. [10] stated that the integration of a DSS aided with knowledge management function helps in the human experts to make consistent decisions.

2.1. Soil and the Measure of its Nutritional Parameters

Research has shown that for a normal functioning and growth, plants need some important nutrients. The expectancy of the plant's nutrients measurement is done by a method called soil sufficiency range. It is described as the range interval with respect to the soil nutrients at which a specific plant will deliver potentially as required. The individual plant species and the specific nutrient therefore determine the width of this range. Nutrient levels below or exceeding the sufficiency range consequently dictates the performance of the plants respectively. The sufficiency range can either result to nutrient deficiency or toxicity. Nutrient deficiency often times signals a low percentage of the required nutrient for a healthy and quality performance of the plant. While on the other hand, the soil is toxic when the inherent nutrient requirement for the healthy growth of a plant is considered excess. There exist seventeen chemical elements that have been acknowledged as important for the growths of plants amongst which are nitrogen, phosphorus, potassium, calcium, and so on [11]. These elements are hereby classified as major elements known as macro nutrients and minor elements known as micronutrients. Plant need macronutrients in measurable quantity (>1,000 mg kg-1) and need micronutrients in low quantity (<100 mg kg-1).

2.2. Land Evaluation

In order to minimize risk in farming, land evaluation plays the role of land use to land qualities requirements matching. The use of scientifically standardized method to match the characteristics of land resources for specific uses is known as land evaluation. Hence, knowing the suitability of a piece of land is highly important to determining the use of land resources in harmony to the best carrying ability in the agricultural development. In determining the potential areas for best agricultural development, sustainable and balanced resource data are needed in the course of land evaluation for its suitability. Different methods of land evaluation such as Land Capability Classification [12], Storie Index [13] and Land Suitability Evaluation [14] have been developed. Basically, the rule of land evaluation deals with the discovery of the features of a specific landscape, knowing the requisite for the land desired use type and matching the two to ascertain the degree to which they match. Several approaches are being adopted in land evaluation systems namely parameters totaling system, parameters multiplying system and matching system between land worth and land characteristics with crop necessities. The Storie Index for instance utilizes parameters multiplying system, whereas Land Suitability Evaluation matches land quality and land characteristics with crop necessities.

In the non-parametric evaluation, soils were foremost positioned in suitability classes by matching their characteristics

with the established requirements. The cumulative suitability classes were specified by the most restrictive feature(s) of the soil. Clearly, evaluation of land has ushered result to problem of making soil analysis information helpful to farmers and other land users [15]. Land quality is the intricate feature of lands and consists of one or more land characteristics. In considering land evaluation, the soil, climate and topography are some of the noticeable land qualities which are directly related to plant requirements [16]. Nutrient retention such as the PH, action exchange capacity, drainage, erosion hazard, flood, texture, alkalinity, and soil depth are the main soil characteristics with respect to land evaluation [3 and 15].

2.3. Land Suitability Classification

Soil suitability evaluation involves characterizing the soils in a given area for specific land use type [17]. The process of measurement and classification of land units in accordance to their suitability for a specific use is known as land suitability classification [18]. The utilization of suitability systems has immensely improved agricultural land use in the recent past. These systems in terms of their abilities have shown reasonable efforts in the evaluation and assessment of proper land for a range of crops. Conversely, a major problem to land utilization is triggered when a meager knowledge of soil suitability for agricultural practice is known. In designing a proper land utilization systems and managing practices, and also for environmental knowledge, a dependable prerequisite is soil data which are highly important for a sustainable crop production. Having established the necessity and the usefulness of soil classification and mapping for broad land utilization planning, the paramount significance to the farmer is knowing the potential inherent in growing a specific crop on a piece of land, and proffering a best practice to optimize productivity of the crop [4]. The Food and Agriculture Organization (FAO) Framework [14] combined with the parametric Riquier index [19] was adopted by Aguilar and Ortiz [20] to describe the suitability classes for land ability.

2.4. Land Suitability Evaluation (LSE) for Cassava (Manihot Esculenta Grantz) Plantation

Often, the plantation of cassava is done by poor farmers in the tropics with low inputs. Depletion of the inherent soil nutrient of a soil usually sets in under continuous cultivation. A substantial measure of nutrients are taken up by cassava, mainly potassium-K and needs a huge measure of nitrogen – N and phosphorous – P [21]. Cassava performs optimally on a wider range of soils but most preferably on permeable, friable soils with some organic matter composition and depth of 30-40cm. Not suitable on a waterlogged conditions. A PH level of 6-7 is preferred and the soil clay content should be less than 18%. It has no tolerance for saline conditions. Most of the constraints predominant with soil suitability evaluation for cultivating cassava are soil fertility, poor soil texture, structure and drainage [17].

This research work shall be taking a critical look at the aspect of the most necessary requirements of soil fertility as such the PH level of the soil, available nitrogen, phosphorous, potassium and the organic matter in respect of cassava plantation as highlighted by Isitekhale et al. [17] and Vanlauwe et al. [21]. Presently, there are various well established Agricultural DSSs. An elaborate presentation and literature done by Antonopoulou et al. [22] gave examples such as FASSET (DSS for Wheat), HADSS (DSS for Wheat), EPIC (DSS for Maize and Cowpeas) and so on. Their scope revealed their exclusiveness and limitation in handling decisions. Existing in literature are diverse narrowly oriented Agricultural DSS for management of nutrients, management of pest and scheduling irrigation [23]. Mokarram et al. [24] employed RotBoost a newly proposed ensemble classifier generation technique, which is constructed by combining Rotation Forest and AdaBoost and also adopting FAO method of suitability classes for the classification of land suitability in Shavur plain, which lies in the northern of Khuzestan Province, Southwest of Iran. This provides positive evidence for the utility of machine learning methods in land suitability classification especially Multiple Classifier System (MCS) methods. The outcomes show that RotBoost can generate ensemble classifiers with notably higher prediction correctness than either Rotation Forest or AdaBoost, which is about 99% and 88.5%, using two different performance evaluation measures.

Ramesh and Vardhan [25] identified a suitable data model that achieved a high degree of accuracy and generality in terms of their capability for predicting yield in agriculture. For this purpose, different types of Data Mining techniques were evaluated on different data sets. Gholap [26] made use of dataset collected from private soil testing laboratory in Pune, India. The datasets contain different attributes and their respective values of soil samples taken from three regions of Pune District. The dataset has 10 attributes and an aggregate of 1988 instances of soil samples. It was discovered in this research that J48 demonstrated high level of accuracy in comparison with NBTree and SimpleCart classifier, hence making it a good predictive model. Sally and Geoffrey [27] worked on a classical data mining tool has been adopted in this research work to sort mushrooms by grading into quality grades and achieving a similar accuracy attainable by inspection done by humans.

2.5. Decision Trees and J48 Algorithm

One of the most widely used machine learning algorithm is decision tree. They are potentially popular due to their adaptability to any form of data type. They are classified as supervised machine learning algorithm. The algorithm works by dividing its training data into continuous smaller parts so that patterns can be identified in order to classify the data. A clearer and logical structure like a flow chart representation is then constructed through the knowledge. The algorithm is mostly well suited for many hierarchical categorical distinctions cases. A heuristic known as recursive partitioning is at play in this algorithm. The heuristic employs a divide and conquer approach using attribute values to split the data into smaller and smaller subsets of analogous classes. Structurally, a root node which represents the whole dataset, decision nodes which do the computation and leaf nodes which generate the classification. The decisions that are to be made for the splitting the labeled training data into its classes are learned at the training phase by the algorithm. The classification of an unknown instance works by passing the data through the tree. A comparative analysis of a specific attribute from the input data with an identified constant in the training phase is done at each decision node. At each decision node, selected feature is being compared with the predetermined constant, the decision is hinged on determining if the attribute has greater or lesser degree than the constant, therefore a two splitting is created in the tree. The data haven passed through these successive decision nodes, on getting to the leaf node is being assigned a class [28].

Among others, the four most used decision tree learning algorithms are J48, ID3, CHAID and CART. J48 algorithm was used in this work. J48 is known as an optimized version of C4.5 algorithm. The classification of any specific data item allows the data item to be divided in various levels starting from the root node to the leaf; this is done in a hierarchical approach. The process continues until it gets over the terminal node which cannot be further subdivided. Decision analysis makes use of this tree such that in this tree, every non-leaf node symbolizes a test or decision on the data item. Some certain branch will be chosen depending on the output at the level. The subdivision continues even up to the last level. The simple algorithm of the process of the J48 Decision tree classifier can be found in Kalmegh and Deshmukh [29]. The general approach employed in building a classification model is shown in figure 1.

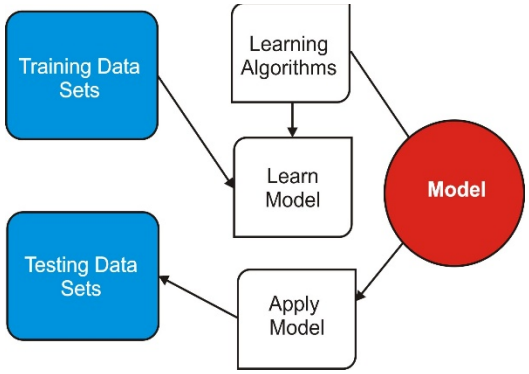


Figure 1: General approach for building a classification model

2.6. Maintaining the Integrity of the Specifications (Heading 2)

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

3. Methodology

Data used for system development was obtained from secondary sources including peer-reviewed published journals on the appropriate soil nutrient components suitable for optimal cassava cultivation. The data consisting of different soil types, cultivated under some nutritional parameters, their evaluated measure of suitability and some other factors shall be pre-processed and cleaned up to be adaptable with the J48 algorithm. The potential degree of suitability of a piece of land was predicted based on the model produced by the generated rules from the J48 decision tree learning algorithm. The system works by providing an interface for the users to input their information concerning the prospective soil. PHP Hypertext Pre-Processor was adopted to handle the server-side development due to its capability to coordinate the request – response cycle from the web browser to the server for processing.

3.1. Data Gathering, Description and Cleaning

The secondary data employed in this project were obtained from reviewed journals on soil suitability for cassava plantation [1, 30, 31 and 32]. The data contain amongst others, different type of soils, their respective chemical properties such as the nutrients composition like the NPK available, the PH values, Organic Matter, and so on. These parameters represent the most likely parameters that influence farmer's decision concerning a given piece of land before cultivation. These data formed the basic source of data used for the analysis where each of them shows the physical and chemical properties of different farmlands. However, the nutritional requirements for cassava plantation were already presented by Howeler [33], and Sys and Debaveye [34]. In order to attain consistency, the data used in this research was calculated and converted to suit the measuring scale given by Howeler [33]. This research work basically considered some of the fundamental parameters for determining soil fertility in respect of cassava plantation. In this context, parameters such as the soil nutrient composition – available NPK, the PH value, necessary for determining crop yield in a piece of farmland. The necessary attributes such as PH, Nitrogen (N), Phosphorous (P), Potassium (K) and Organic Manure were retained for the research while other non-significant attributes were removed. The selected attributes and their possible range of values are shown in Table 1.

Table 1: Description of the featured variables for classification (Adapted from Howeler [33])

Variable	Description	Possible Values
PHV	PH Value of the soil	P1: 4.5-7.0, P2: 3.5-4.49 P3: 0.0-3.49 P4: 7.1 - 8.0
NIT	Available Nitrogen	P1: 0.2 and Above P2: 0.1 - 0.19 P3: Less than 0.1
PHO	Available Phosphorous	P1: 10.0 and Above P2: 4.1 – 9.99 P3: 2.0 - 4.0 P4: 0.0 - 1.99
POT	Available Potassium	P1: 0.1 – 0.25 P2: 0.1 - 0.149 P3: Less than 0.1 P4 : Greater than 0.25
OMA	Organic Matter composition	P1: 2.0 - 4.0 P2: 4.1 and Above P3: 0.0 – 1.99
PREDICTION	Performance Classification	S1: Highly Suitable S2: Moderately Suitable S3: Marginally Suitable N1: Presently Not Suitable

3.2. Data Training

In order to achieve consistency in respect of the workings of the prediction system, 70% of the collected data were used as the training data while the remaining 30% were reserved for model testing. The data was selected and converted into Attribute Relation File Format (ARFF) format using the ARFF converter plug-in in the Waikato Environment for Knowledge Analysis (WEKA) explorer. Decision trees were then generated by WEKA using the J48 algorithm. The IF-THEN rules resulting from the decision tree was inputted into the knowledge-base of the prediction system. The user on interaction via the interface simply makes a request which was utilized in matching up the in-built rules in the knowledge base, and then the prediction followed.

*Corresponding Author: Adewale Opeoluwa Ogunde, Department of Computer Science, Redeemer's University, Osun State, Nigeria, 07036090090, ogunde@run.edu.ng

3.3. Design of the Prediction System

A set of rules needed for the prediction system was integrated in the knowledge base. The knowledge base was programmed to gain some domain experience; it therefore worked as an expert in that wise. The knowledge base on getting input from the user infers programmed rules, that is, rules drawn from the J48 decision tree, stored in the database and hence produces the output. The prediction system is represented in figure 2.

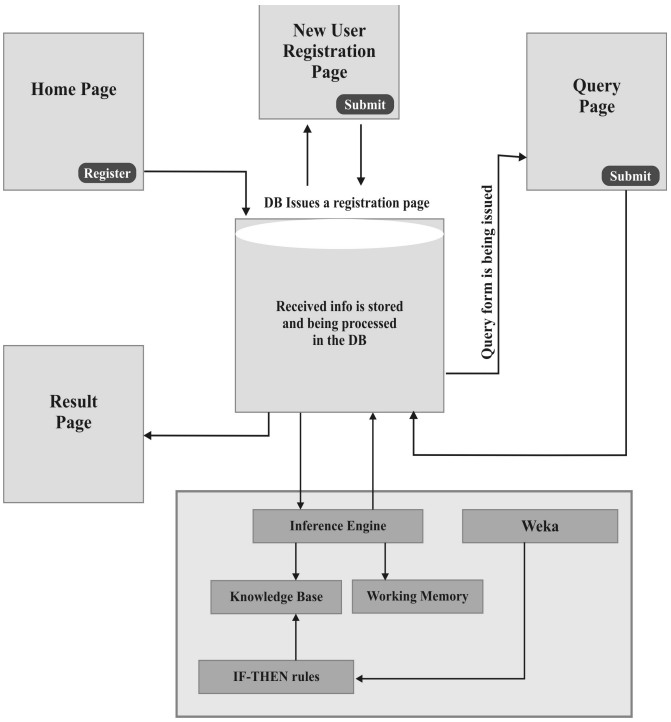


Figure 2: Architecture of the Prediction System

Explanation of major components in the system architecture is hereby provided.

Knowledge Base (KB): The repository of the prediction system domain knowledge is the KB that contains the representation IF-THEN declarative rules. The Waikato Environment for Knowledge Analysis (WEKA) is fondly suitable for the generation of these rules and contains tools for implementation of the knowledge based system.

Working Memory: This component provided the platform for the collection of information inputted by the user. The knowledge residing in the knowledge base is matched up with this information in order to infer new facts. The process continues by entering the new facts into the working memory and consequently, a conclusive fact is reached, which is also entered into the working process.

Inference Engine: This is based on the information in the working memory and the knowledge base. It goes through the rules in order to establish a match existing between their premises and information in-built in the working memory. As soon as a match is found, the conclusive fact is added to the working memory and further proceeds to search for new rules.

End-User: An individual e.g. a farmer seeking knowledge on suitability of available soil initiates a request in order to acquire advice from the prediction system.

4. Implementation and Results

In implementing the system, the following tools were employed: JDK 1.8.0, NetBeans IDE 8.1 and WEKA3.6.9 Explorer for data training and decision tree generation using J48 Algorithm. MySQL version 5.6.2 was used for data management. All algorithms take their input in the form of a single relational table in the ARFF format, which could be read from a file or generated by a database query. The research adopted J48 algorithm from WEKA Explorer for the classification process.

4.1. Model Construction for J48 Decision Tree

The data used in this research work was first preprocessed using the ARFF converter in WEKA after which the preprocessed data was trained. The essence of this is to put the data and its attributes in a format acceptable by WEKA for the classification process. The preprocessing panel of WEKA enabled the import of a data from a database. The data which was in ARFF format was preprocessed using filtering algorithm which was usually used to transform the data from one format to another; for instance the numeric attributes in the data could be transformed into discrete ones. Instance and attributes that were discovered irrelevant were also be deleted. Figure 3 shows the results of the preprocessed data. For instance, for PH values, there are seventeen instances with fifteen distinct values. The data type is numeric with thirteen unique values. Minimum PH value is 4.5. Maximum PH value is 7.3. Mean of the PH data is 5.746 while the standard deviation is 0.696.

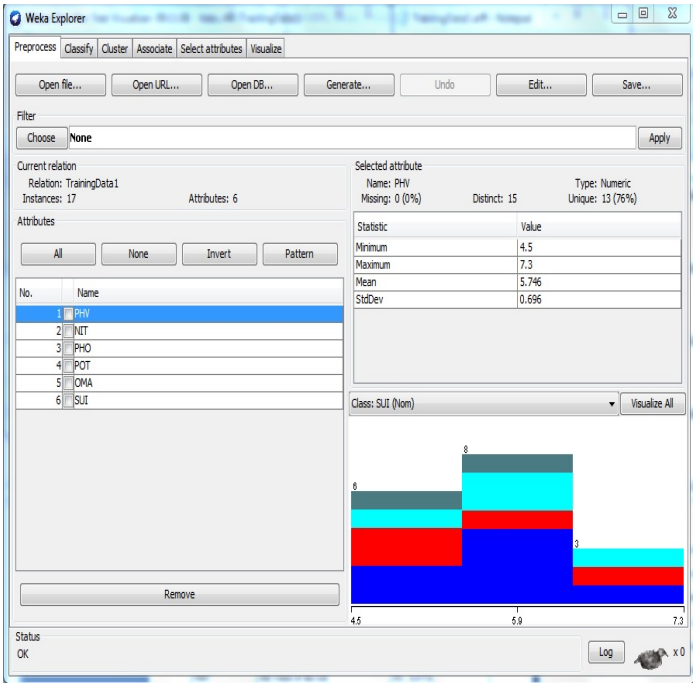


Figure 3: WEKA Interface showing the preprocessed data

After data pre-processing, classification of the data was done using J48 algorithm under the classify panel tab of the WEKA Explorer as shown in Figure 3. J48 had been found to be a very effective classifier for similar works as earlier reported in section 2.4 of this work. It is as an optimized version of C4.5 algorithm.

The classification of any specific data item allows the data item to be divided in various levels starting from the root node to the leaf; this is done in a hierarchical approach. The process continues until it gets over the terminal node which cannot be further subdivided. Decision analysis makes use of this tree such that in this tree, every non-leaf node symbolizes a test or decision on the data item. Some certain branch will be chosen depending on the output at the level. The subdivision continues even up to the last level. Graphical visualization of the model was achieved with a decision tree, which is a powerful technique used in handling real world problems through classifying the problem into a tree formation and applying the control rules over the internal nodes. Two dimensional (2D) plots of current relation were visualized via the Visualize panel tab in Figure 3 in order to have a pictorial view of the model and the resulting model visualization for the six major soil attributes used is displayed in figure 4.

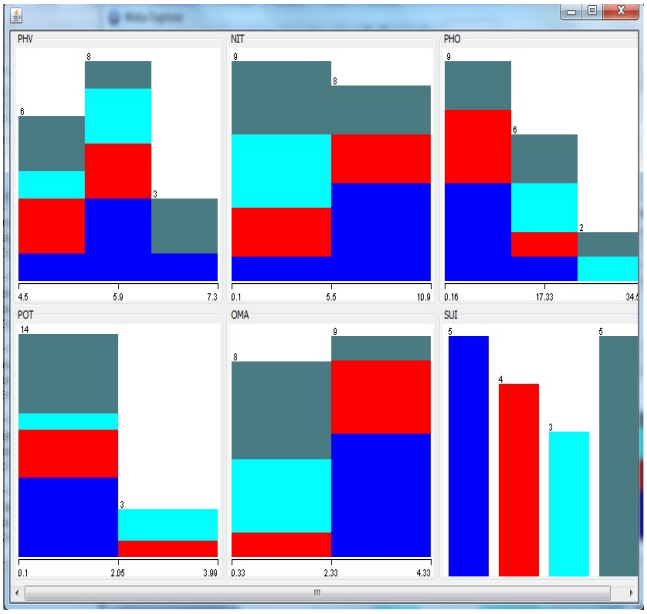


Figure 4: Model visualization of the data classification

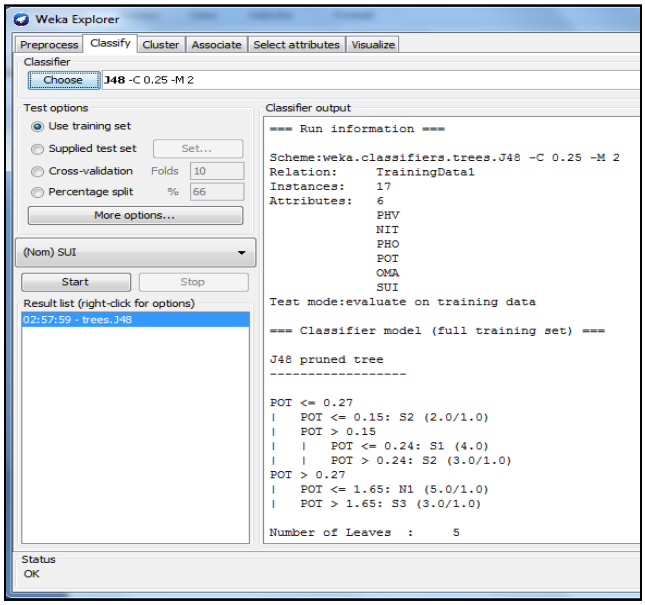


Figure 5: J48 decision rules generated

After chosen the J48 classifier, the test options were set as user training set and the set of rules generated and the result of the classifier are shown in figures 5 and 6 respectively. Figure 5 showed that seventeen instances of the data were classified with a total of six attributes. The number of leaves was 5.

Figure 6 showed that the time taken to build the model was 0.02 seconds with 76.47% of the data reported as correctly classified instances while 23.52% were incorrectly classified. Detailed accuracy by class in terms of true positive rates, false positive rates, precision, recall, f1-measure, ROC Area, confusion matrix and others were also reported.

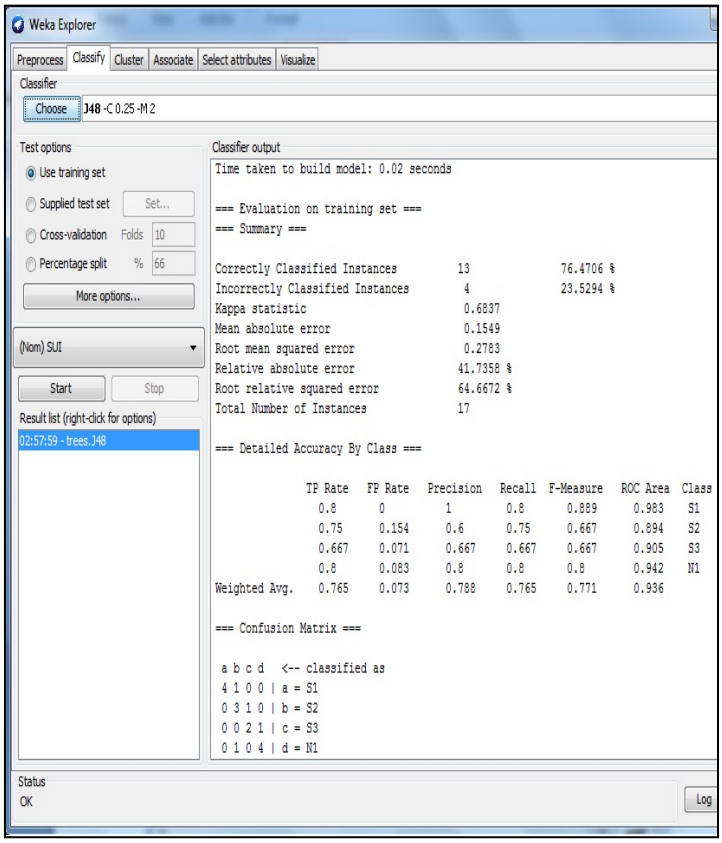


Figure 6: Result of the classifier

In figure 6, the kappa statistic was 0.6837, which showed agreement of prediction with true class. Mean Absolute Error (MAE) was 0.1549, which measured the average magnitude of the errors in the model, without considering their direction. It measured accuracy for continuous variables. It is the average over the verification sample of the absolute values of the differences between the forecasted and the corresponding observation in the model. The value meant that all the individual differences are weighted equally in the average. Root mean squared error (RMSE) was 0.2783, which was a quadratic scoring rule that measured the average magnitude of the error. It is the difference between the predicted and corresponding observed values when they are each squared and then averaged over the sample. The value gave a relatively high weight to large errors. The small errors reported by the model also confirmed that the model is predicting very well. Usually, the RMSE will always be larger or equal to the MAE; the greater difference between them, the greater the variance in the variance in the individual errors in the sample. If the RMSE is equal to MAE, then all the errors are of the same magnitude, which was not the case in this model.

The average true positive rate was 0.765; this represented the correct predictions, which were the number of sample predictions that were truly positive. Also, the average false positive rate was 0.073; this represented the number of samples predicted positive that were actually negative. Recall is the total positive rate (also referred to as sensitivity), that is, what fraction of those that are actually positive were predicted positive, which gave an average of 0.765. Precision is the fraction of those predicted positives that were actually positive, which had an average of 0.788. These very positive values reported by the model implied that the model was predicting with high level of accuracy. The F1 score was 0.771, which showed that recall and precision were evenly weighted. The ROC values and the confusion matrix also scaled very well, supporting the efficacy of the model in predicting new cases of soil suitability. After the classification of the data was done by the J48 classifier, the WEKA classifier tree visualizer was used to generate a decision tree for the model. A portion of the decision tree produced is as shown in figure 7.

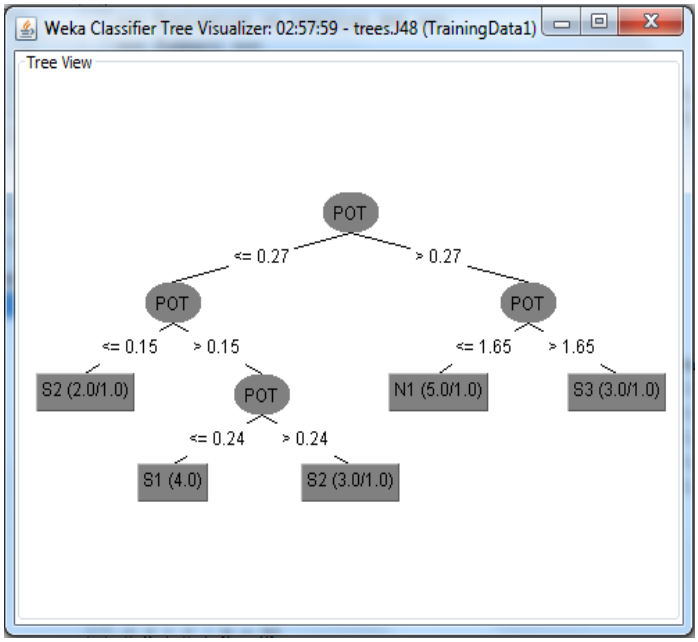


Figure 7: Part of the decision tree produced

4.2. Knowledge Representation and Interpretation of Results

The knowledge represented by the decision tree was extracted and represented in the form of IF-THEN rules as shown in Table 2. For example, the first rule means that If PH Value of the soil is between 4.5 – 7.0, and Nitrogen component of the soil is 0.2 or more, and Phosphorus component is 10.0 or more, while Potassium component is between 0.15 – 0.25, and the organic matter composition of the soil is between 2.0 – 4.0, then the system predicts that the soil is "Highly Suitable" for cassava cultivation.

4.2.1. The Prediction System

The prediction system consists of a graphical user interface that allow the users interact with the system and make necessary predictions concerning the suitability of available soil for the cultivation of Cassava. Firstly, it includes the user's registration page, which provides a platform where new users can input their details and get registered before they can gain access to the prediction system. The user registration page also displays the registration form, which collects the users' details. The second page is the login page, which is an interface where existing users can easily input their login details and can easily gain access to the prediction application. After the user has gained access into the system, the prediction interface then pops up. The prediction

Table 2: A presentation of the IF-THEN rules derived from the decision tree

IF PHV = "4.5 – 7.0" and NIT = "0.2 and above and PHO = "10.0 and Above and POT = "0.15 – 0.25" and OMA = "2.0 – 4.0", THEN PREDICTION = Highly Suitable
IF PHV = "4.5 – 7.0" and NIT = "0.2 and above and PHO = "4.1 – 9.99" and POT = "0.1 – 0.149" and OMA = "4.1 and Above", THEN PREDICTION = "Moderately Suitable"
IF PHV = "4.5 – 7.0" and NIT = "0.2 and above and PHO = "10.0 and Above" and POT = "Greater than 0.25" and OMA = "2.0 – 4.0", THEN PREDICTION = "Moderately Suitable"
IF PHV = "4.5 – 7.0" and NIT = "0.2 and above and PHO = "10.0 and Above" and POT = "Greater than 0.25" and OMA = "0.0 – 1.99", THEN PREDICTION = "Presently Not Suitable"
IF PHV = "4.5 – 7.0" and NIT = "0.1 – 0.19" and PHO = "10.0 and Above" and POT = "0.1 – 0.15" and OMA = "2.0 – 4.0", THEN PREDICTION = "Presently Not Suitable"
IF PHV = "7.1 – 8.0" and NIT = "0.1 – 0.19" and PHO = "10.0 and Above" and POT = "Greater than 0.25" and OMA = "0.0 – 1.99", THEN PREDICTION = "Presently Not Suitable"
IF PHV = "4.5 – 7.0" and NIT = "0.2 and Above" and PHO = "0.0 – 1.99" and POT = "Greater than 0.25" and OMA = "0.0 – 1.99", THEN PREDICTION = "Presently Not Suitable"
IF PHV = "4.5 – 7.0" and NIT = "0.2 and Above" and PHO = "10.0 and Above" and POT = "Greater than 0.25" and OMA = "0.0 – 1.99", THEN PREDICTION = "Marginally Suitable"
IF PHV = "4.5 – 7.0" and NIT = "0.1 – 0.19" and PHO = "10.0 and Above" and POT = "Greater than 0.25" and OMA = "2.0 – 4.0", THEN PREDICTION = "Marginally Suitable"
IF PHV = "4.5 – 7.0" and NIT = "0.1 – 0.19" and PHO = "4.1 – 9.99" and POT = "0.15 - 0.25" and OMA = "2.0 – 4.0", THEN PREDICTION = "Marginally Suitable"
IF PHV = "4.5 – 7.0" and NIT = "0.2 and Above" and PHO = "10.0 and Above" and POT = "Greater than 0.25" and OMA = "0.0 – 1.99", THEN PREDICTION = "Marginally Suitable"
IF PHV = "4.5 – 7.0" and NIT = "0.2 and Above" and PHO = "0.0 – 1.99" and POT = "Greater than 0.25" and OMA = "0.0 – 1.99", THEN PREDICTION = "Presently Not Suitable"
IF PHV = "4.5 – 7.0" and NIT = "0.2 and Above" and PHO = "10.0 and Above" and POT = "Greater than 0.25" and OMA = "2.0 – 4.0", THEN PREDICTION = "Marginally Suitable"
IF PHV = "4.5 – 7.0" and NIT = "0.2 and Above" and PHO = "2.0 – 4.0" and POT = "0.1 – 0.149" and OMA = "2.0 – 4.0", THEN PREDICTION = "Moderately Suitable"

interface shows the actual page where users can input soil parameters and query the system in order to make predictions concerning the suitability of the soil for cassava cultivation. Preset values available for selection by the user in the prediction interface were already described in Table 1. The final outcome of any prediction usually includes one of the following: highly suitable, moderately suitable, marginally suitable and presently not suitable.

4.3. System Testing

The developed system was tested using 30% of the collected data to ensure that the system met the proposed objectives. The results for different runs of the system using input values from the test data set are presented showed that soil could be marginally suitable or moderately suitable or highly suitable or presently not suitable. Figure 8 is a sample case of the situation when the soil was predicted to be highly suitable for cassava cultivation.

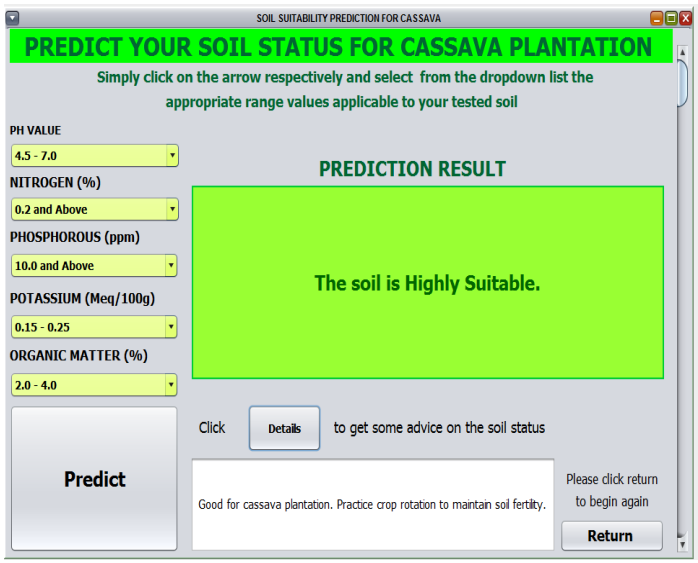


Figure 8: System predicted high suitability of soil

4.3.1. Prediction Accuracy of the System

The class wise accuracy prediction of the developed system was carried out and the results obtained from correctly classified instances and incorrectly classified instance are shown in table 3. Table 3 clearly presents the true positives, false positives and precision correctness for the four prediction classes.

Table 3: Percentage Accuracy for the four class prediction

Prediction Class	True Positive (TP)	False Positive (FP)	Correct Precision (%)
Highly Suitable	0.8	0	100%
Moderately Suitable	0.75	0.154	60%
Marginally Suitable	0.667	0.071	66.7%
Presently Not Suitable	0.8	0.083	0%

From the test results presented in table 3, soil types that were highly suitable in the test data were 100% correctly predicted. The soil types that were moderately suitable in the test data were 60% correctly predicted. The soil types that were marginally suitable in the test data were 66.7% correctly predicted. The soil types that were not suitable in the test data were incorrectly predicted. On the overall, correctly classified instances from the system testing were 76.5% while incorrect classified instances were 23.5%. This proved that the model is very suitable for predicting soil suitability for cassava plantation

5. Conclusions

A study was conducted on the development of a system for predicting the suitability of soil for the cultivation of cassava crop using decision trees. Secondary data was obtained from published journals relating to soil analysis of some farmlands from different geographical locations in Nigeria. The data utilized consists of the predominant parameters for determining typical soil fertility in respect of cassava cultivation. J48 classification algorithm was used to train the data, a good classification of the data was obtained and decision trees were generated. The output of the decision tree was used as the knowledge base for building the prediction system. 30% of the data was then used to test the validity and accuracy of the prediction system. It was deduced from the class-wise accuracy that the true positive rate for obtaining the suitability classes – Highly Suitable, Moderately Suitable, Marginally Suitable and Presently Not Suitable are 100%, 60%, 66.7% and 0% respectively. The percentage accuracy of the data classification was also determined as 76.5% and 23.5% for correctly classified and incorrectly classified instances respectively. In this light, the major objective of developing a web-based decision support system for evaluating agricultural soil suitability for cassava cultivation using a classification mining technique was achieved in this work. The paper provided an improved method of soil evaluation for farmers for cassava plantation with respect to selection of the best available soil for maximum productivity. The decision support system for evaluating agricultural soil suitability for cassava plantation was developed as a web application which made it to be easily accessible by farmers. The developed system had solved a major problem associated with decision making by farmers in determining the potential yield of a piece of farmland. The developed prediction system is esteemed a prospective tool for farmers, soil laboratories and other users in predicting soil suitability for cassava cultivation. Besides, it will drastically boost crop yield and instigate a worthwhile increase in the nation's Gross Domestic Product (GDP). This will on the long run improve the economy, most especially in developing countries of the world. Prospective area of future work will cover the extension of the prediction system to include some other minor soil parameters such as the soil organic carbon, calcium, sodium, magnesium, soil texture, depth and base saturation. Climatic conditions like temperature and rainfall which could also be responsible for good productivity of cassava plantation could be considered. Finally, experiments involving other decision tree algorithms would be carried out in order to make some worthwhile comparisons.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] M.A. Abua, Suitability Assessment of Soil for Cassava Production in the Coastal and Hinterland Areas of South Cross River State – Nigeria, Journal of Soil Science and Environmental Management, 2015, Volume 6 Issue 5, pp 108 – 115.

[2] C. Nwajiuba, Nigeria's Agriculture and Food Security Challenges, Inaugural lecture, No.5, Imo State University, Owerri, Nigeria, 2012.

[3] B.A. Senjobi, Comparative Assessment of the Effect of Land Use and Land Type on Soil Degradation and Productivity in Ogun State, Nigeria, Unpublished Ph.D. Thesis Submitted to the Department of Agronomy, University of Ibadan, Ibadan, 2007, pp 161.

[4] D.O. Aderonke, G.A. Gbadegesin, Spatial Variability in Soil Properties of a Continuously Cultivated Land, African Journal of Agricultural Research, 2013, volume 8 issue 5, pp 475-483.

[5] D.G. Rossiter, Ales: A Frame Work for Land Evaluation Using Microcomputer, Soil Use Manage, 1996, 6, pp 7-20

[6] H. Lin, J. Bourma, J. Wilding, L. Richardson, M. Kutilek, D.R. Nielson, Advances in hydropedology, Advances in Agronomy, volume 85, pp 1-89, 2005, http://cropsoil.psu.edu/people/faculty/lin/ advances2005.pdf

[7] P.I, Ezeaku, Methodologies for Agricultural Land Use Planning: Sustainable Soil Management and Productivity, Great AP Express Publishers Ltd., Nsukka, FAO, 2011.

[8] Y. Adekunle, E.O. Bisong, O.O. Fagbemi, E. E. Obuke, J. Alao, O.D. Maitanmi, Framework Model for A Soil Suitability Decision Support System for Crop Production in Nigeria. American Journal of Engineering Research (AJER), Volume-02, Issue-06, pp-09-13, 2013.

[9] S.A. Liu, H.B. Duffy, R.I. Whitfield, I.M. Boyle, Integration of Decision Support Systems to Improve Decision Support Performance, Knowledge and Information Systems, 2010, 22(3), pp 261-286.

[10] M. Alvarado, A.R. Toral, S. Ayala, Decision Making on Pipe Stress Analysis Enabled by Knowledge-Based Systems, Knowledge and Information Systems—An International Journal, 2007, 12(2), pp 255-278.

[11] K.T. Osman, Soils Principles, Properties and Management, Netherlands: Springer, 2013.

[12] A. A. Klingebiel, P. H. Montgomery, Land Capability Classification. U. S. Dept. Agric. Handbook, 1961.

[13] R. E. Storie, An Index for rating the agricultural values of soils. Calif. Agric. Expt. Sta. Bull. 566, 48, 1933.

[14] FAO, A Framework for Land Evaluation, Food and Agriculture Organization FAO, Rome, Italy, 1976, pp 1.

[15] A. O. Ogunkunle, Soil Survey and Sustainable Land Management. Proceedings of the 29th Annual Conference of the Soil Science Society of Nigeria, 2005.

[16] S. Ritung, A.F. Wahyunto, H. Hidayat, Land Suitability Evaluation With a Case Map Of Aceh Barat District. Indonesian Soil Research Institute and World Agroforestry Centre, Bogor, Indonesia, 2007.

[17] H.H.E Isitekhale, S.I. Aboh, F.E. Ekhomen, Soil Suitability Evaluation for Rice and Sugarcane in Lowland Soils of Anegbetter, Edo State, Nigeria, The International Journal of Engineering and Science (IJES), 2014, volume 3 issue 5, PP 54-62.

[18] N. Hakim, M.Y. Nyakpa, A.M. Lubis, S.G. Nugroho, M.R. Saul, M.A. Diha, B.H. Go, H. Bailey, I.T. Dasar-Dasar, Fundamentals of Soil Science. University of Lampung. Bandar Lampung, 1986, pp 258.

[19] S. Riquier, D. Bramao, I. Comet, A New System of Soil Appraisal in Terms of Actual and Potential Productivity. FAO, Rome, 1970, pp 44.

[20] J. Aguilar, R. Ortiz, Methodology capacity of agricultural land use. III National Congress of Soil Science, Pamplona, Spain, 1992, pp 281-286.

[21] B. Vanlauwe, P. Pypers, N. Sanginga, The Potential of Integrated Soil Fertility Management to Improve the Productivity of Cassava-based Systems. In: Cassava: Meeting of the Challenges of the New Millennium: Proceedings of the First Scientific Meeting of the Global Cassava Partnership, Ghent, Belgium. Institute of Plant Biotechnology for Developing Countries (IPBO), Ghent University, Ghent, Belgium, 2008.

[22] E. Antonopoulou, S. Karetsos, M. Maliappis, A. Sideridis, Web and Mobile Technologies in a Prototype DSS for Major Field Crops, Computers and Electronics in Agriculture, 2010, 70, pp 292 – 301.

[23] M. Alminana, L.F. Escudero, M. Landete, J.F. Monge, A. Rabasa, J. Sanchez-Soriano, WISCHE: A DSS for water irrigation scheduling, Omega 38(6), pp 492–500, 2010.

[24] M.S. Mokarram, F. Hamzeh, A.Z. Aminzadeh, A.Z. Rassoul, Using Machine Learning For Land Suitability Classification, West African Journal of Applied Ecology, Volume 23, No 1, 2015, pp 63-73.

[25] D. Ramesh, V.B. Vardhan, Data Mining Techniques and Applications to Agricultural Yield Data. International Journal of Advanced Research in Computer and Communication Engineering, 2013, 2 (9), pp 3477-3480.

[26] J. Gholap, Performance Tuning Of J48 Algorithm for Prediction of Soil Fertility, Asian Journal of Computer Science and Information Technology, volume 2, issue 8, 2012, pp 251– 252.

[27] J.C. Sally, H. Geoffrey, Developing innovative applications in agriculture using data mining", In the Proceedings of the Southeast Asia Regional Computer Confederation Conference, 1999..

[28] D. John, Twitter Sentiment Analysis, Final Project Report for Higher Diploma in Science in Data Analytics, National College of Ireland, 2014.

[29] S.R. Kalmegh, S.N. Deshmukh, Categorical Identification of Indian News Using J48 and Ridor Algorithm, International Refereed Journal of Engineering and Science (IRJES), Volume 3, Issue 6, 2014, pp 79-84.

[30] O.T. Ande, Soil Suitability Evaluation and Management for Cassava Production in the Derived Savanna Area of Southwestern Nigeria. International Journal of Soil Science, volume 2, issue 6, 2011, pp 142-149.

[31] J. Nwite, J.A. Nwogbaga, G.I. Okonkwo, Assessment of Productivity of Sandy Loam in Abakaliki, Southeastern Nigeria, International Journal of Agriculture and Biosciences, 2015, , 4(2), pp 59-63.

[32] S.H.E Tekwa, S.M. Maunde, Soil Nutrient Status and Productivity Potentials of Lithosols in Mubi Area, Northeastern Nigeria, Agriculture & Biology Journal of North America, 2011, Vol. 2, Issue 6, pp 887.

[33] R.H. Howeler, Cassava Mineral Nutrition and Fertilization in Cassava: Biology, Production and Utilization. CABI Publishing, Wallingford, Oson, UK, 2002, pp 115 – 147.

[34] C. Sys, J. Debaveye, Land evaluation and crop production and calculations. Agric publication, No. 7. General Admin for Development Corporation, Brussels, Belgium, 1991, pp 247.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 51-56 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

Optimized Multi-focus Image Fusion Using Genetic Algorithm

Arti Khaparde*, Vaidehi Deshmukh

Department of Electronics and Telecommunication, Maharashtra Institute of Technology, Pune-411038, India

ARTICLE INFO

Article history:
Received: 06 December, 2016
Accepted: 13 January, 2017
Online: 28 January, 2017

Keywords:
Image fusion
Superimposed
Genetic algorithm

ABSTRACT

Optical imaging systems or cameras have a convex lens with limited depth of field. So, multi-focus images are obtained when such systems are used to capture an image of a particular scene. These images are fused to get all-in-focus image. This paper proposes a new simpler method of multi-focus image fusion. Instead of decomposing input image into blocks in pyramid style, the proposed algorithm considers the complete image for edge detection. The proposed algorithm uses the genetic algorithm (GA) to find out the optimum weights from extracted edges and then fuses the images with the fusion rule based on optimized weights. Experimental results show that this superimposition method performs well; consumes less computation time and thus proves to be suitable for hardware implementation.

1. Introduction

Multi-view information of the scene is obtained from different images captured using different angles and variable focal length of the lens of a camera. In practice, all cameras consist of convex lens and have limited depth of a field. Hence while capturing an image using such camera; certain objects appear sharp whereas others appear blurred depending upon the distance from the lens. Thus multi-focus images with images different objects in focus are obtained [1].

For human perception or machine vision, a well-focused image is preferred. This image can be obtained in optical or computational way. Optical way is to reduce the lens aperture so as to increase the focal length of a lens. But this increases the diffraction and degrades the image resolution. Also small amount of light enters the camera and dark image is produced. De-convolution process when carried out on the image produces the sharp image in a computational way. But, this requires the knowledge of camera specifications and also results of this method are affected by the presence of noise [2]. So, the way out is to combine two or more multi-focus images together.This is multi-focus image fusion.

A single composite image is constructed by integrating the in-focus portions of each image. Such fused images are also useful in biomedical imaging, target identification, microscopic imaging, military operations, machine vision, and object recognition and so on.

A large number of image fusion algorithms have been invented so far. These methods work in spatial domain and transform domain.

Spatial domain methods mainly work upon the image blocks. Sharper image blocks from multi-focus images of the same scene are selected based upon the sharpness measure such as EOG (Energy of Gradient) [3] and spatial frequency [4] and sometimes using focus detection algorithm [5-9]. The main difficulty of these methods is to select the block of optimum size, on which the quality of fused image depends. Too small block gets affected by noise and causes incorrect selection. Too large block size causes sharp and blur pixels to get into the same block which leads to blocking artifacts. Hence the quality of the fused image degrades. In [10], authors have used genetic algorithm to optimize size of the block to obtain better fused image. Proposed fusion techniques apply edge features of an entire image to the genetic algorithm.

In transform based methods, images are decomposed into constituent low resolution images using multi-scale and multi-resolution transforms, due to which computational complexity increases. To reduce this computational complexity, a hybrid fusion technique of DWT and GA is proposed in [11]. This

*Corresponding Author: Arti Khaparde, Dept. of E&TC, Maharashtra Institute of Technology, Pune, India, 411038, arti.khaparde@mitpune.edu.in

method is used to fuse thermal and visual satellite images and outperforms others. Generally, transform based methods consist of three steps viz. decomposition, coefficients fusion and reconstruction. These methods use Laplacian pyramid [12], DCT [3], Wavelet, Curvelet [13], Shearlet [14] and contourlet [15] etc. transforms for decomposition of images. Multi-scale decomposition effectively extracts the visually important information of the images such as lines and details; but it produces halo artifacts near the edges. Hence final fused image has to be sharpened so that edges which are vital for the interpretation of the scene become clear. So, authors in [16, 17] suggest fusion methods which can preserve edges in the final fused image. Morphological toggle contrast operator [16] used to extract edge features which can be used for fusion; takes more time for feature extraction. Edge preserving method proposed in [17] uses decomposition based on weighted least squares filter for the fusion of multi-focus images using focus-measure based fusion rule. The method is fast and effective but the difficulty is to adaptively determine the number of decomposition levels in fusion process. In [18], a fusion method based on saliency detection is proposed. Initially, input images are decomposed into detail and approximation layers by using simple average filter. Then method constructs weight map depending upon focused regions detected using visual saliency. This method works faster in spite of decomposition and produces the fused image with better contrast.

In view of above analysis, authors propose the algorithm which can reduce the computational complexity due to decomposition with simple but fast fusion rule. The proposed novel technique uses three steps to fuse an image

- It finds the edges using which features are extracted.
- These features are used to calculate optimum weights with the help of genetic algorithm.
- And finally images are fused with these weights using the fusion rule of superimposition.

Two fusion techniques are proposed here. They are edge-superimposition, edge-GA-superimposition. This paper presents extended workoriginally presented in ICIS2016 [1].

The paper is organized as follows: Section II explains the proposed fusion techniques. Section III gives the objective evaluation parameters which are used to examine the performance of the proposed method. Section IV deals with the results and Section V is conclusion.

2. Proposed Fusion Techniques

2.1. Edge Superimposition

Quality of the image depends upon the focus and sharpness of objects present in the image. So sharper images have more information than the blurred ones. Sharper images are obtained using fusion; but fusion makes edges blur. Thus to achieve an improvement in the sharpness of the fused images, this algorithm superimposes the edges on the input images. Sobel and Canny operators are used for edge detection. The edges are then averaged to form a new edge image E. This image is superimposed on the first input using weighted addition method.

$$Image_3 = k_1 * Image_1 + k_2 * E \quad (1)$$

$$k_1 = \frac{Avg\ pixel\ value\ in\ Image_1}{Max\ pixel\ value\ in\ Image_1} \quad (2)$$

$$k_2 = (1 - k_1) * 0.1 \quad (3)$$

The intermediate $Image_3$ is superimposed on the second input image to form the final output image.

$$Final = k_3 * Image_2 + k_4 * Image_3 \quad (4)$$

$$k_3 = 1 - \left(\frac{Average\ pixel\ value\ in\ Image_2}{Max\ pixel\ value\ in\ Image_2} \right) \quad (5)$$

$$k_4 = 1 \quad (6)$$

k_4 is taken to be unity as it is used for the intermediate image that is already superimposed.

2.2 Edge-GA-superimposition

Input images are converted into edge images using edge operators. Statistical features in the form of normalized central moments up to order 3 are extracted from each edge image. Total 16 moments are selected. Mean and standard deviation of edges (along x and y directions in case of Sobel) are also calculated. Finally, the feature vector containing the moments, mean and standard deviation is formed and its size is 1 X 20.

The genetic search [1, 11] starts by generating a population of size 100. This population consists of randomly generated binary strings of length 10. Before evaluating the individuals, they are converted from binary to decimal equivalent such that every binary string lies in the range of 0 to 1.This conversion is given as:

$$Ll + (Ul - Ll) * \text{Decimal equivalent} / (2^n - 1) \quad (7)$$

Where, Lower limit (Ll) =0, Upper limit (Ul) =1, n=Number of bits in the binary.The individuals are evaluated using mean squared error as the fitness function. Mathematically, this fitness function is given as:

$$FF = \left\{ \frac{1}{2} \left(\frac{1}{d} \sum_{i=1}^{d} [F_{int}(i) - Img_1(i)]^2 \right) + \left(\frac{1}{d} \sum_{i=1}^{d} [F_{int}(i) - Img_2(i)]^2 \right) \right\} \quad (8)$$

d is the dimension of the input vector which is 1 x 40. $F_{int}(i)$ is the intermediate fused image obtained by using the equation:

$$F_{int}(i) = W_{int1} * Img_1 + W_{int2} * Img_2 \quad (9)$$

Img_1 and Img_2 are the input images, W_{int1} is the weight individual being evaluated and:

$$W_{int2} = (1 - W_{int1}) \quad (10)$$

The value of the fitness function should be as small as possible. However, it can also be transformed to make the fitness function a maximization fitness using the function:

$$F(x) = 1 / (1 + FF) \quad (11)$$

The individuals having the highest fitness values are selected. Different methods exist for the selection process such as- Roulette

wheel selection, Rank based selection, and Tournament selection etc. [1, 11]. Here tournament based selection method with a tournament size of 2 is used. Crossover and mutation operations with probability 0.7 and 0.002 respectively are performed on the selected individuals in the next step. Uniform crossover and bit-flip mutation is used. These modified individuals replace the existing population. The process is repeated for 25 generations. It is experimentally observed that this search is converging in 20-23 generations. After the last generation, the optimum candidate is selected; which serves as first weightW_1.The second weight is obtained $W_2=1-W_1$. (12)

On obtaining the two weights, each is assigned to the input images. Final fused image is obtained as a result of the addition of these optimally weighted images as:

$$F_{final} = W_1 * Img_1 + W_2 * Img_2 \quad (13)$$

The block diagram for the process is described in Figure 1:

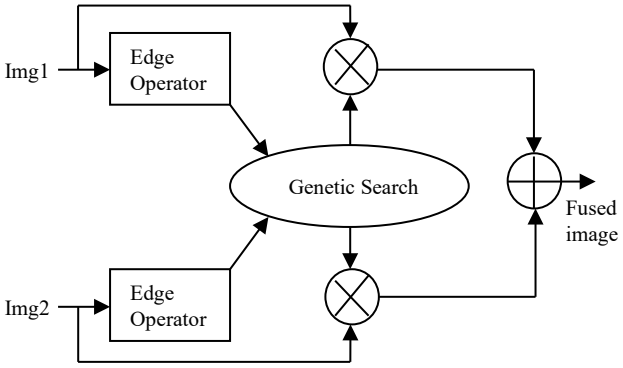


Fig. 1 Flow Diagram for Proposed method

3. Evaluation of the fused images:

The basic requirement is that the output images should contain all the valid and useful information present in the source images without introducing any form of distortion. This can be verified by comparing the resultant images visually, but, these methods, though very powerful, are subjective in nature. Hence, objective statistical parameters are usedfor the evaluation of fused images[15,19-24]. These parameters are defined as below.

3.1 Root Mean Squared Error:

$$RMSE = \sqrt[2]{\frac{1}{mn}\left[\sum_{i=0}^{m-1}\sum_{j=0}^{j-1} Img_1(i,j) - Img_2(i,j)\right]^2} \quad (14)$$

m, n: size of the input images Img_1 and Img_2

3.2 Peak Signal to Noise Ratio:

Peak signal to noise ratio gives the ratio of the maximum power of a signal and the power of corrupting noise.

$$PSNR = 20\log_{10}\left(\frac{Max}{\sqrt{MSE}}\right) \quad (15)$$

Max: Maximum possible pixel value, as 8 bits are used to represent a pixel, Max=255

3.3 Structural similarity index:

SSIM is used to model any image distortion as a combination of correlation losses, radiometric and contrast distortions. Higher the value of SSIM, more similar are the images. It can have a maximum possible value of 1. It is given by

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (16)$$

Where μ_x = average of x; μ_y = average of y; σ_x^2 = variance of x; σ_y^2=variance of y; σ_{xy}= covariance of x and y.

$$c_1 = (k_1 L)^2, c_2 = (k_2 L)^2 \quad (17)$$

Where L = dynamic range of the pixel-values; $k_1 = 0.01$ and $k_2 = 0.03$ by default.

3.4 Entropy:

Entropy is used to give the amount of information contained in an image. Mathematically it is given as:

$$E = -\sum_{i=0}^{L-1} P_i \log_2 P_i \quad (18)$$

P_i is the probability of occurrence of a pixel value in the image, L is the number of intensity levels in the image. Increase in the value of entropy after fusion indicates that the information contained in the image has increased.

3.5 Mutual Information:

It is an indicator of the information obtained from the source images and the quantity that is conveyed by the fused image.

$$MI = MI_{F,I1} + MI_{F,I2} \quad (19)$$

$$MI_{F,I1} = \sum P_{F,I1}(f,i1) \log_2 \frac{P_{F,I1}(f,i1)}{P_F(f)\, P_{I1}(i1)} \quad (20)$$

$$MI_{F,I2} = \sum P_{F,I2}(f,i2) \log_2 \frac{P_{F,I2}(f,i2)}{P_F(f)\, P_{I1}(i2)} \quad (21)$$

Where $MI_{F,I1}$ denotes the mutual information between the fused image and the first input image, $MI_{F,I2}$ denotes the mutual information between the second input and the fused image, $P_{F,I1}(f,i1)$ and $P_{F,I2}(f,i2)$ are the joint histograms of fused image, input 1and fused image, input 2 respectively, $P_F(f), P_{I1}(i1)$ and $P_{I1}(i2)$ are the histograms of the fused image, input 1 and input 2 respectively.

3.6 Run Time:

It gives a measure of the time required to execute the algorithm in seconds.

3.7 Image Quality Index:

Mathematically [9], $IQI = \frac{4\sigma_{xy}\bar{x}\bar{y}}{(\sigma_x^2+\sigma_y^2)[(\bar{x})^2+(\bar{y})^2]}$ (22)

where $\bar{x}, \bar{y}$ are the means of x and y respectively, σ_{xy} is the covariance of x and y, σ_x^2, σ_y^2 are the variances of x and y respectively.

4. Experimental Results

4.1 Experimental settings

All algorithms are coded using Python and implemented on Intel dual core, i5, 3.14GHz processor. Edges of input images are extracted using Sobel and Canny operators in both the fusion techniques. DWT-GA which uses HARR Wavelets for the decomposition is implemented for the reference [1, 11]. These algorithms are tested on different types of images from a database as well as on real time images captured by a camera (canon 60D, Tamron-SP 90 Di VC-macro), focusing on different object in every image. Different evaluation parameters mentioned previously are used for analyasis of results.

4.2 Results for database image sets

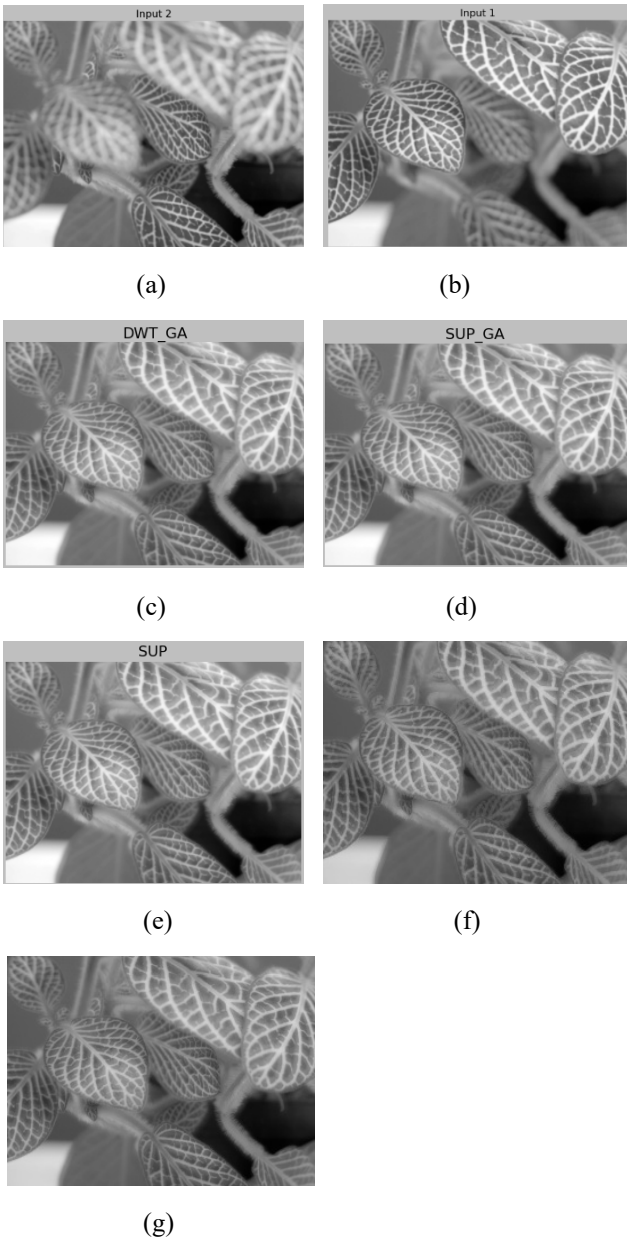


Fig. 2 (a)-(b): Data base input images, (c)DWT_GA (d)SUP_SOB (e) SUP_CA,(f) SUP_GA,(g) CA_GA

4.3 Results

Quantitative evaluation for different methods is carried out. Table I and II compares results for different parameters for Database and real time images respectively. They represent the average value of the respective evaluation parameter when the proposed algorithm is applied to database images and 20 real time images. Fig. 2 and Fig. 3 show results for sample input images taken from database and real time image sets respectively. These results are obtained after applying the proposed method to these samples. It was observed that proposed methods give better results compared to those obtained by DWT_GA[11]. The analysis also shows that for database images Sobel operator gives better performance whereas for real time images Canny operator can be preferred.

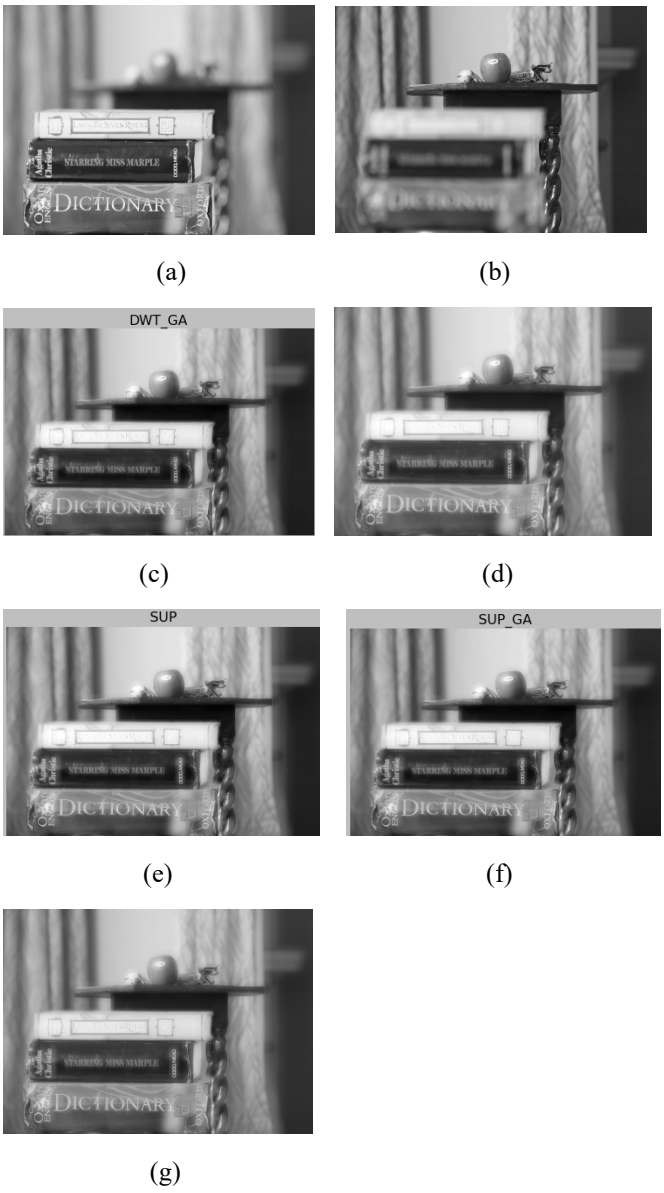
TABLE I. COMPARISION OF RESULTS FOR DATABASE IMAGE SETS

Method		***Edge-Superimposition***		***Edge-GA-Superimposition***	
Parameter	***DWT_GA***	***SUP_SOB***	***SUP_CA***	***SOB-GA***	***CA_GA***
PSNR	**34.409**	34.066	**34.865**	**34.409**	34.086
SSIM	**0.936**	0.934	**0.941**	**0.936**	0.929
ENTROPY	7.359	**7.371**	7.361	**7.359**	7.314
MI	**7.212**	7.185	**8.243**	**7.212**	6.880
IQI	**0.891**	**0.886**	0.878	**0.891**	0.862
TIME(SEC)	0.242	**0.008**	**0.007**	0.229	0.773

TABLE II. COMPARISION OF RESULTS FOR REAL TIME IMAGE SETS

Method		***Edge-Superimposition***		***Edge-GA-Superimposition***	
Parameter	***DWT_GA***	***SUP_SOB***	***SUP_CA***	***SOB_GA***	***CA_GA***
PSNR	21.076	20.768	21.365	21.076	**21.805**
SSIM	0.807	0.803	0.802	0.806	**0.810**
ENTROPY	7.337	7.262	7.401	7.336	**7.446**
MI	5.192	**5.363**	5.141	**5.192**	5.082
IQI	0.601	0.571	0.565	**0.601**	0.588
TIME(SEC)	0.253	0.012	**0.010**	0.241	0.782

4.4 Results for real time image sets



(a) (b) (c) (d) (e) (f) (g)

Figure 3: (a)-(b)Realtimeinput images, (c)DWT_GA, (d)SUP_SOB, (e)SUP_CA, (f)SUP_GA,(g) CA_GA

TABLE III: PERFORMANCE ANALYSIS OF PROPOSED EDGE BASED METHODS

Parameter	Images	
	From Database	Real time
PSNR	SUP-CA	CA-GA
SSIM	SUP-CA	CA-GA
ENTROPY	SUP-SOB	CA-GA
MI	SUP-CA	SUP-SOB
IQI	SOB-GA	SOB-GA
TIME	SUP-CA	SUP-CA

5. Conclusion

The follwing Table III gives us the consolidated perfomance of proposed techniques which work well with respect to DWT method. From above table it can be concluded that proposed edge based techniques work better than DWT based fusion techinque [11]. Edge GA based superimpostion method gives better performance for real time images as the weights are calculated and optimized using image statistics as compared to the constant values taken in edge superimposition method. The added advantage is its simplicity as decomposition and reconstruction of images is not required.As a result the method cosumes less run time compared to DWT based GA [11]. This property along with reduced computational complexity makes it simpler for hardware implementation and paralleling processing. Subjective analysis shows that the contrast of the images fused by the proposed method can be further improved. For real time images Canny based Genetic algorthm gives better performance but the computational time required is large. Thus, this algorithm can be further extended to fuse real time images in a quick and efficient manner which can be used in variety of applications.

References

[1] Arti Khaparde, Maitreyi Abhyankar and Vaidehi Deshmukh, “Spatial domain decision based image fusion”, 2016 IEEE/ACIS 15th International Conference on Computer and Information Science, ICIS 2016 – Proceedings, 23 August 2016, Article number 7550766; DOI: 10.1109/ICIS.2016.7550766.

[2] V. Aslantas, R. Kurban, “Fusion of multi-focus images using differential evolution algorithm”, Expert Systems with Applications vol.37, 2010, pp 8861–8870.

[3] Liu Cao; Longxu Jin; Hongjiang Tao; Guoning Li; ZhuangZhuang; Yanfu Zhang, “Multi-Focus Image Fusion Based on Spatial Frequency in Discrete Cosine Transform Domain”, IEEE Signal Processing Letters, Year: 2015, Volume: 22, Issue: 2, pp. 220- 224.

[4] Y. Zhang, et al., Multi-focus image fusion based on cartoon-texture image decomposition, Optik - Int. J. Light Electron Opt. (2015), http://dx.doi.org/10.1016/j.ijleo.2015.10.098

[5] Xiaoli Zhang, XiongfeiLi, ZhaojunLi, YuncongFeng, “Multi-focus image fusion using image-partition-based focus detection”, Signal Processing, vol.102, (2014), pp. 64–76

[6] Said Pertuz, Domenec Puig, Miguel Angel Garcia, Andrea Fusiello, “Generation of All-in-focus Images by Noise-robust Selective Fusion of Limited Depth-of-field Images”, IEEE Transactions on Image Processing, Year: 2013, Volume: 22, Issue: 3, Pages: 1242 – 1251.

[7] Baohua Zhang, Xiaoqi Lu, Haiquan Pei, He Liu, Ying Zhao and Wentao Zhou,“Multi-focus image fusion algorithm Based on focused region extraction”, Neurocomputing.

[8] X. Zhang, et al., “A new multifocus image fusion based on spectrum comparison”, Signal Processing 2016, http://dx.doi.org/10.1016/j.sigpro.2016.01.006i

[9] Jinsheng Xiao, Tingting Liu, Yongqin Zhang, Baiyu Zou, Junfeng Lei and Qingquan Li, “Multi-focus image fusion based on depth extraction with inhomogeneous diffusion equation”, Signal Processing,

[10] Veysel Aslantas, Rifat Kurban, “Extending depth-of-field by image fusion using multi-objective genetic algorithm”, 7th IEEE International Conference onIndustrial Informatics, 2009. INDIN 2009, pp. 331-336.

[11] Chaunté W. Lacewell, Mohamed Gebril, Ruben Buaba, Abdollah Homaifar, “Optimization of Image Fusion Using Genetic Algorithms and Discrete Wavelet Transform”, Radar Signal & Image Processing, IEEE, 2010

[12] Chuanzhu Liao, Yushu Liu, Mingyan Jiang, “Multi-focus image fusion using Laplacian Pyramid and Gabor filters”, Applied Mechanics and Materials, vol. 373-375, 2013 , pp 530-535.

[13] Shutao Li, Bin Yang, “Multifocus image fusion by combining curvelet and wavelet transform”, Pattern Recognition Letters 29, 2008, pp:1295–1301.

[14] G. Guorong; X. Luping; F. Dongzhu, "Multi-focus image fusion based on non-subsampled shearlet transform", IET Image Processing, Year: 2013, Volume: 7, Issue: 6, pp: 633 – 639.

[15] Y.Yang; S.Tong; S.Huang; P.Lin,"Multifocus Image Fusion Based on NSC T and Focused Area Detection", IEEE Sensors Journal, Year: 2015, Volume: 15, Issue: 5, pp. 2824 – 2838

[16] Xiangzhi Bai, Fugen Zhou, BindangXue, "Edge preserved image fusion based on multi-scale toggle contrast operator", Image and Vision Computing, vol.29, 2011, pp 829–839.

[17] Yong Jiang, Minghui Wang, "Image Fusion using multi-scale edge-preserving decomposition based on weighted least squares filter", IET image Processing, 2014, vol. 8, Issue 3, pp.183-190.

[18] Bavirisetti DP, Dhuli R, Multi-focus image fusion using multi-scale image decomposition and saliency detection, Ain Shams Eng J, 2016, http://dx.doi.org/10.1016/j.asej.2016.06.011

[19] Wei-We Wang, Peng-Lang Shui, Guo-Xiang Song,"Multifocus Image Fusion In Wavelet Domain", Proceedings of the Second International Conference on Machine Learning and Cybernetics, November 2003.

[20] Tanish Zaveri, Mukesh Zaveri," A Novel Region based Image Fusion Method using High-boost Filtering", Proceedings of the 2009 IEEE International Conference on Systems, Man and Cybernetics, San Antonio, TX, USA, October 2009

[21] Hugo R. Albuquerque, Tsang Ing Ren, George D. C. Cavalcanti,"Image Fusion Combining Frequency Domain Techniques Based on Focus", IEEE 24th International Conference on Tools with Artificial Intelligence, 2012.

[22] Wencheng Wang, Faliang Chang, "A Multi-focus Image Fusion Method Based on Laplacian Pyramid", Journal Of Computers, Vol. 6, No. 12, December 2011, pp. 2559-2566

[23] Yifeng Niu, Lincheng Shen, Lizhen Wu and Yanlong Bu," Optimizing the Number of Decomposition Levels for Wavelet-Based Multifocus Image Fusion", Proceedings of the 8th World Congress on Intelligent Control and Automation, Jinan, China, July 2010

[24] Zhou Wang, Alan C. Bovik, "A Universal Image Quality Index", IEEE Signal Processing Letters, March 2002.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 57-68 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

A Novel Approach for Designing Mobile Native Apps

Sasmita Pani*, Jibitesh Mishra

College of Engineering and Technology, Bhubaneswar, India

ARTICLE INFO

Article history:
Received: 07 December, 2016
Accepted: 14 January, 2017
Online: 28 January, 2017

Keywords:
Content Design
Architecture Design
Conceptual Design
Personification Design

ABSTRACT

Mobile devices are differed from desktop based systems in terms of particular execution environment, constrained resources, and high mobility requirement. To overcome these shortcomings, various agile based methodologies are developed for native mobile applications such as Mobile-D, Scrum etc. These agile techniques are based on various phases and these phases begin from exploring, initializing and implementing the mobile apps. But these techniques are not focusing on elaborating design model for mobile native apps. The aim of the paper is to provide a layered approach or layered model for design of mobile native apps which can be used as a framework for developing mobile native apps. Any mobile native app developer can use this sequential approach or design model for design and development of mobile native apps. This design model gives a standard or framework, based on which generic native mobile apps can be designed and developed. This paper also shows an empirical analysis among the web app design models with the proposed design model for mobile native app development.

1. Introduction

Mobile devices are the pervasive based systems embedded with various sensors and powerful processors which can provide information about any domain like agriculture, health care system and learning system. There are various platforms of mobile phones like ios phone, windows phone, android phone where mobile native apps are developed. The agile software methodology is used to develop any mobile apps but this approach is not taken consideration about design modeling of generic native apps under mobile domain. Generally the agile process models are using HTML, HTML5, JavaScript, XML and CSS for design of mobile native apps. These agile process models doesn't provide design model for mobile native apps. There also exists various web app design models such as UWE, WebE, OOHDM and OOWS for designing web apps but not for mobile native apps. We have developed a design model consisting of seven design phases such as content design, architecture design, conceptual design, user interface design, navigation design, presentation design and personification design. The content design describes about how the contents of a generic mobile native app will be organized, constructed and visualized on the screen of a mobile device. The architecture design focuses the specific way of describing the various components needed for developing generic mobile native apps. The conceptual design specifies about how the generic mobile native app will build semantically and consistently so that user can get meaningful information from the mobile apps. The user interface design identifies various widgets, views, input control elements, layouts etc. which enables the user interactions with the native mobile apps in a ubiquitous manner. The navigation design describes about how a user will navigate from one screen to another screen in a native mobile apps. The presentation design focuses on various transition and animation mechanisms for efficient visualization of mobile native apps. The personification design discusses about integration of social apps with the mobile native apps so that user can get personalized information and can be benefited.

Here the research process is to identify and provide a design model for generic mobile native apps which incorporates various design elements like content, architecture, user interface, transition animations and social media tools etc and provide an empirical analysis among the web app design models with the proposed mobile native app design models.

The rest of the paper is organized as follows. Section 2 discusses related works on web app development process models. Section 3 provides the research approach for developing design model for mobile native apps. Section 4 discusses design elements and proposes a design model for mobiles native apps. Section 4 also gives an empirical study about the existing web app models

*Corresponding Author: Sasmita Pani, College of Engineering and Technology, Bhubaneswar, India, +919776216142, susmitapani@gmail.com

and the proposed mobile native app design model. This section also discusses the different design phases for the development of mobile native apps. Section 5 provides the development practices of different design phases in mobile native apps. Section 6 provides conclusion and future work respectively.

2. Background Study

2.1. Web App Design models or Process Models

WebE [1], is a web development methodology employs agile process model when the web app development process is considered as an immediacy and continuous evolution process. This methodology follows incremental process model when the web app development process is developed over a long period of time. The WebE process framework consists of various phases such as customer communication, planning, modeling, construction, delivery and evaluation. Customer communication phase includes business analysis and formulation as steps to build a web app. Business analysis determines the business or organizational context for the web app. The business analysis includes motivation, usefulness and specifies users of the web app to be developed. The formulation phase involves requirement gathering activity which includes all stakeholders for describing the problem that the web app is to be solved. The planning phase consists of task definition and a timeline schedule for the time period used for the development of web app. The modeling phase includes analysis model and design model. The analysis model provides a basis for web app design. It analyses contents, user's interaction, usage scenarios and a particular environment for a web app to be developed. The design model employs various design activities such as content design, aesthetic design, architecture design, interface design, navigation design and component design.

The object oriented hyper design method (OOHDM) [2], includes hypermedia applications in four phases i.e. conceptual design, navigational design, abstract interface design and implementation. These phases are performed in a mix of incremental, iterative and prototype-based development styles. During each activity a set of object- oriented models are given to describe about how a particular design concern is built or enriched from previous iterations. In the conceptual phase, a class scheme is built out of sub-systems, classes and relationships. Navigational design is considered a critical step in the design of a hypermedia application. Each navigational model is built as a view over conceptual models. It is expressed into two schemas the navigational class schema and navigational context schema. OOHDM provides ADV (Abstract Design View) approach provides the way in which different navigational objects will look and which interface objects will activate the navigation. The abstract interface design activity specifies interface objects at abstract level through which the user will perceive. In this implementation phase, the designer will actually implement the design. In this phase, the designer has to decide how the conceptual and navigational objects will be stored. It can be done by mapping of object oriented model to relational model.

WebML [3], is modeling language for designing web sites at the conceptual level. WebML supports XML syntax which enables the high-level description of a web site. WebML consist of four orthogonal perspectives or aspects for specifying a web site. These are structural model, hypertext model, presentation model and personalization model. From the structural design point of view, it undergoes in two phases such as requirement gathering and data modeling phases. The requirement gathering phase includes objective of the web site, style guideline of the web content, constraints due to legacy data. The hypertext design aspect considers two ways of hypertext design models such as composition model and navigational model. The composition model also can be defined as by designing the "hypertext at large" since it specifies pages, units and provides hyperlink among the pages through units. The composition model specifies nodes for making up the hypertext contained in the web site. This model also defines the pages that compose the hypertext and the content units those make up a page. Six types of content units can be used to compose pages: data, multi-data, index, filter, scroller and direct units. The presentation model specifies the layout and graphic appearance of pages, independently of the output device and of the rendition language that is used generating web pages. It adds to each page a presentation style after all pages are sufficiently stable. It is concerned with the actual look and feel of the pages identified by composition modeling. From the personalization perspective, it enables two ways of design approach for designing web which includes user and group design. In the user and group modeling, user and groups are explicitly modeled in the structure schema.

UWE [4], stands for UML based web engineering, is composed of different models at different phases such as requirements engineering, analysis, design, and implementation of the development process which are used to represent different views of the web application that includes content view, navigation view and presentation view. UWE includes UML diagrams to represent structural aspect of different views and behavioural aspect of web apps. Requirement modeling is done at two steps. At first a rough description of the functionalities is produced by UML use case diagram. In this step, UWE provides three types of use cases such as navigation, process, and personalized use cases. Navigation use cases are used to model typical user behaviour when interacting with a web application. Process use cases are used to describe business tasks that end users will perform with the web apps. Personalized use cases are used to provide customization services to user with in web apps. In a second step, a more detailed description of the functionalities is developed by UML activity diagrams that give the responsibilities and actions of the stakeholders. The navigation model comprises the navigation space model and the navigation structure model. The first specifies which objects can be visited by navigation through the web application and the second specifies how these objects are reached. The navigational space model specifies two modeling elements such as navigation class and navigation association. These elements are termed as page and link in web terminology. The presentation model, is based on the navigation model provides an abstract view of the user interface of a web app. The presentation model only describes the basic structure of the user interface by including UI elements such as text, images, anchors, forms. These UI elements are used to present the navigation nodes. The presentation model also

specifies how each node or page is presented to a user and how the user can interact with them.

OOWS is an object oriented web solution method [5], which provides a strategy for going from the problem space, is presented as conceptual models to the solution space that is the final software product in an automatic way. The OOWS web app development methodology is composed of two main phases such as conceptual modeling and solution development. The conceptual modeling phase is again decomposed into three sub phases sub steps such as functional requirements elicitation, classic conceptual modeling, navigation and presentation modeling. In the functional requirement elicitation phase, conceptual schema is built by applying use cases and scenarios of university department by taking it as case study. In the classic conceptual modeling sub phase, structural model, functional and dynamic models are used to capture the system structure and behaviour. The navigational modeling is done in two phases which includes user identification and categorization step and navigational diagram specification step. This user identification and categorization step provides a user diagram which specifies the system kind of users and the interaction among users. The diagram also specifies the accessibility of each kind of user to the system information and functionality. The navigational diagram specification phase provides a navigation model to capture the navigational requirements of web apps by defining a navigational view or navigational map for each kind of relevant users of the system. The presentation modeling provides various presentation patterns such as information paging pattern, circularity pattern and layout patterns which are used to specify navigation features and capture the essential requirements for the web app development. The specified information will be used by the model compiler to generate the web pages interface. The solution development is the second main step in the OOWS approach which is determined by the target platform and is chosen a specific architectural style.

WSDM [6], stands for web site design method, is an audience driven methodology which does the requirement analysis from audience point of view unlike the other web apps development methodologies like OOHDM, WebML, OOWS, UWE and WebE. The WSDM is a user centric web app development approach and it integrates with V-model for testing of each phase to enhance the maintainability and effectiveness of the web site. WSDM phases include mission statement phase, audience modeling phase, conceptual design phase, implementation design phase and implantation phase. The mission statement phase defines the subject, purpose and specifies the intended users and declares them as target audience. It ensures that the designer should clearly establish the borders of the design. The audience modeling phase provides users a general indication of the audiences involved in the web app. The users identified in the mission statement are taken as a starting point and classified into different audience classes based on their information and functional requirements.

3. Research Approach

In order to explore the issues around mobile native app design specification processes and what characteristics are typically included in these design specifications, we have established three research questions (RQ1to RQ3). These are as follows.

RQ1. How the design phases are developed from various participants for developing mobile native apps?
This question RQ1 was established to develop the design phases through sending questionnaires to different mobile app developers of different software industries. Based on the responses we have obtained design elements and design phases for mobile native app development.

RQ2. How the design models are defined and modeled in mobile native apps considered in existing process models under web app domain for effective design?
In order to define the design models and modeling them in developing mobile native apps, RQ2 was established to examine the current process trends into designing web applications through empirical study, with a particular focus developing the design phases in mobile native apps through mapping the current web app design models and its attributes into those of mobile native apps.

RQ3.How the different attributes or elements under mobile native app development are identified in order to give effective design for mobile native apps.
For specifying the design elements and modeling them in semantic way, RQ3 was established to identify and understand the attributes or elements that lean in different design phases using different specifications under mobile domain.

We undertook an extensive set of questionnaires (60) based on these research questions and send these questionnaires to different mobile app software developers of different software industries of India having international offices as well. The questions were placed in Google forms and sent to different mobile app developers of different software industries. After getting their responses, we are able to find out different attributes or elements used in different design phases under mobile native app domain. All these sending questionnaires and response processes were conducted over a period of 6 weeks, primarily during October/November 2015.

4. Proposed Design Model for Mobile Native Apps

Although different mobile app development methodologies describes various phases such as preparation, design, code generation, testing and commercialization of the product for developing mobile apps but they don't emphasize on design modeling of mobile native apps. There is no specific design model for mobile native generic apps so that any user can design the generic mobile native apps and based on the design model, development for mobile native apps can be done. Here our purpose is to develop a design model for mobile native apps which is built by going through the best practices in various software industries.

Some of the questions and responses obtained from the mobile app developers are given below.

4.1. Questionnaires through Google Forms

How are the contents organized for designing mobile native apps?

Ans. At first pages are created and XML layouts are used.
No of respondents for this answer specified above = 7
No of respondents not for this answer=3

Is the sensing mechanism used in architecture design or other design?

a. Yes

b. No

No of respondents for a = 8
No of respondents for b =2

Which design phase is used for moving one screen to another screen?
Ans. Navigation design
No of respondents for this answer specified above = 9
No of respondents not for this answer=1

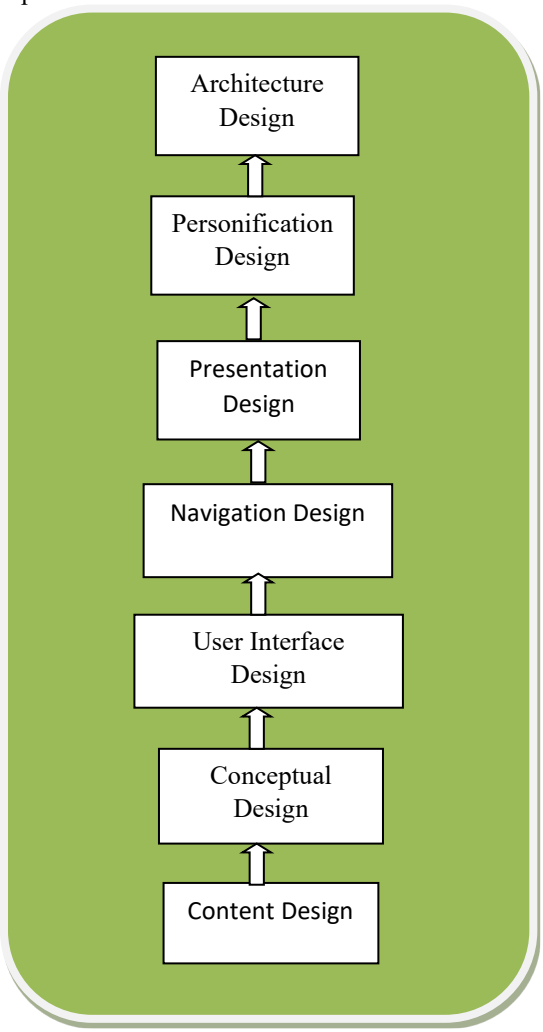


Figure 1. Proposed design model for mobile native apps

After collection of data from different software industries and through the response of different mobile app developers, the design model is built which is consisting of various phases such as content design, architecture design, structural design, user interface design, navigation design, presentation design and personification design for mobile native apps which is shown in fig.1. Also based on the responses from the mobile app developers, we have found out various design elements or attributes. Some of the design attributes are shown below.

a. Different sensors
b.Keyboard
c.WLAN
d.User's role and profile
e.User's perceiveness

4.2. An Empirical Analysis among the other Web App Design Model with the Proposed Mobile Native App Model

We have studied the existing design models for web app development such as WebE, UWE, and WebML etc. and identified the design phases of those models as discussed in section related work. These design models are used only for designing and developing web apps, but not for mobile native apps. Thus we have proposed and build a design model for mobile native app development which is consisting of seven phases and we have done an empirical analysis among the web app development design models with our proposed design models.

	Design Phases						
	Content Design	Conceptual design	User Interface Design	Navigation Design	Presentation Design	Personification Design	Architecture Design
UWE	No	Yes	It is included in the presentation model	Yes	Yes	No	No
WebML	No	It is included in the structural model	It is included in the composition model.	Yes	Yes	Yes	No
OOHDM	No	Yes	It is included in abstract interface design	Yes	Yes	No	No
WSDM	No	Yes	It is included in the presentation design sub phase of implementation design phase.	It is included in the navigation modeling sub phase of conceptual design phase.	It is included in the presentation modeling sub phase of implementation design phase.	No	No
OOWS	No	It is included in the classic conceptual modeling phase	No	It is included in the conceptual modeling phase	It is also included in the conceptual modeling phase	No	It is included in the solution development phase.
WebE	Yes	It is included in the architecture design phase	Yes	Yes	Yes	No	No
Proposed mobile app design model	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 1. An empirical analysis among the web app design model with the proposed mobile native app design model

In the above table 1 provides an empirical analysis of the design phases between the proposed mobile native app design model and the web app design model.

The content design phase is present only in UWE and WebE design model and our proposed mobile app design model, which defines the content to be designed. This design phase is not present in any other web app design models such as OOHDM, WebML, OOWS and WSDM etc.

The architecture design is given only in the OOWS web app design model and our proposed mobile app design model. The architecture design is given in the solution development phase of OOWS web app design model. The conceptual model is provided in the all web app design models as mentioned above in table 1 and our proposed mobile app design model. The architecture design phase in WebE, defines the template which provides consistent hypermedia structure for web app development and hence it is included in the structural model. In OOWS, the conceptual model is included in the classic conceptual modelling sub phase of conceptual modelling phase. The user interface design is present in all the web app design models except OOWS and our proposed mobile app design model. In UWE and WebML, it is included in the presentation model and composition model. In OOHDM and WSDM, it is included in the abstract interface design and presentation design sub phase of implementation phase. The navigation design is also present in all the web app design models and our proposed mobile app design model. In WSDM, it is provided in the navigation modelling sub phase of conceptual design phase. In OOWS, it is included in the conceptual modelling phase. The presentation design is present in all the web app design models and our proposed mobile app design model. In WSDM, it is included in the presentation modelling sub phase of implementation design phase. In OOWS, it is included in the conceptual modelling phase. The personification design is present only in WebML and our proposed mobile app design model.

By far the most significant observation from the responses of questionnaires sending to the mobile native app software developers of different software industries is that identifying the "design phases" or "design aspects" for developing mobile native generic apps. Further these design phases are built in XML language and android studio framework for developing mobile native apps.

4.3. Identifying attributes leaning different design phases in mobile native apps

The device specific attributes includes different sensors, memory, keyboard, gestures and screen resolution. The attributes sensors, memory and keyboards are configured in the system software irrespective of platform and hence it reflects in the architecture design for the generic mobile native app development. The gestures and screen resolution attributes enhances the usability of mobile native apps for the end-user and these attributes are concerned in developing the interface for user in mobile generic native apps. The mobility specific attributes include WLAN, GPRS, EDGE etc and these are configured in the hardware of any platform and hence these are considered in the architecture design of mobile native generic apps.

The user specific attributes include user's role, profile, user's perceiveness and user's usefulness. All these attributes are considered in developing user interface for mobile native apps and hence these attributes reflects in the user interface design of mobile native generic apps. The contents are the views or pages which are designed depending upon all the user specifications and hence all these user specific attributes are concerned in developing content of mobile native generic apps. All the user's specification attributes are considered in developing conceptual design because user specification is the main aspect for building mobile native generic apps. While developing the navigation, it concerns about user's perceived ease of use and ease of usefulness towards using the native application in a mobile device. Hence these two attributes such as user's perceiveness and user's usefulness are reflected in the navigation design for developing mobile native apps. The social specific attributes are used for improving the social interactions of user's using the mobile native apps. Hence these attributes are used for optimizing the social tools providers to the users to enhance personification in the mobile native apps.

Different specification under Mobile domain	*Attribute's name*	*Design phases*
Device specific	Different Sensors	Architecture design
	Memory	Architecture design
	Keyboard	Architecture design
	Gestures	User Interface design
	Screen resolution	User Interface design
Mobility Specific	WLAN	Architecture design
	GPRS	Architecture design
	EDGE	Architecture design
	Bluetooth	Architecture design
	GPS	Architecture design
User specific	User's role and profile	User Interface design

	User's perceiveness	User Interface design
	User's usefulness	User Interface design
	User's perceiveness	Navigation design
	User's usefulness	Navigation design
	User's role and profile	Content design
	User's perceiveness	Content design
	User's usefulness	Content design
	User's role and profile	Conceptual design
	User's perceiveness	Conceptual design
	User's usefulness	Conceptual design
Social specific	Blogs	Personification design
	Content Hosting	Personification design
	Social Networking	Personification design
	Podcasting	Personification design
	Wikis	Personification design

Table 2. Attributes in Developing Design Phases of Design Model

4.4. Content Design

Content design plays an important role in developing mobile native apps. It is because the content need to appear on which screen of mobile native apps and hence it need to be organized and constructed very well. At first the contents for generic mobile native apps are organized in each page and each page corresponds to each xml or java script or html file and java activity file depending upon the different platforms of mobile native apps. In android platform, each page corresponds to XML and java activity file which are shown in fig.2.

4.5. Conceptual Design

The agile software methodology is used to develop any mobile apps but this approach is not taken consideration about semantic modelling of data under the generic mobile domain. Here the approach is to build mobile app in a systematic and semantic way so that the native mobile app user can get the required meaningful information. The conceptual model are built or designed with ontology in the mobile native apps domain to describe the concepts and axioms that determine the semantics of the concepts in that domain.

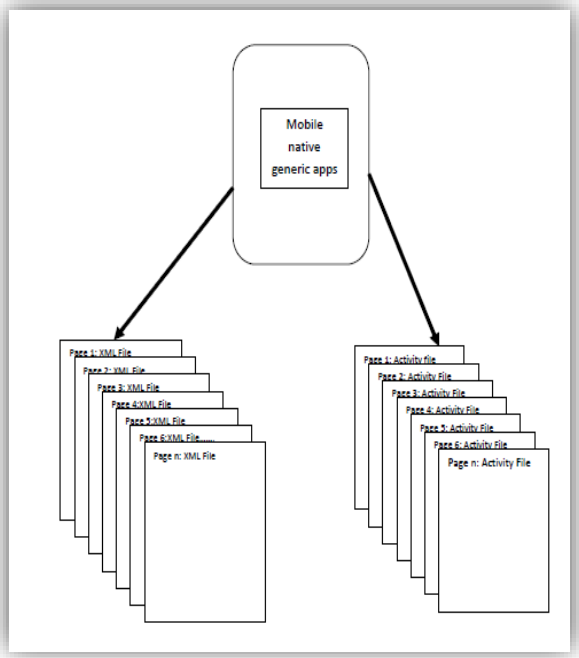


Figure 2. Content Design in Mobile Native Apps

To define the meaning of representing such model to be applied in the semantic web, few mark up languages are specifically available. Web ontology language or OWL [7], is the modern modelling language available to create and represent the ontology having classes and relationships between the classes, clearly describes its meaning so that mobile native application can be developed in a structured, semantic and consistent manner. The conceptual design will work as hidden layer in the mobile native apps so that native app user can get meaningful information. Here we have taken the user specific element or attribute and it has three sub attributes such as based on role, based on preferences and based on usefulness for building mobile native generic app. We have given focus to the user aspect element for designing it semantically because it plays an important role on development of mobile native generic app design model. We have modelled this user aspect element using OWL DL and established class axioms which are shown below.

4.5.1 Building Class Axioms through Making Subclass

For example the user specific element is represented as class and the attributes are represented as subclasses in ontology which is referred as sub class. Here it is done in OWL modelling

language in protégé 5.0 beta framework [8], which is shown below.

```
<!--
http://www.semanticweb.org/ontologies/2015/6/untitled-
ontology-176#BasedOnRole -->

    <owl:Class
rdf:about="http://www.semanticweb.org/ontologies/2015/
6/untitled-ontology-176#BasedOnRole">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/ontologies/20
15/6/untitled-ontology-176#UserSpecific"/>
    </owl:Class>
Similarly the attribute based on preference can be
made subclass to the class user specific in the
ontology for building the semantics.

        <!--
http://www.semanticweb.org/ontologies/2015/6/untitled-
ontology-176#BasedOnPreferences -->

    <owl:Class
rdf:about="http://www.semanticweb.org/ontologies/2015/
6/untitled-ontology-176#BasedOnPreferences">
        <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/ontologies/20
15/6/untitled-ontology-176#UserSpecific"/>
    </owl:Class>
```

4.5.2 Building Class Axioms through Making Classes Disjoint

To build the ontology consistently and semantically, an instance of one class cannot be instance of another class and hence it is achieved by structuring them as disjoint classes which is modelled through OWL DL language.

```
<rdf:Description>
        <rdf:type
rdf:resource="&owl;AllDisjointClasses"/>
        <owl:members rdf:parseType="Collection">
            <rdf:Description
rdf:about="http://www.semanticweb.org/ontologies/2015/
6/untitled-ontology-176#BasedOnPreferences"/>
            <rdf:Description
rdf:about="http://www.semanticweb.org/ontologies/2015/
6/untitled-ontology-176#BasedOnRole"/>
            <rdf:Description
rdf:about="http://www.semanticweb.org/ontologies/2015/
6/untitled-ontology-176#BasedOnUsefuleness"/>
        </owl:members>
    </rdf:Description>
```

4.6 User Interface Design

The user interface design takes into account various elements for designing mobile native apps. The elements are specific to platform for a mobile generic native app development. In android platform [9], the elements for designing user interface are in mobile native are

- Different layouts
  - Linear layouts
  - Relative layouts
  - Frame layouts
  - Table layout
  - Grid layout
- Widgets
  - Buttons
    - Radio buttons
    - Toggle buttons
    - Image buttons
  - Text views
  - Image views
  - Web views
  - Progress bars
  - Check box
  - Rating bars
  - Date and time pickers
  - Menu
- Containers
  - Scroll view
  - Grid view
  - Radio group
  - List view

4.6.1 Designing Layouts

The layouts for designing mobile native apps should be matched or wrapped with screen size of the mobile device which is referred as match_parent or wrap_content as shown below.

```
   <?xml version <?xml version="1.0" encoding="utf-8"?>
<RelativeLayout
xmlns:android="http://schemas.android.com/apk/res/andr
oid"
    android:layout_width="match_parent"
android:layout_height="match_parent">
</RelativeLayout>
```

4.6.2 Designing Widgets

4.6.2.1 Designing Textviews

To view text we can use the tag <TextView> and to provide input during run time we can use the tag <EditText> for designing mobile native apps. The attributes of the textviews are height, width, id and text etc. The height and width of the text view and edit text can either be matched with the parent layout or enclosed around the content of the text.

```
<TextView
android:layout_width="wrap_content"
        android:layout_height="wrap_content"
    android:text="Hello World!"
    android:id="@+id/textView" />[Here id identifies
the textview anywhere in the code]
      <EditText
android:layout_width="wrap_content"
android:layout_height="wrap_content"
 android:text="Hello World!"
 android:id="@+id/EditText" />
```

Here if more than one textview and edittexts are added, then they should be labeled according to their occurrence in the mobile native apps.

4.6.2.3 Designing Buttons

There exists various buttons such as general buttons, toggle buttons, radio buttons and image buttons for designing mobile native apps. The buttons are used for initializing an event or action corresponding to the respective stimuli or input. The attributes defined for buttons are height, width, id, source and text. The attribute "text" specifies text displayed on the button. The attributes are defined within the specified button tag like <Button> and they are represented as follows.

```
<Button
         android:layout_width="wrap_content"
        android:layout_height="wrap_content"
        android:text="Press Button"
        android:id="@+id/button"
        />
     <RadioButton
       android:layout_width="200dp"
        android:layout_height="wrap_content"
        android:text="Yes"
        android:id="@+id/radioButton"
      />
    <ImageButton
      android:layout_width="wrap_content"
        android:layout_height="wrap_content"
        android:id="@+id/imageButton"
        android:src="@drawable/thumb1small"[src :
attribute is used for accessing the image on the
button]
```

4.6.3 Designing Containers

The containers are used to present group of elements in a single view or page. The elements like group of radio buttons, checkboxes, list of items, textviews etc. are represented through containers. The different types of containers are scrollview, listview and radiogroup etc. and the attributes specified same as above with some new attributes such as discussed below.

```
<ScrollView
xmlns:android="http://schemas.android.com/apk/res/andr
oid"
        android:id="@+id/content"
        android:layout_width="match_parent"
        android:layout_height="match_parent">
     <TextView style="?android:textAppearanceMedium"
android:lineSpacingMultiplier="1.2"
            android:layout_width="match_parent"
            android:layout_height="wrap_content"
            android:text="@string/lorem_ipsum"
            android:padding="16dp" />
 </ScrollView>
```

Here <ScrollView> is the container tag with in which any widgets or group of widgets can be defined and it reduces the code complexity by grouping the like elements together with in container.

Note: - For every activity_XML file, there is a corresponding main activity java file or individual java activity file which contains all the required packages, methods, references and instances to perform the various actions and the corresponding listener methods irrespective to mobile native apps.

The containers are used to present group of elements in a single view or page. The elements like group of radio buttons, checkboxes, list of items, textviews etc. are represented through containers. The different types of containers are scrollview, listview and radiogroup etc. and the attributes specified same as above with some new attributes such as discussed below.

```
<ScrollView

xmlns:android="http://schemas.android.com/apk/res/andr
oid"
        android:id="@+id/content"
        android:layout_width="match_parent"
        android:layout_height="match_parent">
     <TextView style="?android:textAppearanceMedium"
android:lineSpacingMultiplier="1.2"
            android:layout_width="match_parent"
            android:layout_height="wrap_content"
            android:text="@string/lorem_ipsum"
            android:padding="16dp" />
 </ScrollView>
```

Here <ScrollView> is the container tag with in which any widgets or group of widgets can be defined and it reduces the code complexity by grouping the like elements together with in container.

Note: - For every activity_XML file, there is a corresponding main activity java file or individual java activity file which contains all the required packages, methods, references and instances to perform the various actions and the corresponding listener methods irrespective to mobile native apps.

4.7 Navigation Design

The navigation mechanism provides traversing between the different screens or pages or views with in a mobile native app. The navigation is performed in various ways such as top down navigation, front navigation, back navigation etc. and navigation in any direction if possible. The navigation mechanism is specific to various platforms. In android platform, all the navigation mechanisms are handled by navigation controller. The navigation drawer provides custom navigation interface. The navigation drawer is defied within the tag <android.support.v4.widget. DrawerLayout>, is shown below.

```
    <android.support.v4.widget.DrawerLayout>
xmlns:android=http://schemas.android.com/apk/res/andro
id
android:id="@+id/drawer_layout"
android:layout_width="match_parent"
    android:layout_height="match_parent">
</android.support.v4.widget.DrawerLayout>
```

Note: - For every activity_XML file, there is a corresponding main activity java file or individual java activity file which contains all the required packages, methods, references and instances to perform the various actions and the corresponding listener methods irrespective to mobile native apps.

4.8 Presentation Design

For providing interactive interfaces or environment presentation design is used in generic mobile native apps. The presentation design focuses on presenting a view in a native mobile app. It includes various transitions _animations to provide a view or a page in the mobile app. For presentation design transition and animations are used in various ways such as zoom in, zoom out, fade in, fade out, and card flip. For fade in operation. It is designed in XML and shown below.

```
<com.example.animationdemo.TouchHighlightImageButton[T
his button is used for displaying image instead of
text]
 android:id="@+id/thumb_button_1"
 android:layout_width="100dp"
                android:layout_height="75dp"
                android:layout_marginRight="1dp"
                android:src="@drawable/thumb1"
                android:scaleType="centerCrop"

android:contentDescription="@string/description_image_
1" />[For zooming image_1]
```

Note: - For every activity_XML file, there is a corresponding main activity java file or individual java activity file which contains all the required packages, methods, references and instances to perform the various actions and the corresponding listener methods irrespective to mobile native apps.

4.9 Personification Design

The personification design allows the mobile native app user to share information with another user through social media tools such as Facebook, Twitter, LinkedIn and Google plus. It can be designed using different layouts, views and make it more interactive we can use different buttons and textviews. The following is shown below.

```
    <ScrollView
xmlns:android=http://schemas.android.com/apk/res/androi
d
android:layout_width="fill_parent"
    android:layout_height="fill_parent"
      android:background="@color/grey_light">
```

Here “ScrollView” is used to encapsulate the contents that will be displayed in this particular page or view. The elements used here are linear layout, relative layout, buttons and text views. Scroll view is used here to adjust the list of users within the display, when a mobile native app user tries to connect with its friends or users which overfills the display size of the mobile device.

```
   <LinearLayout
  android:layout_width="match_parent"
  android:layout_height="wrap_content"
  android:id="@+id/buttonLayout"
  android:layout_below="@+id/card"
  android:layout_alignParentLeft="true"
  android:layout_alignParentRight="true"
   android:gravity="center"
   android:background="@color/grey_light">
     <Button
android:layout_width="match_parent"
                android:layout_height="match_parent"
                android:text="Share"
                android:id="@+id/share"
                android:padding="8dp"
                android:background="@color/dark"
                android:layout_weight="1"
                android:textColor="#ffffffff"/>
        </LinearLayout>
    </ScrollView>
```

4.10 Architecture Design

The architecture design concerns about different mechanisms such as mobility, sensors, input, output and memory mechanisms. This architecture design shown in fig.3, incorporates elements WLAN, GPRS, GPS, and browsers for mobility mechanisms in designing mobile native apps. It is also taking concern about different sensors such as location sensor, touch sensor, proximity sensor and temperature sensors etc. for sensing mechanism in designing the architecture of mobile native generic apps. The elements includes in the architecture design are only used for improving the versatility of the mobile native apps and these are configured into the hardware of the different platforms.

The elements for mobility mechanisms WLAN, GPRS, EDGE and Bluetooth. These elements are hardcoded into the hardware of the mobile device through establishing connections via adapters and sockets irrespective of any platform. The functionality of these mobility elements can be achieved by defining events and listeners for particular operation in mobile native generic apps. These are configured in the mobile device through protocols and their features are accessed by applications developed for this purpose irrespective of platform. There are various types of sensors such as position sensor, motion sensor, environment sensor and location based sensors (GPS). Mostly these sensors are built in mobile devices of any platform. The functionality of sensors are quite similar to the mobility mechanisms. Sensors are used in generic mobile native apps to perform specific operations like sensing temperature, getting location services, finding position of the mobile device and adjusting the orientation of the device.

The sensor framework is a part of hardware package irrespective of platform for any mobile native app. The sensor package consist of various classes and interfaces such as Sensor, SensorEvent, Sensormanager and SensorEventListner for utilizing sensor mechanisms. These interfaces include different methods for performing different actions and events in any platform, in mobile native apps.

Mobile devices incorporate memory mechanisms such as internal memory and external memory irrespective of platform. These memory mechanisms are configured by the mobile device manufacturer for running operating systems of various platforms and mobile native apps. The input and output mechanisms are provided by the system application of the mobile device irrespective of any platform for providing UI to the native app user. The functionality of the input and output mechanisms can be modified based on client requirements for developing mobile

native apps. The architecture design can also be included in other design phases of the proposed mobile native app design model.

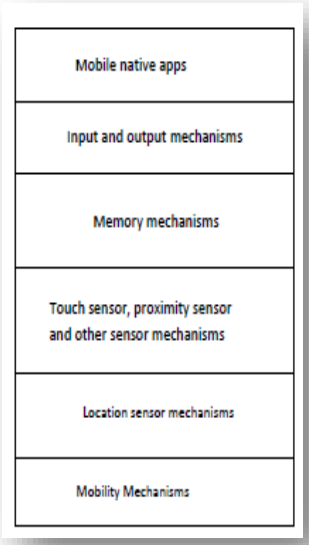


Figure 3. Architecture Design in Mobile Native Apps

5. Development Practices

Here the design phases such as user interface design, navigation design, presentation design and personification design can be developed and built irrespective of any platform. Here we have used android studio framework as the development environment supported by java language, text editor and the testing process is carried using an android emulator. We have already designed the phases using XML text editor as explained above.

5.1 Developing User Interface

The user interface is built and developed through different layouts, views, widgets and wrapped through various containers. These design elements are developed through java programming languages in android studio frame work. Further, these design elements are implemented in genymotion emulator for validation purpose.

To develop widgets like edittext and buttons, which are designed previously in xml editior as shown in section 4.5, the widgets are defined in "EditText" and"Button" which is shown in figure 1. When we perform an action in button "change", the onclicklistener( ) is activated and performs the necessary action for displaying the page or screen in mobile native apps.

The particular code segment comprises methods such as onCreate( ), onView( ),onClick( ) and corresponding listener methods onClickListner( ) and onViewlistner( ) for developing UI and many actions.

5.2 Developing Navigation

To develop the navigation mechanism, a navigation drawer is

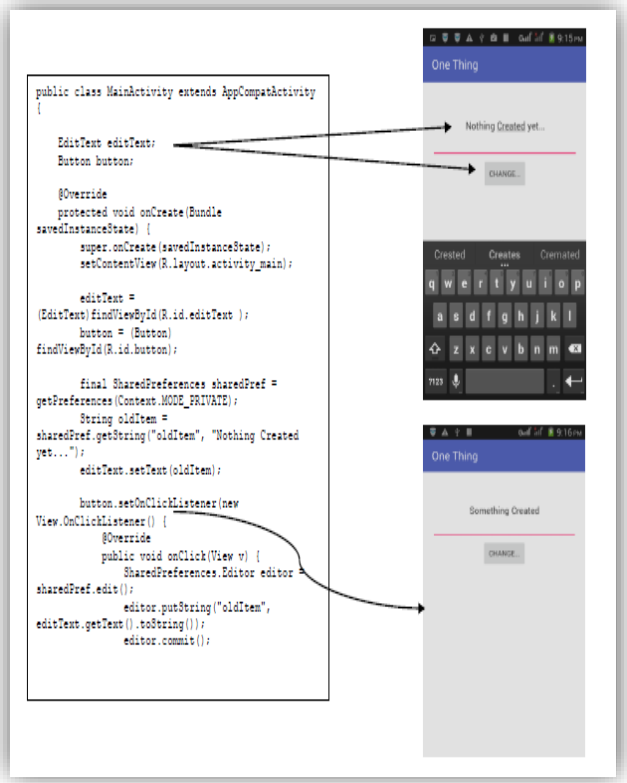


Figure 4. Developing User Interface in Mobile Native Apps

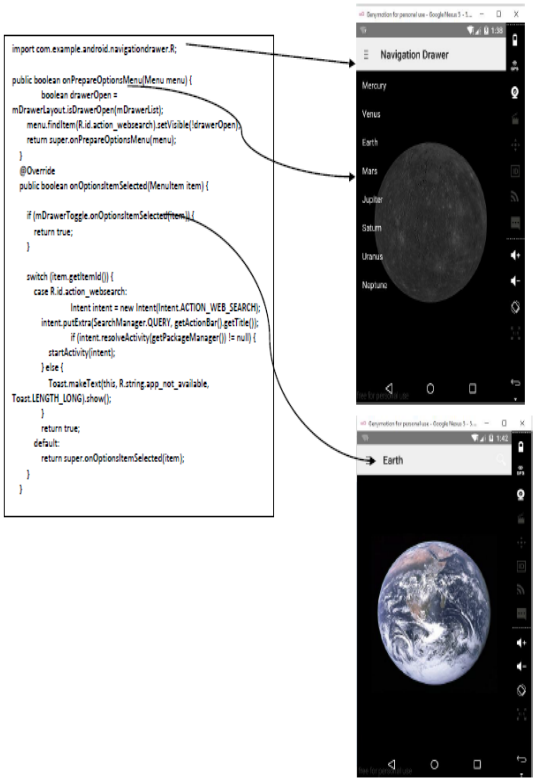


Figure 5. Developing Navigation in Mobile Native Apps

defined in the java activity file. The function of navigation drawer is to provide list of items present in the native mobile apps so that native mobile app user can select the desired option or item.

The navigation drawer holds the list of items through defining the method"onprepareoptionmenu( )" where the items present in the mobile native apps. For selection of the desired item by the native app user,"onoptionsitemselected ( )" method is used which is shown in fig.5.

5.3 Developing Presentation

For presenting a view in or a number of views in the generic native mobile apps, various transition animation operations are used. The transition animation includes zoom in, zoom out, fade in and fade out. The duration of the transition animation is defined by the mobile native app developers based on the client requirements.

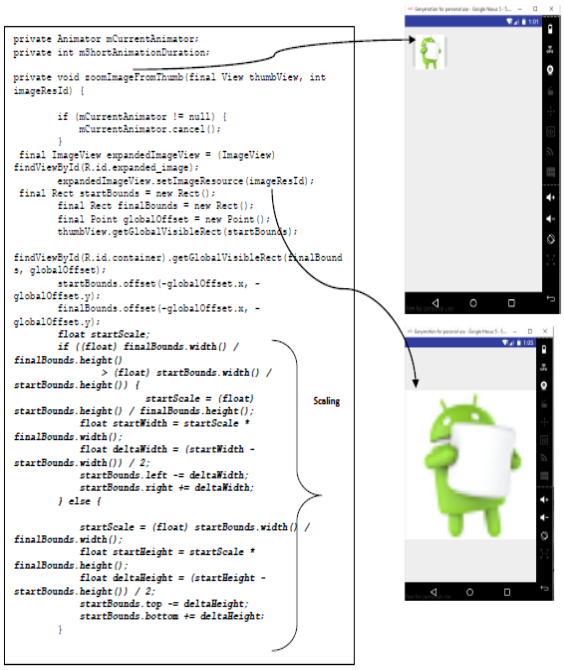


Figure 6. Developing Presentation in Mobile Native Apps

For transition operation "mCurrentAnimator" is used and its duration is set through "mShortAnimationDuration" which is shown in fig.6. For zoom-in operation the method 'zoomImageFromThumb()'is used as shown in the above code. For scaling the image into high resolution," startScale" is defined and "xy" offset are set to expand the image where the image is zoomed according to the screen size of the mobile device for building the presentation design in mobile native apps.

5.4 Developing Personification

For achieving customization and personification in generic native mobile apps, the mobile native app user should be connected with other users through the social media tools so that mobile native app user can share and find its desired information about any domain. For sharing and finding information about any domain, the native app user must be connected with the social apps. It can be by clicking on the "share" button as shown in fig.7, which prompts the user to the log in page of the mobile native apps. From the login page, which is shown in fig.8 the native app user logged into the desired social media tools and can find and share information about a generic domain.

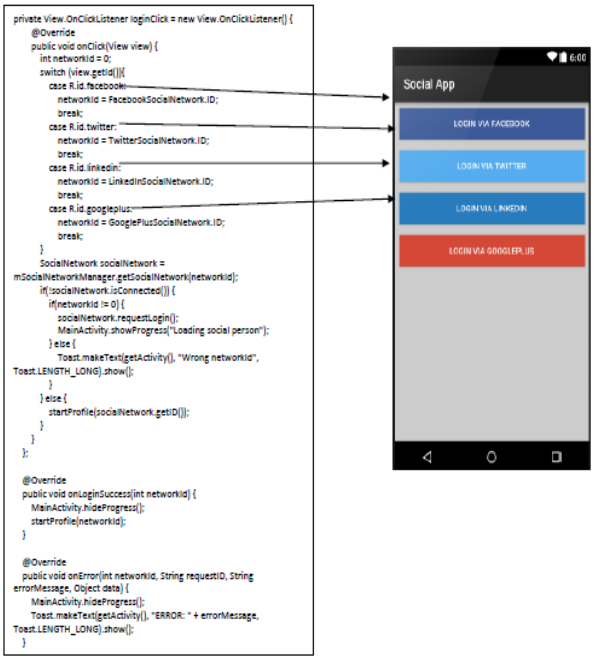


Figure 7. Developing Personification in Mobile Native Apps

6. Discussion

Here we have sent questions through Google forms to mobile native app developers of different software industries. After gating their responses, we have identified and proposed seven design phases for developing mobile native apps. The design model will work as a base for developing any mobile native app irrespective of any domain. At first, the user will analyze the content and generate the content for developing mobile native apps. For retrieving meaningful information, the users have to build ontology for a mobile native app. For building ontology, the user has to establish class axioms and property axioms among the classes for developing mobile native apps irrespective of any domain. The next phase is the user interface design, here the user will develop interfaces through using any platform for mobile native app customers. Here, we have done the user interface in android platform.

The next phase is the navigation design, here the user will build the navigation mechanism among the various activities through using any platform. Similarly the user will develop personification design by connecting social media tools through any platform. The architecture design provides the hardware components needed for building the mobile device and achieving mobility in it.

7. Conclusion and Future Work

In this paper we have proposed a design model for mobile native apps. This proposed model consists of seven phases as discussed above. We have tried to establish semantic relationship

between the attributes of mobile native apps in structural design model. In this conceptual design model, the attributes of mobile apps are designed semantically and structurally. We have find mobility mechanisms, input mechanisms and various sensors will be considered in the architecture design.

We have also identified the user interface design elements navigation design elements and presentation design element and social apps. These design elements can be designed in any framework like ios, windows and android etc and here we have taken android framework for design of these elements. In the personification design, any user can find and share its information about generic domain in mobile native generic apps. This proposed model will work as a generic standard for design and development of mobile native generic apps irrespective of any domain. Any user can design and develop the mobile native app in any domain following these phases of the proposed design model.

This model will solve various requirements of the users. If further the requirements of user changes, then changes will occur in all the design phases of mobile native apps and accordingly better design model will be built and developed. In the personification design, future work can be done on establishing chatting capability in the social media tools.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] A.Ginigi, S. Murugesan, "Web engineering: an introduction", MultiMedia, IEEE, **8**(1), 14-18, 2001.

[2] D. Schwabe,G.Rossi," Developing hypermedia applications using OOHDM, In Workshop on Hypermedia Development Process, Methods and Models, Hypertext, **98**, 1998.

[3] S.Ceri,P. Fraternali,A. Bongio," Web Modeling Language (WebML): a modeling language for designing Web sites", Computer Networks,**33**(1), 137-157,2000.

[4] N.Koch,A.Knapp,G.Zhang," UML-based web engineering", In Web Engineering: Modelling and Implementing Web Applications ,157-191,2008,Springer London.

[5] O.Pastor,J.Fons,V.Pelechano,"OOWS: A method to develop web applications from web-oriented conceptual models", In International Workshop on Web Oriented, Software Technology (IWWOST), 65-70, 2003.

[6] A.Mushtaha,R.Tolba,"Integrating the V-Model into the Web Development Process".

[7] http: //www.w3.org/TR/owl-guide/

[8] http: //www.google.co.in/?gws_rd=ssl#q=OWL+5.0+beta+framework+

[9] http:// developer.android.com/guide/topics/ui/declaring-layout.html

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 69-75 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

Recent Trends in ELM and MLELM: A review

R. Manju Parkavi[*,1], M. Shanthi[1], M.C. Bhuvaneshwari[2]

[1]Department of ECE, Kumaraguru College of Technology, 641035, India

[2]Department of EEE, PSG College of Technology, 641004, India

ARTICLE INFO

Article history:
Received: 08 December, 2016
Accepted: 15 January, 2017
Online: 28 January, 2017

Keywords :
Extreme learning machine (ELM)
Artificial Neural Network (ANN)
MLELM
Deep Learning

ABSTRACT

Extreme Learning Machine (ELM) is a high effective learning algorithm for the single hidden layer feed forward neural networks. Compared with the existing neural network learning algorithm it solves the slow training speed and over-fitting problems. It has been used in different fields and applications such as biomedical engineering, computer vision, remote sensing, chemical process and control and robotics. It has better generalization stability, sparsity, accuracy, robustness, optimal control and fast learning rate This paper introduces a brief review about ELM and MLELM, describing the principles and latest research progress about the algorithms, theories and applications. Next, Multilayer Extreme Learning Machine (MLELM) and other state-of-the-art classifiers are trained on this suitable training feature vector for classification of data. Deep learning has the advantage of approximating the complicated function and mitigating the optimization difficulty associated with deep models. Multilayer extreme learning machine is a learning algorithm of an Artificial Neural Network (ANN) which takes to be good for deep learning and extreme learning machine. This review presents a comprehensive view of these advances in ELM and MLELM which may be worthy of exploring in the future.

1. Introduction

As early as in the 1940s, arithmetician Pitts and psychologist McCulloch have put forward neurons mathematical model (MP model) from the arithmetical logic view (McCulloch and Pitts 1943) which opened the prelude of artificial neural network experimental work. Neural network with parallel and distributed information processing network framework has a strong nonlinear mapping ability and robust self-learning, adaptive and fault tolerance characteristics. Due to the recent popularity of deep learning, two of the most widely studied artificial neural networks these days are auto encoders and Boltzmann machines. An auto encoder with a single hidden layer as well as a structurally restricted version of the Boltzmann machine, called a restricted Boltzmann machine, have become popular due to their application in layer-wise pretraining of deep Multilayer perceptron's. Table.1 describes about Difference in Conventional learning method and Biological learning method. The progress of machine learning and artificial intelligence is based on the coexistence of three necessary conditions: superlative computing environments, rich and large data, and very effective learning techniques (algorithms). The Extreme Learning Machine as an evolving learning technique gives efficient unified solutions to generalize feed-forward networks including but not limited to (each supports single- and multi-hidden-layer) neural networks, Radial Basis Function (RBF) networks and kernel learning. Single hidden layer feed forward networks (SLFNs), is one of the most accepted feed forward neural networks, have been greatly studied from each theoretical and application aspect for their learning capabilities and fault-tolerant dependability [1–6]. However, most accepted learning algorithms for training SLFNs are still relatively not quick since all the parameters of SFLNs need to be tuned through repetitive procedures and these algorithms may also stuck in a local minimum.

Recently, a new rapid learning neural algorithm for SLFNs, named extreme learning machine (ELM) [7, 8], was developed to enhance the efficiency of SLFNs. Different from the existing

[*]R.ManjuParkavi, 4 Mahendrakumar building, Burgur road, anthiyur, erode. 9894768880, vprmanju@gmail.com

learning algorithms for neural networks (such as Back Propagation (BP) algorithms), which may face problems in manually tuning control parameters (learning rate, learning epochs, etc.) and/or local minima, ELM is automatically employed without iterative tuning, and in theory, no intervention is required from users. Furthermore, the learning speed of ELM is exceptionally fast compared to other traditional methods. Unlike current single-hidden-layer ELM methods, in this paper, we indicate that with general hidden nodes (or called sub network nodes) added to existing networks, MLELM can be used for classification, dimensions reduction, and feature learning.
In addition, Autoencoder (AE) based on deep network also can get better features and classification results. From this point of view, deep learning is suitable for feature extraction of images. However, the learning speed of deep learning is not fast. Due to this reason, similar to deep learning, MLELM stacks Extreme Learning Machine Autoencoder (ELM-AE) to create a Multilayer neural network which has much faster learning speed with better performance [14-16].

Table 1. Difference in Conventional learning method and Biological learning method

Conventional Learning Methods	Biological Learning
Very sensitive to network size	Stable in a wide range (tens to thousands of neurons in each module)
Difficult for parallel implementation	Parallel implementation
Difficult for hardware implementation	"Biological" implementation
Very sensitive to user specified parameters	Free of user specified parameters
Different network types for different type of applications	One module possibly for several types of applications
Time consuming in each learning point	Fast in micro learning point
Difficult for online sequential learning	Nature in online sequential learning
"Greedy" in best accuracy	High speed and high accuracy
"Brains (devised by conventional learning methods)" are chosen after applications are present	Brains are built before applications

2. Extreme Learning Machine

Extreme learning machine was planned for generalized single hidden layer feed forward neural network [9] where the hidden layer need not be neuron alike. Unlike other neural networks with back propagation [10], the hidden nodes in ELM are randomly generated, whereas the activation functions of the neurons are nonlinear piecewise continuous. The weights between the hidden layer and the output layer are calculated analytically. The general architecture of single layer ELM is shown in Figure 1.

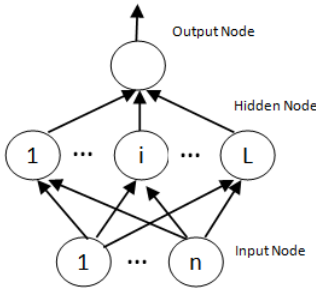


Figure 1.Structure of single layer feed-forward neural network

There are two training phases in ELM algorithm: feature mapping and output weights solving. ELM feature mapping: Given input data x, the output function of ELM for generalized SLFNs is

$$f(x) = \sum_{i=1}^{L} \beta_i h_i(x) = h(x)\beta \quad (1)$$

Where $h(x) = [h_1(x), \dots, h_L(x)]$ is the output vector of the hidden layer and $\beta = [\beta_1, \dots, \beta_L]^T$ denotes the output weights between the hidden layer of L hidden nodes and the output layer. The procedure of getting h is called ELM feature mapping which maps the input data from RD to the feature space RL. In real applications, **h** can be described as

$$h_i(x) = g(a_i, b_i, X), a_i \epsilon R^D, b_i \epsilon R \quad (2)$$

Where g(a,b,x) is an activation function satisfying ELM universal approximation capability theorems [11]. In fact, any nonlinear piecewise continuous functions (e.g. Gaussian, Sigmoid, etc.) can be used as activation function h. In ELM, the parameters of h are randomly produced based on a continuous probability distribution.

ELM output weights solving: In the second phase, given a training sample set, $(X_i, t_i)_i^n = 1$ with $t_i = [0, \dots 0, 1_i, 0, \dots, 0_m]^T$the class indicator of xi, ELM aims to minimize both the training error and the Frobenius norm of output weights. This objective function used for both binary and multi-class classification tasks, can be expressed as follows,

$$\underset{\beta,\xi}{\text{Min}} \frac{\omega}{2} \sum_{i=1}^{n} \|\xi\|_2^2 + \frac{1}{2}\|\beta\|_F^2,$$

$$s.t. \quad \beta h(X_i) = t_i - \xi_i, \ \forall \in 1,2, \dots, n \quad (3)$$

where n is the number of samples and denotes the training error of the i-th sample, ω is a regularization parameter which trades off the average of output weights and training error and denotes the Frobenius norm. The optimization problem in (3) can be efficiently solved. Specifically, according to the Wood bury identity [12] the optimal β can be analytically obtained as

$$\beta^* = \begin{cases} \left(H^T H + \frac{I_L}{\omega}\right) H^T T & if\ L \le k \\ H^T \left(HH^T + \frac{I_n}{\omega}\right)^{-1} T & otherwise \end{cases} \quad (4)$$

Where In and IL are identity matrices, and H is the hidden layer output matrix (randomized matrix) which is defined in (5)

$$H = \begin{bmatrix} h(x_1) \\ . \\ . \\ . \\ h(x_n) \end{bmatrix} = \begin{bmatrix} h_1(x_1) & \dots & h_L(x_1) \\ . & . & . \\ . & . & . \\ . & . & . \\ h_1(x_n) & \dots & h_L(x_n) \end{bmatrix} \quad (5)$$

2.1. ELM-Based Autoencoder

Apart from the ELM-based SLFNs, the ELM theory has also been applied to build an autoencoder for an MLP. Concept wise, the autoencoder functions are of feature extractor in a multilayer learning framework. It uses the encoded outputs to approximate the original inputs by minimizing the reconstruction errors.

Mathematically, the autoencoder maps the input data x to a higher level representation, and then uses latent representation y through a deterministic mapping $y = h\theta(X) = g(A * X + b)$, parameterized by $\theta = \{A, b\}$, where g( ) is the activation function, A is a d' × d weight matrix and b is a bias vector. The resulting latent representation y is then mapped back to a reconstructed vector z in the input space

$$z = h\theta\ '(y) = g(A'\ y + b) \text{ with } \theta' = \{A', b'\}.$$

Using randomly mapped outputs as inherent representation y, one can easily build the ELM-based autoencoder as shown in [10]. Then, the reconstruction of x can be regarded as an ELM learning problem, in which A is obtained by solving regularized least mean square optimization. However, due to the use of penalty in the original ELM, the extracted features by the ELM autoencoder in [10] tend to be dense and may have redundancy. In this case, a more sparse solution is preferred.

In classifier design, used for classification of heart stroke level with the help of 13 inputs, is derived from the standard database and 5 target values of kind of disease, each of which represents a class Algorithm, describes the steps of the proposed solving system [17]. As we can see from the algorithm steps, the output binary values are changed to [-0.9 +0.9] make the output more observable, the two selected parameters that effect the performance of the network are the constant value C, which is selected here to be 10000, and number of hidden neurons L, which is set to 100 for most of the works. The biggest computation effort will be done in the inversion of the matrix G. Whenever the dimensions are big, the system needs more time and even specific inverse computation algorithms, like the Moore-Penrose generalized inverse which has been worked by Samet and Miri to increase the speed of computation.

2.2. ELM Variants

Though achieving great victory on both theoretical and practical aspects, basic ELM cannot efficiently handle large-scale learning tasks because of the limitation of memory and the intensive computational cost of the inverse of larger matrices. To decrease the runtime memory, many variants including Online Sequential ELM (OS-ELM) [20] and Incremental ELM (I-ELM) [21] have been proposed. OS-ELM can decrease the requirement of runtime memory because the model is trained based on each chunk. I-ELM can diagnose memory problem by training a basic network with some hidden nodes and then adding hidden node to the conventional network one by one. To rest the computational cost incurred by these operations, many variants containing partitioned ELM [22] and parallel ELM [23], have been recommended. Recently, a variant of OS-ELM also called parallel OS-ELM (POS-ELM) [24] and parallel regularized ELM (PR-ELM) [25] are recommended to reduce training time and memory requirement instantaneously.

3. Multilayer Extreme Learning Machine

Multilayer ELM is an advanced machine learning approach based on the architecture of artificial neural network and is inspirited by deep learning and extreme learning machine. Deep learning was first suggested by Hinton and Salakhutdinov (2006) who in their process used deep structure of multilayer autoencoder and recognized a multilayer neural network on the unsupervised data. In their method, first they used an unsupervised training to obtain the parameters in each layer. Then, the network is fine-tuned by supervised learning. Hinton et al. (2006), who suggested the deep belief network, outperforms the conventional multilayer neural network, SLFNs, Support Vector Machines (SVMs), but it has slow learning speed. Working in this way, recently Kasun et al. (2013) proposed multilayer ELM which performs unsupervised learning from layer to layer, and it is not necessary to iterate during the training process and hence, it does not spend a long time in the training phase. Compared to other existing deep networks, it has a better or comparable performance.

Figure 2 shows the structure of MLELM with ELM-AE, output weights β^1 of (a) ELM-autoencoder (AE) is denotedwith respect to input data x are the layer 1 weights of MLELM . (b) Output weight matrix is denoted as β^{i+1} in ELM-AE, with respect to i-th hidden layer and output hi of MLELM are the i + 1th layer weights of MLELM; (c) Regularized least squares are used for output layer weight calculation of MLELM.

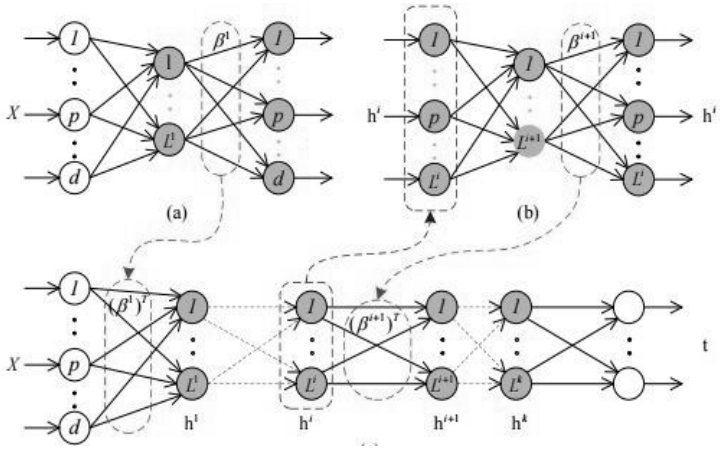


Figure 2. ELM-AE and Multilayer ELM

3.1. Extreme Learning Machine- Autoencoder

Autoencoder is an unsupervised neural network. The outputs and inputs of the autoencoder are similar. Like ELM, ELM-AE has n input layer nodes, single hidden layer of L nodes and n output layer nodes. In spite of many resemblances between these two, there are two major differences that exist between them which are as regards,

i. . ELM is a supervised neural network while ELM-AE is an unsupervised one and its output is similar as the input.

ii. Input weights and biases of the hidden layer are random in case of ELM, but they are orthogonal in ELM-AE.

Depends upon the number of hidden layer nodes, the ELM-AE [18] can be classified and shown as the following three categories,

(i) Compressed representation (n > L): In compressed representation, features of training dataset need to be presented from a higher-dimensional input signal space to a lower-dimensional feature space.

(ii) Equal dimension representation (n = L): In this module representation of features, the dimension of input signal space and feature space needs to be equal.

(iii) Sparse representation (n < L): It is just the reverse of compressed representation where features of training dataset need to be applied from a lower-dimensional input signal space to a higher dimensional feature space.

The multilayer ELM is extensively faster than deep networks because iterative tuning mechanism is not require in case of MLELM and got better or similar performance compared to deep networks. It is also known that in ELM, for L hidden nodes and N training examples (x j,y j), the following (6) holds

$$g_L = (x_j) = \sum_{i=1}^{L} \beta_i g_i (x_j, w_i, b_i) = y_j, j = 1, \dots, N \quad (6)$$

In order to perform unsupervised learning, few alterations have been done in ELM-AE whose working principle is similar to regular ELM, which are represented as follows,

1) The output data and the input data remain same for every hidden layer. Hence, for every input data X: Y = X

(2) To improve the performance of ELMs, we need to consider the weights and the biases of the random hidden nodes to be present orthogonal and can be represented as follows

$$H = g(w \cdot x + b), wT \cdot w = I \text{ and } bT \cdot b = 1 \quad (7)$$

(3) The output weight $\boldsymbol{\beta}$ is decided based on the following conditions,

i. if $\boldsymbol{n > L}$ then,

$$\beta = \left(\frac{I}{C} + H^T H\right)^{-1} H^T X \quad (8)$$

ii. if n = L then,

$$\beta = H^{-1} X \quad (9)$$

iii. if $\boldsymbol{n < L}$ then,

$$\beta = H^T \left(\frac{I}{C} + H^T H\right)^{-1} X \quad (10)$$

Where C is a scale parameter which adjusts analytical and experiential risk. ELM-AE is used for training the parameters in each layer of MLELM. The general formula representing MLELM is described as follows,

$$H^n = g((\beta^n)^T H^{n-}) \quad (11)$$

For n = 0, the 0th hidden layer or the first layer is considered to be the input layer X. The final output matrix Y can be acquired by computing the results between the last hidden layer and the output layer using the regularized least squares model work.

3.2. Hierarchical Extreme Learning Machine Variants

Currently, ELM is also extended to Multilayer structures, i.e. Multilayer ELM (MLELM) [19], hierarchical ELM [26] and another reference work hierarchical local receptive fields ELM (H-LRF-ELM) [27]. Similar to other deep networks, MLELM [19] performs unsupervised learning in layer-by-layer. In different to deep networks, MLELM does not require fine-tuning using back propagation which will decrease the computational cost in learning process. H-ELM [26] is a new ELM-based hierarchical learning frame work for Multilayer perception. For feature extraction, H-ELM work as unsupervised Multilayer encoding. Unlike the demanding layer-wise training of deep learning, the layers of H-ELM are learned in a forward manner. Therefore, it has better learning efficiency than the deep learning methods. LRF-ELM was first suggested by Huang et al. [27].

In this model, the connections between the input and hidden nodes are sparse and bounded by local receptive fields (LRF). LRF-ELM is learning the local structures and generates more meaningful representations at the hidden layer when dealing with image processing and like tasks. LRF-ELM can be enhanced to Multilayer architecture called hierarchical LRF-ELM (H-LRF-ELM). The layers of H-LRF-ELM can be classified into two parts, namely, the feature extractor and the ELM classifier. After all, ELM-based Multilayer networks seem to provide better accuracy, performance and efficiency than other deep networks.

3.3. Hierarchical Extreme Learning Machine

The proposed H-ELM is built in a multilayer manner, H-ELM training architecture is structurally divided into two separate phases: 1) unsupervised hierarchical feature representation and 2) supervised feature classification. H-ELM, as well as its advantages against the existing DL and multilayer ELM (MLELM) algorithms. Before unsupervised feature learning, the input raw data should be transformed into an ELM random feature space, which can help to exploit hidden information among training samples. The universal approximation capability of the ELM is exploited for the design of autoencoder and moreover, sparse constraint is added upon the autoencoder optimization and therefore, we term it as ELM sparse auto encoder.

4. Discussion

The field of deep neural networks, or deep learning, is expanding rapidly, and it is impossible to discuss everything in this thesis. Single layer ELM, Multilayer ELM, autoencoder and Boltzmann machines, which are main topics of this survey, are certainly not the only neural networks in the field of deep neural networks. However, as the aim of this survey is to provide a brief overview and introduction to deep neural networks.

Although deep neural networks have shown extremely competitive performance in various machine learning tasks, the theoretical motivation for using them is still debated. Compared to the feed forward neural networks such as autoencoders and Multilayer ELM, it is difficult to evaluate Boltzmann machines.

Even when the structure of the network is highly restricted, the existence of the intractable normalization constant requires using a sophisticated sampling-based estimation method to evaluate Boltzmann machines.Table.2 presented the performance comparison of MLELM method with other deep learning algorithm. MLELM achieves best training accuracy and Training Time.

Table 2. Performance comparison of MLELM and other method

Learning method	Testing Accuracy	Training Time
Multilayer ELM [Huang, et al 2013]	99.03±0.04	444.7sec
Deep Belief Networks (DBN)	98.87	20580sec
Deep Boltzmann Machines	99.05	68246sec
Stacked Auto Encoders (SAE)	98.6	>17 hours
Stacked Denoising Auto Encoders (SDAE)	98.72	>17 hours

5. Applications

Neural network is extensively used in artificial intelligence, data mining, image classification and other applications. Single hidden layer feed forward neural network has a strong training ability. Extreme learning methods is proposed to overcome the disadvantage of a single hidden layer feed forward neural network, improve the learning ability and generalization efficiency.

These recent success of deep neural networks in both academic research and commercial applications may be attributed to several recent breakthroughs Layerwise pretraining of a Multilayer [13] showed that a clever way of initializing parameters can easily overcome the difficulties of optimizing a large, deep neural network. As ELM algorithm does not require large-scale operations, there is no need to adjust the hidden layer of the network and compared with the conventional computational intelligence techniques, such as BP algorithm the whole learning process is necessary only one iteration and the training speed improved more than 10 times.

5.1. ELM for fault detection in Analog Integrated Circuits

Test generation algorithm based on ELM, similar algorithm is cost effective and low-risk for analog device under test (DUT). This method uses test patterns from the test generation algorithm to support DUT and then samples output responses of the DUT for fault classification and detection. Active test generation algorithms cannot save time effectively and it lags in accuracy when the number of impulse-response samples decreases. Due to this computational complexity and classification theory of methods, ELM-based algorithm is used [28].

5.2. Image Processing

In H-ELM framework, several H-ELM-based feature extraction and classification algorithms are developed for practical computer vision applications, such as object(image) detection,object recognition, and object tracking. The obtained results are quite promising, and further confirm the generality and capability of the H-ELM.

Applications of MLELM have recently been presented in sparse learning and classification, computer vision, clustering learning, image processing, high-dimensional and large-data applications, etc.

5.3. Car Detection

Car detection is a classical application in computer vision. The proposed MLELM-based car detection algorithm and one can see that a sliding window is used to extract a fixed-size image patch, which are grouped and input into the H-ELM for training and testing.

5.4. Fault Diagnosis of Power Transformer

The power transformer, one of the key power equipment, is mainly composed of hydrocarbon oil and paper insulation. Various methods are applied to interpret dissolved gas-in-oil analysis (DGA) results, they may sometimes fail to diagnose exactly. The incipient fault identification accuracy of artificial intelligence (AI) is varied with input variable variation. Therefore, selection of input variable is prime research area. Principle component analysis using RapidMiner software is adapted to IEC TC10 and related datasets is used to identify most relevant input variables for incipient fault classification. Extreme learning machine (ELM) is implemented to classify the incipient faults of power transformer.New application of ELM to the real-time fault solving system for rotating machinery. The structure is successfully applied on recognizing fault patterns coming from the Windmill [29].

5.5. Gesture Recognition

Gesture recognition is also an important topic in computer vision due to its wide ranges of applications, such as human–computer interfaces, sign language interpretation, and visual surveillance. The different images are directly fed into the AIOS-ELM, H-ELM-based classifier to recognize each of the gesture [30].

5.6. Online Incremental Tracking

Online incremental tracking refers to the tracking method that aims to learn and adapt a representation to reflect the appearance changes of the target. The appearance changes of the specified target include pose variation and shape deformation. Apart from these, extrinsic illumination change, camera motion, viewpoint, and occlusions inevitably cause large appearance variation. Modelling such appearance variation is always one of the nature problems of tracking. For this purpose H-ELM is used for incremental tracking method.

5.7. Multiclass Applications

SVM is used only for binary classification. In order to solve problem in multiclass classification, ELM is preferred. ELM,

which is of higher scalability and low computational complexity, not only unifies other popular learning algorithms but also provides a unified solution to different practical applications (e.g., regression, binary, and multiclass classifications).

5.8. Medical Applications

Extreme Learning Machine (ELM) algorithm is used to model these factors such as age, sex, serum cholesterol, blood sugar, etc. The proposed system can replace costly medical checkups with a warning system for patients of the probable presence of heart based disease. Brain-computer interface (BCI) is a technology that is used for communication with a computer or to control devices with EEG signals .The main technologies of BCI areto extract the preprocessed EEG features and to classify ready processed EEG. BCI is used for many fields, such as medicine and military.

Currently, many methods have been proposed for EEG classification, including decision trees, local Backpropagation algorithm, Bayes classifier, *K*-nearest neighbors (KNN), support vector machine (SVM, and ELM . Most of them are shallow neural network algorithms where the capabilities achieve approximating the complex functions have certain restrictions, and there is no such restriction in deep learning. Deep learning [31] is a multilayer perceptron artificial neural network algorithm. DELM in EEG classification and makes use of two BCI competition datasets to test the performances of DELM.

5.9. 3D Applications

3D shape features play a critical role in graphics applications like 3D shape matching, recognition, and retrieval. Various 3D shape descriptors have been developed over the last two decades.

The Convolutional Autoencoder Extreme Learning Machine (CAE-ELM) can be used in practical graphics applications, such as in 3D shape completion. Optical acquisition devices most often generate incomplete 3D shape data because of occlusion and unfavorable surface reflectance properties.

5.10. Security

Privacy protection is a significant issue such that confidential images must be encrypted in corporations. Nevertheless, decryption and then classifying millions of encrypted images becomes a heavy burden to computation. Multilayer Extreme Learning Machine is proposed for encrypted image classification framework that is able to precisely classify encrypted images without decryption.

5.11. Fault Diagnosis of Wind Turbine Equipment

Fault diagnosis is critical in the wind turbine generator system to reduce the maintenance cost and to avoid unplanned interruption. The data generated from wind turbine generator system is always of high-dimensional and dynamical. To avoid these problems, a new fault diagnosis method of multiple extreme learning machines is proposed. It enables both representational feature learning and fault classification. Feature learning includes data preprocessing, feature extraction and dimensional reduction. Unlike the other greedy layer wise learning scheme adopted in back propagation; it does not need iterative fine-tuning of parameters. The results show that the evolved diagnostic framework accomplishes the best performance among the compared approaches in terms of accuracy and effectiveness in several faults detection in wind turbines [32].

6. Conclusion

This paper demonstrates the overall review of ELM and MLELM algorithm, especially emphasizing on its variants and applications. Our aim is to introduce a valuable tool for applications to the researches, which provides more accurate results and spend less calculation time in the classification or regression problems than the existing methods, such as BP neural networks and LS-LVM. MLELM training scheme, which is based on the universal approximation capability of the original ELM. The proposed H-ELM achieves high level representation with layerwise encoding, and outperforms the original ELM in various simulations. Moreover, compared with other ELM training methods, the training of MLELM is much faster and achieves higher learning accuracy.There are also some open problems of the ELM algorithm to be diagnosed. The following problems remain open and may be worth absorbing the attentions of researchers in the future,

1. ELM with sub network nodes architecture [33] has not attracted much research attentions. Recently, many works have been proposed for supervised/unsupervised dimension reduction or representation learning, one type of problem.

2. More applications may be needed to check the generalization ability of MLELM, especially in some areas with mass data.

References

[1] Adamos, D. A., Laskaris, N. A., Kosmidis, E. K., &Theophilidis, G. (2010). NASS: an empirical approach to spike sorting with overlap resolution based on a hybrid noise-assisted methodology. *Journal of neuroscience methods*, *190*(1), 129-142.

[2] An, L., &Bhanu, B. (2012, September). Image super-resolution by extreme learning machine. In *2012 19th IEEE International Conference on Image Processing* (pp. 2209-2212). IEEE.

[3] Avci, E. (2013). A new method for expert target recognition system: Genetic wavelet extreme learning machine (GAWELM). *Expert Systems with Applications*, *40*(10), 3984-3993.

[4] Avci, E., & Coteli, R. (2012). A new automatic target recognition system based on wavelet extreme learning machine. *Expert Systems with Applications*, *39*(16), 12340-12348.

[5] Bai, Z., Huang, G. B., Wang, D., Wang, H., & Westover, M. B. (2014). Sparse extreme learning machine for classification. *IEEE transactions on cybernetics*, *44*(10), 1858-1870.

[6] Balbay, A., Avci, E., Şahin, Ö.,&Coteli, R. (2012). Modeling of drying process of bittim nuts (Pistaciaterebinthus) in a fixed bed dryer system by using extreme learning machine. *International Journal of Food Engineering*, *8*(4).

[7] Balbay, A., Kaya, Y., & Sahin, O. (2012). Drying of black cumin (Nigella sativa) in a microwave assisted drying system and modeling using extreme learning machine. *Energy*, *44*(1), 352-357.

[8] Baradarani, A., Wu, Q. J., &Ahmadi, M. (2013). An efficient illumination invariant face recognition framework via illumination enhancement and DD-DTCWT filt.

[9] Barea, R., Boquete, L., Ortega, S., López, E., & Rodríguez-Ascariz, J. M. (2012). EOG-based eye movements codification for human computer interaction. *Expert Systems with Applications*, *39*(3), 2677-2683.

[10] Huang, G. B., Zhou, H., Ding, X., & Zhang, R. (2012). Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *42*(2), 513-529.

[11] Basu, A., Shuo, S., Zhou, H., Lim, M. H., & Huang, G. B. (2013). Silicon spiking neurons for hardware implementation of extreme learning machines. *Neurocomputing*, *102*, 125-134.

[12] Bazi, Y., Alajlan, N., Melgani, F., AlHichri, H., Malek, S., &Yager, R. R. (2014). Differential evolution extreme learning machine for the classification of hyperspectral images. *IEEE Geoscience and Remote Sensing Letters*, *11*(6), 1066-1070.

[13] Belkin, M., &Niyogi, P. (2003). Laplacianeigenmaps for dimensionality reduction and data representation. *Neural computation*, *15*(6), 1373-1396.

[14] Wang, Y., Xie, Z., Xu, K., Dou, Y., & Lei, Y. (2016). An efficient and effective convolutional autoencoder extreme learning machine network for 3d feature learning. *Neurocomputing*, *174*, 988-998.

[15] Roul, R. K., Asthana, S. R., & Kumar, G. (2016). Study on suitability and importance of multilayer extreme learning machine for classification of text data. *Soft Computing*, 1-18.

[16] Wang, W., Vong, C. M., Yang, Y., & Wong, P. K. (2016). Encrypted image classification based on multilayer extreme learning machine. *Multidimensional Systems and Signal Processing*, 1-15.

[17] Ismaeel, S., Miri, A., &Chourishi, D. (2015, May). Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis. In *Humanitarian Technology Conference (IHTC2015), 2015 IEEE Canada International* (pp. 1-3). IEEE.

[18] Decherchi, S., Gastaldo, P., Leoncini, A., & Zunino, R. (2012). Efficient digital implementation of extreme learning machines for classification. *IEEE Transactions on Circuits and Systems II: Express Briefs*, *59*(8), 496-500.

[19] Decherchi, S., Gastaldo, P., Leoncini, A., & Zunino, R. (2012). Efficient digital implementation of extreme learning machines for classification. *IEEE Transactions on Circuits and Systems II: Express Briefs*, *59*(8), 496-500.

[20] Lemme, A., Freire, A., Barreto, G., &Steil, J. (2013). Kinesthetic teaching of visuomotor coordination for pointing by the humanoid robot icub. *Neurocomputing*, *112*, 179-188.

[21] Kavukcuoglu, K., Ranzato, M. A., &LeCun, Y. (2010). Fast inference in sparse coding algorithms with applications to object recognition. *arXiv preprint arXiv:1010.3467*.

[22] Tang, J., Deng, C., & Huang, G. B. (2016). Extreme learning machine for multilayer perceptron. *IEEE transactions on neural networks and learning systems*, *27*(4), 809-821.

[23] Le, Q., Sarlós, T., & Smola, A. (2013, June). Fastfood-approximating kernel expansions in loglinear time. In *Proceedings of the international conference on machine learning*.

[24] Liang, N. Y., Huang, G. B., Saratchandran, P., &Sundararajan, N. (2006). A fast and accurate online sequential learning algorithm for feedforward networks. *IEEE Transactions on Neural networks*, *17*(6), 1411-1423.

[25] Li, B., Li, Y., &Rong, X. (2013). The extreme learning machine learning algorithm with tunable activation function. *Neural Computing and Applications*, *22*(3-4), 531-539.

[26] Li, L., Liu, D., & Ouyang, J. (2012). A new regularization classification method based on extreme learning machine in network data. *J. Inf. Comput. Sci*, *9*(12), 3351-3363.

[27] Li, Y., Li, Y., Zhai, J., & Shiu, S. (2012). RTS game strategy evaluation using extreme learning machine. *Soft Computing*, *16*(9), 1627-1637.

[28] Zhou, J., Tian, S., Yang, C., & Ren, X. (2014). Test generation algorithm for fault detection of analog circuits based on extreme learning machine. *Computational intelligence and neuroscience*, *2014*, 55.

[29] Malik, H., & Mishra, S. (2015, December). Extreme learning machine based fault diagnosis of power transformer using IEC TC10 and its related data. In *2015 Annual IEEE India Conference (INDICON)* (pp. 1-5). IEEE.

[30] Cambria, E., Huang, G. B., Kasun, L. L. C., Zhou, H., Vong, C. M., Lin, J., ... & Leung, V. C. (2013). Extreme learning machines [trends & controversies]. *IEEE Intelligent Systems*, *28*(6), 30-59

[31] Ding, S., Zhang, N., Xu, X., Guo, L., & Zhang, J. (2015). Deep extreme learning machine and its application in EEG classification. *Mathematical Problems in Engineering*, *2015*

[32] Yang, Z. X., Wang, X. B., & Zhong, J. H. (2016). Representational Learning for Fault Diagnosis of Wind Turbine Equipment: A Multi-Layered Extreme Learning Machines Approach. *Energies*, *9*(6), 379.

[33] Yang, Y., & Wu, Q. J. (2016). Multilayer extreme learning machine with subnetwork nodes for representation learning.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 76-83 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

Frame Filtering and Skipping for Point Cloud Data Video Transmission

Carlos Moreno, Ming Li

Department of Computer Science, California State University, Fresno, 93740, USA

ARTICLE INFO

Article history:
Received: 18 December, 2016
Accepted: 19 January, 2017
Online: 28 January, 2017

Keywords:
Filtering
Frame Skipping
Point Clouds

ABSTRACT

Sensors for collecting 3D spatial data from the real world are becoming more important. They are a prime research area topic and have applications in consumer markets, such as medical, entertainment, and robotics. However, a primary concern with collecting this data is the vast amount of information being generated, and thus, needing to be processed before being transmitted. To address the issue, we propose the use of filtering methods and frame skipping. To collect the 3D spatial data, called point clouds, we used the Microsoft Kinect sensor. In addition, we utilized the Point Cloud Library to process and filter the data being generated by the Kinect. Two different computers were used: a client which collects, filters, and transmits the point clouds; and a server that receives and visualizes the point clouds. The client is also checking for similarity in consecutive frames, skipping those that reach a similarity threshold. In order to compare the filtering methods and test the effectiveness of the frame skipping technique, quality of service (QoS) metrics such as frame rate and percentage of filter were introduced. These metrics indicate how well a certain combination of filtering method and frame skipping accomplishes the goal of transmitting point clouds from one location to another. We found that the pass through filter in conjunction with frame skipping provides the best relative QoS. However, results also show that there is still too much data for a satisfactory QoS. For a real-time system to provide reasonable end-to-end quality, dynamic compression and progressive transmission need to be utilized.

1. Introduction

This paper is an extension of work originally presented in the International Conference on Internet and Multimedia Technologies 2016 as part of the World Congress on Engineering & Computer Science 2016 [1]. We extend our work by implementing an additional technique to achieve a better quality of service (QoS). Originally, we relied on filtering methods to lower the cost of network transmission for point cloud data (PCD). In this work, we explore the use of a frame skipping technique in addition to filtering to improve the frame rate and other QoS metrics.

The collection of 3D spatial data from the real world for video streaming is a method of communication that has experienced growth in recent years. This area provides research challenges and contains applications in consumer markets, such as medical, entertainment, and robotics [2]. Providing good efficiency and high quality for video streaming requires:

(i) a hardware sensor that is able to capture spatial information in all three dimensions;

(ii) encoding and decoding algorithms; and

(iii) data quality assurance mechanisms that meet application needs.

Hardware sensors have been in development for the past decade to allow for 3D image acquisition from a real world environment. A variety of techniques can be used to develop these sensors, such as time-of-light (TOF), stereo, lasers, infrared light, and structured light [3]. Due to their different approaches in acquiring depth data, there is a varying cost and size for these sensors. For example, the Velodyne spinning LiDAR system is expensive, costing thousands of dollars, making it, and sensors similar to it, impractical for many projects [4]. In contrast, there has been a rise in low cost solutions to collect RGB-D data, such as Microsoft's Kinect sensor [5]. For this study, we used the Kinect sensor to collect the 3D spatial data and the Point Cloud Library (PCL) to process them.

*Corresponding Author: Carlos Moreno, Department of Computer Science, California State University, Fresno, 93740, USA
Email: mmxzbnl@mail.fresnostate.edu

The primary concern with these hardware sensors is the rate and volume of information that is being generated. Depending on the quality of the sensor and the physical environment being sensed, a single capture can range from thousands to millions of points. The Kinect sensor, specifically, generates 307,200 points per frame (ppf), which leads to a data rate of about 45 megabytes per second (MBps) [6]. Such a data rate is impractical to achieve over a network for real-time video streaming, especially when juxtaposed with the average American household bandwidth of 12.6 megabits per second (Mbps) [7].

Therefore, in order to overcome the network bandwidth bottleneck, the use of filtering methods is considered a viable solution. Filtering allows us to modify or remove data from a dataset according to certain criteria. Numerous methods for filtering data have been developed over the past years, originating from different services and application needs. Given the variation in filtering algorithms, a comparative study is necessary that is able to compare and contrast different filters for use in PCD video streaming.

While filtering the PCD does significantly reduce the data rate for video transmission, it is still impractical for any real world application. Therefore, in addition to studying the effects of different filtering methods, we require the use of a frame skipping technique. While traditional frame skipping is used when buffer overflow occurs, we propose a different use of the scheme [8]. Rather than transmitting every frame that the hardware sensor generates, we check for how similar the frame is to the previously transmitted one. Frames that are deemed too similar are skipped. Using such a technique allows the application to send only the PCD that is significantly different visually for the user, thereby saving transmission costs. Furthermore, as explored in [9], the use of frame skipping techniques is beneficial from the end-user's perspective.

In this paper, we have implemented and conducted experiments to evaluate four filters: pass through filter, voxel grid filter, approximate voxel grid filter, and bilateral filter. In addition, we tested the frame skipping technique both in isolation and with the use of filtering. Results show that, overall, the pass through filter with frame skipping provides the best QoS relative to the other options. Yet, despite the reduction in the number of points being transmitted, the data rate is still too high. We conclude that, in conjunction with filtering methods, a PCD video streaming service will require further techniques such as dynamic compression and progressive transmission.

The rest of this paper is organized as follows. Section 2 discusses the works related to this paper, including a discussion on other uses of the Microsoft Kinect, PCL, and frame skipping. Section 3 reviews the architecture behind the Kinect and PCL, as well as introducing the octree data structure. Section 4 explores the different filtering methods used in the paper. Section 5 discusses the frame skipping technique in the context of our work. Section 6 explains the experiment setup while section 7 details the QoS metrics developed for evaluation purposes. Section 8 provides an analysis of the experiment results. Lastly, sections 9 and 10 summarize the work in this paper and offer a conclusion.

2. Related Works

Microsoft's Kinect was originally meant for entertainment purposes, but since its inception, it has been integrated in many other fields. For example, [10] has integrated the Kinect with Simulink to allow for real-time object tracking. Reference [11] shows that the hardware sensor is applicable to the medical field through its use in a virtual rehabilitation system to help stroke victims regain balance. The Kinect's inherit issues, as described in [12], can be transformed into useful information for use in creating automatic foreground segmentation. Lastly, [13] shows a unique example of utilizing the Kinect in the music field.

PCL has also experienced multiple uses. Reference [14] uses it to develop efficient facial registration processes, encompassed into the Digital Face-Inspection (DFI) system, to help in the dental field. Reference [15] explains how indoor robots also benefit from PCL by allowing for real-time, general object recognition. An important data structure used in PCL is the octree, and it too has implications in other fields, as shown in [16] and [17].

Frame skipping as a technique has seen considerable use in video encoders. For example, [8] proposes a dynamic frame skipping scheme coupled with an adaptive sliding window that is able to reduce the frame rate to better match the available bandwidth in live video streaming. The work in [18] takes this idea further by calculating the optimum frame allocation for the sliding window. A different approach is proposed in [19] by utilizing motion information, buffer state information and changes in the video scene. Lastly, [20] tackles the problem of real-time mobile application in a stereo video setting, and proposes the use of frame skipping as a hybrid approach.

3. System Architecture

Microsoft's Kinect is a peripheral hardware sensor originally developed for the interactive use of Microsoft's Xbox 360 video game console. The RGB-D sensor was developed with three main functions in mind: 3D image detection, human skeleton tracing, and audio processing. This functionality, in addition to its relatively inexpensive price, has drawn researchers and developers to utilize the Kinect for other purposes, such as robot vision and healthcare systems. The sensor itself consists of an RGB camera, an infrared (IR) emitter, an IR camera, and an array of microphones. The Kinect can achieve a maximum frame rate of 30 frames per second (fps) with a resolution of 640 x 480. Each color channel uses 8 bits while the depth data is represented in 16 bits [21]. Therefore, for a single frame consisting of 307,200 points, this Kinect raw data is represented in about 1.46 MB.

The Kinect raw data can be converted to point clouds. To process these point clouds, we require the use of an external

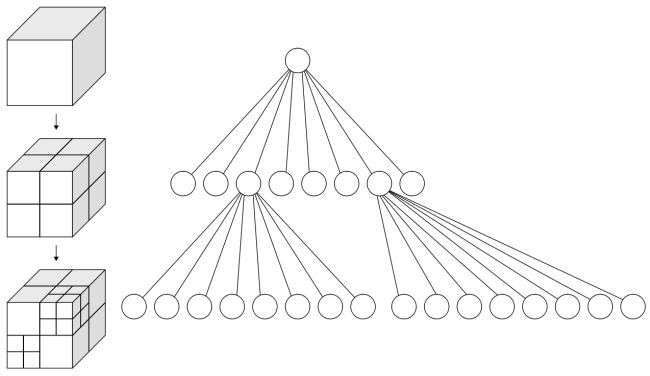


Fig. 1. A visual representation of a voxel (left) and its corresponding octree (right).

library. One such library is PCL, an open-source, fully templated C++ library. It incorporates many algorithms for point clouds and 3D geometry, such as filtering, feature estimation, visualization, segmentation, and more [22]. The PCD for the Kinect is about 10 MB in size; much larger than the raw RGB-D data. If we aimed for a target frame rate of 30 fps, this would result in a data rate of 300 MBps for PCL, compared to the 45 MBps of raw Kinect data. Processing and transmitting at either data rate is impractical.

In PCL, one of the fundamental data structures used to represent a point cloud is the octree. An octree recursively divides a point cloud into eight smaller sections, called voxels (see figure 1). These voxels are 3D cubes that encapsulate a subset of points from the point cloud. The resolution of the frame is determined by the depth level of the octree, i.e. a higher resolution translates to more voxels and a deeper octree.

Hardware and software aside, what makes reliable video streaming particularly challenging is finding the balance between the data rate and the network bandwidth. Having a high-quality hardware sensor and efficient software for processing is wasted if the network bandwidth is unable to deliver the data in real-time. Therefore, for this paper, we assume that the network is in such a state: the bandwidth is much less than the data rate. Moreover, we are using the following configurations for two computers: (i) a wired desktop PC with an Intel i7-6700 processor, GTX 745 GPU, and 16GB DDR3 RAM; and (ii) a wireless laptop with an Intel i7-2630QM processor and 8GB DDR3 RAM.

4. Filtering Methods

The filters used for this comparative study come from PCL. There is an extensive list of available filters; however, not all of these are practical or suited for data transmission in real-time. Narrowing down this list, we determined that the following four filters seem most applicable to real-time video streaming: pass through, voxel grid, approximate voxel grid, and bilateral. These filters each take a point cloud as input and will output a new, filtered point cloud.

4.1. Pass Through

The pass through filter passes the input points through constraints based on a particular field. It iterates through the entire point cloud once, performing two different operations. First, it removes non-finite points. Second, it removes any points that lie outside the specified interval for the specified field. For example, a programmer is able to set the field as the z-dimension (depth) and set the limit so that the filter removes any points that are half a meter away from the sensor.

4.2. Voxel Grid

The voxel grid filter assembles a 3D voxel grid over the entire input point cloud. Visually, this can be represented as a set of cubes being placed over the entire point cloud. For each individual voxel, the points that lie within it are down-sampled with respect to their centroid. This approach has a few drawbacks: (i) it requires a slightly longer processing time as opposed to using the voxel center; (ii) it is sensitive to noisy input spaces; and (iii) it does not represent the underlying surface accurately [23].

4.3. Approximate Voxel Grid

The approximate voxel grid filter attempts to achieve the same output as the voxel grid filter. However, rather than using the method described above, it sacrifices accuracy for speed. Rather than carefully determining the centroid and down-sampling the points, this filter makes a quick approximation of the centroid through the use of a hashing function.

4.4. Bilateral

The bilateral filter preserves the edges in a frame and reduces the noise of a point cloud. This is performed by replacing the intensity value for each point in the frame by the weighted average of intensity values from nearby points, based on a Gaussian distribution. These weights depend on the Euclidean distance and the differences in ranges (such as color and depth). For further information and a detailed explanation of the bilateral filter, see [24].

5. Frame Skipping

In general, frame skipping refers to the process of selecting a key reference frame to interpolate other frames. Rather than transmitting all video frames, the technique is able to skip certain frames that share a large number of similar data. By selecting a reference frame that is transmitted, the decoder is able to interpolate and fill in the skipped frames. The aim for frame skipping algorithms is to minimally affect the video quality and to significantly reduce the transmission size [25].

Frame skipping is typically caused by buffer overflow. This phenomenon frequently occurs after the encoding of an I-frame. As discussed in [26], this is due to the relatively large number of bits entering the buffer in contrast to the bits moving out. Interestingly, as noted in [25], frame skipped videos tend to be nearly as tolerable to packet loss as their raw video stream counterparts.

In this paper, we use frame skipping to refer to a simpler process. Rather than using interpolation, or some other method of estimation, our experiment simply displays the same frame once more. This keeps our system relatively complex-free and allows us to focus on the way the frame skipping algorithm selects which frames to skip. It does so by comparing the octrees of the two point cloud frames. Using PCL's octree change detector data structure, we are able to compare the voxels of the two octrees recursively. This will return a vector containing the indices of the voxels that differ. To calculate a similarity percentage (SP), we use:

$$SP = \frac{\#\ of\ points - \#\ of\ different\ points}{\#\ of\ points} * 100 \quad (1)$$

Frames that have a SP exceeding our similarity threshold are deemed too similar and therefore skipped.

6. Experiment Setup

In order to conduct a survey of different filtering algorithms, we needed to setup an experiment that is able to measure certain characteristics and compare these measurements. Because this is in the context of video streaming, we developed a threaded client/server application (see figure 2) that tests each filter and the possible benefits of frame skipping.

The client program is designed for generating, skipping, filtering, and transmitting the point clouds. It is connected to the Kinect hardware sensor to capture the RGB-D data from the real world environment. The responsibilities of the client program are divided into four threads:

Table 1 (Summary of QoS Metrics)

Filter	Filter Percentage	Branch Similarity	Point Similarity	Color Similarity
No Filter	0.00%	100.00%	100.00%	100.00%
Pass Through	88.07%	13.26%	99.77%	98.81%
Voxel Grid	79.82%	62.46%	78.20%	99.42%
Approximate Voxel Grid	73.27%	74.96%	75.33%	99.19%

Table 2 (Summary of Frame Similarity metric)

Filter	Without frame skipping	With frame skipping
No Filter	98.85%	92.01%
Pass Through	96.32%	87.74%
Voxel Grid	97.92%	92.26%
Approximate Voxel Grid	97.15%	90.93%

(i) convert the raw RGB-D data from the Kinect sensor into PCD using PCL and store this into buffer A;

(ii) access the PCD from buffer A, calculate the similarity percentage against the most recently transmitted PCD, skip the frame if the percentage exceeds the threshold, otherwise store the PCD in buffer B;

(iii) access the PCD from buffer B, filter it, and store the filtered PCD in buffer C; and

(iv) access the PCD from buffer C, fragmentize it, and transmit the datagrams to the server program using UDP.

The server program is designed for receiving the point clouds from the client and visualizing them to the monitor for display. Similar to the client, the server disperses its responsibilities to threads:

(i) receive datagrams containing PCD from the client program, defragmentize, and store the PCD in a buffer; and

(ii) access the PCD from the buffer and visualize it using PCL's visualization functionality.

We ran a total of ten tests: a base case which used no filter nor frame skipping, each of the filters, frame skipping, and each filter with frame skipping. These tests were all conducted in a wired-to using a wired connection to the Internet, while the server is receiving data through a wireless connection.

To test the relative effectiveness of each filter and the possible benefits of using a frame skipping technique in this experiment, we had to develop ways of measuring the QoS. We therefore created a set of QoS metrics that captured information for each test. This allows us to compare them in a quantitative fashion.

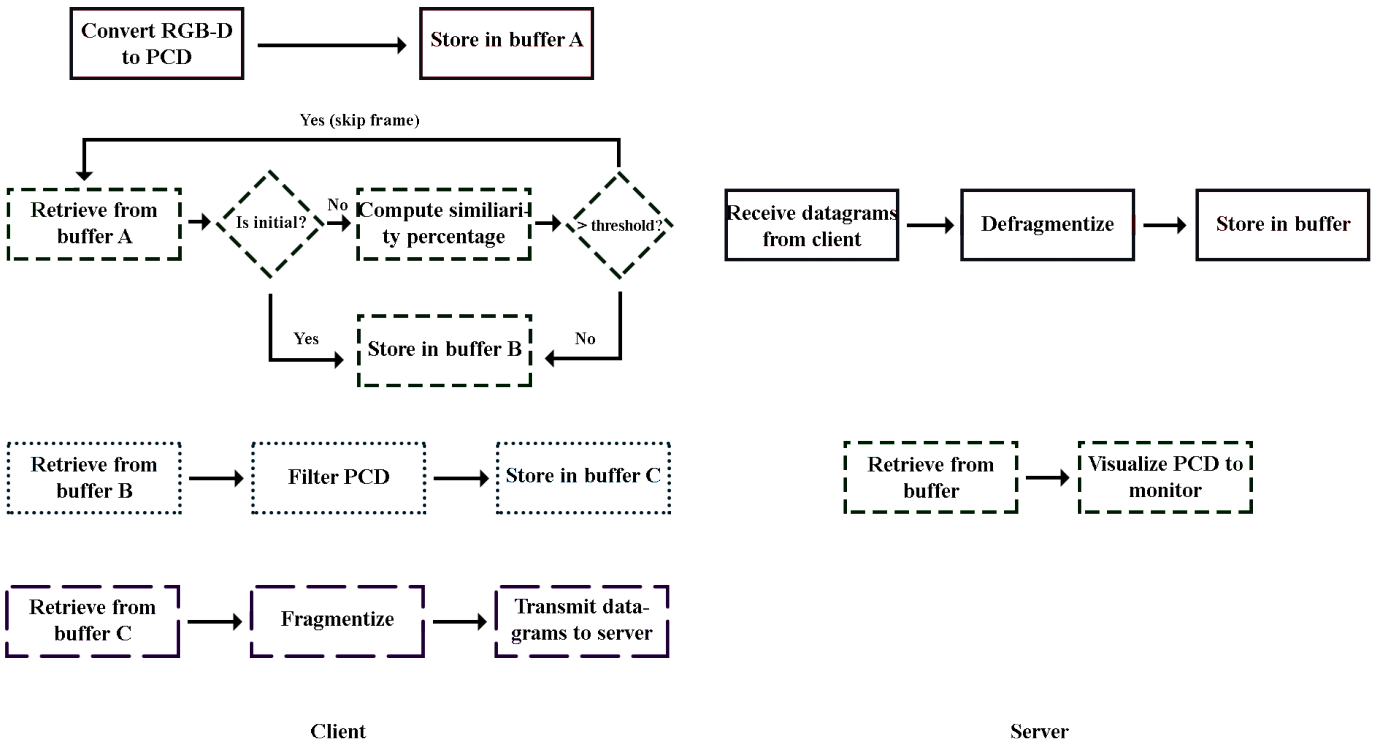


Fig. 2. The experiment flowchart of the threaded client/server applications. The client program (left) consists of four threads. The server program (right) consists of two threads.

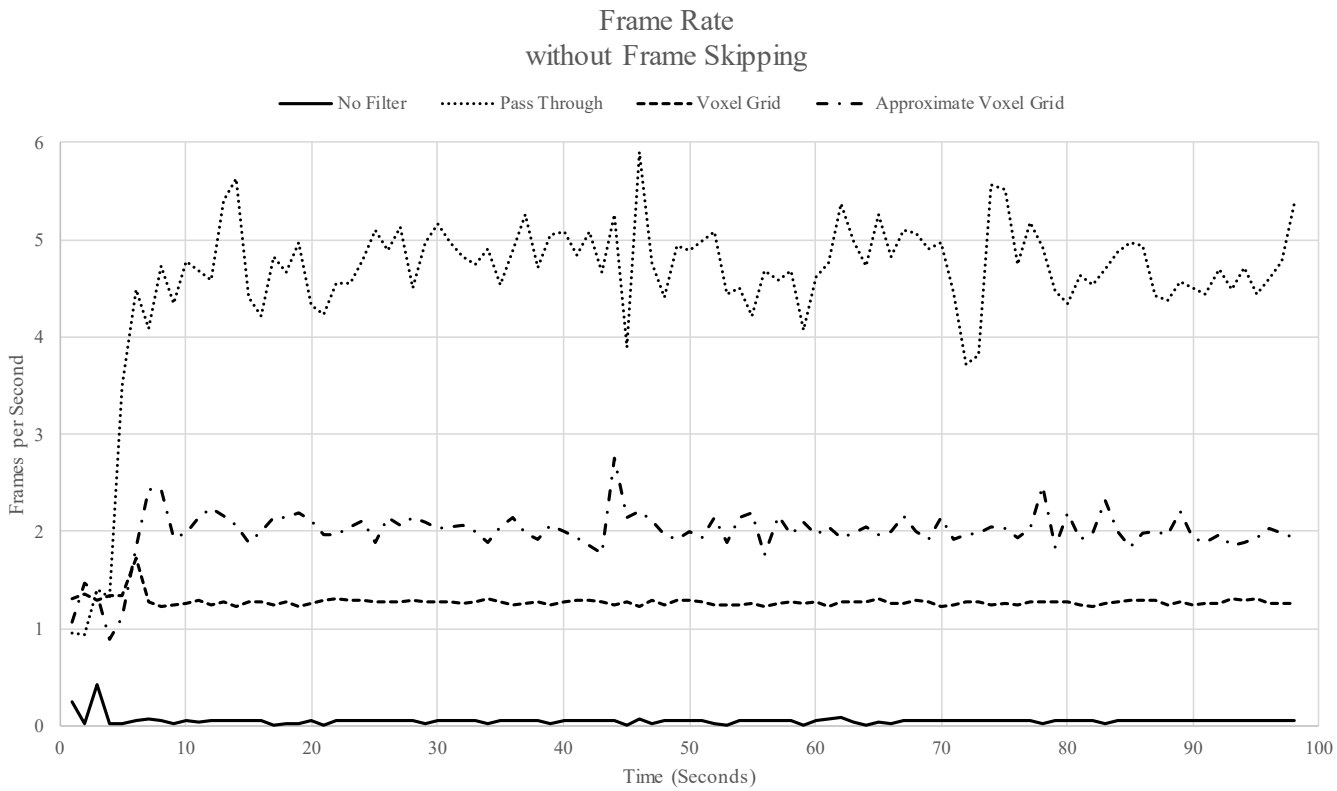


Fig. 3. A scatter plot that compares the frame rate for different filters in a wired-to-wireless network environment. The pass through filter maintains the highest fps of the four cases. This data was collected over a period of 500 frames.

wireless network environment, meaning that the client program is QoS Metrics

QoS is critical for the success of live video streaming. The quality of a video stream is, in part, determined by its *fps*. Video streams with a high fps usually result in a smooth visual experience for the users; low fps, on the other hand, tends to result in a "choppy" stream and poor user experience.

Another important characteristic of video streaming is the overall processing time required for each frame. To measure the QoS for this aspect, we used a *processing time metric*, which is the overall summation of the individual phases. This means, it adds up the time it takes for:

(i) determining if the point cloud should be skipped;

(ii) point cloud filtering;

(iii) point cloud transmission;

(iv) receiving point clouds; and

(v) point cloud visualization.

While the above two metrics work well in determining the QoS for video streaming in general, they do not take into account how well the filters are performing. Therefore, we also developed a *filter percentage metric*, meaning how many of the points in the original point cloud were filtered out. To calculate this, we found the difference in the number of points in the filtered point cloud from the original point cloud, and divided that by the original number of points.

Despite these three metrics, there is still a lack of measurement of how well the video stream performs visually. For that, we developed a set of three additional metrics:

(i) *branch similarity*, which compares the two branch structures of the octrees against each other;

(ii) *point similarity*, which measures how well two point values match up; and

(iii) *color similarity*, which calculates the similarity in the color values.

These visual QoS metrics are calculated by comparing the filtered PCD against the raw PCD.

In addition to the above three, we also required an additional visual QoS metric:

(iv) *frame similarity*, which compares how similar the current frame appears against the previous frame.

This allows us to quantitatively measure the video stream in terms of its visual differences.

7. Results

The tests that used the bilateral filter were disregarded due to the large processing time required. A single point cloud required over half an hour to filter, making it impractical for any live video streaming application.

Using the pass through filter provided the highest frame rate compared to the other filters (see figure 3). Moreover, it also achieved the lowest processing time. These can be explained due

to the pass through filter filtering the most points relative to the others (see table 1). With fewer points, the filtered point clouds from the pass through filter are smaller in data size, which allows for a higher frame rate and lower overall processing time.

In terms of the visual QoS metrics, the pass through filter appears to have a relatively low branch similarity. However, this is due to the nature of the filter. It effectively removes a large portion of the original point cloud, which drastically changes the underlying octree data structure. While the octree might be different, the pass through filter still maintains the highest similarity for the points (excluding the control case which uses no filter). See figure 5 for a visual depiction of the filter effects.

In regards to frame skipping, the results show that using a frame skipping technique provides a better end-to-end QoS. Comparing the five tests that did not use this scheme against the five that did shows that the benefits of the using the frame skipping technique. In terms of the fps, we found that the addition of frame skipping increased the frame rate for all filters (see figure 4). Specifically, the best combination arose from the pass through filter and frame skipping technique, which also achieved the lowest processing time. Moreover, by utilizing this technique, we found that the resulting PCD video stream was more responsive to changes (see table 2).

8. Summary

Among the four filtering methods allowed by PCL, the pass through filter results in the best scores for the QoS metrics. It removes the unnecessary background data, which reduces the point cloud data size and allows for a better experience in PCD video streaming. If the whole frame is required, however, the best filter is the approximate voxel grid, which outperforms the (normal) voxel grid filter in all QoS metrics. The addition of a frame skipping scheme had positive benefits overall. The frame rate increased for all filter options, although at the cost of a slight increase in processing time.

Although the use of filters and frame skipping reduces the original PCD size, the highest average frame rate that was achieved is merely 6.41 fps. Such a low frame rate cannot be considered to provide excellent end-to-end QoS. Therefore, while filtering and frame skipping improve the QoS additional techniques will be required.

For that purpose, we propose three additional techniques. First, compression, which will allow the data size to become even smaller, translating to a higher frame rate. Second, because of the network behavior that causes bandwidth fluctuation, a static compression ratio might work at certain bandwidth rates, but not all; instead, we need a dynamic compression algorithm that adjusts the compression ratio as a response to the bandwidth. Third, a progressive transmission scheme allows us to transmit the PCD layer-by-layer, in which each additional layer provides more details for the frame; the number of layers sent depends on the bandwidth and dynamically adjusts as the network changes.

9. Conclusion

Collecting 3D spatial data for PCD video streaming provides research challenges due to the high volume and high velocity data rate from the hardware sensors. Using Microsoft's Kinect sensor to collect the RGB-D data and PCL to process them, we were able

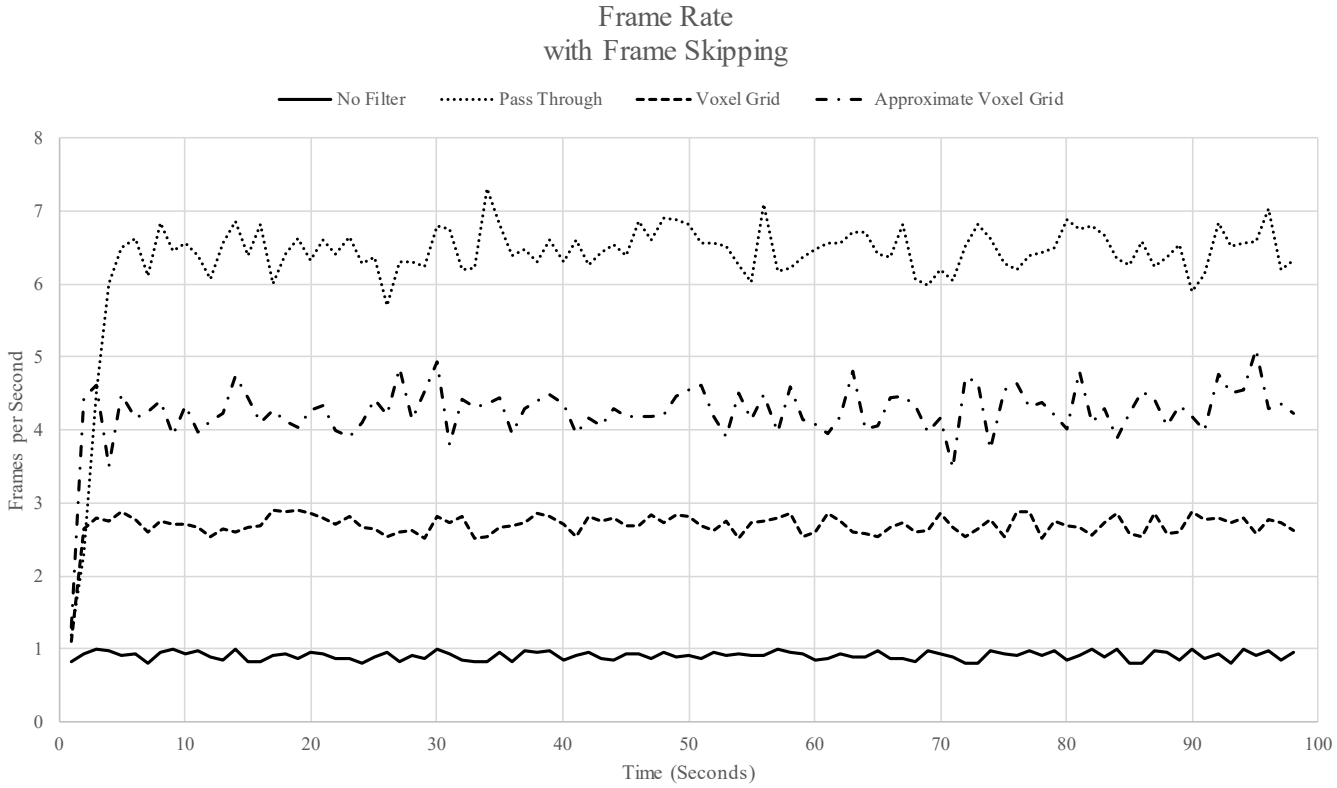


Fig. 4. A scatter plot that compares the frame rate for different filters using the frame skipping technique. The pass through filter maintains the highest fps of the four cases. This data was collected over a period of 500 frames.

(a) no filter

(b) pass through filter

(c) voxel grid filter

(d) approximate voxel grid filter

Fig. 5. A visual comparison of the four filter cases used in the experiment.

to compare and contrast different filtering methods to be used with the PCD. Filtering is a requirement due to the high data rate compared to the low bandwidth (300 MBps vs. 12.6 Mbps). In addition, we explored the use of a frame skipping technique that provides a better QoS experience for the end-user.

Using a threaded client/server application, we were able to survey the different filtering algorithms and the use of the frame skipping scheme by measuring a set of QoS metrics. Our results show that, in a live network environment, the pass through filter in conjunction with frame skipping achieves the highest scores in these metrics. Yet, utilizing only filters will not achieve a desirable video streaming experience. To do so, we propose the use of three additional techniques: octree compression, dynamic compression, and a progressive transmission scheme. Using these techniques provide further improvements by:

- compressing the octree data structure on the client-side to reduce the data rate
- dynamically adjusting the compression ratio in response to network bandwidth to support a reliable end-to-end QoS
- selectively transmitting high importance layers of the octree to lower the data rate

Conflict of Interest

The authors declare no conflict of interest.

References

[1] C. Moreno, M. Li, "A comparative study of filtering methods for point clouds in real-time video streaming," Proceedings of the World Congress on Engineering and Computer Science, San Francisco, CA, 2016.

[2] M. Miknis, R. Davies, P. Plassmann, A. Ware, "Near real-time point cloud processing using the PCL," 2015 International Conference on Systems, Signals and Image Processing, London, 2015.

[3] J. Fu, D. Miao, W. Yu, S. Wang, Y. Lu, S. Li, "Kinect-Like Depth Data Compression," in IEEE Transactions on Multimedia, **15**(6), 1340-1352, 2013.

[4] J. Kammerl, N. Blodow, R. B. Rusu, S. Gedikli, M. Beetz, E. Steinbach, "Real-time compression of point cloud streams," 2012 IEEE International Conference on Robotics and Automation, Saint Paul, MN, 2012.

[5] R. B. Rusu, Z. Marton, N. Blodow, M. Dolha, M. Beetz, "Towards 3D point cloud based object maps for household environments," Robotics and Autonomous Systems, **56**(11), 927-941, 2008.

[6] F. Nenci, L. Spinello, C. Stachniss, "Effective compression of range data streams for remote robot operations using H.264," 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, Chicago, IL, 2014.

[7] D. Belson, J. Thompson, J. Sun, R. Möller, M. Sintorn, G. Huston, "The state of the Internet," Akamai, Cambridge, MA, Tech. Rep., 2015.

[8] Z. Zhang, H. Shi, S. Wan, "Dynamic frame-skipping scheme for live video encoders," 2010 International Conference on Multimedia Technology, Ningbo, 2010.

[9] Y. Qi, M. Dai, "The effect of frame freezing and frame skipping on video quality," 2006 International Conference on Intelligent Information Hiding and Multimedia, Pasadena, CA, 2006.

[10] J. Fabian, T. Young, J. C. P. Jones, G. M. Clayton, "Integrating the Microsoft Kinect with Simulink: real-time object tracking example," in IEEE/ASME Transactions on Mechatronics, **19**(1), 249-257, 2014.

[11] C. L. Lai, Y. L. Huang, T. K. Liao, C. M. Tseng, Y. F. Chen, D. Erdenetsogt, "A Microsoft Kinect-based virtual rehabilitation system to train balance ability for stroke patients," 2015 International Conference on Cyberworlds, Visby, 2015.

[12] T. Deng, H. Li, J. Cai, T. J. Cham, H. Fuchs, "Kinect shadow detection and classification," 2013 IEEE International Conference on Computer Vision Workshops, Sydney, NSW, 2013.

[13] M. F. Lu, J. S. Chiang, T. K. Shih, S. Wu, "3D sphere virtual instrument with Kinect and MIDI," 2015 8th International Conference on Ubi-Media Computing, Colombo, 2015.

[14] C. T. Hsieh, "An efficient development of 3D surface registration by Point Cloud Library (PCL)," 2012 International Symposium on Intelligent Signal Processing and Communications Systems, New Taipei, 2012.

[15] Q. Zhang, L. Kong, J. Zhao, "Real-time general object recognition for indoor robot based on PCL," 2013 IEEE International Conference on Robotics and Biomimetics, Shenzhen, 2013.

[16] F. Ouyan, T. Zhang, "Octree-based spherical hierarchical model for collision detection," 2012 10th World Congress on Intelligent Control and Automation, Beijing, 2012.

[17] J. He, M. Zhu, C. Gu, "3D sound rendering for virtual environments with octree," IET International Conference on Smart and Sustainable City 2013, Shanghai, 2013.

[18] Y. Zhou, H. Ma, Y. Chen, "A frame skipping transcoding method based on optimum frame allocation in sliding window," 2010 2nd International Conference on Signal Processing Systems, Dalian, 2010.

[19] S. Bhattacharyya, E. Piccinelli, "A novel frame skipping method in transcoder, with motion information, buffer fullness and scene change consideration," 2009 17th European Signal Processing Conference, Glasgow, 2009.

[20] I. L. Jung, T. Chung, K. Song, C. S. Kim, "Efficient stereo video coding based on frame skipping for real-time mobile applications," IEEE Transactions on Consumer Electronics, **54**(3), 1259-1266, 2008.

[21] A. Jana, Kinect for Windows SDK Programming Guide, Packt, 2012.

[22] R. B. Rusu, S. Cousins, "3D is here: Point Cloud Library (PCL)," 2011 IEEE International Conference on Robotics and Automation, Shanghai, 2011.

[23] S. Orts-Escolano, V. Morell, J. García-Rodríguez, M. Cazorla, "Point cloud data filtering and downsampling using growing neural gas," The 2013 International Joint Conference on Neural Networks, Dallas, TX, 2013.

[24] C. Tomasi, R. Manduchi, "Bilateral filtering for gray and color images," 6th International Conference on Computer Vision 1998, Bombay, 1998.

[25] K. W. Lim, J. Ha, P. Bae, J. Ko, Y. B. Ko, "Adaptive frame skipping with screen dynamics for mobile screen sharing applications," IEEE Systems Journal, **PP**(99), 1-12, 2016.

[26] P. Feng, Z. G. Li, L. Keng Pang, G. N. Feng, "Reducing frame skipping in MPEG-4 rate control scheme," 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, 2002.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 84-89 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

Design and Fabrication of a Dielectrophoretic Cell Trap Array

Logeeshan Velmanickam, Dharmakeerthi Nawarathna*

Department of Electrical and Computer Engineering, North Dakota State University, Fargo, ND, 58102-6050

ARTICLE INFO

Article history:
Received: 01 December, 2016
Accepted: 08 January, 2017
Online: 28 January, 2017

Keywords:
Dielectrophoresis
Cell isolation
Electrode cell trap array
Electric field gradient

ABSTRACT

We present a design and fabrication of an integrated micro-fabricated dielectrophoretic (DEP) cell trap array in a microfluidic channel. The cell trap array is capable of isolating target cells in high-throughput manner and producing cell clusters of tunable cell numbers. In this work, we have used commercially available polystyrene beads to show the concept. Bead clusters of various sizes were successfully produced using DEP force (attractive or repulsive). We have found that the number of beads in a cluster depends on the frequency of electric field and the concentration of beads in the mixture.

1. Introduction

This report is an extension of work that was originally presented in 2016 IEEE International Conference on Electro Information Technology (EIT) [1]. Cell isolation is one of the basic steps of the devices that are commonly used in the applications such as Point-Of-Care diagnosis, food pathogen screening, and environmental monitoring. [2]. It is desirable that the cell isolation needs to be simple and high-throughput. Among the currently available techniques for cell isolation, FACS (fluorescence activated cell sorting), MACS (magnetic activated cell sorting), laser micro dissection, manual cell picking are the methods that are the most widely used. [3]. FACS is the most commonly used method but, to use FACS to isolate target cells, complete knowledge about the surface proteins of the target cells and other non-target cells in the cell mixture is needed. In FACS, first, a fluorescence dye is selectively attached to a surface protein of the target cell and then the separation of target cells is through the fluorescence [4, 3]. To use MACS in cell isolation, prior knowledge about the target cells are also needed. Briefly, in MACS, the target cells are labeled with magnetic beads and they are separated from the cell mixture by applying a magnetic force [5]. Therefore, the prior knowledge about the surface proteins of the cells in the mixture is required for both FACS and MACS methods [3]. Unfortunately, the information of target cells' surface proteins is not always available and therefore, both of these methods can only be used to isolate limited number of target cells from cell mixtures. In laser micro-dissection, typically, the operator observes the tissue sample through a microscope and the target cell population area is separated by using laser cutting [6]. In manual cell picking, micro-pipettes are used to pick-up the target cell from a cell mixture by applying a negative pressure [2]. Due to the manual operation, laser micro dissection and manual cell pick up methods are low-throughput and not applicable in most of the applications. These critical limitations in existing cell isolation techniques prevent advancing many important areas of biomedical engineering such as Point-Of-Care diagnosis, food pathogen screening, and environmental monitoring. Therefore, there is an urgent need to develop high-throughput, label-free cell isolation technique.

To address this critical need, we have developed a high-throughput and label-free technique for cell isolation. The technique is based on the dielectrophoresis (DEP) and microfluidics. First, we briefly describe DEP and microfluidics below.

2. Theory

Dielectrophoresis is a motion of suspensoid particles relative to the suspended medium resulting from polarization forces produced by an inhomogeneous electric field [7-11], which is widely used in many biomedical applications such as medical diagnostics, cell therapeutics and molecular separation [14]. In

* Corresponding Author: Dharmakeerthi Nawarathna,
Email: dharmakeerthi.nawara@ndsu.edu

particular, most of the biomedical materials including cells, DNA and proteins experience frequency dependent DEP force [7-9, 12].

Mathematically, the time-average DEP force acting on a dielectric particle such as cell or bead that are in a non-uniform external electric field can be represented as

$$F_{DEP} = 2\pi\, r^3\, \varepsilon_m\, Re\{fcm\}\, \nabla|E|^2 \qquad (1)$$

where r is the radius of the particle, ε_m is suspended medium permittivity, $Re\{fcm\}$ is real part of the Clausius-Mossoti (CM) factor, ∇ is the vector operator, and $\boldsymbol{E}$ is the r.m.s of the external electric field [11,13]. CM factor depends on the conductivity and permittivity of the particle $(\sigma_p, \varepsilon_p)$ and medium $(\sigma_m, \varepsilon_m)$, at the applied frequency (ω). Since the CM factor changes with frequency, DEP can be represented as negative (repulsive) $(-0.5 \leq Re\{fcm\} < 0)$, where polarized particles move towards the lowest field strength region), (attractive) positive $(0 < Re\{fcm\} \leq 1)$, where the polarized particles repelled from the lowest field strength region and move to regions of highest field gradient) or zero-force $(Re\{fcm\} = 0)$ [11-13].

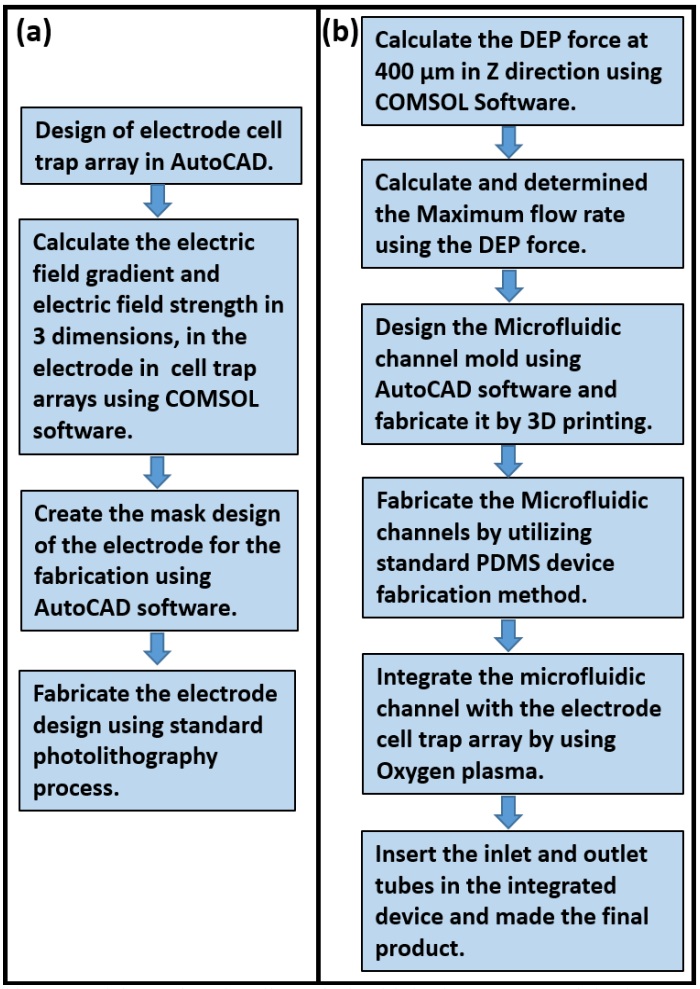


Figure 1: Dielectrophoretic cell trap device fabrication steps. (a) Fabrication steps of the DEP cell trap arrays, (b) fabrication steps of the microfluidic channel and integration method of the electrode cell trap array and microfluidic channel.

The σ_p of the cell or bead can be written as the sum of bulk conductivity (σ_{pbulk}) and surface conductivity (K_S), which can be represented as $\sigma_p = \sigma_{pbulk} + (K_S/r)$. Depending on the cell type and stage of the cell growth, surface conductivity value can be vary from cell to cell [9]. This will produce cell type or stage dependent DEP force on cells. Therefore, target cells can be separated from other cells using this unique feature.

3. Microfluidics

Microfluidics facilitates the handling of fluid flow through micrometer channels [23]. Microfluidics flow channels have been using in various applications [23,24]. The inclusion of microfluidic devices has led to increase the throughput, sensitivity and decrease the cost of applications such as flow cytometry, patterned three-dimensional cultures and drug delivery systems [23,24].

This is mainly due to the fact that Microfluidics requires only small volumes of samples (typically µL) and reagents, produces little waste, offers short reaction and analysis times [25].These unique characteristics make microfluidics particularly useful in biology and medicine [25]. Moreover, fabrication cost is considerably less as it uses the standard semiconductor manufacturing techniques such as photolithography, e-beam evaporation and etching techniques.

In our work we have integrated this microfluidics channels to flow cell/beads sample over the designed electrodes. Below, we describe the design, fabrication and the integration method of our device.

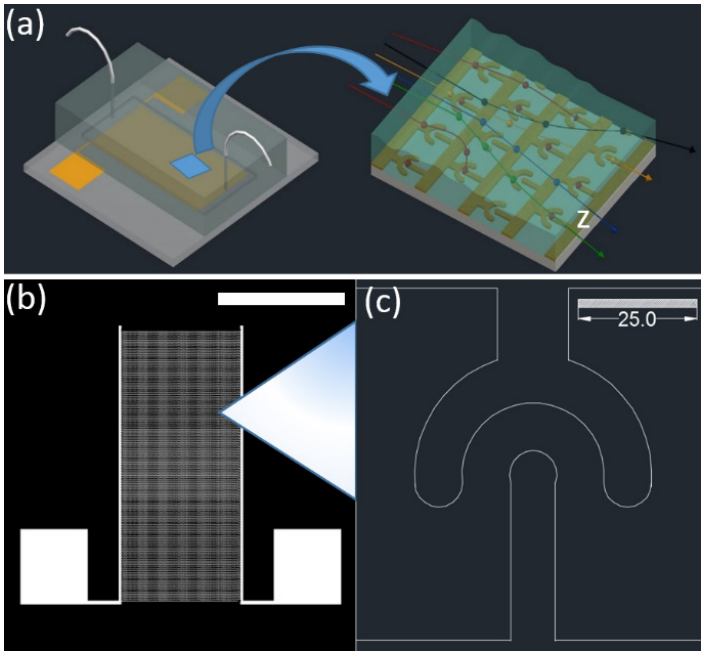


Figure 2: Design of the cell isolation device (a) Conceptual view of our cell isolation device, left side image is the magnified view of the device during the experiment. The electrodes are shown in gold color. Target cells (indicated in red) will be trapped over electrodes and non-target cells (indicated in green, yellow black and blue) will flow into the waste outlet of the device. (b) AutoCAD drawing of the electrode cell trap array (scale bar indicates 1cm), (c) enlarged image of one electrode cell trap (scale bar is in µm).

4. Design of the Dielectrophoretic Cell Isolation Device

The main steps of the fabrication are indicated in the Figure 1. Briefly, cell isolation device has two components. They are: (1) micro-scale DEP cell traps (electrodes) fabricated on glass wafer. Below, we will discuss design of cell traps using electric-current simulations (using COMSOL software), and (2) Microfluidics flow channels designed to flow cell sample over the DEP traps. The conceptual view of the cell isolation device is shown in "Figure. 2(a)" .We will discuss the design of the channel including, fabricating the mold, microfluidic channels and other important experimental parameters below.

4.1. Design and fabrication of electrodes.

Based on our prior experience on design electrodes [26], we came up with an electrode design that generate large DEP force on target cells. Since the DEP force depends on the electric field gradient (see equation 1), the electrode must be capable of generating high electric field gradient (∇E^2).

To estimate the expected electric field gradients from our electrodes, first, electrodes were drawn to a scale using AutoCAD software (see "Figure. 2(b and c)") and imported into COMSOL software. We then used the COMSOL software to calculate the electric fields and electric field gradients generated by the electrodes. This calculation was performed using AC/DC electric current (ec) module and frequency domain studies of COMSOL.

In the simulation, we have assumed that the electrode material is gold and the device is filled with phosphate buffered saline (PBS) solution. We then applied a 1 Vpeak-peak and 100 kHz frequency to the electrodes and design was meshed using free triangular extremely fine mesh with maximum element size= 5 µm and minimum element size= 0.026 µm. Then the electric field gradient and normalized electric field strength pattern were calculated in X-Y plane. Finally, variations of electric field gradient were plotted along contours to see the variation of the field gradient in the device. From these calculations, maximum electric field gradient is 1.09 x1016 V^2/m^3 and this electric field gradient will generate µN DEP force on cells and beads [1]. In comparison with other forces acting on the cell such as viscous drag and buoyancy (these forces are sub µN for the flow rate and cells that we use), DEP force is significantly large. Therefore, these electrodes are suitable for high-throughput cell isolation applications.

We then fabricated the electrodes using standard photolithography process following by metal deposition and lift-off process [15]. The images of the fabricated electrodes were shown in "Figure. 4(a,b, and c)".

4.2. Design and fabrication of microfluidics channels

The proper design is needed to achieve the maximum recovery of the target cells. To capture target cells without losing, we calculated the maximum flow rate that we will use in experiments.

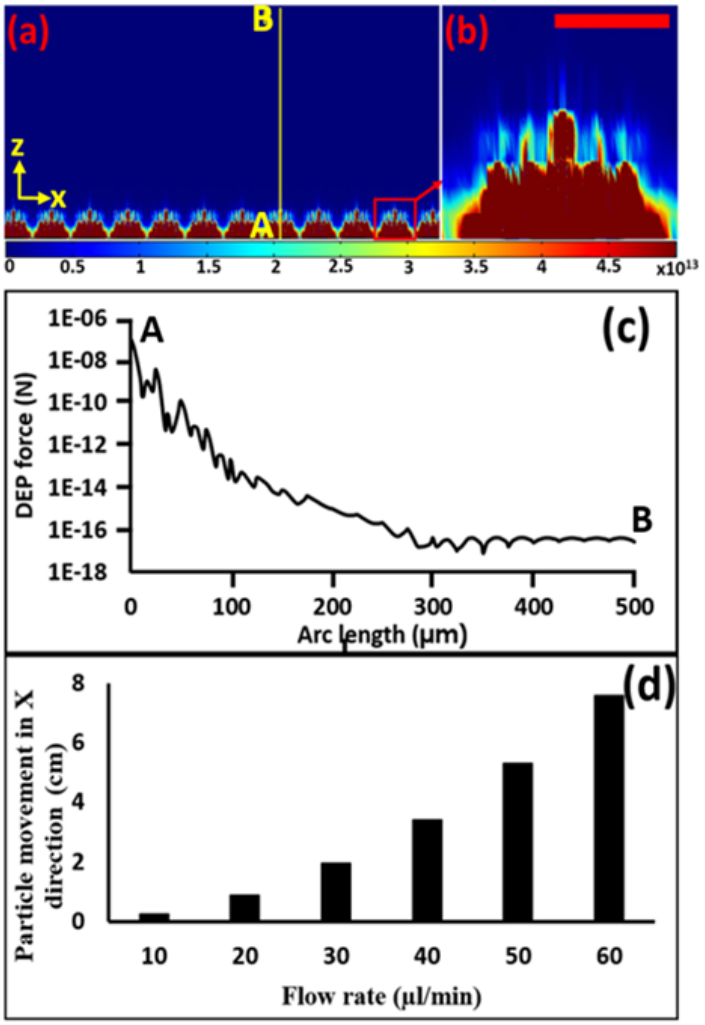


Figure 3: Calculated Electric field gradient in z direction. (a) Electric field gradient pattern in x-z plane in V^2/m^3, (b) enlarged image of the electric field gradient of a single electrode cell trap in V^2/m^3 (scale bar indicates 25µm), (c) variation of DEP force along the contour A-B in "Figure. 2(a)", (d) the variation of flow rate vs. distance travelled by the cell (at z=400 µm) in the x direction when it travel from 400 -0 µm under the influence of DEP.

To estimate the flow rate, first, we calculated the DEP force on a 20µm diameter cell (or bead) in the z direction along the contour A-B ("Figure. 3(a)"). We used $2\pi r^3 \varepsilon_m Re\{fcm\} = 4.4535 \times 10^{-24} (F.m^2)$, and calculated the DEP force using equation (1) and plotted in "Figure. 3(c)" [18-20]. For more stringent conditions, we have calculated the critical dimensions of the device needed to capture the cells that are about 400 µm above the electrodes. We have used the DEP force at 400 µm as $2.51 \times 10^{-17} N$ and assumed all other forces acting (e.g. buoyancy and weight) on the cell at 400 µm in the vertical direction are zero. We selected the particle speed to be 1.5 times faster than the flow speed of the fluid and calculated the fluid drag force [28]. Using this information, we calculated the horizontal distance travelled by the cell during trapping. Then we varied the flow rate and calculated distance travelled for each flow rate. The results are indicated in the "Figure. 3(d)". From this calculation, at 30 µl/min flow rate, the target cell will move about 1.9 cm, which is less

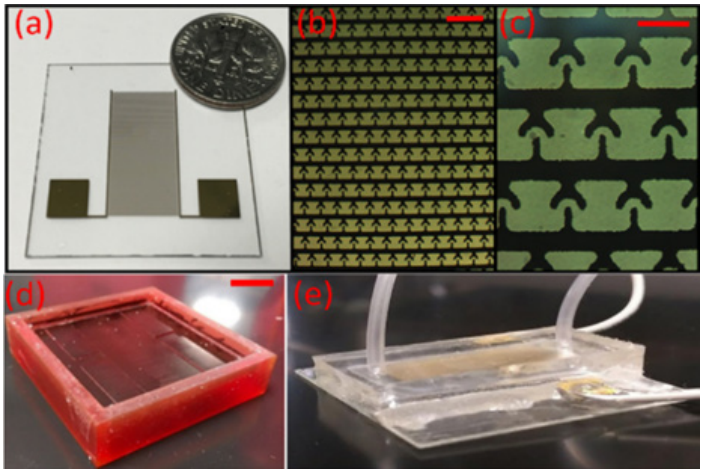


Figure 4: Fabrication of the device. (a) a picture of a fabricated electrode, (b) 10X zoomed view of the electrode under the microscope, showing the electrode array (scale bar indicates 200µm) (c) 30X zoomed image of the electrode showing individual traps (scale bar indicates 80µm), (d) Picture of the mold used for fabricating microfluidic channels (scale bar indicates 1.5cm), (e) Picture of the final version of the cell separation device.

than the length of our electrode dimension (2 cm). Therefore, theoretically, all the target cells that are below 400µm will be captured in the device.

To fabricate the microfluidics channel, we have utilized the standard microfluidics channel fabrication using PDMS (Polydimethylsiloxane) [16-17]. Briefly, microfluidics channels were designed in AutoCAD and fabricated a mold (3D printing: Proto Labs, Inc.2600 Niagara Ln N, Plymouth, MN 55447), "Figure. 4(d)" indicate the mold used in fabricating the microfluidics channel. We then mixed Sylgard 184 elastomor with curing agent (10:1) and poured over the mold. Finally, the mold with PDMS was cured by keeping on a hotplate (at 80°C) for 3-4 hours. After curing, PDMS was peeled off from the mold and inlet and outlet holes were drilled using a biopsy punch. Then, PDMS channels and electrodes were bonded using oxygen plasma [27]. Finally, the micro-bore tube with inner diameter 0.04 inches and outer diameter 0.07 inches were glued using the epoxy glue at the inlet and the outlet of the microfluidic device for the purpose of flowing the cell mixture. Final version of the device is indicated in "Figure. 4(e)". We then used the fabricated microfluidics device in the experiments and demonstrated the DEP based cell-trapping. Details about those experiments is discussed below.

5. Experimental Setup

Our first experiment were focused on to validate the cell isolation using positive DEP and negative DEP. We then studied how these forces can be used to precisely control the number of cells in a trap. We started with polystyrene beads and determined the experimental conditions (electric field magnitude and frequency) needed to isolate beads. We then use these conditions as the starting point for isolating cells. We used 500 nm diameter green fluorescence polystyrene beads to represent the trapping of smaller cells (e.g.: bacteria) and 20 µm diameter polystyrene beads to represent mammalian cells.

The bead samples were prepared in following manner. 1 ml sample was prepared with 500 nm green fluorescence beads by adding 20 µl of beads solution and 980µl of 0.01XPBS solution (conductivity= 0.03 S/m). We then mounted the microfluidics device on the fluorescence microscope stage, bead sample was directly pipetted on to the electrodes or loaded to the syringe and connected to the syringe pump (flow rate= 30 µl/min, Braintree Scientific INC, BS-300). When the beads are over the electrodes, an external electric field was applied (frequency =100 kHz, and magnitude= 5 V_{pp}, Tektronix AFG 3021B)

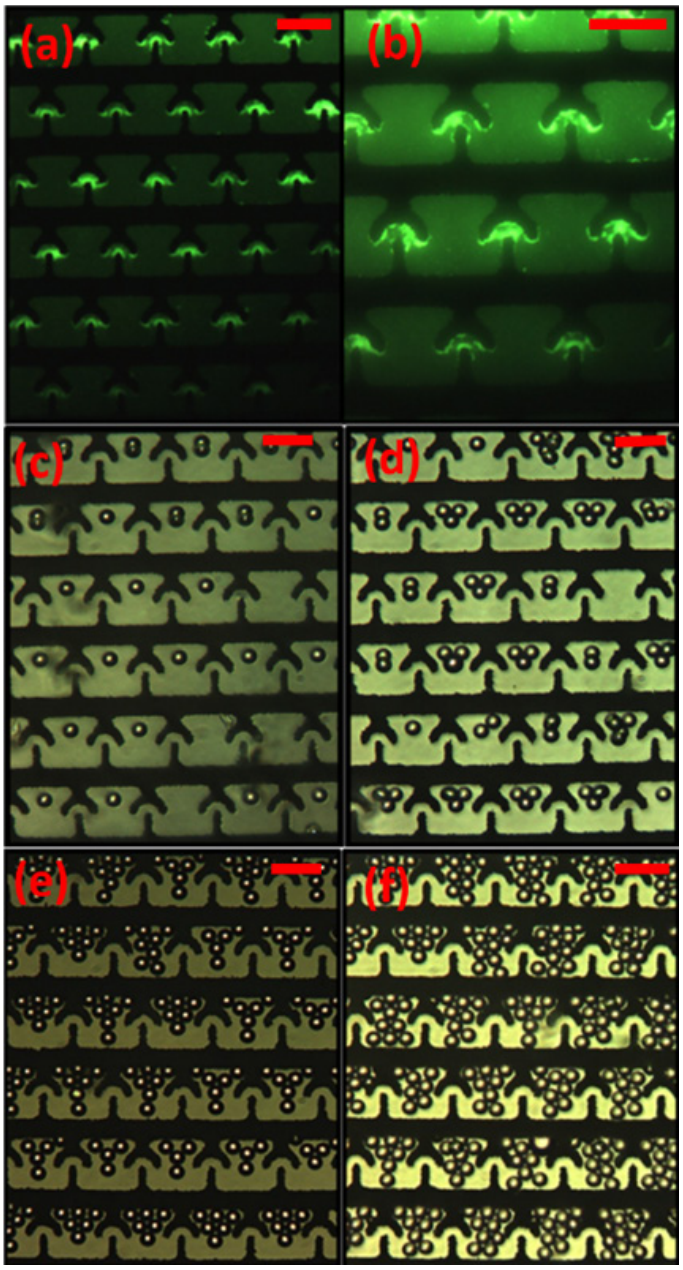


Figure 5: Experimental results. (a and b) results of the experiment with 500 nm green fluorescence polystyrene beads under the fluorescence microscope with 20X and 30X magnification respectively,(c,d,e and f) results of the experiment with 20 µm polystyrene beads under the microscope with 20X magnification. Scale bars indicate 80µm.

and observed the beads trapping. The pictures of trapped beads are shown in ("Figure. 5(a, and b)"). Similarly, we repeated this experiment for 20 µm beads and studied the beads trapping with (repulsive) negative and (attractive) positive DEP. The starting point of our experiments with 20 µm beads was 1 MHz (frequency) and 1.5 Vp-p (magnitude). For the 20 µm beads, we flowed the sample (30 µL/min), but it was unable to trap multiple beads in a single-trap. Therefore, for trapping multiple beads we use the static condition (sample flow rate=0 µL/min). The "Figure. 5(c,d,e,and f)" indicate the some of our results that demonstrate how we used DEP (negative) to vary the number of beads in a trap. We will discuss this in detail below.

6. Results

The "Figure 5(a, and b)" shows the trapping of 500 nm beads using positive DEP (100 kHz and 5 Vp.p). The Figure 5(b) is the magnified view of the Figure 5(a) that demonstrate the DEP based trapping. We then varied the frequency of the electric field from 100 kHz to 15 MHz but did not observe the trapping through negative DEP. For 20 µm beads, we observed that it is indeed possible to trap beads using only negative DEP. Further, with negative DEP, it is also possible to vary the number of beads trapped by simply changing frequency and the concentration of beads. "Figure 5(c, d, e and f)" indicates some of the results from those experiments. Briefly conditions that we used for trapping 20 µm beads are summarized in the table 1 below. Our conclusions from this study are summarized below.

Concentration of beads(beads/ml) $x10^3$	Applied Frequency (MHz)	Average number of beads per trap	Standard deviation
90	1	1	0.071
180	8	3	1.000
540	15-18	6	1.035
1440	20-22	8	1.760
1800	24-25	10	2.407

Table 1: Trapping results of 20 µm beads using negative DEP.

7. Conclusion

We have successfully designed and fabricated an electrode for efficient cell trapping. Experiments were performed using two sizes of beads to demonstrate the usage of negative and positive DEP in trapping and clustering beads. Clustering can be used to count number of cells in the sample or culturing the trapped cells. For 500 nm beads, we did not see controllable trapping with both negative and positive DEP. For 20 µm beads, negative DEP produced controllable trapping. Our next step is to use the cells and demonstrate cell trapping with our device. Finally, we will use the device to develop assays based on cell trapping [21-22]. The results from those experiments will be published in forthcoming articles.

Acknowledgment

We are grateful for the financial support from Richard Offerdahl research grant. We would like to thank Vidura Jayasooriya, Bharat Verma, Darrin Laudenbach and Jenna Pender in the Department of Electrical and Computer Engineering at the North Dakota State University for their help in developing 3D AutoCAD models and fabrication of the microfluidic device (PDMS) molds. We also thank the staff of the Center for Nanoscience and Engineering at the North Dakota State University for their help in fabricating electrodes.

References

[1] L.Velmanickam, and K.Nawarathna, "Dielectrophoretic cell isolation in microfluidics channels for high-throughput biomedical applications", in IEEE International Conference on Electro Information Technology (EIT), Grand folks, ND, USA, 2016.

[2] I.L.Doh and Y-H Cho, " A continuous cell separation chip using hydrodynamic dielectrophoresis (DEP) process", Sensors and Actuators A 121,2005.

[3] A. Gross, J. Schoendube, S. Zimmermann, M. Steeb, R. Zengerle, and P. Koltay, "Technologies for Single-Cell Isolation." International Journal of Molecular Sciences, vol. 16,16897-16919, 2015.

[4] W.A. Bonner, H.R Hulett, R.G Sweet and L.A Herzenberg, "Fluorescence Activated Cell Sorting", the review of scientific instruments, vol. 43,404-409, 1972

[5] S. Miltenyi, W. Muller, W. Weichel, and A. Radbruch, " High Gradient Magnetic Cell Separation With MACS", Cytometry, vol.11, 231-238, 1990.

[6] F. Fend, and M. Raffeld, "Laser capture microdissection in pathology", Journal of Clinical Pathology, vol.53, 666-672, 2000.

[7] R. Pethig, "Dielectrophoresis: Status of the theory, technology and applications", Bio microfluidics, vol.4, 022811, 2010.

[8] P. R.C.Gascoyne, X. B Wang, Y. Huang and F. F.Becker, "Dielectrophoresis Separation of Cancer Cells from Blood", IEEE Transactions on Industrial Applications, vol. 33, 670-678, 1997.

[9] A. Nakano and A. Ros, "Protein Dielectrophoresis: Advances, Challenges and Applications", Electrophoresis, vol. 34, 1085-1096, 2013.

[10] N. Demierre, T. Braschler, R. Muller, P. Renaus, "Focusing and continuous separation of cells in a microfluidic device using lateral dielectrophoresis", Sensors and Actuators B, vol. 132, 388-396, 2008.

[11] I. Ermolina and H. Morgan. "The electrokinetic properties of latex particles: Comparison of electrophoresis and dielectrophoresis", Journal of Colloid and Interface Science, vol. 285, 419-428, 2005.

[12] L. Zheng, J. P.Brody, P. J.Burke, "Electronic manipulation of DNA, proteins, and nanoparticles for potential circuit assembly", Biosensors and Bioelectronics, vol.20, 606-619, 2004.

[13] Y. Huang, S. Joo, M. Duhon, M. Heller, B. Wallace, and X. Xu, "Dielectrophoretic Cell Separation and Gene Expression Profiling on Microelectronic Chip Arrays", Analytical Chemistry, vol.74, 3362-3371, 2002.

[14] Z. R. Gagnon, "Cellular dielectrophoresis: Applications to the characterization, manipulation, separation and patterning of cells", Electrophoresis, vol. 32, 2466-2487, 2011.

[15] H. Lu, M. A. Schmidt, & K. F. Jensen "A microfluidic electroporation device for cell lysis" Lab on a Chip, vol. 5, 23-29, 2004.

[16] S. H. Kim, Y. Ciu, M. J. Lee, S-W. Nam, D. Oh, S. H. Kang, Y. S. Kim, and S. Park, " Simple fabrication of hydrophilic nano-channels using the chemical bonding between activated ultrathin PDMS layer and cover glass by oxygen plasma", Lab Chip, vol.11, 348-353, 2011.

[17] S. Bhattacharya, A. Datta, J. M.Berg, and S. Gangopadhyay, " Studies on Surface Wettability of Poly(Dimethyl) Siloxane (PDMS) and Glass Under Oxygen-Plasma Treatment and Correlation With Bond Strength", Journal Of Micromechanical System, vol.14, 590-597, 2005.

[18] D. Nawarathna, T. Turan, and H. Kumar Wickramasinghe, "Selective probing of mRNA expression levels within a living cell", Applied Physics Letters, vol.95, 083117, 2009.

[19] Hao Zhou, Matthew A. Preston, Robert D. Tilton, Lee R. White, "Calculation of the electric polarizability of a charged spherical dielectric particle by the theory of colloidal electrokinetics", Journal of Colloid and Interface Science, 285, 845-856, 2005.

[20] Ronald Pethig, "Review Article—Dielectrophoresis: Status of the theory,technology, and applications", BioMicrofluidics, vol.4, 022811, 2010.
[21] A. Rosenthal, A. Macdonald and J.Voldman, " Cell patterning chip for controlling the stem cell Microenvironment", Biomaterials, vol.28, 3208-3216, 2007.
[22] R.S. Kane, S.Takayama, E. Ostuni, D.E. Ingber, and G.M Whitesides, " Patterning proteins and cells using soft lithography", Biomaterials, vol.20, 2363-2376, 1999.
[23] D.J.Beebe, G.A.Mensing, and G.M.Walker, "Physics and applications of microfluidics in biology", Annu.Rev.Biomed.Eng, vol.4, 261-286, 2002.
[24] A.L.Paguirigan, and David J.Beebe, "Microfluidics meet cell biology, bridging the gap by validation and application of microscale techniques for cell biological assays", BioEssays, vol.30.9, 811-821, 2008.
[25] D.B.Weibel and G.M.Whitesides, "Applications of microfluidics in chemical biology", Science Direct, vol.10, 584-591, 2006.
[26] V.Jayasooriya and D.Nawarathna, " Design of Micro-interdigitated Electrodes and Detailed Impedance Data Analysis for Label-free Biomarker Quantification" Electroanalysis, vol 28, DOI: 10.1002/elan.201600364, 2016
[27] Xiong, Liangcai, Peng Chen, and Quansheng Zhou. "Adhesion promotion between PDMS and glass by oxygen plasma pre-treatment. Journal of Adhesion Science and Technology, vol.28, 1046-1054, 2014.
[28] Jacques Magnaudet and Dominique Legendre, " The viscous drag force on a spherical bubble with a time-dependent radius", Physics of fluid, Vol.10, 550-554, 1998.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 90-96 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

Configuration/Infrastructure-aware testing of MapReduce programs

Jesús Morán[1*], Bibiano Rivas[2], Claudio de la Riva[1], Javier Tuya[1], Ismael Caballero[2], Manuel Serrano[2]

[1]*University of Oviedo, Department of Computing, 33394, Spain*

[2]*University of Castilla-La Mancha, Institute of Technology and Information Systems, 13051, Spain*

ARTICLE INFO

Article history:
Received: 15 December, 2016
Accepted: 20 January, 2017
Online: 28 January, 2017

Keywords:
Software testing
Functional testing
MapReduce programs
Big Data Engineering
Hadoop

ABSTRACT

The implemented programs in the MapReduce processing model are focused in the analysis of large volume of data in a distributed and parallel architecture. This architecture is automatically managed by the framework, so the developer could be focused in the program functionality regardless of infrastructure failures or resource allocation. However, the infrastructure state can cause different parallel executions and some could mask the faults but others could derive in program failures that are difficult to reveal. During the testing phase the infrastructure is usually not considered because commonly the test cases contain few data, so it is not necessary to deploy a parallel execution or handle infrastructure failures, among others potential issues. This paper proposes a testing technique to generate and execute different infrastructure configurations given the test input data and the program under test. The testing technique is automatized by a test engine and is applied to real world case studies. As a result, the test engine generates and executes several infrastructure configurations, revealing a functional fault in two programs.

1. Introduction

The massive data processing trends have brought to light several technologies and processing models in the *Big Data Engineering* field [1]. Among them, *MapReduce* [2] can be highlighted as it permits the analysis of large data based on the "divide and conquer" principle. These programs run two phases in a distributed infrastructure: the *Mapper* and the *Reducer*. The first one divides the problem into several subproblems, and then the *Reducer* phase solves each subproblem. Usually, *MapReduce* programs run on several computers with heterogeneous resources and features. This complex infrastructure is managed by a framework, such as *Hadoop* [3] which stands out due to its wide use in the industry [4]. Other frameworks as for example Apache Spark [5] and Apache Flink [6] among others also use the *MapReduce* programming model.

From point of view of the developer, a *MapReduce* program can be implemented only with *Mapper* and *Reducer*, regardless of the infrastructure. Then the framework that manages the infrastructure is also responsible to, over several computers, automatically deploy, run the program and lead the data processing between the input and output. Among others, the framework divides the input into several subsets of data, then processes each one in parallel and re-runs some parts of the program if necessary.

Although that the program can be implemented abstracting the infrastructure, the developer needs to consider how the infrastructure configuration could affect the program functionality. A previous work [7] detects and classifies several faults that depend on how the infrastructure configuration affects the program execution and produces different output. These faults are often masked during the test execution because the tests usually run over an infrastructure configuration without considering the different situations that could occur in production, as for example different parallelism levels or the infrastructure failures [8]. On the other hand, if the tests are executed in an environment similar to the production, some faults may not be detected because it is common that the test inputs contain few data, and in these cases *Hadoop* does not parallelize the program execution. There are some tools to enable the simulation for some of these situations (for example computer and net failures) [9, 10, 11], but it is difficult to design, generate and execute the tests in a deterministic way because there are a lot of elements that need fine grained simulation, including the infrastructure and framework.

*Corresponding Author: Jesús Morán, Viesques edificio departamental 1 Gijón (Spain) moranjesus@lsi.uniovi.es

The main contribution of this paper is a technique that can be used to generate automatically the different infrastructure configurations for a *MapReduce* application. The goal is to execute test cases with these configurations in order to detect functional faults. Given a test input data, the configurations are obtained based on the different executions that can happen in production. Then each one of the configurations is executed in the test environment in order to detect functional faults of the program that may occur in production. This paper extends the previous work [12] and the contributions are:

1. A combinatorial technique to generate the different infrastructure configurations, taking into account characteristics related to the *MapReduce* processing and the test input data.
2. Automatic support by means of a test engine based on MRUnit [13] that allows the execution of the infrastructure configurations.
3. Evaluation of the approach detecting failures in a two real world programs.

The rest of the paper is organized as follows. In Section 2 the principles of the *MapReduce* paradigm are introduced. The related work about software testing in *MapReduce* paradigm is presented in Section 3. The generation of the different configurations, the execution and the automatization of the tests are defined in Section 4. In Section 5 it is applied to a two case studies. The paper ends with conclusions and future work in Section 6.

2. MapReduce Paradigm

The function of the *MapReduce* program is to process high quantities of data in a distributed infrastructure. The developer implements two functionalities: *Mapper* task that splits the problem into several subproblems and *Reducer* task that solves these subproblems. The final output is obtained from the deployment and the execution over a distributed infrastructure of several instances of *Mapper* and *Reducer*, also called tasks. Hadoop (or other framework) automatically carry out the deployment and execution. First, several *Mapper* tasks analyse in parallel a subset of input data and determine which subproblems these data need. When the execution of all *Mappers* are finished, several *Reducers* are also executed in parallel in order to solve the subproblems. Internally *MapReduce* handles <key, value> pairs, where the key is the subproblem identifier and the value contains the information to solve it.

To explain *MapReduce* let us suppose a program that calculates the average temperature per year from historical data about temperatures. This program solves for each year one subproblem, so the year is the identifier or key. The *Mapper* task receives a subset of temperature data and emits <year, temperature of this year> pairs. Then *Hadoop* aggregates all values per key. Therefore, the *Reducer* tasks receive subproblems like <year, [all temperatures of this year]>, that is all temperatures grouped per year. Finally, the *Reducer* calculates the average temperature. For example, in Figure. 1 an execution of the program considering the input is detailed: year 2000 with 3º, 2002 with 4º, 2000 with 1º, and 2001 with 5º. The first two inputs are analysed in one *Mapper* task and the remainder in another task. Then the temperatures are grouped per year and sent to the *Reducer* tasks. The first *Reducer* receives all the temperatures for the years 2000 and 2002, and the other task for the year 2001. Finally, each *Reducer* emits the average temperature of the analysed subproblems: 2º in the year 2000, 4º in 2002 and 5º in 2001. This program with the same input could be executed in another way by the framework, for example with three *Mappers* and three *Reducers*. Regardless of how the framework runs the program, it should generate the expected output.

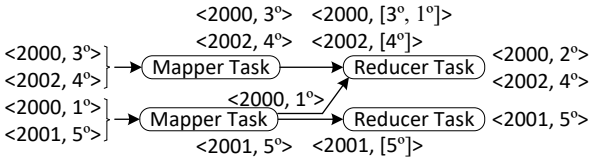


Figure. 1 Program that calculates the average temperature per year

In order to optimize the program, a *Combiner* functionality can be implemented. This task is run after the *Mapper* and the goal is to remove the irrelevant <key, value> pairs to solve the subproblem. In *MapReduce* there are also other implementations such as for example Partitioner that decides for each <key, value> pair which *Reducer* analyses it, Sort that sorts the <key, value> pairs, and Group that aggregates the values of each key before the *Reducer*.

An incorrect implementation of these functionalities could cause a failure in one of the different ways in which *Hadoop* can run the program. These faults are difficult to detect during testing because the test cases usually contain few input data. In this way it is not necessary to split the inputs and therefore the execution is over one *Mapper*, one *Combiner* and one *Reducer* [2].

3. Related Work

Despite the testing challenges of the *Big Data* applications [14, 15] and the progresses in the testing techniques [16], little effort is focused on testing the *MapReduce* programs [17], one of the principal paradigms of *Big Data* [18]. These large-scale programs have several issues and challenges to measure and assure the quality [19]. A study of Kavulya et al. [20] analyses several *MapReduce* programs and 3% of them do not finish, while another study by Ren et al. [21] places the number between 1.38% and 33.11%.

Many of the works about testing of the *MapReduce* programs focus on performance [22, 23, 24] and to a lesser degree functionality. A testing approach for *Big Data* is proposed by Gudipati et al. [25] specifying several processes, one of which is about *MapReduce* validation. In this process Camargo et al. [26] and Morán et al. [7] identify and classify several functional faults. Some of these faults are specific of the *MapReduce* paradigm and they are not easy to detect because they depend on the program execution over the infrastructure. One common type of fault is produced when the data should reach the *Reducer* in a specific order, but the parallel execution causes these data to arrive disordered. This fault was analysed by Csallner et al. [27] and Chen et al. [28] using some testing techniques based on symbolic execution and model checking. In contrast to the previous works, the approach of this paper is not focused on the detection of only one type of fault, it can also detect other *MapReduce* specific faults. To do this, the test input data is executed over different infrastructure configurations that could lead to failures.

Several research lines suggest injecting infrastructure failures [29, 30] during the testing, and several tools support their injection [9, 10, 11]. For example, the work by Marynowski et al. [31] allows the creation of test cases specifying which computers fail

and when. One possible problem is that some specific *MapReduce* faults could not be detected by infrastructure failures, but require full control of *Hadoop* and the infrastructure. In this paper, the different ways in which *Hadoop* could run the program are automatically generated from the functional point of view, regardless of the infrastructure failures and *Hadoop* optimizations.

Furthermore, there are other approaches oriented to obtain the test input data of *MapReduce* programs, such as [32] that employs data flow testing, other based on a bacteriological algorithm [33], and [34] based on input domain together with combinatorial testing focused on ETL (Extract, Transformation and Load). In this paper, given a test input data, several configurations of infrastructure are generated and then executed in order to reveal functional faults. The test input data of this approach could be obtained with the previous testing techniques.

The functional tests can be executed directly in the production cluster or in one computer with *Hadoop*. Herriot [35] can be used to execute the tests in a cluster while providing access to their components supporting, among others, the injection of faults. Another option is to simulate a cluster in memory with the MiniClusters libraries [36]. In the unit testing, JUnit [37] could be used together with mock tools [38], or directly by MRUnit library [13] adapted to the *MapReduce* paradigm. These test engines only execute one infrastructure configuration and usually without parallelization. In this paper a test engine is implemented by an MRUnit extension that automatically generates and executes the different infrastructure configurations that could occur in production. The test engine proposed extends MRUnit because in Hadoop is very usual to develop java programs [21], and the java programs usually employs JUnit libraries [39]. MRUnit put together JUnit with mocks, reflection and other tools in order to simplify the execution of the test case for the *MapReduce* programs.

4. Generation and Execution of Tests

The generation of the infrastructure configurations for the tests are defined in Section 4.1, and a framework to execute the tests in Section 4.2.

4.1. Generation of the test scenarios

To illustrate how the infrastructure configuration affects the program output, suppose that the example of Section 2 is extended with a *Combiner* in order to decrease the data and improve the performance. The *Combiner* receives several temperatures and then they are replaced by their average in the *Combiner* output. This program does not admit a *Combiner* because all the temperatures are needed to obtain the total average temperature. The *Combiner* is added in order to optimize the program, but injects a functional fault in the program. Figure. 2 represents three possible executions of this program with the same input (year 1999 with temperatures 4º, 2º and 3º) that could happen in production considering the different infrastructure configurations.

The first configuration executes one *Mapper*, one *Combiner* and one *Reducer* and produces the expected output. The second configuration also generates the expected output executing one *Mapper* that processes the temperatures 4º and 2º, another *Mapper* for 3º, two *Combiner*, and finally one *Reducer*. The third configuration also executes two *Mapper*, two *Combiner* and one *Reducer*, but produces an unexpected output because the first *Mapper* processes 4º and the second *Mapper* the temperatures 2º and 3º. Then one of the *Combiner* tasks calculates the average of 4º, and the other *Combiner* of 2º and 3º. The *Reducer* receives the previous averages (4º and 2.5º), and calculates the total average in the year. This configuration produces 3.25º as output instead of the 3º of the expected output. The program has a functional fault only detected in the third configuration. Whenever this infrastructure configuration is executed the failure is produced, regardless of the computer failures, slow net or others. This fault is difficult to reveal because the test case needs to be executed in a completely controlled way under the infrastructure configuration that detect it.

Given a test input data, the goal is to generate the different infrastructure configurations, also called in this context *scenarios*. For this purpose, the technique proposed considers how the *MapReduce* program can execute these input data in production. First, the program runs the *Mappers*, then over their outputs the *Combiners* and finally the *Reducers*. The execution can be carried out over a different number of computers and therefore the *Mapper-Combiner-Reducer* can analyse a different subset of data in each execution. In order to generate each one of the *scenarios*, a combinatorial technique [40] is proposed to combine the values of the different parameters that can modify the execution of the *MapReduce* program. In this work the following parameters are considered based on previous work [7] that classifies different types of faults of the *MapReduce* applications:

- *Mapper* parameters: (1) Number of *Mapper* tasks, (2) Inputs processed per each *Mapper*, and (3) Data processing order of the inputs, that is, which data are processed before other data in the *Mapper* and which data are processed after.
- *Combiner* parameters for each *Mapper* output: (1) Number of *Combiner* tasks, and (2) Inputs processed per each *Combiner*.
- *Reducer* parameters: (1) Number of *Reducer* tasks, and (2) Inputs processed per each *Reducer*.

The different *scenarios* are obtained through the combination of all values that can take the above parameters and applying the constraints imposed by the sequential execution of *MapReduce*. The constraints considered in this paper are the following:

1. The values/combinations of the *Mapper* parameters depend on the input data because it is not possible more tasks than data. For example, if there are three data items in the input, the maximum number of *Mappers* is three.
2. The values/combinations of the *Combiner* parameters depend on the output of the *Mapper* tasks.
3. The values/combinations of the *Reducer* parameters depend on the output of the *Mapper-Combiner* tasks and another functionality executed by *Hadoop* before *Reducer* tasks. This other functionality is called Shuffle and for each <key, value> pair determines the *Reducer* task that requires these data, then sorts all the data and aggregates by key.

Suppose the program of Figure. 2 to illustrate how the parameters are combined and how the constraints are applied. The input of this program contains three records, and these data constrain the values that the *Mapper* parameters can take because the maximum number of *Mapper* tasks is three (one *Mapper* per each <key, value> pair). The first *scenario* is generated with one *Mapper*, one *Combiner* and one *Reducer*. For the second *scenario*

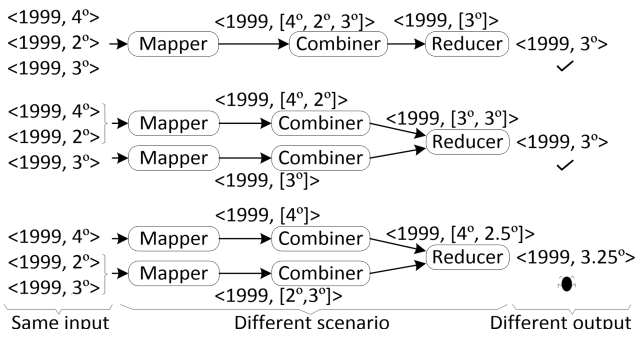


Figure. 3 Different infrastructure configurations for a program that calculates the average temperature per year with Combiner task

the parameter "Number of *Mapper* tasks" is modified to 2, where the first *Mapper* analyses two <key, value> pairs, and the second processes one pair. The third *scenario* maintains the parameter "Number of *Mapper* tasks" at 2, but modifies the parameter "Inputs processed per each *Mapper*", so the first *Mapper* analyses one <key, value> pair and the other *Mapper* processes two pairs. The *scenarios* are generated by the modification of the values in the parameters in this way and considering the constraints

4.2. Execution of the test scenarios

The testing technique proposed in the previous section is focused in the generation of *scenarios* that represent different infrastructure configurations according to the characteristics of the *MapReduce* processing. The test cases are systematically executed in these scenarios according to the framework described in Figure. 2.

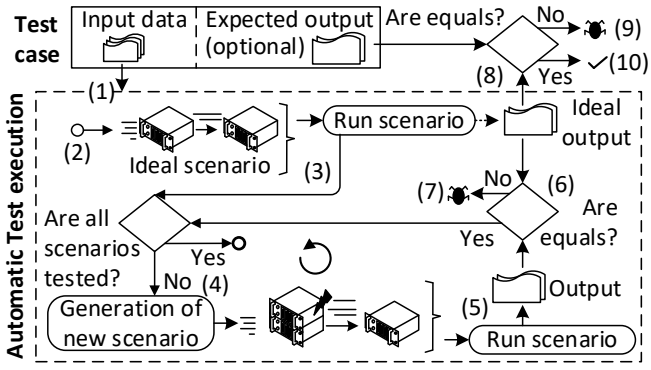


Figure. 2 General famework of test execution

The framework takes as input a test case that contains the input data and optionally the expected output. The test input data can be obtained with a generic testing technique or one specifically designed for *MapReduce*, such as MRFlow [32]. Then, the *ideal scenario* is generated (1) and executed (2, 3). This is the *scenario* formed by one *Mapper*, one *Combiner* and one *Reducer* which is the usual configuration executed in testing. Next, new *scenarios* are iteratively generated (4) and executed (5) through the technique of the previous section. The output of each *scenario* is checked against the output of the *ideal scenario* (6), revealing a fault if the outputs are not equivalent (7). Finally, if the test case contains the expected output, the output of *ideal scenario* is also checked against the expected output (8), detecting a fault when both are not equivalent (9, 10).

Given a test case, the *scenarios* are generated according to the previous section, then they are iteratively executed and evaluated following the following pseudocode:

```
Input: Test case with:
   input data
   expected output (optional)
Output: scenario that reveals a fault
(0)  /* Generation of scenarios (section 4.1)*/
(1)  Scenarios ← Generate scenarios from input data
(2)  /* Execution of scenarios */
(3)  ideal scenario output ← Execution of ideal
                                              scenario
(4)  ∀ scenario ∈ Scenarios:
(5)    scenario output ← Execution of scenario
(6)    IF scenario output <> ideal scenario output:
(7)      RETURN scenario with fault
(8)  IF ideal scenario output <> expected output:
(9)    RETURN ideal scenario
(10) ELSE:
(11)   RETURN Zero faults detected
```

For example, Figure. 2 contains the generation and execution of a program that calculates the average temperature per year in three *scenarios* considering the same test input: year 1999 with temperatures 4º, 2º and 3º. The first execution is the *ideal scenario* that produces 3º as output through one *Mapper*, one *Combiner* and one *Reducer*. Then the second *scenario* that contains two Mappers and two Combiners is executed and also produces 3º. Finally, a third *scenario* with two *Mappers* and two *Combiners* is executed, but with different information in the *Mappers* than the second *scenario*, and produces 3.25º as output. This temperature is not equivalent to the 3º of the *ideal scenario* output. Consequently, a functional fault is revealed without any knowledge of the expected output of the test case.

This approach is automatized by means of a test engine based on MRUnit library [13]. This library is used to support the execution of each *scenario*. In MRUnit the test cases are executed in the *ideal scenario*, but this library is extended to generate other *scenarios* and enable parallelism. In order to support the execution of several *Mapper*, *Combiner* and *Reducer* tasks, MRUnit is extended providing support for advanced functionalities as for example customized *Partitioners*.

5. Case Studies

The following two real world programs are used as case studies in order to evaluate the proposed approach: (1) the Open Ankus recommendation system [41], and (2) the *MapReduce* program described in I8K|DQ-BigData framework [42]. Each case study is detailed in the below sections.

5.1. Open Ankus recommendation system

This recommendation system is part of a machine learning library implemented in the *MapReduce* paradigm. The system predicts and recommends several items (books, films or others) to each user based on the personal tastes saved in the profile. One functionality checks the accuracy of the recommendations based on the points predicted by the systems against points assigned by the users for each item. This functionality has a MultipeInputs [43] design that consists in two different *Mapper* implementations: one receives the points predicted by the system and the other the points assigned by the user, but both *Mappers* emit data to the same *Reducer* implementation. The *Mappers* tasks receive from all users and all items the points predicted and the points assigned, and then the *Mapper* aggregates these points for each user-item pair. The

Reducer tasks receive for each user-item all points predicted by the system and all points assigned by the user, and then calculate the accuracy of the predictions.

For this program a test case is obtained using the MRFlow testing technique based on data flow adaptation to the *MapReduce* programs [7]. The test input data contain two predictions and two user assignments for one item: (1) the system predicts that Carol could assign 0 points to Don Quixote item, (2) Carol assigns 0 points to Don Quixote, (3) later the system detects a change in the Carol taste and predicts that Carol could assign 10 points to Don Quixote, and (4) Carol assigns 10 points to Don Quixote. These data are passed to the test engine saved in two files, one for predictions and other for assignments. The expected output is 100% of accuracy in the predictions.

The procedure described in Section 4 is applied on the previous program using the previous test case as input. As a result, a fault is detected and causes a failure when some inputs are processed before others. In this case, the program should check the points assigned by the user against their predictions, but could check it against other predictions and then a wrong accuracy could be obtained. In the previous test case, a failure occurs when the system checks the prediction 0 against the 10 points assigned by the user instead to check against the 0 points assigned by the user. The bottom of Figure. 4 represents one *scenario* that reveals the failure. This *scenario* starts with one *Mapper* to analyse the predictions and other two *Mappers* for the points assigned, 0 and 10 respectively. In this *scenario* the 10 points assigned by the user are analysed before the 0 points also assigned by the user. The *Reducer* task receives several points predicted and assigned by the user, and then checks the first prediction against the first points assigned by the user, and so on. In this *scenario* the *Reducer* task receives: (1) the predictions 0 and 10 points, and (2) the points assigned by the user, 10 and 0. This *Reducer* task generates a wrong accuracy because checks the 0 points predicted against the 10 points assigned by the user instead the 0 points. As output, the system did not predict well, but given the test input data the system should have predicted perfectly.

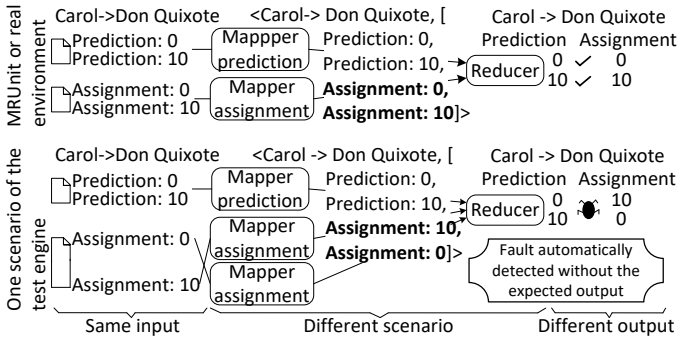


Figure. 4 Execution of the Open Ankus test case in different scenarios

The testing technique proposed in this paper detects the fault with the previous test case. However, the following test environments do not detect the fault: (a) *Hadoop* cluster in production with 4 computers, *Hadoop* in local mode (simple version of *Hadoop* with one computer), and (c) MRUnit unit testing library. These environments mask the fault because only execute the test cases in one *scenario* represented in the top of Figure. 4. This *scenario* is the *ideal scenario* with only two *Mappers* due a MultipleInputs design of the program: one for predictions and other for points assigned by user. In the *ideal scenario* there is no parallelization for predictions and points assigned by user, then the fault is masked.

The test engine proposed in this paper detects the fault because executes the test case in several *scenarios* that could happen in production. In contrast with the other environments, this test engine does not need the expected output to reveal the fault. At first point, the test engine obtains the output from the *ideal scenario* and then checks if the other *scenario* produces an equivalent output or not. For example, in the previous test case, for the same input some *scenarios* produce one output (the system predictions are perfect) and other *scenarios* produce different output (the system predictions are wrong).

5.2. 18K|DQ-BigData framework

This program measures the quality of the data exchanged between organizations according to part 140 of the ISO/TS 8000 [44]. The program receives (1) the data exchanged in a row-column fashion, together with (2) a set of mandatory columns that should contain data and (3) a percentage threshold that divides the data quality of each row in two parts: the first part is maximum if all mandatory columns contain data and zero otherwise, and the second part of the data quality is calculated as the percentage of the non-mandatory columns that contain data. The output of the program is the data quality of each row, and the average of all rows.

Over the previous program, a test case is obtained using again a specific *MapReduce* testing technique based on data flow [7]. The test input data and the expected output of the test case contain two rows represented in Table. 1. Row 1 contains two columns (Name and City), and only one column has data, so the data quality is 50%. Row 2 contains data in all columns, so the data quality is 100%. The total quality is 75%, which is the average of both rows.

<table>
<tr><th colspan="2">Input</th><th colspan="2" rowspan="3">Expected output</th></tr>
<tr><td colspan="2">Data quality threshold: 50%</td></tr>
<tr><td colspan="2">Mandatory columns: “Name”</td></tr>
<tr><td rowspan="2">Row 1</td><td>Name: Alice</td><td rowspan="2">50%</td><td rowspan="3">75% (Average)</td></tr>
<tr><td>City: (no data)</td></tr>
<tr><td rowspan="2">Row 2</td><td>Name: Bob</td><td rowspan="2">100%</td></tr>
<tr><td>City: Vienna</td></tr>
</table>

Table. 1 TEST CASE OF THE I8K|DQ-BIGDATA PROGRAM

The procedure described in Section 4 is applied on the previous program using the previous test case as input. As a result, a fault is detected and reported to the developer. This failure occurs when the rows are processed in different *Mappers* and only the first *Mapper* receives the information related to the mandatory columns and the data quality threshold, because *Hadoop* splits the input data into several subsets. Without this information, the *Mapper* cannot calculate the data quality and does not emit any output. The bottom of Figure. 5 represents the *scenario* that produces the failure. There are two *Mappers* that process different rows. The first *Mapper* receives the data quality threshold (value of 50%), the mandatory column (“Name”) and the two columns of row 1 with only data in one column, so the *Mapper* emits 50% as data quality of row 1. The second *Mapper* processes only row 2, but no other information about the mandatory columns or data quality threshold, so this *Mapper* cannot emit any output. Then the *Reducer* receives only the data quality of row 1 and emits an incorrect output of the average data quality.

This fault is difficult to detect because it implies the parallel and controlled execution of the program. Moreover, this fault is not revealed by the execution of the test case in the following environments: (a) *Hadoop* cluster in production with 4 computers, *Hadoop* in local mode (simple version of *Hadoop* with one computer), and (c) MRUnit unit testing library. These environments do not detect the fault because they only execute one *scenario* that masks the fault. Normally these environments run the program in the *ideal scenario* that is formed by one *Mapper*, one *Combiner* and one *Reducer*, and then the fault is masked due to a lack of parallelism.

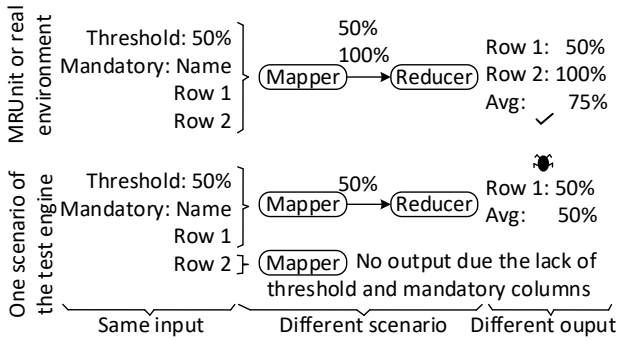


Figure. 5 Execution of the 18K|DQ-BigData test case in different scenarios

The test engine proposed in this paper executes the test case in the different *scenarios* that can occur in production with large data and infrastructure failures. In contrast with the other environments, the test engine proposed does not need the expected output to detect faults. For example, in this case study the fault is revealed automatically because the outputs of the different *scenarios* are not equivalent to each other. The execution of some *scenarios* obtains an average quality of 75%, whereas the execution of other *scenarios* obtains 50%. These outputs are not equivalent, and the test engine detects automatically a fault despite the unknown expected output.

After the detection and report of the fault during the test phase, the developer fixed the program and then the test case passed.

6. Conclusions

A testing technique for *MapReduce* applications is described and automatized as a test engine that generates and executes different infrastructure configurations for a given test case. This test engine can detect automatically functional faults related to the *MapReduce* paradigm without the expected output. In general, these design faults are difficult to detect in test/production environments because the execution is performed without parallelization or infrastructure failures. This testing technique is applied in test cases with little data in two real world programs. As a result, functional faults are revealed automatically.

As future work the generation of the infrastructure configurations could be improved by the extension of the testing technique in order to select efficiently the configurations that are more likely to detect faults. The current approach is *off-line* because the tests are not carried out when the program is in production. As future work we plan to extend the approach to *on-line* testing, in order to monitor the functionality with the real data when the program is executed in production and detect the faults automatically.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was supported in part by PERTEST (TIN2013-46928-C3-1-R), project funded by the Spanish Ministry of Science and Technology; TESTEAMOS (TIN2016-76956-C3-1-R) and SEQUOIA (TIN2015-63502-C3-1-R), projects funded by the Spanish Ministry of Economy and Competitiveness; GRUPIN14-007, funded by the Principality of Asturias (Spain); CIEN LPS-BIGGER project; and ERDF funds.

References

[1] ISO/IEC JTC 1 – Big Data, preliminary report 2014, ISO/IEC Std., 2015.

[2] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," in Proc. of the OSDI - Symp. on Operating Systems Design and Implementation. USENIX, 2004, pp. 137–149.

[3] Apache Hadoop: open-source software for reliable, scalable, distributed computing, https://hadoop.apache.org, accessed: 2017-01-16.

[4] Institutions that are using Apache Hadoop for educational or production uses, http://wiki.apache.org/hadoop/PoweredBy, accessed: 2017-01-16.

[5] Apache Spark: a fast and general engine for large-scale data processing, https://spark.apache.org, accessed: 2017-01-16.

[6] Apache Flink: Scalable batch and stream data processing, https://-flink.apache.org/, accessed: 2017-01-16.

[7] J. Morán, C. de la Riva, and J. Tuya, "MRTree: Functional Testing Based on MapReduce's Execution Behaviour," in Future Internet of Things and Cloud (FiCloud), 2014 International Conference on, 2014, pp. 379–384.

[8] K. V. Vishwanath and N. Nagappan, "Characterizing cloud computing hardware reliability," in Proceedings of the 1st ACM symposium on Cloud computing. ACM, 2010, pp. 193–204.

[9] AnarchyApe: Fault injection tool for hadoop cluster from yahoo anarchyape, https://github.com/david78k/anarchyape, accessed: 2017-01-16.

[10] Chaos Monkey, https://github.com/Netflix/SimianArmy/wiki/Chaos-Monkey, accessed: 2017-01-16.

[11] Hadoop injection framework, https://hadoop.apache.org, accessed: 2017-01-16.

[12] J. Moran, B. Rivas, C. De La Riva, J. Tuya, I. Caballero, and M. Serrano, "Infrastructure-aware functional testing of mapreduce programs," 2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW), Aug 2016. [Online]. Available: http://dx.doi.org/-10.1109/W-FiCloud.2016.45

[13] Apache MRUnit: Java library that helps developers unit test Apache Hadoop map reduce jobs, http://mrunit.apache.org, accessed: 2017-01-16.

[14] S. Nachiyappan and S. Justus, "Getting ready for bigdata testing: A practitioner's perception," in Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on. IEEE, 2013, pp. 1–5.

[15] A. Mittal, "Trustworthiness of big data," International Journal of Computer Applications, vol. 80, no. 9, 2013.

[16] A. Bertolino, "Software testing research: Achievements, challenges, dreams," in Future of Software Engineering, 2007. FOSE '07, 2007, pp. 85–103.

[17] L. C. Camargo and S. R. Vergilio, "Mapreduce program testing: a systematic mapping study," in Chilean Computer Science Society (SCCC), 32nd International Conference of the Computation, 2013.

[18] M. Sharma, N. Hasteer, A. Tuli, and A. Bansal, "Investigating the inclinations of research and practices in hadoop: A systematic review," confluence The Next Generation Information Technology Summit (Confluence), 2014 5th International Conference -.

[19] J. Merino, I. Caballero, B. Rivas, M. Serrano, and M. Piattini, "A data quality in use model for big data," Future Generation Computer Systems, 2015.

[20] S. Kavulya, J. Tan, R. Gandhi, and P. Narasimhan, "An analysis of traces from a production mapreduce cluster," in Cluster, Cloud and Grid Computing (CCGrid), 2010 10th IEEE/ACM International Conference on. IEEE, 2010, pp. 94–103.

[21] K. Ren, Y. Kwon, M. Balazinska, and B. Howe, "Hadoop's adolescence: an analysis of hadoop usage in scientific workloads," Proceedings of the VLDB Endowment, vol. 6, no. 10, pp. 853–864, 2013.

[22] M. Ishii, J. Han, and H. Makino, "Design and performance evaluation for hadoop clusters on virtualized environment," in Information Networking (ICOIN), 2013 International Conference on, 2013, pp. 244–249. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6496384

[23] Z. Liu, "Research of performance test technology for big data applications," in Information and Automation (ICIA), 2014 IEEE International Conference on. IEEE, 2014, pp. 53–58.

[24] G. Song, Z. Meng, F. Huet, F. Magoules, L. Yu, and X. Lin, "A hadoop mapreduce performance prediction method," in High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC_EUC), 2013 IEEE 10th International Conference on, 2013, pp. 820–825. [Online]. Available: http://-ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6832000

[25] M. Gudipati, S. Rao, N. D. Mohan, and N. K. Gajja, "Big data: Testing approach to overcome quality challenges," Big Data: Challenges and Opportunities, pp. 65–72, 2013.

[26] L. C. Camargo and S. R. Vergilio, "Cassicação de defeitos para programas mapreduce: resultados de um estudo empírico," in SAST - 7th Brazilian Workshop on Systematic and Automated Software Testing, 2013.

[27] C. Csallner, L. Fegaras, and C. Li, "New ideas track: testing mapreduce-style programs," in Proceedings of the 19th ACM SIGSOFT symposium and the 13th European conference on Foundations of software engineering. ACM, 2011, pp. 504–507.

[28] Y.-F. Chen, C.-D. Hong, N. Sinha, and B.-Y. Wang, "Commutativity of reducers," in Tools and Algorithms for the Construction and Analysis of Systems. Springer, 2015, pp. 131–146.

[29] F. Faghri, S. Bazarbayev, M. Overholt, R. Farivar, R. H. Campbell, and W. H. Sanders, "Failure scenario as a service (fsaas) for hadoop clusters," in Proceedings of the Workshop on Secure and Dependable Middleware for Cloud Monitoring and Management. ACM, 2012, p. 5.

[30] P. Joshi, H. S. Gunawi, and K. Sen, "Prefail: A programmable tool for multiple-failure injection," in ACM SIGPLAN Notices, vol. 46, no. 10. ACM, 2011, pp. 171–188.

[31] J. E. Marynowski, A. O. Santin, and A. R. Pimentel, "Method for testing the fault tolerance of mapreduce frameworks," Computer Networks, vol. 86, pp. 1–13, 2015.

[32] J. Morán, C. de la Riva, and J. Tuya, "Testing Data Transformations in MapReduce Programs," in Proceedings of the 6th International Workshop on Automating Test Case Design, Selection and Evaluation, ser. A-TEST 2015. New York, NY, USA: ACM, 2015, pp. 20–25.

[33] A. J. Mattos, "Test data generation for testing mapreduce systems," in Master's degree dissertation, 2011.

[34] N. Li, Y. Lei, H. R. Khan, J. Liu, and Y. Guo, "Applying combinatorial test data generation to big data applications," Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering - ASE 2016, 2016. [Online]. Available: http://dx.doi.org/10.1145/2970276.2970325

[35] Herriot: Large-scale automated test framework, https://wiki.apache.org/-hadoop/HowToUseSystemTestFramework, accessed: 2017-01-16.

[36] Minicluster: Apache Hadoop cluster in memory for testing, https://-hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/-CLIMiniCluster.html, accessed: 2017-01-16.

[37] JUnit: a simple framework to write repeatable tests, http://junit.org/, accessed: 2017-01-16.

[38] Mockito: Tasty mocking framework for unit tests in java, http://mockito.org/, accessed: 2017-01-16.

[39] D. Qiu, B. Li, and H. Leung, "Understanding the api usage in java," Information and Software Technology, vol. 73, pp. 81–100, 2016.

[40] M. Grindal, J. Offutt, and S. F. Andler, "Combination testing strategies: a survey," Software Testing, Verification and Reliability, vol. 15, no. 3, pp. 167–199, 2005.

[41] Open Ankus: Data mining and machine learning based on mapreduce, http://-www.openankus.org/, accessed: 2017-01-16.

[42] B. Rivas, J. Merino, M. Serrano, I. Caballero, and M. Piattini, "I8k| dq-bigdata: I8k architecture extension for data quality in big data," in Advances in Conceptual Modeling. Springer, 2015, pp. 164–172.

[43] Multipleinputs: library to support mapreduce jobs that have multiple input paths with a different inputformat and mapper for each path, https://-hadoop.apache.org/docs/current/api/org/apache/hadoop/mapred/lib/-MultipleInputs.html, accessed: 2017-01-16.

[44] ISO/TS 8000-140, Data quality - Part 140: Master data: Exchange of characteristic data: Completeness, ISO/TS Std., 2009.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 97-107 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTES Journal

ISSN: 2415-6698

Cross layers security approach via an implementation of data privacy and by authentication mechanismfor mobile WSNs

Imen Bouabidi[1], Pr. Mahmoud Abdellaoui[*,2]

[1]WIMCS-ENET'COM, ENIS, Sfax University, Sfax - Tunisia,

[2]WIMCS-ENET'COM, Sfax University, Sfax - Tunisia,

ARTICLE INFO

Article history:
Received: 20 December, 2016
Accepted: 21 January, 2017
Online: 28 January, 2017

Keywords:
Mobile WSNs
Smart Sensors
MAC protocols
Data privacy
Malware and viruses detection
Authentication
Cross layers
Jamming attacks
OMNet++ simulator

ABSTRACT

To implement a new secure network with high mobility and low energy consumption, we use smart sensors. These sensors are powered by micro batteries generally non rechargeable. So, to extend their lifetime, it is necessary to implement new energy conservation techniques. Existing works separate the two features (security, energy conservation) and are interested specifically in only one layer.Consequently, the originality of this work consists to combine together the two features using a crossing between three layers: physical layer, data link layer and network layer. Our proposition consists firstly in developing a new network deployment in hierarchical areas. This model takes place at the network layer. Secondly, implementing an energy efficient and secure MAC protocol providing a secure authentication, data privacy and integrity in a mobile WSN. Finally, implementing an intrusion detection system protecting the physical layer from malware and viruses that threaten it. We have been used OMNet++ for simulation .Our proposed protocol SXMachiavel offered the best performances and more reliability at the mobility rate (can reach 99% compared with XMachiavel, which doesn't exceed 35%), loss packets rate (0.05% for a small network size) and energy consumption (decreases by 0.01% for each exchanged packet).

1. Introduction

To answer the specific questions proposed by academics-researchers and industrialists working in the wireless sensors network field and to have a mobile-secure and efficient WSN, our work is particularly concerned the security and the energy conservation of this network with mobile sensors [1-2].Indeed, we have implemented a cross layers security solution which provides data privacy and authentication for mobile WSN. We define WSN as a special type of ad hoc network, which usually consists of a number of nodes randomly deployed in an area to supervise or monitor diverse phenomena. The main difference between traditional networks and wireless sensors network is the scalability factor. In fact, WSN have thousands of nodes (more than 125000 nodes). This particular feature offers the possibility of an infinity path's number, which solve the network congestion problem and breaking links.

In the literature and up to now, the existing security solutions aren't adopted to the wireless sensors network constraints. For this purpose, security and energy conservationmechanisms must be added to protect the network against intrusions and virus that threaten it. In this context, we propose solutions specific to WSN. These solutions are applied to mobile WSN. As a result, we focus in this work on the major issue of security in wireless sensors network such as: authentication, privacy, data integrity, virus and intrusion detection while maintaining minimal energy consumption. Recall that our work has as objectives: firstly, the organization of network with low energy consumption and high mobility. Secondly, the introduction of security techniques to protect WSN. Finally, the use of smart sensors [3]. These smart sensors stand out over traditional sensors by digital processing thanks to the addition of processor which simplifies network management. All nodes share the same properties; they are autonomous and intelligent sensors with significant storage and computing capacity over traditional wireless sensors.

*Corresponding Author: Pr. Abdellaoui Mahmoud, WIMCS-ENET'COM, (+216) 50429427 & mahmoudabdellaoui4@gmail.com

Existing works in security and energy conservation are separated. They are interested only in the security theme or in the energy conservation theme without making the fusion between the two concepts. There are superficial solutions that aren't explored. Also, the majority of these works exploit the single-layer strategy and each solution deals with a particular layer and forgets the residual of the protocol stack. These works are limited on particular attacks which is insufficient to protect the entire WSN. To surmount these limitations, our solution is installed. It consists in merging both concepts security and energy conservation together in a crossing between three layers: physical layer, data link layer and network layer. This fusion doesn't exist in the existing works which separate the two issues. Indeed, the cross-layers design represents an interesting solution to address the security problem while guaranteeing a low rate of resource consumption.

To provide a complete security solution, our work is divided into three levels. A first level already made and published deals with network layer [4-6]. It consists of a network deployment in hierarchical areas. A second level is interested in the data link layer and especially in MAC sublayer through the development of an energy efficient and secure MAC protocol called SXMachiavel[7]. A last level concentrates on the physical layer by the implementation of an intrusion detection system at the sink level to protect the physical layer against Jamming attacks.

2. Energy conservation and security tools

Security and energy conservation are two essential elements in the wireless sensors network design. Therefore, our main objective consists in combining these two parameters to develop a generic and efficient sensors network. In the literature, these two themes are separated. Existing research works are focused either on the first theme or on the other theme. Consequently, the originality of our work consists in using these two parameters (energy conservation and security) at the same time in order to manufacture a complete security solution that meets the WSN requirements. Our solution is based on the crossing between the three layers:physical layer, data link layer and network layer.Up to now, these two fields are separated. Researchers are interested either in the security theme or in the energy conservation theme without making the fusion between the two. In fact, MAC protocols are divided into four types. Protocols basing on TDMA access: there are classic algorithms of slots reservation. These approaches seem to be complex, and present problems with the scalability factor like TRAMA protocol [8]. Then, protocols using CSMA access mode: these protocols, which use the contention period, are the most popular and represent the majority of MAC protocols proposed for wireless sensors network. But, they suffer from the latency when nodes are in sleeping mode; nodes have to wait until the receiver wakes up before they can forward a packet. SMAC, TMAC [9] and BMAC [10] are examples of this category. The third group is hybrid protocols: in spite of they try to combine points of TDMA and CSMA based protocols, these techniques seem to be complex such as ZMAC protocol [11]. The last type is inter-layers protocols. In order to minimize more the energy consumption, some researchers turn to the crossing between layers strategies and more exactly between MAC and network layers. One of the first inter-layers protocols is MAC CROSS [12]. These approaches are implemented with static sensors networks and they require the information of routing path from upper layer to determine the next hop. Afterward, nodes that aren't on the routing path go to sleep.

Turn now to present some existingsecurity mechanisms. The inherent characteristics of WSN, including wireless communication, random deployment and limited resources make it vulnerable against number of attacks. It is extremely easy for an intrusion to usurp the traffic circulating on the network. So, it comes necessary to protect the network. In the literature, some works were tried to solve security problems in the WSN. Generally, existing works were interested in saving energy, but they didn't address security of MAC protocols; some of them are concentrated in security against specific attacks such as a protection against Jamming attack at the physical layer level [13], or Wormhole attack in network layer [14]. On the other hand, some existing works take care well of network layer. Indeed, this layer is the module responsible for forwarding correctly a data from a point of network to another one. Thus, several authors go to the security of the routing path. Such as SecLEACH and SecPEGASIS protocols, which present the first solutions for securing hierarchical (cluster-based) networks with the dynamic clusters formation [15-16], also inwork [17] authors present a cross layers routing protocol. With the aim of providing an acceptable level of security, other researchers have developed either protocols assuring data confidentiality, such as DiDrip algorithm [18], techniques guaranteeing authentication and data integrity [19], or techniques providing secure data aggregation [20].Another research line turns to the cryptography aspects. The key management is one of these solutions. For such a system works and be secured each user must have a set of secret keys. In work [21], authors have implemented the encryption algorithm AES to protect data privacy in WSN. Other research theme is the intrusion detection system. For example, work [22] showed an implementation of IDS to secure the WSN. A final research line concerns the cross-layers strategies. The design cross layers represents an interesting solution to remedy the security problems while guaranteeing a low energy consumption. Work [23] explored the benefits of cross layers approach to overcome the single-layer protocols.

Our literature review of existing security and energy conservation techniques for WSN concludes with a comparison of the characteristics of each existing solution. This comparison shows that security is a complex problem. Each approach is characterized by its own needs and constraints that are required to achieve the desired security level. Indeed, existing security solutions are expensive in energy, memory space and don't offer the possibility of discovering new attacks. Moreover, the majority of works are based on the assumption that the WSN is static (all the used nodes are fixed). To remedy these limitations and to

reinforce the security level in the WSN some of them can be corrected and improved in order to manufacture a complete solution. It is *the cross layers approach*. At this stage, our solution settles down. It consists of a crossing between the three layers: network layer, data link layer and physical layer to provide a mobile and secure network. A detailed description of our contribution is explained in the next section.

3. Proposed Approach

As we have already said, our work deals with the security and energy conservation problems in the wireless sensors network. In the literature, *few works* are interested in the security aspect. According to our studies; we can notice that existent works in security and energy conservation terms present some drawbacks as following as: existing solutions are expensive in energy, in memory space and didn't offer the possibility of discovering new attacks. To overcome these limits our solution settles down. It consists of a crossing between three layers to provide a secure WSN. It has as objectives: firstly, the organization of a WSN with high mobility and low energy consumption. Secondly, the introduction of security techniques to protect the WSN. These security mechanisms should take in consideration the wireless sensors network constraints and limitations (limited memory capacity and computing). Finally, the use of intelligent sensors. Our contribution is based on three steps: the first step is interested in the network layer by developing a hierarchical network deployment that meets the requirements of this network. The second step is concentrated in the MAC sublayer through the development of an energy efficient and secure MAC protocol. The third step deals with the physical layer by implementing an intrusion detection system in order to protect physical layer against Jamming attacks.

3.1. Network deployment in hierarchical areas

Recall that the first step of our work deals with the network layer by constructing a new model specific and adapted to the WSN. In fact, the network layer is the module responsible for routing information to the right destination through a given connection network. Like ad hoc networks, wireless sensors networks are characterized by the absence of pre-existing fixed infrastructure. So, to ensure network connectivity, each node must participate in the routing, enabling it to discover existing paths and reach the other nodes of the network.

Our network model consists of a sensors network deployment in hierarchical areas. Indeed, the network structuring is one of the primordial tools to design the energy in each node which results in its lifetime prolongation. Hierarchical organization is a technique that consists in partitioning the network into subsets to facilitate its management and especially the routing which is realized at several levels. The literature has several contributions in the hierarchization techniques that can be classified into two types: areas and clusters. Clusters are defined as a set of nodes that have a node called a Cluster-Head (CH). This node acts as a relay between nodes in the same cluster and the base station or gateway. Generally, this node possesses higher energy resources compared to the other nodes. Clusters models suffer from some weaknesses. Indeed, the main disadvantage of the clustering approach is that it is mainly based on CH. So, if this node goes down all the approach becomes unuseless. To remedy this problem, the second approach appears. It is based on network partitioning into areas. In this context, we proposed a new network deployment in hierarchical areas. In our network model, areas are defined according to it carried radio which is equal to 400m and that can reach 8km by using plugging (SKY 65336, SKY 65337) and according to the number of jumps. This model presents some advantages: firstly, during the deployment phase all nodes have the same role and the same energy level. Then after, knots having important resources execute more complex tasks and knots having limited resources execute simple tasks. Also, this network model presents a large cover and high connexion and can improved reliability [4-6].

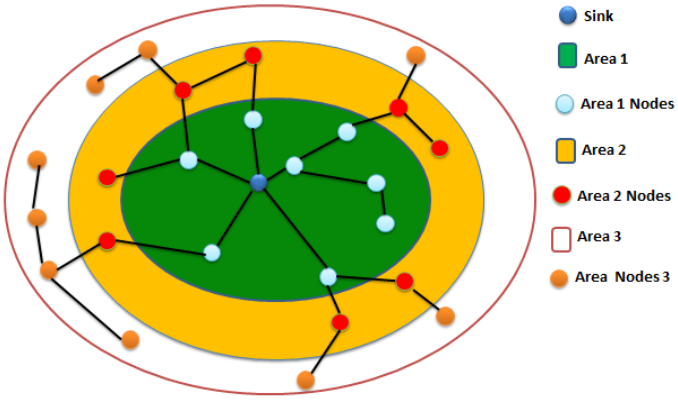


Figure1. Network topology

3.2. Secure Energy Efficient MAC protocol for mobile WSN

The data link layer and especially the MAC sublayer is mean subject of this step. The Media Access Control serves as an interface between software part that controls the link between nodes and physical layer. It controls the access management of multiple stations in a shared medium. In fact, each station listens to the medium before transmitting. If it is free, then the knot can send his data. If not, this knot either it passes in sleeping mode, or it waits its role. In absence of prevention mechanisms, this medium will be vulnerable against attacks and more exactly denial of sleep attacks. The adequate solution to fight against these attacks is the duty cycling technique. Among the existing energy conservation solutions, we choose the XMachiavel protocol. According to studies which were made, this protocol is the least energy consuming.We implemented this protocol in the existing MAC sublayer. Also, we developed an intrusion detection system in physical layer.These two methods increase network performances in security term. To validate our solution we used the OMNet++ simulator. Our work consists in implementing and adapting the XMachiavel protocol in our simulator. We supplied more details in the next section.

3.2.1. XMachiavel Protocol

Recall to modify and to ameliorate the XMachiavel protocol whose objective is to acquire a mobile and secure wireless sensors network. Consequently, we have chosen to use the existent MAC this protocol. According to the studies made and validated by simulations, this protocol is the most economical in energy. We implemented this protocol in the existing MAC sublayer using the OMNet++ simulator. OMNet++ doesn't support this protocol, so we have implemented and adapted the XMachiavel to our simulator. Figure 2 presents a description of its header.

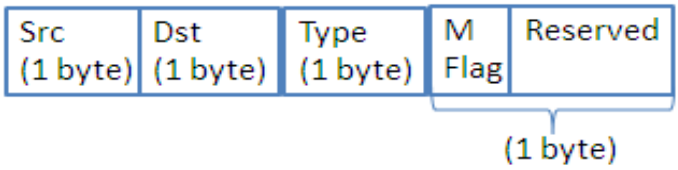


Figure 2. XMachiavel header [24]

- Src =Source node identifier,
- Dst = destination node identifier,
- Type: indicate whether a packet is a preamble, ACK or data,
- M Flag: used by the mobile node when it sends a data packet.

The header and preambles given in Figures 2, 4 and 5 improve the processing time; we present now a theoretical explanation of transmission and propagation time in a preamble MAC protocols

To include the effect of the preamble sampling technique, we present in this stage an adaptation and extension of classical CSMA analysis.A unicast DATA messagewill be preceded by a preamble of length T_P, and followed by an ACK message, as shown in Figure 3. If the ACK is not received, the message is repeated at a latertime [25].

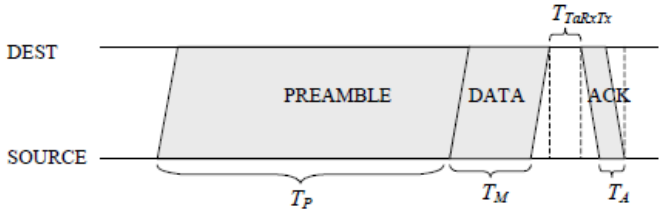


Figure 3. Data - ACK transaction using a wake-up preamble [25]

Transmission attempts at each node follow a random Poisson process with a mean rate g (mean inter arrival 1 / g). Therefore, the number of total transmission attempts will be the sum of the attempt of each node and correspond to a Poisson process of rate (N + 1) g, where N is the number of neighbors within a node. We will compute the throughput, the delay and the power consumption of the protocol in function of the mean rate g of the transmission attempts [25].

To calculate the power consumption of our protocol, wemust count the power consumption due to preamble sampling, the power consumption of listening to the medium when it is busy and the power consumption of transmitting [25].

Note that each node at a time t can take one of the following states: sleep state, channel listening state or transmission / reception state. The idle period starts at the end of the transmission of a packet and ends at the start of the next transmission. The duration of the idle period is a random variable $I = X + T_{TaRxTx}$, where X is the random time between the

end of the last transmission and the next arrival. Its cumulative distribution is [25]:

$$P(X <= x) = 1 - P(X > x) = 1 - e^{-(N+1)gx} \quad (1)$$

Giving a mean of

$$E[X] = 1/(N+1)g \quad (2)$$

Therefore, the mean duration of an idle period is

$$E[I] = 1/(N+1)g + T_{TaRxTx} \quad (3)$$

Turn now to the second state: the transmission/reception.This state is called a busy period. This period starts when the transmission starts and ends when the last interfering packet ends. Its duration is a random variable $_M B = Y + T$, where Y is the random time between the start of the first packet and the start of the last interfering packet.The mean busy period duration is presented by the following formulation [25]:

$$E[B] = T_M + T_P + T_{TaRxTx} - (1 - e^{-(NgTTaRxTx)})/Ng \quad (4)$$

Note that the proportion of the time when the medium is busy with non-persistent CSMA with preamble sampling is hence

$$b = E[B] / E[I] + E[B] \quad (5)$$

To simplify computations, we assume that a node will not stop listening as soon as the medium become idle, but will continue to listen until the next preamble sampling time. The proportion of listening periods will be equal to the proportion of busy medium given in formula (5).

Note that the processing time and the transmission time depend on the length of the frame. The longer the frame, the longer the processing and transmission times. According to Figure 3, we notice that the preamble length T_P is greater than the ACK length T_A .Consequently, the processing time of a preamble packet is more important than the processing time of an ACK packet.

However, XMachiavel protocol presents two weak points which are: first of all, it supports a low mobility. Then, it isn't secure: packets are sent unencrypted over the network. So, they can be easily intercepted by intruders. Also, it doesn't support any authentication, data privacy and integrity or intrusion detection mechanisms [26]. So, we are supposed to secure the packets exchange by introducing: firstly, an authentication mechanism. Secondly, a hash function guaranties the data integrity and finally, encryption process providing data privacy in the mobile WSN while increasing the mobility rate**.** A description of our improvement is detailed in the next section.

3.2.2 SXMachiavel protocol

Our solution consists in improving the existing MAC protocol XMachiavel by adding security mechanisms. A first part of our improvement is proposed at publication in [7]. It consists of:

- Increasing the mobility rate,

- Authentication mechanism by using a new field in each sent packet,
- Implementing of SHA1 as a hash function to provide the data integrity.

Firstly, we start by the authentication mechanism. It consists of adding a new field in each packet. We proceed by the following manner:

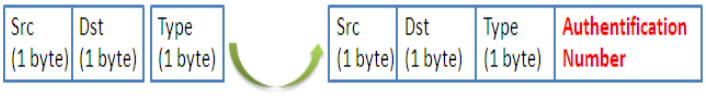


Figure 4. Preambles.

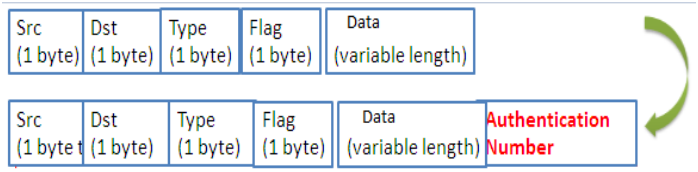


Figure5. Data Packet.

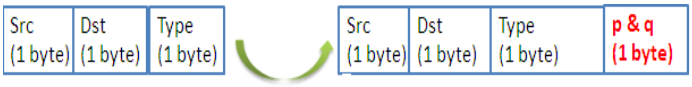


Figure 6. Acknowledgment

The authentication number is calculated as follows:

- A source node before sending a packet (preamble or data) generates two prime numbers p and q randomly [27].
- It calculates then the function f (n) = (p-1) * (q-1)
- The function f (n) corresponds to the AuthNum value (which will be added as a new field in each sent packet).
- The receiver calculates p and q from the received AuthNum value.

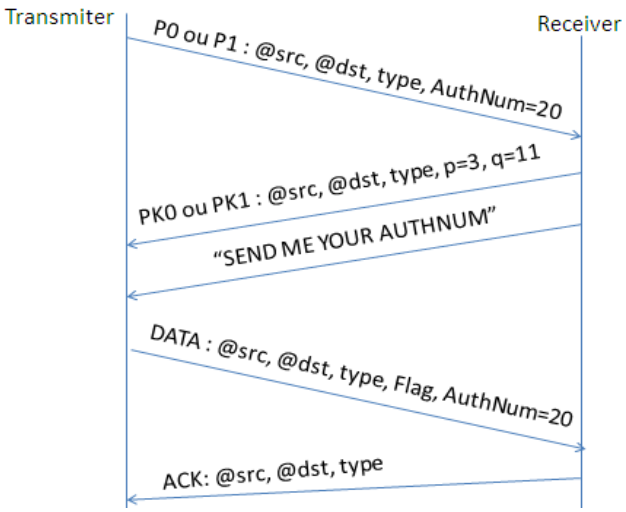


Figure 7. Sequence diagram of authentication process between two nodes

Two communicating nodes must authenticate firstly, so a sender node transmit an encrypted packet containing his identity which consist in the authentication number (AuthNum). Then, the receiver decrypts the packet and calculates the p and q values from the AuthNum value. Afterward, it sends an ACK packet and the message ‘’SEND ME YOUR AUTHNUM”. The transmitter then calculates AuthNum from the received values p and q. If the both values of AuthNum are equals, the source authenticates the destination and responds with a data packet containing the authentication number. Later, the destination node verifies the two value of AUTHNUM,if both values are equals, the sender will be authenticated. If not, the packet will be rejected.

Note that by adding authentication mechanism, the level of energy consumed in the network increases because the exchanges number between the transmitter and the receiver has increased. This increase is small compared to that obtained by the encryption mechanism. This will be approved by the simulation results presented in the next part.

With the aim of improving more the security level, the second step consists of: firstly, replacing the SHA1 hash function by a more recent and efficient hash function called SHA3 to provide the data integrity in the network [28]. Secondly, using the AES algorithm to guarantee the data privacy in mobile WSN [29]. Recall that the existent protocol XMachiavel doesn’t provide any security threats. Indeed,SHA3 is a new cryptographic hash function built on a principle completely different from SHA1 and SHA2 [28]. In previous work [7], we used SHA1. We present in the next section a comparison between SHA1 and SHA3 in term of loss packets rate and energy consumption. This function presents some security problems (collision attack). Weaknesses discovered on SHA1 raise fears of a fragility of SHA2 which is built on the same plan. SHA3 hash function uses the sponge constructionin which data is "absorbed" into the sponge and then the result is "squeezed" out. In the absorbing phase, message blocks are XORed into a subset of the state[28], which is then transformed as a whole. In the "squeeze" phase, output blocks are read from the same state subset, alternated with state transformations. The size ‘*r*’ of the state part that is written and read is called the "rate". The site ‘*c*’ of the part that is untouched by input/output is called the "capacity". The capacity determines the security of the scheme. The maximum security level is half the capacity. The hash function guarantees the data integrity. With this function, it is possible to determine the digest from the original message [28]. But, it is difficult to find another message that verifies the same function and gives the same digest. The receiver recalculates the digest and compares the two values. If they are different, the message is modified.

Turn now to the encryption method, we have used the AES algorithm. The AES (Advanced Encryption Standard) is a symmetric encryption standard to replace the Data Encryption Standard (DES), which has present limits in the face of current attacks. It operates on 128-bit blocks which it transforms into 128-bit encrypted blocks by a sequence of Nr operations or "rounds", from a 128, 192 or 256-bit key. Depending on its size, the number of rounds differs: 10, 12 and 14 rounds, respectively [29].The encryption process is described as following as:

- *BYTE_SUB (Byte Substitution)*: It is a non-linear function operating independently on each block from a so-called substitution table.
- *SHIFT_ROW:* It is a function that performs offsets (typically it takes 4-byte 4-bit input and shifts 0, 1, 2 and 3 bytes for bits 1, 2, 3 and 4 respectively) [30].
- *MIX_COL*: It is a function which transforms each input byte into a linear combination of input bytes and can be expressed mathematically by a matrix product on the Galois field (28).

The decryption process consists of applying inverse operations in reverse order and with sub keys also in reverse order.

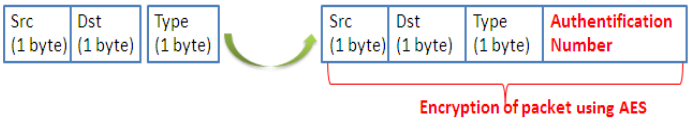


Figure 8.Encryption of Preambles using AES.

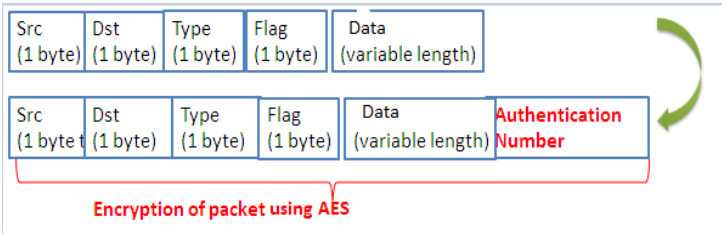


Figure 9. Encryption of data packets using AES.

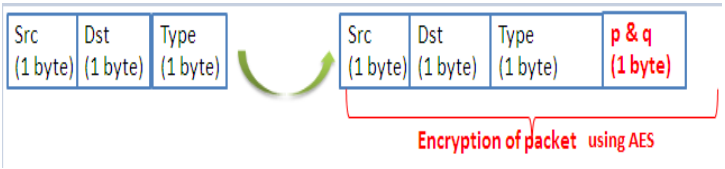


Figure 10. Encryption of acknowledgements using AES.

To secure the existing MAC protocol XMachiavel, we proceeded by the following way:

- ✓ Authentication mechanism that guaranties the authenticity of the network.
- ✓ A recent hash function called SHA3 that ensures the exchanged data integrity [31]. In our previous work, we have applied the hash function on all packets exchanged, whether preambles, ACKs or data [7]. But, according to simulations that were made, we observed the increase in the energy consumption rate. In addition, preambles and ACKs contain less important information's than data so; there is no need to apply the hash on these packets. As result, we gain at the energy consumption level.
- ✓ A symmetric encryption algorithm AES that guarantees data privacy in WSN. We apply this algorithm on all packets exchanged (preambles, ACKs and data) in order to limit the appearance of malware and viruses.

In transmission, the source node follows the following steps:

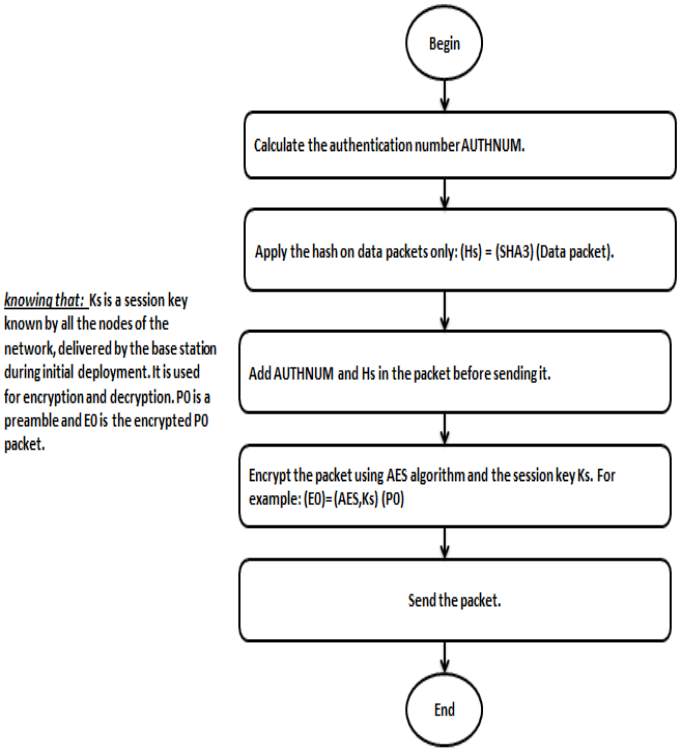


Figure 11. Security process followed by each node during the transmission

In reception, the destination node follows the following steps:

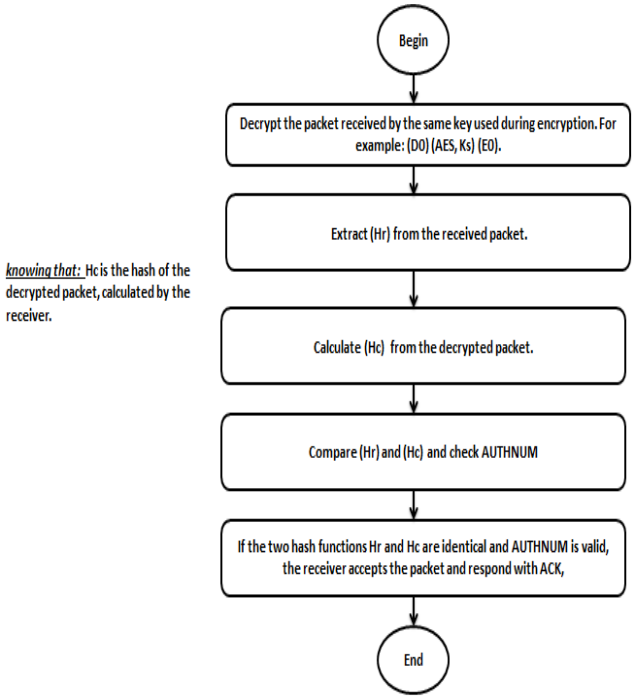


Figure 12.Security process followed by each node during the reception.

3.3. Implementation of an IDS to protect physical layer against Jamming attacks

As we have already said, our contribution consists in using a cross between the three layers: physical layer, data link layer and network layer to provide secure WSN. We have previously

specified our security solutions with regard to both data link and network layers. These approaches begin with the development of a new network deployment model in hierarchical areas. This model takes place at the network layer level and its goal is to provide an efficient and mobile WSN network. Next, we are focused on the data link layer and more specifically the MAC sublayer through the implementation of an energy-efficient MAC protocol that ensures secure authentication and data privacy in the WSN. Finally, in order to provide a complete and specific security solution to WSN, we address to protect the physical layer. The wireless sensors network is considered one of the communication technologies that have emerged over the years. Among the WSN features, we quote the wireless communication. As a result, the physical channel of this network is distinct from other traditional networks. For example, the fluctuations caused by an unstable wireless channel are more severe. These characteristics are considered as gaps that must be eliminated in the network design. Consequently, the security of the physical layer represents an interesting attitude. Indeed, the physical layer is the lowest layer in the OSI model. Its main role is to transmit and to receive the stream of unstructured raw bits on a physical medium.

In the last step of our work, we focus on how to ensure the security of the physical layer in the wireless sensors network. Existing security solutions are often implemented in the upper layers. In addition, the coexistence of the secure physical layer must be taken into consideration. Recall that this layer is vulnerable to attacks and specifically Jamming attacks. These malware and viruses attack wireless communication. Indeed, an attacker tries to interfere with the radio frequency used by the sensor nodes in the network. So and to eliminate these anomalies, we proposed as solution the implementation of an intrusion detection system. Despitethe IDSs are expensive in energy and the WSN is characterized by low memory capacity and limited energy, we chose to apply this IDS only at the sink node level. This node is responsible for detecting the anomaly and then informing the other nodes of the intrusion existence using a broadcast message or an alert and finally taking the prevention decision.

In the security field, we called intrusion any attempt to violate the security policy of a system. Precisely, it is the violation of the security services namely: data privacy, integrity or authentication. The network should continue to operate despite the appearance of unknown behavior likely to impede the proper network functioning. As example of intrusion we present Jamming attacks that threat the physical layer. In fact, the Jamming is a well-known attack that target wireless communication. Indeed, an adversary attempts to interfere with the radio frequency (interference) used by the sensor nodes in the network.

In Jamming, the nodes do not have access to the medium and cannot communicate because of the radio interference. However, a network without access to the medium is a network out of order so, the Jamming is a denial-of-service attack. This radio interference will have a direct influence on the values of several network parameters. Subsequently, increasing or decreasing the values of these parameters will be an indication of the existence of a Jamming attack. Then, these parameters can be used in the detection of these anomalies. These parameters include: PDR, PSR, ECA, LPR and CST [13].

-*Packet Delivery Ratio (PDR*): The PDR is defined as the ratio between the number of packets successfully delivered to a destination node and the number of sent packets by the sender node.

-*Packet Send Ratio (PSR):* The PSR is defined as the ratio between the number of packets sent successfully by a node and the number of packets it intends to send (messages in queue).

-*Loss Packet Rate (LPR*): The LPR is defined as the ratio between the number of packets lost and the total number of packets over a given period.

-*Energy Consumption Amount (ECA):* The ECA parameter is defined as the amount of energy consumed by a node for a specified period of time.

-*Carrier Sensing Time (CST):* In Medium Access Control (MAC) protocols, such as the Carrier Sense Multiple Access (CSMA) protocol, each node tries to detect when the media is free. So that it can then send its own packets. The period during which the node must wait for the carrier to become free is called Carrier Sensing Time (CST). This period is calculated as the average time elapsed between the moment when a node is ready to send its packet and the moment when the medium is released so that the node can send its packet.

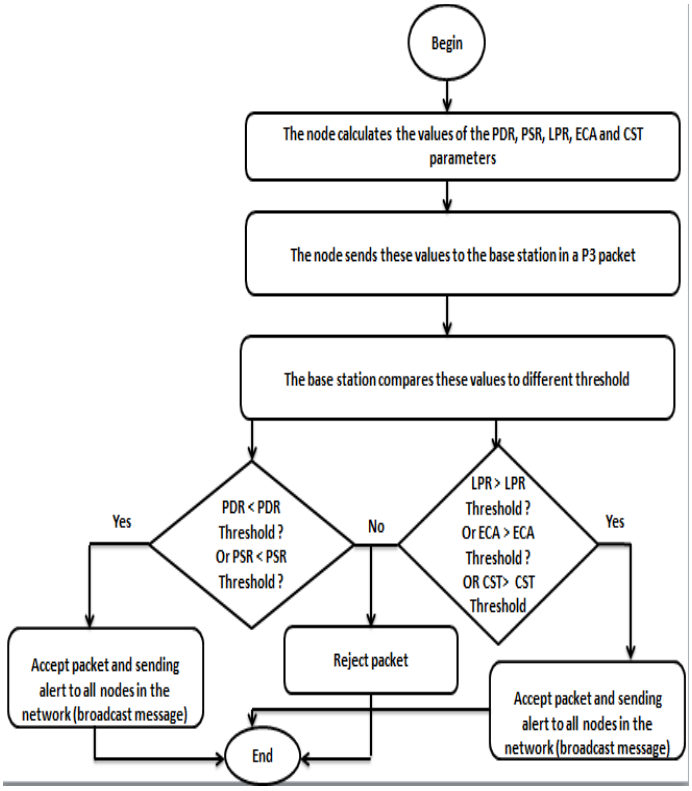


Figure 13. IDS process.

Existing works for detecting Jamming attacks use the combination of two, three, or up to four parameters. Such as the work [32] which use the four parameters PDR, BPR, RSSI and CCA to calculate the interference index. This calculation is done by each node. The CCA is a variable that counts the number of times the transmitter finds the channel busy trying to send a packet. The major disadvantage of this method is the necessity of complicated computation at the nodes, which is costly in energy terms. Other researchers use the parameters BPR, PDPT and SNR to calculate the Jamming index [21]. This method is able to solve the complex computational problem at the nodes, because all the computation is done by the base station. In our solution, we kept this solution by increasing the number of parameters to five. These parameters are: PDR, PSR, LPR, ECA and CST. Each node performs the calculation of each parameter for a given period and then sends these values to the base station which compares these values to the different thresholds. We proceeded as follows:

In our solution, we add a new packet P3 which containing the values of these parameters. Each node must encrypt the packet using AES and the session key Ks before sending it to the base station. The P3 fields are showed in Figure 14.

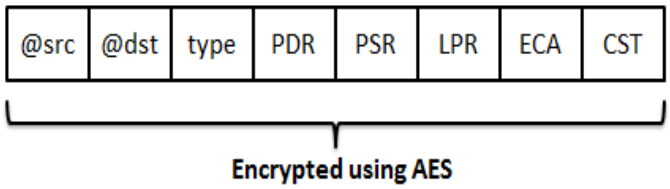


Figure 14. P3 fields

4. Simulation Results & Discussion

To evaluate the performances of both protocols XMachiavel and SXMachiavel, we use the simulation through OMNet++. For the case of our proposition, we used firstly SHA1 hash function; secondly, we change it to SHA3. We evaluate three metrics which are: mobility rate, loss packets rate and energy consumption.Before beginning the simulation, some parameters must be adjusted. First, we used a wireless sensors network with a topology of 500m * 500m with a random deployment of nodes. These nodes may be fixed or mobile nodes. For the XMachiavel protocol case, the percentage of fixed nodes is larger than the percentage of mobile nodes (see Table 1). On the other hand, the mobile nodes are more numerous than fixed nodes for the SXMachiavel protocol case. After each simulation duration(equal to150s), we incremented the number of nodes and we calculated the different metrics.

Table 1 detailed our simulation parameters.

4.1. Mobility rate

Before calculating the mobility rate, we present a histogram which indicates the partition of fixed and mobile nodes in the network for both protocols XMachiavel and SXMachiavel (using SHA1 or SHA3).

Table 1. Simulation parameters

Network size	*500m*500m*
Nic Type	*NicXMachiavel ; Nic SHA1SXMachiavel; Nic SHA3SXMachiavel*
Mobility Type	*ConstSpeedMobility*
Thermal noise	*-121dbm*
Carrier Frequency	*868e+6HZ*
Header length	*24 bits*
Queue length	*5*
Simulation Time	*150s*
Bit rate	*15360 bps*
Number of mobile nodes	*Variable (from 0 to 199)*
Number of fixed nodes for XMachiavel case	*Variable (from 0 to 149)*

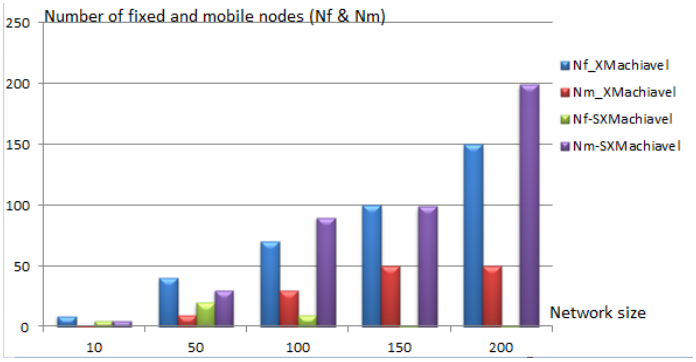


Figure 15. Number of fixed and mobile nodes in the network

Recall to: the mobility rate is the report between the number of mobile nodes and the total number of nodes in the network.

Mobility rate = Number of mobile nodes / Total number of nodes.

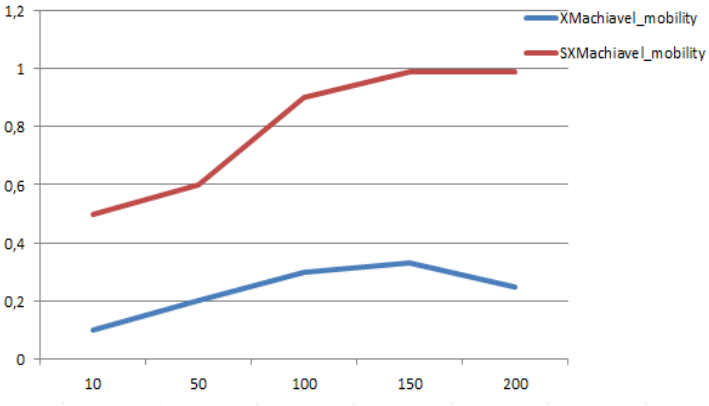


Figure 16.Variation of mobility rate according to network size

According to the Figure 16, we notice that mobility rate is different for both protocols. Recall that our objective is the developing of an efficient and secure WSN to combine the two metrics (energy conservation and security). So, we have increased the mobility rate in our approach. In the XMachiavel protocol, the number of fixed nodes is more than the number of mobile nodes (which doesn't exceed 50 nodes). But, it doesn't the case for our proposition which has as goal providing a totally mobile network. For this reason, we have increased after each simulation duration (= 150s) the total number of mobile nodes. From a network size 200, all the nodes will be mobile except the sink which remains

fixed. So, the mobility rate reached 99%. Subsequently, we achieved our goal which is providing a totally mobile network.

4.2. Loss packets rate

Recall to: the loss packets rate is the report between the number of lost frames and the total number of sent frames. To realize a correct simulation, we specified, at first, the simulation conditions (see table 1). Furthermore, we increased, every time, the number of knots to show the resistance of protocols with scaling factor. The formula is presented as following as:

Loss packets rate = Counts loss frames / Number of transmitted frames**.**

To evaluate the loss of packets, we performed several simulations between the non-secure protocol XMachiavel and our secure protocol SXMachiavel. The first simulation is done between XMachiavel and SXMachiavel with only the authentication mechanism. The results of this simulation are shown in the Figure17.

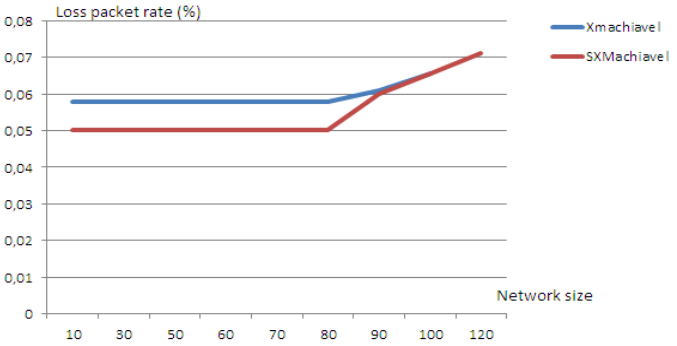


Figure 17**.**Variation of loss packets rate according to network size

The loss packets rate is 0.05% for a small network size (from 10 to 80 knots): a non-significant loss. Our network is completely secured (implementation of three security mechanisms as follow as: authentication - hash and encryption. Then, the risk of intrusion appearance is low. This explains the lack of packet loss (0% loss) for a small network (see Figure 18).

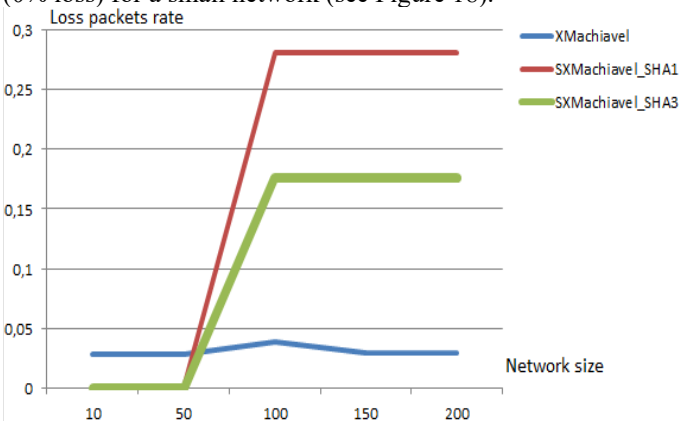


Figure 18**.**Variation of loss packets rate according to network size

As become famous in Figure 18, the loss packets rate increases with the number of nodes in the network. For a small network size (10 to 50), our proposed approach provides the best results (0%loss), but when the network size becomes vast, XMachiavel protocol shows the lowest loss packets rate. We explain these results by adding security mechanismsin SXMachiavel and the increase of number of packets exchanged. Indeed, the addition of security technologies increases the data transfer betweennodes and subsequently the loss of packets also increases. Turn now to compare between SXMachiavel protocol using SHA1 and SHA3, we note that the loss packets rate is higher in case of SHA1. Which demonstrate the fragility of this hash function to security attacks and shows the effeciency of SHA3 in terms of loss packets rate.

4.3. Energy Consumption

Recall to: Energy consumption = Transmission energy + processing energy + energy consumed by units.

We notice that the processing energy and the energy consumed by units are neglected. Thus, we consider that the transmission energy presents the total of energy consumption.

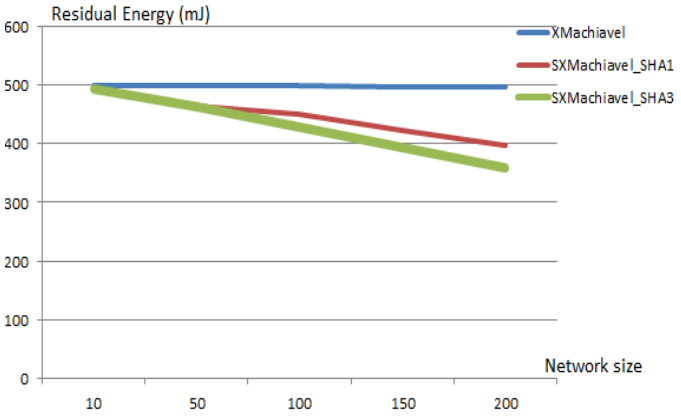


Figure 19**.**Variation of residual energy according to network size

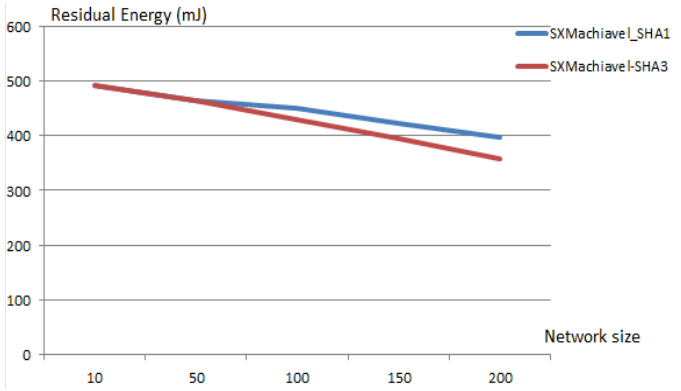


Figure 20**.**Variation of residual energy according to network size

Figures 19 and 20 show the variation of residual energy, according to the network size. According to these Figures, we notice that the existing non-secure protocol XMachiavel is less energy consuming; due to the absence of security mechanisms in this protocol. On the other hand, the energy consumption is higher for the case of SXMachiavel with SHA3; due to the addition of several security mechanisms which increase the total energy consumption, namely: an authentication mechanism, a hashing mechanism and an encryption mechanism. Consequently, it'spossible to exhaust the energetic resources of the nodes.

Unfortunately, there is always a trade-off between efficiency and energy consumption.

5. Conclusion

The implementation of a sensors network application must take into account the constraints that characterize the sensors nodes and differentiate them from other wireless networks types. These specifities, such as wireless communication-random deployment-resource limitation or unmonitored environment, have prompted the need to protect these networks by eliminating security vulnerabilities from the WSN.

The main challenges of our work are revealed in the cancelling at the security and energy conservation problems. These problems have required the emergence of numerous research studies proposed many solutions to ensure the reliability system, but it is based on the separation of both themes using either one or the other. Consequently, the originality of this work consists in combining the two features together (security and energy conservation) using a crossing between three layers: physical layer, data link layer and network layer. In the other hand and in our work, we have proposed a contribution which has consisted in using a cross layers security solution to obtain a secure-mobile network. This solution has started with the construction of a network deployment model specific to WSN. This model has consisted of sensors nodes partition in hierarchical areas; these areas are defined according to the radio range and the number of jumps. Then, this model can be generalized on any type of sensors (static, mobile or intelligent sensors that are differentiated from other sensors by adding the processor allowing data intelligent processing). Finally, we have secured this model. Stubborn that the proposed security solution has been appeared at: Firstly, in developing a new energy efficient and secure MAC protocol which provides a secure authentication- privacy & integrity. Secondly, implementing an intrusion detection system to protect the physical layer against jamming attacks. Simulations using OMNet++ take place to validate our solution and to estimate the performances of both protocols in terms of mobility rate, loss packets rate and energy consumption. The results showed that SXMachiavel is more successful in terms of mobility rate (can reach 99% compared with XMachiavel which doesn't exceed 33%), therefore and concerning the packets loss, the simulation result showed that it advances from 0.05% for a small network size. But, we deducted that the energy consumption increase by the addition of security techniques (residual energy decreases by 0.01% for each exchanged packet).

Acknowledgment

This work has been accomplished at WIMCS-Team research, ENET'COM, Sfax-University. Part of this work has been supported by MESRSTIC Scientific Research Group-Tunisia.

References

[1] M.Mezghani, R.Gatgout, G.Ellouze, A.Grati, I.Bouabidi, M.Abdellaoui, "Multitasks generic Plateform via WSN",International Journal of Computer Networks and Communications (IJCNC), vol.4, No.6, June 2011, pp.54-67. DOI: 10.5121/ijidps.2011.2406.

[2] Pr.M.M.Abdellaoui,''Multitasks-Generic-Intelligent-Efficiency-Secure WSNs and theirs Applications'', LAMBERT Academic Publishing (LAP), 2017, Part 3 : Secure Wireless Sensors Network, pp.142-185.

[3] S.Athmani, A.Bilami, "Protocole de sécurité pour les réseaux de capteurs sans fil", master memory, Hadj Lakhder-Batna University, Algeria, 15 juillet 2010, pp .1-96.

[4] Bouabidi Imen, Pr. Abdellaoui Mahmoud, ''Hierarchical organization by crossing between different layers for WSN energy saving'', International Conference on Advanced Technology & Sciences (ICAT'14), Antalya, Turkey , August 12-15 , 2014.

[5] Bouabidi Imen, Pr. Abdellaoui Mahmoud, ''Hierarchical organization by crossing between different layers for WSN energy saving'', 15th Internationnal conference on Sciences and Techniques of Automatic control & computer engineering-STA'2014, Tunisia, December 21-23 , 2014, pp.1-4.

[6] Bouabidi Imen, Pr. Abdellaoui Mahmoud, ''Hierarchical organization with a cross layers using smart sensors for intelligent cities'', SAI Intelligent Systems Conference, London, United Kingdom, November10-11 , 2015, pp. 446-451.

DOI: 978-1-4673-7606-8/15/$31.00 ©2015 IEEE.

[7] Bouabidi Imen, Pr. Abdellaoui Mahmoud, ''Energy Efficient Cross-layer MAC Protocol and Secure Authentication via an implementation of data confidentiality and integrity in WSN»,Future Technologies Conference 2016 (FTC), San Francisco, United States, December 6-7 , 2016. http://wwwpub.iaea.org/MTCD/publications/PDF/http://SAIconference.com/Conferences/FTC2016.

[8] V.Rajendram,K.Obraczka, J.J.Garcria-luma-Aceves, " Energy-Efficient, collision free Medium Access Control for Wireless Sensor Networks", Sensys'03, Los Angelos, California, USA, November 5-7, 2003, pp.1-12. DOI:1-58113-707-9/03/0011.

[9] S.Khatarkar, R. Kamble "Wireless Sensor Network MAC Protocol : SMAC and TMAC", Indian Journal of Computer Science and Engineering (IJCSE),vol.4, Aug-Sep 2013, pp.304 - 310.

DOI:ISSN: 0976-5166.

[10] B.Narain,et al,"Energy Efficient MAC Protocols for Wireless Sensor Networks: A survey", International Journal of Computer Science &Engineering Survey (IJCSES), vol.2, No.3, August 2011, pp.121-131.

DOI: 10.5121/ijcses.2011.2309.

[11] A.Warrier, J.Min, I.Rhee,''Z-MAC : a hybrid MAC for Wireless Sensor Networks'', pp.1-2.

http://conferences.sigcomm.org/sigcomm/2005/poster-123.pdf.

[12] C.Suh, Y. Ko and D.Son, " An Energy Efficient Cross-Layer MAC Protocol for Wireless Sensor Networks", Springer-Verlog Berlin Heidelberg, 2006, pp.410-419.

[13] A. Makke,''Détection d'attaques dans un systéme WBAN de surveillance médicale à distance'', phd diploma, Paris Descartes university, France, 30 May 2014, pp. 1-163.

[14] M.Rmayti, et al.,''Détection d'attaques Whormhole dans les réseaux MANETs en utilisant la théorie des graphes'', conference paper, december 2014,pp.1-16.

[15] L.B.Oliveira, et al., ''SecLEACH-A Random Key Distribution Solution for Securing Clustered Sensor Network'', FAPESP Under grant 2005/005579-9, pp.1-8.

http://www.cs.cmu.edu/~hcwong/Pdfs/secleach.pdf.

[16] A.N.Kulkarni, A.S.Tavildar,''Design and Performance Assessment for Energy Aware security Enhancing Strategy for PEGASIS protocol for MWSN'', International Conference on Information Processing (ICIP) Vishwakarna Institute of Technology'', December 16-19, 2015, pp.1-6.

[17] R.Singh, A.K.Verma,''Energy Efficient cross layer based adaptive threshold routing protocol for WSN'', International Journal of Electronics and communications-ELSEVIER, vol. 7 , February 2016, pp.166-173.

[18] U.Senthil Kumaran, P.Tlango, ''Secure authentication and integrity techniques for randomized secured routing in WSN'', Springer Science+Business Media Network, 27 August 2014, pp.1-9.

[19] S. Ghormare, V. Share,''Implementation of data confidentiality for providing High Security in Wireless Sensor Network'', IEEE Sponsored 2nd International Conference on Innovation in Information, Embedded and Communication Systems (ICIIECS), 2015, pp.1-5.
DOI: 978-1-4799-6899-6818-3/15/$31.00.

[20] P.Mohanty, M.R.Kabat, ''Energy Efficient structure-free data aggregation and delivery in WSN'', Egyptian Information Journals, ScienceDirect, vol.17, Issue 3, november 2016, pp.273-284.

[21] S. Misra , R. Singh, S. V. Rohith Mohan. "Information Warfare-Worthy Jamming Attack Detection Mechanism for Wireless Sensor Networks Using a Fuzzy Inference System", Sensors journal, 2010, pp. 3444-3479.

DOI:10.3390/s100403444.

[22] P.Inverardi, et al., ''Distributed IDSs enhacing Security in Mobile Wireless Sensor Networks'', Procceding of the 20th International Conference on Advanced Information Networking and Applications (AINA'06).
DOI: 1550-445X/06-$20.00, 2006, pp.1-5.

[23] D. E. Boubiche., ''Une approche inter-couches(Cross layer) pour la sécurité dans les RCSF'', PhD diploma in computer Sciences, Batna University, Algeria, pp.1-165.

[24] R. Kuntz,"Medium Access Control facing the dynamic of Wireless Sensor Networks", Phd diploma, University of Strasbourg, september 27 , 2010, pp.1-183.

[25] A. El-Hoiydi,''Spatial TDMA and CSMA with Preamble Sampling for Low power ad hoc wireless sensor networks'', 2007, pp.1-8.

http://www.mics.org/getDocum.pdf?docid=236&docnum=1.

[26] Best practices for EH&S software strategy planning, Verdantix, enablon, chapter1 Practice questions Mastering the Basis of security.

http://enablon.com/reports/best-practices-ehs-software-strategy-planning?campaign=G_D_Verdantix_New&gclid=CNa7ztXzxtECFU6dGwod43IMkg.

[27] H.Souilah, A.Baadache,''Coping with spoofed PS-Pool Based DOS Attack in IEEE 802.11 Networks'', pp.57-62.

http://ceur-ws.org/Vol-1256/paper5.pdf.

[28] https://en.wikipedia.org/wiki/SHA-3.

[29] https://en.wikipedia.org/wiki/Advanced_Encryption_Standard

[30] F. Ahmed,''Strongest AES with SBox Bank and Dynamic key MDS Matrix (SDK-AES),''International Journal of computer and Communication Engineering, vol. 2, No. 4, July 2013, pp.1-5.

[31] S.J.Chang, el al.,''Third Round Report of the SHA3 cryptographic Hash Algorithm competition'', National Institute of standards and Technologiy, U.S.Departement of commerce, pp.1- 84.

[32] H. I.Reyes, N. Kaabouch, "Jamming and Lost Link Detection in Wireless Networks with Fuzzy Logic", International Journal of Scientific & Engineering Research , vol. 4, Issue 2, February 2013, pp.1-7.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 1, 108-120 (2017)
www.astesj.com

ASTES Journal
ISSN: 2415-6698

Mobi-Sim: An Emulation and Prototyping Platform for Protocols Validation of Mobile Wireless Sensors Networks

Omina Mezghani[1,2], Pr. MAHMOUD ABDELLAOUI [2, *]

[1] *ISI, National Engineering School of Sfax (ENIS), Tunisia*

[2] *WIMCS, National Engineering School of Electronics and Tele-Communication (ENETCOM), Tunisia*

ARTICLEINFO

Article history:
Received: 10 December, 2016
Accepted: 14 January, 2017
Online: 28 January, 2017

Keywords :
Mobile WSNs
WSN simulator tools
Mobi-Sim
Clustering
Routing

ABSTRACT

The objective of this paper is to provide a new simulator framework for mobile WSN that emulate a sensor node at a laptop i.e. the laptop will model and replace a sensor node within a network. This platform can implement different WSN routing protocols to simulate and validate new developed protocols in terms of energy consumption, loss packets rate, delivery ratio, mobility support, connectivity and exchanged messages number in real time. To evaluate the performance of Mobi-Sim, we implement into it two popular protocols (LEACH-M and LEACH sink-mobile) and compare its results to TOSSIM. Then, we propose another routing protocol based on clustering that we compare it to LEACH-M.

1 Introduction

Wireless sensors networks represent a very interesting multidisciplinary field providing tremendous benefit for a number of domains in daily life. It is compounded by a set of sensor nodes characterized by a low-energy consumption in both hardware and software to maximize the network lifetime [1-2]. These characteristics make WSN suitable to be used in several applications like military, domotic, health, environment... In spite of its various usage in numerous applications, WSN confronts many issues like energy, self-organization, sensors localization, fault tolerance, security, mobility(since many applications require the presence of mobile entities in the network) [3]. So, WSNs field is under rigorous research to develop new algorithms, techniques and protocols to make more efficient and reliable the mobile WSN. In fact, nodes are often expected to operate for long periods of time. So, the design of protocols for mobile sensors network requires adequate methods for efficient data routing and minimal energy consumption. Then, researchers need to test and validate their proposed approaches before their implementation in a real system. So, simulation is the common way used by researchers to test and validate their developed prototypes in various forms (energy, connectivity, QoS ...) before any implementation. There are many simulation tools, out of which they are dedicated towards traditional networks and just have been adapted to be used with WSN, other support only fixed nodes in the network and are unable to model mobile nodes behavior. Therefore, implementing new protocols under these simulators is often difficult and they cannot model all the required performance criteria (energy, loss packets rate, connectivity ...).

This paper presents a new simulator 'Mobi-Sim' that emulate a sensor node at a laptop in order to study the dynamic behavior of mobile nodes in real time without any scenario for mobile sensors movement programmed in advance. In fact, each laptop running the Mobi-Sim application can replace a node and behave as a sensor within a network. We can imagine that this ambition poses very difficult problems to solve at all levels in a surveillance system:

- The exchange of data in the network,
- The routing algorithms process,
- Battery level monitoring in real time,

*Corresponding Author: Pr. Mahmoud Abdellaoui, WIMCS-ENET'COM, (+216) 50 42 94 27 & mahmoudabdellaoui4@gmail.com

https://dx.doi.org/10.25046/aj020113

- The system reliability in terms of connectivity between moving nodes

To evaluate the performance of Mobi-Sim, we implement into it two popular routing protocols (LEACH-M and LEACH sink-mobile) to compare its result with those obtained by TOSSIM. In fact, in the previous work [4], we just implemented and tested the flooding protocol into our proposed simulator.

On the other hand, since many applications require the presence of mobile entities and nodes are often expected to operate for long periods, new challenges for WSN management are envisaged. So, the design of new routing protocols adequate for mobile sensor networks is very demanded. For this reason, we describe in this paper a new routing protocol based on clustering that supports mobility. We validate it by the developed Mobi-Sim simulator because the existing simulators do not model all the required performance criteria (energy + loss packets rate + connectivity + number of events).

2 Simulation Tools

In this section, we present an overview of the existing simulators for both real and non real time systems to find out their limits and their problems. These simulators are discussed according to different scales: network deployment & planning, data routing and performance metrics.
To overcome these simulator's limits, we propose a new simulator "Mobi-Sim" which answer at all expectations and all requirements of researchers and designers. Notice that protocol engineering requires an unavoidable phase which is the evaluation of the new developed protocols [5]. Several tools are offered for researchers and designers, each has its advantages, but also its disadvantages.

- **Ptolemy II** addresses the modeling, simulation and design of real time embedded systems. It offers different models of simulation paradigms such continuous time, data flow and discrete events. It is based on java packages. But, this simulator implies some limits in terms of scalability issues. In fact, simulation of several nodes remains a challenging problem [6].

- **IDEA** is a systemC simulator for WSN. It implements a power module for energy consumption modelisation. It can also model the delivery packet rate and transmission latency. But, it fails to simulate mobile sensor nodes in the network [7-8].

- **True Time toolbox** is a Matlab library dedicated to real-time system. It is a simulation tool for WSN. It proves its ability and flexibility to support different physical layers. It also provides a graphical environment that make easy the designing and the implementation of a variety of sytems. But, in this simulation tool the step from simulation to production code is not large. Therefore, execution of experiments is time consuming [9-10].

- **MULE** is an hybrid simulator. It performs the execution of events in real time where physical motes are connected to PC. This simulator is dedicated for fixed WSN and cannot model mobile sensors behaviour. Therefore, this tool does not offer a graphical user interface [6].

- **COOJA** is a simulator mainly developed for wireless networks. It is useful when simulating sensors which are running the Contiki Operating System . COOJA is a cross-layer simulator. It operates at three abstraction levels. The networking level is the highest one, at which routing protocols can be implemented by users. The developed program code which is tested at the operating system level on COOJA could be run directly on Contiki OS without modifications. The third level is the machine code instruction. In this level, the emulation is done at a bit level [9]. Simulated sensor nodes can be from different type in terms of hardware and software aspects including sensors which don't use the Contiki operating system. But, it has some disadvantages. First, each node is simulated solely at one of the supported level. Second, cross-layer behavior exhibits increased granularity in simulation outcomes which is not available with TOSSIM [9-10]. Third, due to its extensibility, COOJA has relatively low efficiency. Therefore, it Supports a limited number of simultaneous node types, so it has to be restarted if this number exceeds the limit.

- **SENSE** is a discrete event simulator written in C++, written in CompC++ (an extension to C++). SENSE has modules for battery models, MAC Layer protocols, application, network and physical layer models. It also includes a set of SSR Protocols such as Self-Selective Routing (SSR), Self-Healing Routing (SHR) and Self-Selective Reliable Path (SRP). The iNSpect tool is used by SENSE to allow simulation visualization in order to analyze the simulation output and perform simulation animation showing message flows [11-12]. SENSE can be modified easily and is fast compared to NS-2. But, the need for direct integration into developed applications makes the use of NS-2 and SENSE unwanted. Therefore, SENSE don't support mobile nodes and cannot model mobility behavior of sensor nodes.

- **TOSSIM** is a TinyOS simulator. TinyOS is a sensor-based network operating system [13].It allows to simulate the sensor behavior (sending/receiving messages, information processing, ...) within a WSN. The TinyOS execution mode is directed by events and is adapted to discrete events simulator. However, TOSSIM did

not allow the execution of more than one application on the network and to estimate important behaviors including CPU and energy characteristics[14-15]. An extension of TOSSIM called **PowerTOSSIM** is added to model the energy consumption [16]. For the consumption values, the designers relied on the Mica2 type nodes. The designers relied on the different components consumption in a node according to their states. So, it is necessary to know the state of each component in a node during the simulation. Thanks to the simulation model based on TinyOS, the components state is known immediately since their changes correspond to events in TinyOS and thus in TOSSIM. However, several components in a sensor node are sometimes abstracted in TOSSIM by a single component [6]. Therefore, this tool does not offer an ergonomic user interface and does not support the mobility of nodes.

- **OMNet++** is a discrete simulator implemented in C ++. Getting started with this simulator is quite simple thanks to a clear design of the simulator. It also provides a powerful graphical user interface for debug animation and management. In addition, a framework for mobility has been added and it can be used to model WSNs. However, there is a serious lack of available protocols in the library compared to other simulators such as NS-2. On the other hand, there is a compatibility problem because researchers develop their problems separately which make its combination difficult and may lead a high number of bugs [14].
- **NS-2** is a discrete event simulator. NS-2 proposes four levels of abstraction which allow to adapt this simulator to different interests. Indeed, some will want low-level information, for example to study the effect of radio signals propagation in an environment while others study routing protocols and only need information at the network level. NS-2 was originally intended for wired networks, which certainly explains the simplicity of the radio propagation models in this simulator. Despite the different levels of abstraction, NS-2 is mainly used by researchers who are interested in routing protocols and/or medium access protocols. The open source NS-2 code allows each one to add his stone to the building. NS-2 research remains a very used simulator. However, the simulation is not real time; It is considered virtual. Therefore, it is used to simulate traditional networks and it is not targeted to WSN which make it behave differently than WSNs. In fact, packets format, MAC protocols, energy models and the sensing hardware are not the same with those found in most sensors. Another problem is the tracing system. In fact, it has a poor graphical support i.e. without GUI. So, users have to manipulate text commands of the electronic devices [11]. The same for **NS-3**, it does not provide any graphical tool for analyzing or visualizing network simulation. Therefore, although the NS3 models are written in C++ like NS2, the NS2 models are not compatible to thus of NS3 and including them into a NS3 simulation is not possible [5].
- **J-Sim** has been developed in Java. It offers a framework for WSN simulation built on ACA, INET and Wireless protocol stack. It encompasses the following components:

  1. Sink, Sensor and target Nodes.
  2. Sensor Channels and Wireless Communication channels
  3. Physical media such as Seismic channels, mobility models and power models.

  J-sim implements three WSNs protocols: Localization, Geographic Routing and Directed Diffusion. The design of J-sim makes difficult adding new protocols or node components to users because it was not originally dedicated for WSNs. Therefore, the execution time is much greater than thus of NS-2 [6].

As discussed below, a large number of network simulators are available. But, they present several drawbacks in term of network deployment, data routing and performance metrics. First, in some simulators the user must manipulate command lines to plan its network and define its architecture. Indeed, He must resort to deploy sensor nodes one by one which is painful for the large scale networks. On the other hand, the architecture of a mobile network is likely to change over time since nodes can change their positions. This may not be applied because the simulator takes as input a topology file describing the network architecture which remains the same throughout the simulation. In terms of scalability, several simulators support a limited number of simultaneous node i.e. small networks and cannot simulate a large network. Second, for data routing, the majority of simulators have low efficiency in term of extensibility. In fact, they present a difficulty to add new protocols or node components within this simulator. It is also very difficult to visualize all the exchanges between sensors composing the network. Finally, among the existing simulators cited below, there isn't a complete simulator and they fail to model all the required performance metrics needed by a designer to validate its data routing protocol in a mobile WSN.

As conclusion, practical issues such as mobility modeling , energy modeling, loss packets rate modeling, sensors network specific models, large scale network support, real time simulation, real system implementation, extensibility to add new protocols, graphical display to see simulation parameters,... are important way to be found in a developed simulator.

3 Proposed simulator: Mobi-Sim

In its most general sense, the term ”system” refers to a set of real or abstract units that interact to serve a common purpose. Modeling is used to create models for each element in the system environment in order to interpret an experiment as well as to predict how the system evolves and how it behaves in a given situation. This modeling serves to validate an operation by simulation and to refine the specifications [4]. Indeed, modeling and simulation have seen an unprecedented progress in various sectors of industrial engineering for the design of new systems in order to study their characteristics as widely as possible, which increasingly encourages methods for the development of modeling and simulation tools.

Indeed, according to the existing simulators limitations, presented in the previous section, mainly because they don't support the presence of mobile nodes in a network, do not model all the performance metrics intended by a designer to validate his new approach, present a difficulty for users to implement their new protocols, do not offer real-time simulation and the simulation in others does not model a real sensor... So, to accelerate the design process, a complete-fast and powerful simulation environment that meets all these requirements may be essential for the designers.

In this context, we are interested to ”monitoring systems” based on a mobile Wireless Sensors Network. The main idea of our Mobi-Sim simulator is to emulate a sensor node on a laptop. Indeed, a laptop will react (with the components that surround it in its environment) as a sensor. The proposed testbed, launching a simulation, is shown in Figure 1.

Figure 1: Proposed testbed

Our simulation system is able to get closer to the reality of things, to reproduce the behavior of the elements which belong the network in real time, to simulate a network containing fixed nodes only, to simulate a network containing mobile nodes and to simulate a hybrid network (consisting of fixed & mobile nodes). It is dedicated for a mobile wireless sensors network to deploy a large number of laptops as mobile nodes. Therefore, it has an efficient extensibility to easily integrate new routing protocols and it presents a generic platform to implement any routing protocol. In fact, a laptop including Mobi-Sim will replace a sensor node that can executes any existing algorithm in its memory. For the performance metrics, Mobi-Sim presents a set of performance criteria for a routing protocol namely the energy consumption, the delivery packets rate, nodes connectivity and packets loss rate. The Mobi-Sim characteristics are detailed in the next section.

3.1 Mobi-Sim development environment

Mobi-Sim template is a complete working project emulating a sensor node behavior in a WSN on a laptop. This sensor may be a stationary node operating either in a fixed network or in a hybrid network (containing both fixed and mobile sensors); as well as it may be a mobile node communicating by wireless means. This application is developed with an ergonomic interface that help users, specially for novices, to simulate and validate their proposed approaches. In fact, to provide researchers to add their own proposed protocols (by writing them and implementing them into the Mobi-Sim template), this template has tools that generate empty modules in which the eventual protocol code should be integrated. Mobi-Sim is programmed using the Microsofts .NET framework. It is developed in the language C#.Net providing a graphical user interface (GUI) for users to interact with the different tab of this application in online mode. This GUI allows the deployment, the supervision, the control, the managing and the tracking of the different sensor nodes automatically in the network based on their ability in the network to communicate together.

3.2 Operating mode

The developed Mobi-sim operating mode includes the node deployment, neighbor discovery, periodic data exchange and energy consumption supervision. Node deployment may be done manually by users if it is a fixed node (static deployment) or randomly if it is mobile node (random deployment). In a static deployment, the user can define for each fixed sensor node its eventual neighbors and define its level in the network topology as shown in Figure 2. Then, it should start to exchange data with its neighbors according to the defined protocol.

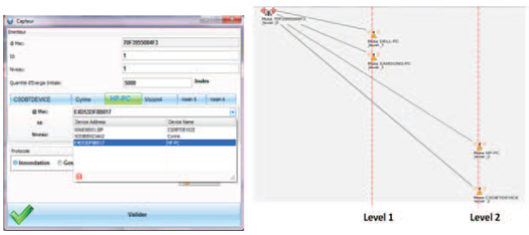


Figure 2: Static WSN deployment

If sensor nodes are mobile, they should be deployed using a random deployment as shown in Figure 3. The mobile sensor should discover automatically its neighbors with which it can communicate. The node discovery algorithm is based on Hello messages. Each node is defined by its MAC address (@MAC of the laptop implementing the node).

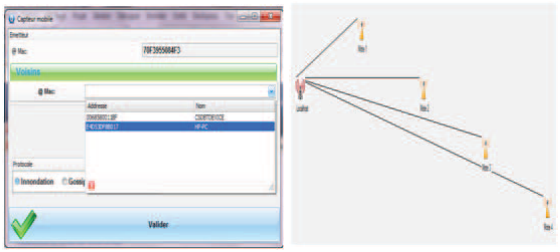


Figure 3: Mobile WSN architecture

The implemented functionalities in Mobi-Sim are shown in Figure 4.

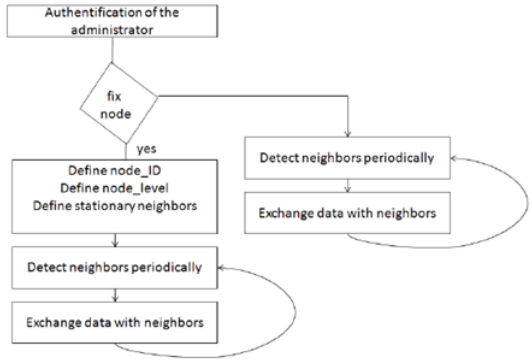


Figure 4: Mobi-Sim operating mode

3.3 Energy Model

The basic entity of a network topology is the sensor node. In Mobi-Sim, a laptop is represented as a node. In this simulator, the power component serves for power management. The consumed energy by a sensor node is essentially due to the following operations: data capture, processing and communication. The radio module is the component of the sensor node which consumes the most energy (about 94%)because it is responsible to ensure the communication between the nodes [17-18]. There are four states in the radio components (transmitter and receiver): active, receive, transmit and sleep.

The amount of energy required for both sending and receiving data from node j to another node, computed as in [19], are given by:

$$E_{Tx}(j,k,d) = E_{Trans} \times j \times k \times d^{\alpha} \quad (1)$$

$$E_{Rx}(j,k) = E_{Receive} \times j \times k \quad (2)$$

where E_{Trans} is the energy dissipated to transmit k bit along a distance d which separate the sensor node j to another sensor node. While $E_{Receive}$ is the amount of energy consumed by a sensor node to receive k-bits message. α is the path loss exponent due to wireless channel transmission [19-20]. In Mobi-Sim, these parameters are left to be set by users. Using this radio model, the transmission cost and receiving cost for a k-bit message with a distance d are: - Transmission:

$$E_{Tx}(k,d) = E_{elec} \times k + E_{amp} \times k \times d^2 \quad (3)$$

- Receiving:

$$E_{Rx}(k,d) = E_{elec} \times k \quad (4)$$

where: E_{elec} is the radio electronics dissipation, E_{amp} is the amplifier dissipation and d is the distance between transmitter and receiver. Mobi-Sim implements two type of sensor nodes: MICAZ and Telosb. A Simple_Power component has been implemented these four states. The power component response to the control signal from networking components. It can switch from one state to another depending on its operating mode. In the proposed energy framework, the parameters modeling such as initial energy are defined as attributes by user. Figure 5 shows a description of the various modules included in our framework. We implement the EnergySource and DeviceEnergyModel classes in this framework to represent energy sources and device energy models. Device_energy_model_container defines sensor nodes type (MICAZ or Telosb). Mobi-Sim provide users to add other types and define their attributes. In each node, device_energy_models consume energy from the energy_source. The energy source give notice for device_energy_models when it is completely drained. The proposed Mobi-Sim simulator energy framework provides functions that allow external simulation modules to send requests to the node's Device_energy_models. For example, a routing protocol may demand information about the battery level by querying the energy source on the node.

4 Mobi-Sim: Simulation & Evaluation

In this paper, we add other hierarchical protocols into our Mobi-Sim simulator: the LEACH mobile sink protocol and LEACH-M. In fact, hierarchical routing protocols have proved an efficient energy [21-26]. In the Hierarchical protocols, sensor nodes are dynamically organized into partitions called clusters. For each cluster, a cluster head (CH) is elected. The cluster heads have to collect and aggregate the data; thus reducing exchanged data packets and saving energy [27]. Instead of transmitting data directly towards the sink through long distances, sensor nodes have to transmit their data to the cluster to optimize energy consumption [28]. The implemented protocol treats the two aspects of mobility in a network: LEACH-M includes mobile nodes when LEACH sink-mobile include mobile sink.

Thus, we simulated them according to the previous cited criteria in [4] to show our simulator efficiency face to other popular protocols. We have also add another performance criterion which is sensor nodes connectivity. Then, we present another approach for data routing in a mobile WSN based on clustering that we simulate by Mobi-Sim.

4.1 LEACH-M Evaluation

LEACH-M is a protocol based on a dynamic network partitioning into a set of clusters. The protocol implementation involves three operating phases: a phase

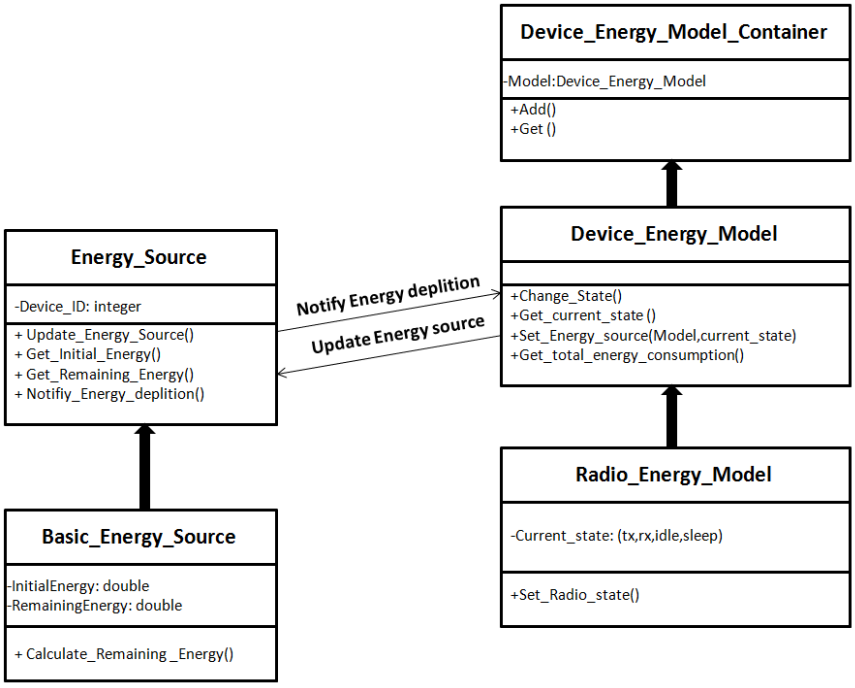


Figure 5: Class diagram of the proposed Mobi-Sim energy framework

of announcement and clusters creation, a scheduling phase and a transmission phase [28].

- *Announcement phase and clusters creation:* During this phase, the base station announces a new round. The different nodes self-elect to be CHs according to an election probability and a predetermined percentage of CHs selection in advance [5% to 15%]. This election can be described as follows:
  if $n \in G$:

$$\frac{P}{1 - P \times (r mod \frac{1}{P})} \tag{5}$$

  Where P and r denote respectively the percentage of the nodes wanting to be CH and the current round, G represents the set of nodes which have not be CHs during the last $\frac{1}{P}$ iterations. n is a random value generated by a node in the range [0,1]. If n is less than a threshold T (n), the node is CH candidate and then should informs its neighbors about its election. Thus, an announcement containing its identifier as CH is broadcast via a CSMA/CA MAC protocol to avoid probable collisions and interferences between the adjacent CHs. The nodes decide to join the nearest CH. In case of equality, the choice of belonging to a CH is made randomly. This is done by sending a join packet by the sensor node to the chosen CH, the latter sends an acknowledgment message for confirmation.

- *Scheduling phase*: Once the clusters are formed, each CH moves from a simple member node (MN) to a coordinator for information transmission within its group. Based on the task scheduling method, it implements the TDMA protocol and assigns to each of its MNs a slot time during which the node can communicate its information. All of these time intervals constitute a frame which duration differs depending on the number of nodes in the cluster.

- *Transmission phase*: It is a phase for gathering information. Based on the TDMA method for accessing the MAC layer, the MNs communicate their data to the corresponding CHs for a predefined period of time. The collected data at each CH are aggregated and then transmitted directly to the base station. Outside the allowed time for transmission, each node has the ability to standby to save its resources. In this protocol, the data processing at each cluster is done locally and each CH coordinate the exchanges with the other MNs. The network has the capacity to self-reorganize during the CHs election phase. Each node has the possibility to be elected CH and inversely, each CH can become a simple MN that can belong to a cluster. The CH election is based on the residual energy. It also performs data aggregation from MNs. Unlike the MNs, the CHs are in permanent activity since they communicate either with all the nodes belonging to their clusters or they exchange data with the base station.

The LEACH-M routing protocol make the following assumptions:
- The network uses N homogeneous nodes deployed

randomly and are able to communicate with The base station in mono-hop.
- In each cluster, except for the initialization phase, the new CH election is carried out by the CH of the previous round according to its residual energy.

Initially, the base station broadcasts a message to determine the number of clusters as well as the probabilities 'P' associated to the different nodes wishing to become CHs. In LEACH-M, the initial number of CHs corresponds to 10% of the nodes total number within a network. Each node receiving a message from the base station must rebroadcast it to its neighbors. It must generate a random number nb_a that will be compared to the probability P. If the probability P is less than nb_a then the node is a CH candidate, otherwise it waits new CH notifications from other CHs.

Once elected as CH, the node broadcasts its new state to all its neighbors by sending an advertisement message 'ADV' containing its identifier ID. Other nodes select their clusters based on their positions. Each MN must inform and join a CH in their coverage. Based on the scheduling method, the protocol implements the medium accessing method TDMA by assigning for each MN a time slot for communication. After having received, processed and transmitted its data, each node can pass to the Idle mode in order to optimize its energy. The different CHs perform data aggregation, data fusion and then transmit it to the base station. Then, they proceed, depending on the residual energy of each node, to the new CHs election. The old CHs inform the new CHs of their new state and transmit them the list of identifiers of all the MNs. In the following, we present and analyze the simulation results and experiments obtained during the implementation of this protocol on two simulators: Mobi-Sim and TOSSIM.

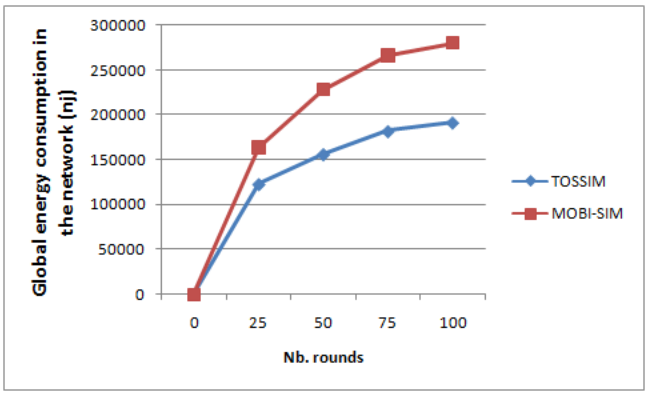


Figure 6: Average consumed energy in the network by LEACH-M

As illustrated in figure 6, the results of the LEACH-M implementation show that the energy consumed by sensor nodes is greater in Mobi-Sim than in TOSSIM. The obtained result is due to sensors nodes mobility which can change their positions over the time which make their supervision a difficult task for TOSSIM. In fact, unlike Mobi-Sim, TOSSIM cannot reproduce mobility behaviors and then mobile sensor nodes (which change their initial position) are not able to detect and to send their data packets toward other sensors. So, TOSSIM loses the control at the moving nodes. These nodes will behave as disconnected nodes whose neighbors could not receive their data packets which will influence into the total rate of the energy consumption in the network. In against, Mobi-Sim could model and reproduce the behavior of mobile nodes as well as could supervise the exchanged data packets between mobile sensors whatever their position as long as they are in the sensing field. Since sensor nodes are able to be supervised while moving and they still able to communicate together, the exchanged data packets rate would be higher for Mobi-Sim (see figure 7) as well as the connectivity rate i.e. the number of active nodes (see figure8). Therefore, if the exchanged data packets rate increases, energy consumption will also be very important (because radio communication module dissipates the major amount of energy within a sensor node).

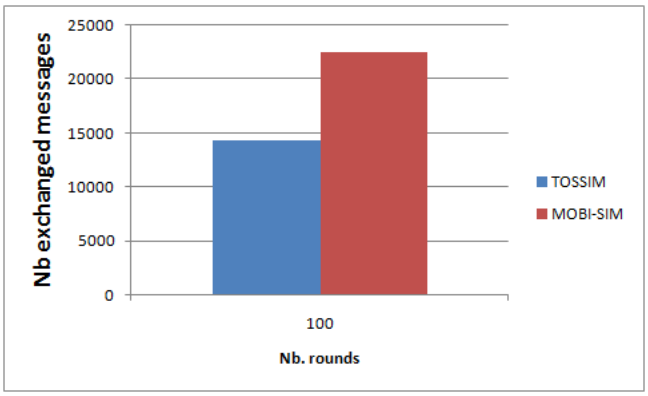


Figure 7: Exchanged messages number by LEACH-M

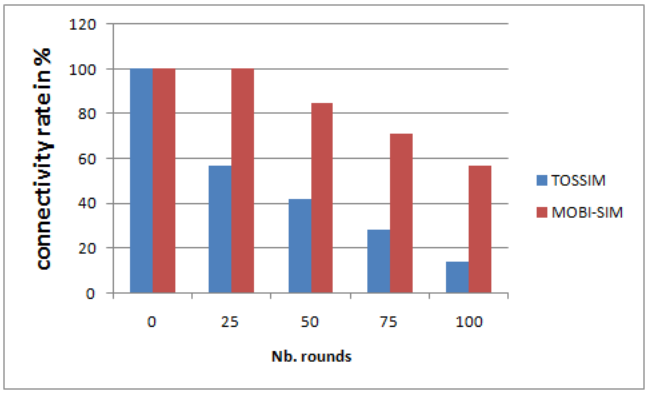


Figure 8: Connectivity rate by the LEACH-M in both simulators: Mobi-Sim and TOSSIM

The next performance criterion is the lost packet rate. The packets lost number is the difference between the number of the generated packets from sensor nodes within a network and the number of the received one at the sink. Figure 9 shows that the loss data packets rate is higher in TOSSIM than Mobi-Sim. This is mainly due to sensors mobility. In fact, when leaving an area to enter in another, TOSSIM consider this sensor node as died node which cannot deliver its data packets to its neighbors.

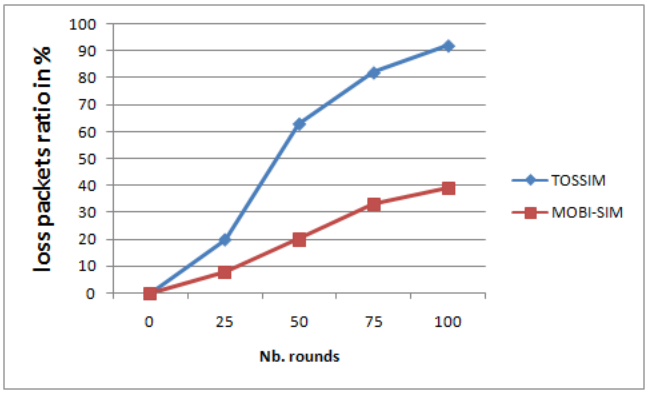


Figure 9: Loss packets rate by LEACH-M

Mobi-Sim presents better performances on the popular LEACH-M protocol and can model well the mobile sensor nodes behavior than TOSSIM in a mobile WSN.

4.2 LEACH Sink-Mobile Evaluation

The LEACH sink-mobile communication architecture consists to form clusters and to elect for each cluster a cluster header as a gateway to the base station. CHs are randomly elected (5% of the total nodes number) based on a probability function according to a specific election algorithm. Nodes member transmit their data packets to their CH. Then, each CH receiving data aggregates and transmits it to the BS directly or via other CHs. For the topological model, the BS is mobile and can move from an area to another one. LEACH Sink-Mobile is based on the following points:

- Sensor nodes are fixed,
- More than one mobile base stations,
- Each elected CH, broadcasts a message to other members in its cluster. MNs have to join this CH in order to send it their collected data,
- If a sensor node fail to recover the new CH ID, it waits this ID from the old CH after sending it its data,
- If the CH fails to find an itinerary to send data to the BS, it tries to join another CH in another cluster. Otherwise, it randomly selects a sensor node in the border to ask for a path to the BS [27].

The simulated result in Figure 10 exposed the considerable variation in the network lifetime for the LEACH sink-mobile protocol on two simulators: Mobi-Sim and TOSSIM. The energy consumption in the WSN using Mobi-Sim is greater because mobile sinks stay active as a whole and can be followed even if they move. This was not the case in TOSSIM which loses the control on mobile entities when they change their position.

The main metric in a routing protocol is its ability to find itineraries to the base station in order to deliver collected data from sensor nodes to it. On failure sensor nodes try to resolve this problem by retransmission. But, because TOSSIM don't support the presence of mobile entities in the network, any moving node will be considered as died node. For thus, sensor nodes could not locate the sink and the sink could not communicate with other nodes. So, this is why the number of exchanged messages is less in TOSSIM than Mobi-Sim as shown in figure 11.

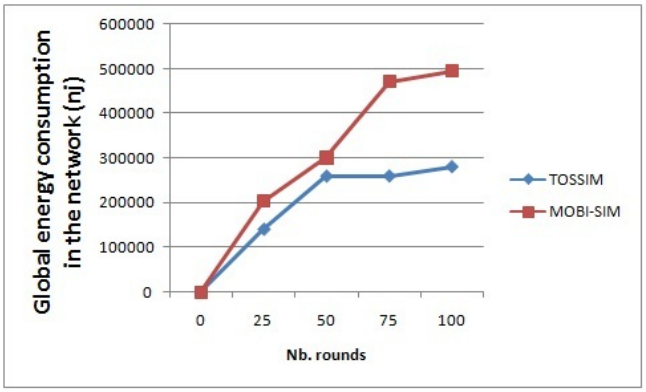


Figure 10: Energy consumption by LEACH sink-mobile

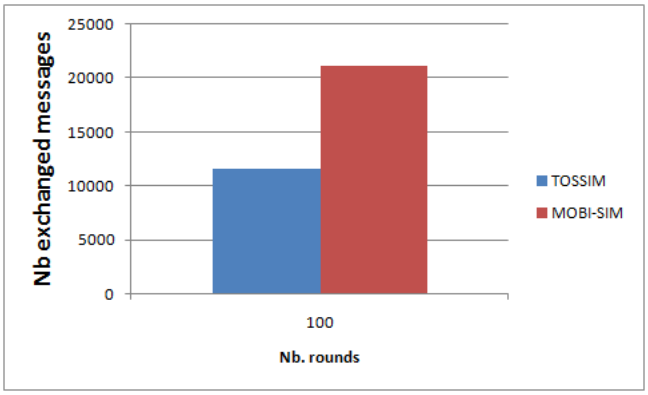


Figure 11: Exchanged packets number with LEACH sink-mobile

Therefore, since the mobile sink is considered as died node when changing its position, sensor nodes try to find it by retransmission and then they are likely to deplete their energy resource which may decrease the connectivity rate over the time and increase data loss packets. Inversely, Mobi-Sim supports the presence of mobile nodes and could follow it even if it is moving. This is explained in Figure 12 which show more connectivity rate respectively in Figure 13 which show less loss data packets rate in Mobi-Sim than TOSSIM.

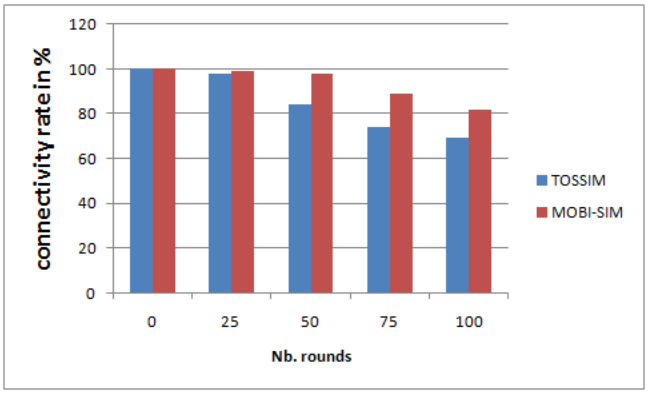


Figure 12: Connectivity rate by LEACH sink-mobile

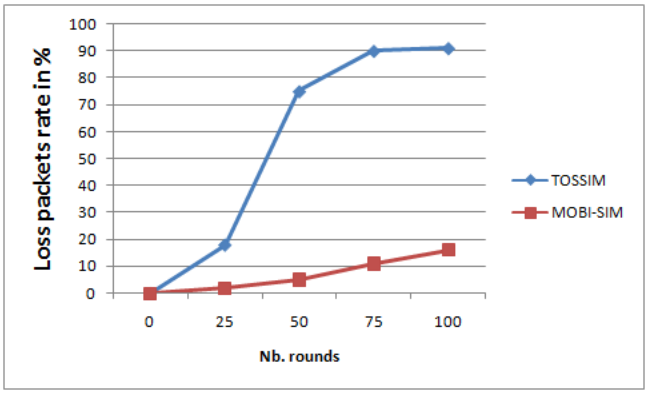


Figure 13: Loss packets rate in LEACH sink-mobile

4.3 Proposed Approach Evaluation

The proposed hierarchical routing protocol is a protocol based on a dynamic partitioning of the network into a set of clusters. Our goal is to ensure optimal operation in the network by minimizing the energy consumption and the delays when routing the information. This protocol is dedicated to operate with mobile sensors network and to support the frequent topology change.

To implement the proposed routing protocol, we make the following assumptions:

- The network uses N homogeneous nodes deployed randomly.
- The base station is fixed.
- There are few fixed nodes as reference points and a large number of mobile sensor nodes.
- In each cluster, the new CH election is carried out by the reference points periodically according to the energy for which the residual energy of each node within the same cluster is maximum and to the mobility for which the CH should have a low mobility rate.

This protocol implementation involves three phases: Clusters formation, Cluster heads election and data transmission.

- *Cluster formation*: During this phase, the reference points announce a new round to start the organization process. The different reference points broadcast an announcement to launch the clustering process. According to their adjacency, sensor nodes are organized into clusters.
- *CH election phase*: Once the clusters are formed, each reference point elects a coordination center role for each group. In each group, the transmission of information should be from member nodes toward the CH. This CH is elected based on energy criteria i.e if it has the maximum amount of energy and based on mobility if it has a low mobility rate compared to other members within a cluster.
- *Transmission phase*: The CH collects and aggregates data in order to transmit it to its corresponding anchor. Then, the anchor transmits its data packets directly to the sink or via other anchors. In this protocol, data processing in each cluster is done locally and the role of each CH is to coordinate the exchanges with other members. Each node has the possibility to be elected as CH and inversely, each CH can become a simple member that can belong to a cluster.

The operating mode of a sensor node in our proposed routing protocol is shown in Figure 14.

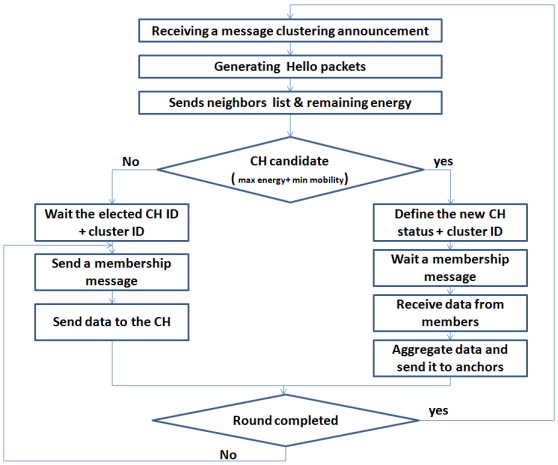


Figure 14: proposed routing protocol implementation

The proposed approach is the third protocol which is simulated in both Mobi-Sim and TOSSIM simulators. Like the previous results, Figure 15 show that the number of the exchanged packets is higher in Mobi-Sim than TOSSIM because unlike TOSSIM it can model and reproduce the behavior of mobile nodes.

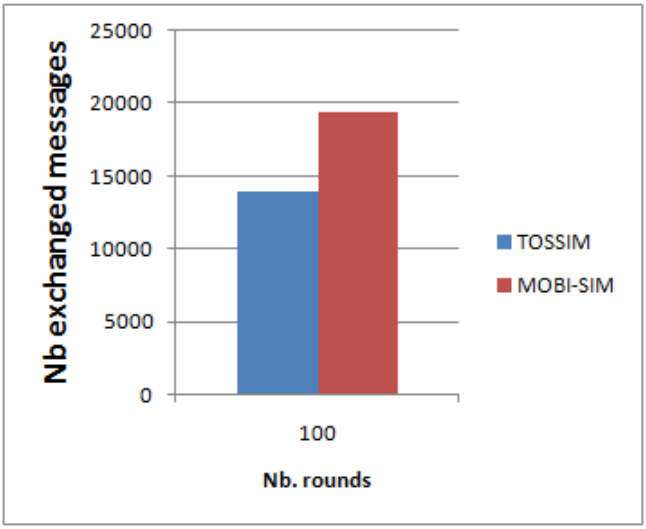


Figure 15: Exchanged packets number of the proposed approach

Figure 16 shows the average of consumed energy in the network over time. Since the number of exchanged packets in Mobi-Sim is greater than TOSSIM, the consumed energy is also higher for this simulator.

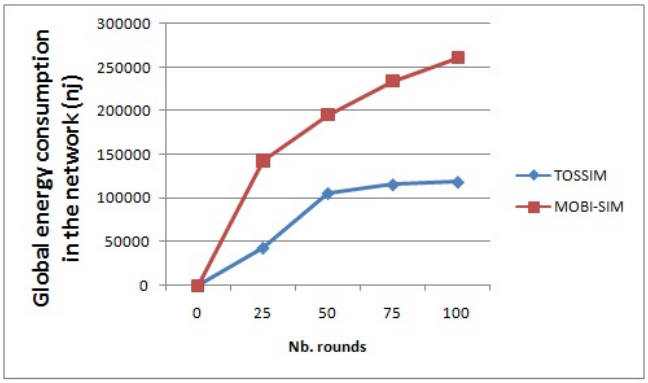


Figure 16: Energy consumption of the proposed approach

Like other protocols, TOSSIM cannot follow mobile nodes running a routing method dedicated for mobile WSN. In fact, the obtained results in Figure 17 show that the loss packets rate is greater in TOSSIM than Mobi-Sim. As TOSSIM loses the control on mobile nodes, the number of non-receiving nodes is increased. Then, the lost packets rate is increased in TOSSIM than Mobi-Sim.

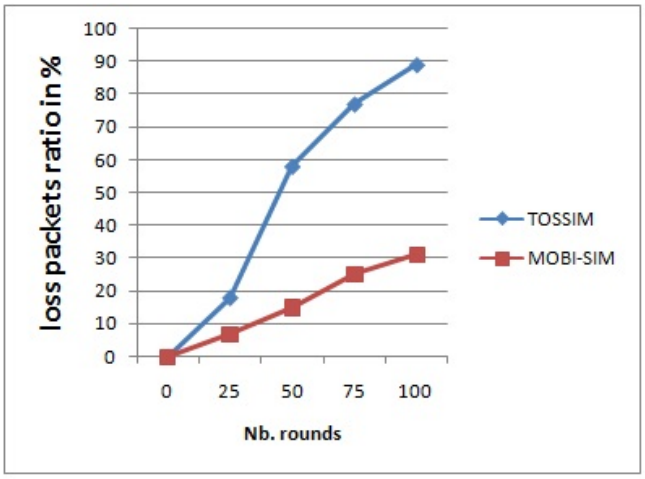


Figure 17: Loss packets rate of the proposed approach

The same for connectivity rate, Mobi-Sim supports the presence of mobile nodes and could track it even if it change its initial position. Thus, Figure 17 shows more connectivity rate in Mobi-Sim than TOSSIM. This justifies the previous results.

Therefore, to evaluate the performance of our proposed approach based on clustering, we simulate it and LEACH-M together on Mobi-Sim. Figure 19 shows the global energy consumption by sensor nodes within a mobile WSN with respect to different rounds. The obtained results show that LEACH-M do not provide an efficient energy consumption. The reasons behind this is that LEACH-M is a centralized protocol in which the base station is responsible for clusters formation, CH election and sensor nodes organization. However, our proposed protocol is distributed i.e. sensors nodes communicate together to perform sensors organization and clusters formation. Therefore, our approach which is multi-hop routing method, LEACH-M is a mono-hop protocol and CHs should communicate their data packets via long distances. So, it is evident from Figure 19 that there is a significant difference in the energy consumption between LEACH-M and our proposed protocol in the network, which directly impacts on the performance of the network or network lifetime.

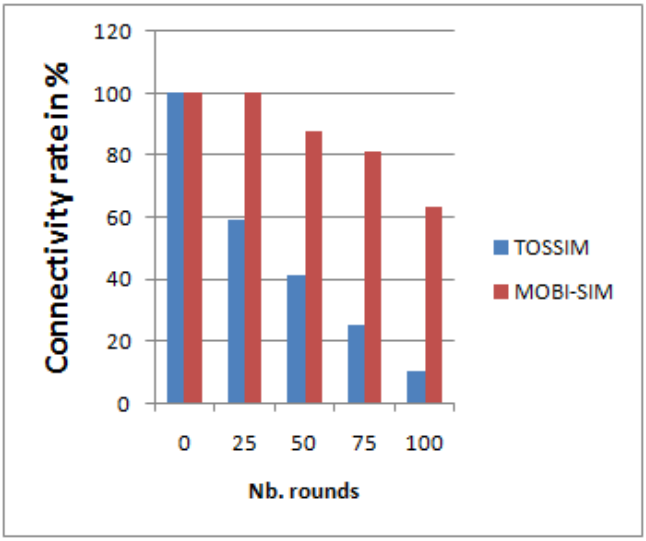


Figure 18: Connectivity rate of the proposed approach

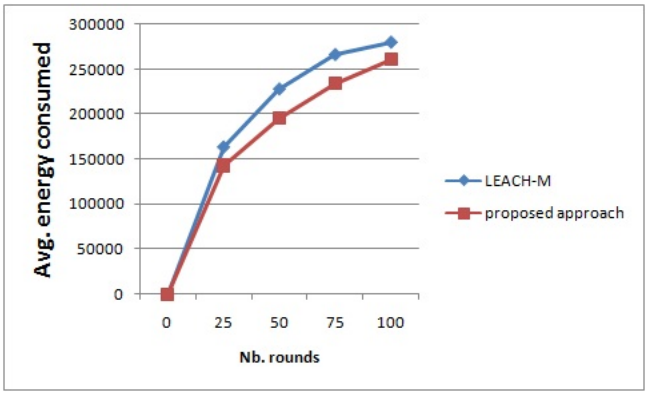


Figure 19: Sensors nodes energy consumption of our proposed protocol compared to LEACH-M

Figure 20 shows the loss packets rate. The obtained results show that the proposed protocol outperforms LEACH-M in terms of data delivery ratio. This is mainly due to the efficient CH election method which is based on the combination of two criteria: remaining energy and low mobility. In fact, sensor node with low mobility rate and maximal amount of energy in its battery is elected as CH. But, LEACH-M is elected initially basing on the remaining energy and on a generated random number compared to a probability threshold. So, data packets in our approach are more likely to achieve the base station than in LEACH-M.

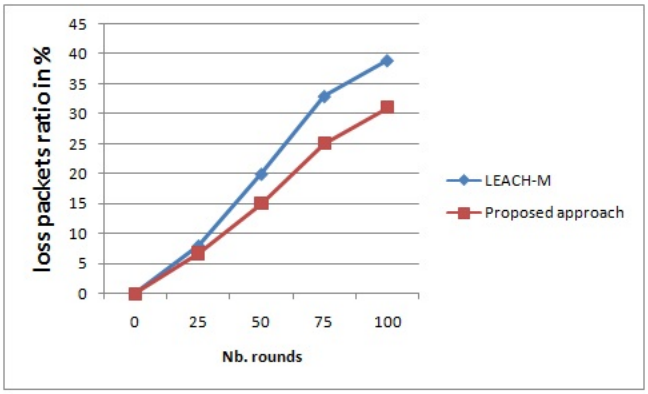


Figure 20: Loss packets rate of the proposed approach compared to LEACH-M

Figure 21 shows the connectivity rate between sensor nodes over time. The obtained results show that the number of active nodes over time is higher in our proposed approach than LEACH-M. This is behind many reasons: First, LEACH-M is based on negotiation i.e. when the CH fails to recover sent data by a sensor node in its time slot, it sends a request for two data frames then it considers it as mobile node. So, an extra traffic will be added which give chance for energy depletion which may quicker disable sensor nodes. But, in the proposed approach, the elected CH broadcasts an announcement message to inform neighbors by its state and then waits for joining message and data packets. Second, CH in the proposed protocol is elected if it has a low mobility. This offers a more efficient method.

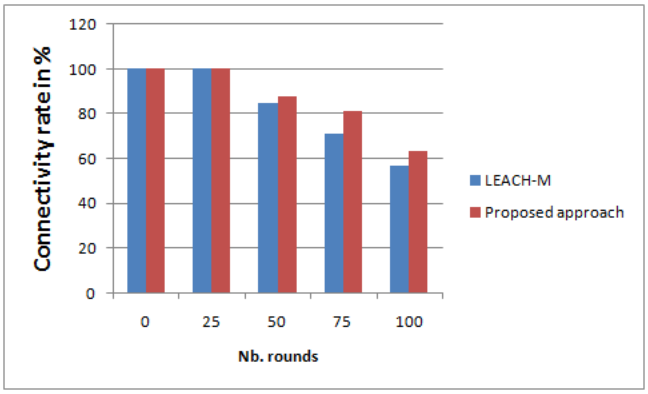


Figure 21: Connectivity rate of the proposed approach compared to LEACH-M

By all the performed simulations discussed above, the proposed simulator Mobi-Sim outperforms TOSSIM to support and to well reproduce the mobile sensor nodes behavior. Therefore, it efficiently model the loss data packets, the connectivity, the energy consumption and the exchanged data packets in real time. This simulator has also shown its efficiency when we implemented in previous work the flooding protocol as well two other popular routing protocols: LEACH-M and LEACH sink-mobile. Therefore, the proposed routing protocol based on clustering has demonstrated its performance compared to LEACH-M in terms of network lifetime, energy consumption, loss packets rate and sensor nodes connectivity.

5 Conclusion

The main objective of this paper is to provide a new simulation framework that can implement different WSN routing protocols to deliver the most accurate result. We adopt a model for mobile WSN that is able to reproduce mobile sensors behavior by implementing this sensor on a laptop. It is offered to researchers to evaluate the performance of their proposed approach before any implementation in a real system. We tested our model on two different popular routing protocols for better understanding. Simulation results have demonstrated that Mobi-Sim is suitable for mobile WSN and is open to add new approaches. Therefore, we proposed and simulated on Mobi-Sim a new routing protocol based on clustering. This proposition has demonstrated its efficiency in term of energy consumption and connectivity between sensor node as well as it demonstrated its robustness against loss packets.

Programming model is a huge step to deploy more applications in WSNs domain. There are currently many research works that addresses several parameters related to the sensors, such as: their positioning, energy, processing and data routing. We plan also to propose a routing protocol that supports nodes mobility and providing theirs localizations.

Acknowledgment This work has been accomplished at WIMCS-Team research, ENETCOM; Sfax University. Part of this work has been supported by MESRSTIC Scientific Research Group- Tunisia.

References

1. Pr. M. M. Abdellaoui, "Multitasks-Generic-Intelligent-Efficiency-Secure WSNs and theirs Applications", LAMBERT Academic Publishing (LAP), 2017, Part 2: Intelligent Wireless Sensors Network, pp. 108-141.

2. M. Mezghani, R. Gatgout, G. Ellouze, A. Grati, I. Bouabidi, M. Abdellaoui, "Multitasks generic platform via WSN", International Journal of Distributed and Parallel Systems (IJDPS), vol. 2, No. 4, July 2011, pp. 54-67.

3. O. Mezghani, M. Abdellaoui, "An Efficient Localization Method For Mobile Nodes Tracking in Wireless Sensors Network", IEEE International Conference on Communications, May 21-25, 2017, Paris. (submitted at publication).

4. O. Mezghani, M. Abdellaoui, "Mobi-sim: A Simulation Environment for Mobile Wireless Sensors Network", 3rd International Conference on Control-Decision and Information Technologies (CODIT2016), April 6- 8, 2016 , Malte.

5. S. P. Singh, S.C. Sharma, "A Survey on Cluster Based Routing Protocols in Wireless Sensor Networks", International Conference on Advanced Computing Technologies and Applications (ICACTA), March 26-27, 2015 , Mumbai, India.

6. G. Leelavathi, K. Shaila, K. R. Venugopal, L. M. Patnaik, "Design Issues On Software Aspects And Simulation Tools For Wireless Sensor Networks", International Journal of Network Security & Its Applications (IJNSA), vol.5, No.2, March 2013, pp. 47-64.

7. D. Navarroa, W. Dua, F. Mieyevillea, L. Carrela, "Hardware and Software System-Level Simulator for Wireless Sensor Networks", International Journal of Procedia Engineering, vol. 5, 2010, pp. 228-231.
DOI:10.1016/j.proeng.2010.09.089.

8. W. Du, F. Mieyeville, D. Navarro, I. O. Connor, "IDEA1: A validated SystemC-based system-level design and simulation environment for wireless sensor networks", EURASIP Journal on Wireless Communications and Networking, No. 1, 2011, pp. 1-20.
DOI: 10.1186/1687-1499-2011-143.

9. I. A. Qutaiba, A. Abdulmaowjod, M. M. Hussein, "Simulation & Performance Study of Wireless Sensor Network (WSN) Using MATLAB", Iraq Journal of Electrical and Electronic Engineering, vol. 7, No. 2, 2011, pp. 112-119.

10. K. Satish,S. J. Rane, D. Vishwakarma, "A Simulation Study of Behaviour of Wireless Motes With Reference To Parametric Variation", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, vol. 1, No. 2, August 2012, pp. 91-95.

11. B. Musznicki, P. Zwierzykowski, "Survey of Simulators for Wireless Sensor Networks ", International Journal of Grid and Distributed Computing, vol. 5, No. 3, 2012, pp. 23-49.

12. P. Levis, S. Madden, J. Polastre, R. Szewczyk, K. Whitehouse, A. Woo, D. Gay, J. Hill, M. Welsh, E. Brewer, D. Culler, "Tinyos : An operating system for sensor networks", Werner Weber, Jan M. Rabaey, Emile Aarts, editors, Ambient Intelligence, 2005, pp. 115-148.

13. P. Levis, N. Lee, M. Welsh, D. Culler, "Tossim : accurate and scalable simulation of entire tinyos applications", 1st international conference on Embedded networked sensor systems (SenSys 03), 2003, New York, NY, USA.

14. V. Shnayder, M. Hempstead, B. Chen, G. W. Allen, M. Welsh, "Simulating the power consumption of large-scale sensor network applications", 2nd international conference on Embedded networked sensor systems (SenSys 04), November 3-5, 2004, New York, NY, USA.

15. H. Sundani, H. Li, V. K. Devabhaktuni, M. Alam, P. Bhattacharya, "Wireless Sensor Network Simulators. A Survey and Comparisons", International Journal Of Computer Networks (IJCN), vol. 2, No. 5, 2011, pp. 249-265.

16. T. Voigt, R. Sauter, V. Reynolds, "Towards Comparable Simulations of Cooperating Objects and Wireless Sensor Networks", 1st International Workshop on Performance Methodologies and Tools for Wireless Sensor Networks, October 23, 2009, Pisa, Italy.

17. G. Dimitriou, P. Kikiras, G. Stamoulis, I. Avaritsiotis, "A tool for calculating energy consumption in wireless sensor networks", Advances in Informatic, 2005, pp. 611-621.

18. G. Mathur, P. Desnoyers, P. Chukiu, D. Ganesan, P. Shenoy,"Ultra-low power data storage for sensor networks", Journal ACM Transactions on Sensor Networks (TOSN), vol. 5, No. 4, 2009, pp. 96-101.

19. G. S. Kumar , V. Paul, K. P. Jacob, "Mobility Metric based LEACH-Mobile Protocol", 16th International Conference on Advanced Computing and Communications (ADCOM 2008), December 14-17, 2008.

20. A.Shahzad, B.Sardar, "An Intelligent Routing protocol for VANETs in City Environments", 2nd International Conference on Computer, Control and Communication (IC4), Febrary 17-18, 2009, Pakistan.

21. K. Chan, H. PishroNik, F. Fekri, "Analysis of Hierarchical Algorithms for Wireless Sensor Network Routing Protocols", Wireless Communications and Networking Conference, IEEE, March 13-17, 2005.

22. X. Huang, H. Zhai, Y. Fang, "Lightweight Robust Routing in Mobile Wireless Sensor Networks", Military Communications Conference (MILCOM 2006), IEEE, October 23-25, 2006.

23. Y. Miao, H. Jason, L. Renato, "Mobility Resistant Clustering in Multi- Hop Wireless Networks ", Journal of Networks, vol. 1, No. 1, May 2006, pp. 12-19.

24. S. Madani, D. Weber, S. Mahlknecht, "A Hierarchical Position Based Routing Protocol for Data CentricWireless Sensor Networks", 6th IEEE International Conference on Industrial Informatics (INDIN 2008), July 13-16, 2008, Korea.

25. F. Tashtarian, A. Haghighat, M. Honary, H. Shokrzadeh, "A New Energy-Efficient Clustering Algorithm for Wireless Sensor Networks", 15th International Conference SoftCOMon Software, Telecommunications and Computer Networks, September 27-29, 2007, Portsmouth, UK.

26. X. Zou, B. Ramamurthy, S. Magliveras, "Routing techniques in wireless ad hoc networks classification and comparison", 6th world multiconference on systemics, cybernetics, and informatics (SCI 2002), July 14-18, 2002, Orlando, Florida, USA.

27. O. Mezghani, M. Abdellaoui, "Improving Network Lifetime with Mobile LEACH Protocol for Wireless Sensors Network", 15th International conference on Sciences and Techniques of Automatic control & computer engineering (STA2014), December 19-21, 2014, Hammamet, Tunisia.

28. G. Renugadevi, M. G. Sumithra, "An Analysis on LEACH-Mobile Protocol for Mobile Wireless Sensor Networks", International Journal of Computer Applications, vol. 6, No. 21, March 2013, pp. 38-42.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 121-126 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

Expression of balance function during exposure to stereoscopic video clips

Fumiya Kinoshita[1], Masaru Miyao[2], Masumi Takada[3], Hiroki Takada[4*]

[1] *Institute of Innovation for Future Society, Nagoya University, 464-8601, Japan*

[2] *Graduate School of Information Science, Nagoya University, 464-8601, Japan*

[3] *Faculty of Nursing and Rehabilitation Department, Chubu Gakuin University, 501-3993, Japan*

[4] *Department of Human & Artificial Intelligence Systems, Graduate School of Engineering, University of Fukui, 910-8507, Japan*

ARTICLE INFO

Article history:
Received: 20 December, 2016
Accepted: 21 January, 2017
Online: 28 January, 2017

Keywords :
Visually Induced Motion Sickness,
Stabilometry,
Stereoscopic Image,
Stochastic Differential Equation.

ABSTRACT

Recently, with the rapid progress in image processing and three-dimensional (3D) technologies, stereoscopic images are available not only on television but also in theaters and gaming consoles. Asthenopia and visually induced motion sickness (VIMS) are well-known phenomena experienced by users while viewing video clips, playing immersive video games, and other such activities. In previous studies, we showed evidence that peripherally viewing plays a role in the pathogenesis of VIMS and described the anomalous sway with the use of mathematical models. Stochastic differential equations are known to be a mathematical model of body sway. In this study, we discuss the evolution in potential functions to control the standing posture during exposure to stereoscopic video clips.

1. Introduction

Current 3D display systems include stereoscopy, integral photography, the differential binocular vision method, volumetric display, and holography [1]. With the rapid progress in image processing and stereoscopic technologies, three-dimensional (3D) images have become available not only on television but also in theaters, on game machines, and elsewhere. Unpleasant symptoms such as asthenopia, dizziness, and nausea have been observed in some individuals viewing 3D videos [2]. While the symptoms of general motion sickness include dizziness and vomiting, the phenomenon of visually-induced motion sickness (VIMS) is not fully understood. Currently, there is insufficient knowledge regarding the effects of stereoscopic images on the living body, and basic research is thus important [3].

Contradictory messages originating from different sensory systems, or the absence of a sensory message that is expected in a given situation, are thought to lead to a feeling of sickness. Spatial localization of self-becomes unstable and produces discomfort. Researchers agree that there is a close relationship between the vestibular and autonomic nervous systems both anatomically and electrophysiologically. Motion sickness is considered to be caused by excess signals from the vestibular nuclei to the hypothalamus. This strongly indicates that the equilibrium system is associated with the symptoms of motion sickness [4] and provides a basis for the quantitative evaluation of motion sickness based on body sway, an output of the equilibrium system.

Stabilometry is a useful test of body equilibrium for investigating the overall equilibrium function. Stabilometry methods are presented in the standards of the Japanese Society for Equilibrium Research and international standards [5]. Stabilometry is a simple test in which 60-second recording starts when body sway stabilizes [6]. Objective evaluation is possible by computer analysis of the speed and direction of the sway, enabling diagnosis of a patient's condition [7].

In previous studies, subjective exacerbation and deterioration of equilibrium function were observed after peripheral viewing of a 3D video clip [8, 9]. A persistent influence has been observed while subjects view a poorly depicted background element peripherally, which generates depth perception that contradicts daily life. In this study, we examined the effect of viewing 3D video clips on equilibrium function systems using mathematical analysis.

*Hiroki Takada, *3-9-1 Bunkyo, Fukui, Fukui 910-8507, Japan*,
Tel: +81-776-27-8795
E-mail: takada@u-fukui.ac.jp

2. Material and Method

Sixteen healthy male subjects (mean age ± standard deviation: 22.2 ± 0.7 years) participated voluntarily in the study. We ensured that the body sway was not affected by environmental conditions. We used an air conditioner to adjust the temperature to 25 °C in the exercise room. The experiment was explained to all subjects and written informed consent was obtained in advance.

Experiments were performed in a dark room to avoid irritation from sources other than the video. 2D or 3D video clips were shown on a 3D display (KDL 40HX80R; SONY) placed 2 meters away from the subject. In the video clip used in the experiment, a sphere moved around the screen (Figure 1). A comparison was then made with subjects who were asked to simply gaze at a point 2 meters in front of them at eye level where no video clips were displayed. The experiments were carried out in random order. Each experiment was carried out on a separate day.

The subjects stood without moving on the detection stand of a stabilometer (Wii Balance Board; Nintendo) in the Romberg posture with their feet together for 30 seconds before the sway was recorded. Each sway of the center of pressure (COP) was then recorded at a sampling frequency of 20 *Hz* during the measurement; subjects were instructed to maintain the Romberg posture for the 60 seconds. The *x*-*y* coordinates were recorded with subjects' eyes open for each sampling time, and stabilograms were constructed from the time sequences. The following quantitative indices were calculated:

- Area of sway: Area of a region surrounded (enveloped) by the circumferential line of sway on the *x*-*y* coordinates. An increase in the value represents a more unstable sway;
- Total locus length: Total extended distance of movement of the COP within the measurement time period. An increase in the value represents a more unstable sway;

The area of sway and total locus length are analytical indices of stabilograms that were used in previous studies. We used these based on the definitions established by the Japanese Society for Equilibrium Research.

Figure 1: The video clip shown to the subjects on a 3D display.

3. Mathematical Models of Body Sway

In stabilograms, variables *x* (right designated as positive) and *y* (anterior designated as positive) are regarded to be independent [10]. A linear stochastic differential equation (Brownian motion process) has been proposed as a mathematical model to describe body sway [11, 12, 13]. To describe the individual body sway, we show that it is necessary to extend the following nonlinear stochastic differential equations:

$$\frac{\partial x}{dt} = -\frac{\partial}{\partial x}U_x(x) + \mu_x w_x(t), \quad (1)$$

$$\frac{\partial y}{dt} = -\frac{\partial}{\partial y}U_y(y) + \mu_y w_y(t), \quad (2)$$

where $w_x(t)$ and $w_y(t)$ are pseudorandom numbers produced by white Gaussian noise [14]. The following formulas describes the relationship between the distribution in each direction, $G_x(x)$ and $G_y(y)$, and the temporal averaged potential constituting the stochastic differential equations (SDEs):

$$U_x(x) = -\frac{\mu_x^2}{2}\ln G_x(x) + const., \quad (3)$$

$$U_y(y) = -\frac{\mu_y^2}{2}\ln G_y(y) + const. \quad (4)$$

The variance of stabilograms depends on the temporal averaged potential function (TAPF) with several minimum values when it follows the Markov process without abnormal dispersion. SDEs can represent movements within local stability with a high-frequency component near the minimal potential surface, where a high density at the measurement point is expected.

4. Numerical Analysis

A histogram of each stabilogram was obtained. The mean of each stabilogram was set to be (0, 0) by statistical processing. We compared histograms that were composed of all subjects' stabilograms with eyes open. The TAPFs while viewing 2D and 3D video clips were determined from the histograms using Eq. (5, 6). The TAPFs were regressed by the following degree 4 polynomial (Figures 2).

$$\widehat{U_x}(x) = a_x x^4 + b_x x^3 + c_x x^2 + d_x x \quad (5)$$

$$\widehat{U_y}(y) = a_y y^4 + b_y y^3 + c_y y^2 + d_y y \quad (6)$$

The following SDEs were derived from the mathematical model of the body sway (1, 2) into which was substituted Eq. (5, 6).

$$\frac{\partial x}{dt} = -(4a_x x^3 + 3b_x x^2 + 2c_x x + d_x) + \mu_x w_x(t) \quad (7)$$

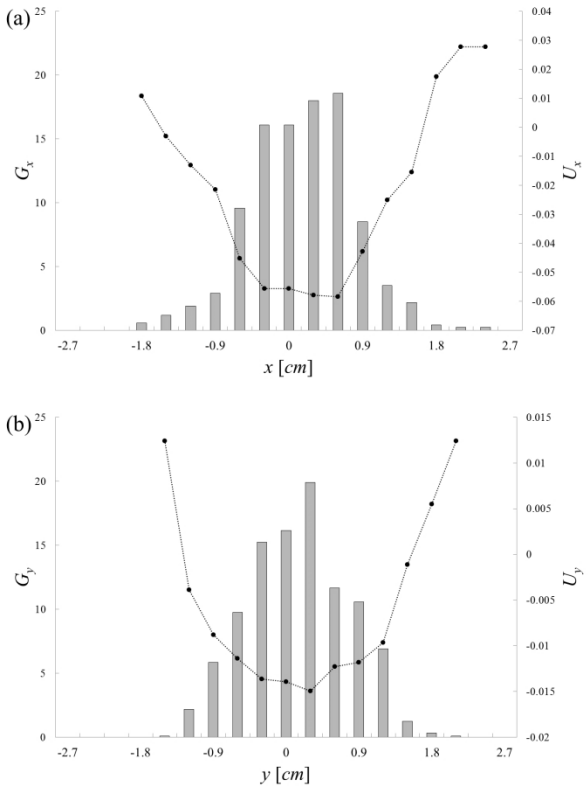


Figure 2: A typical example of TAPF derived from stabilograms during exposure to the 3D video clip: *x* direction (a), *y* direction (b).

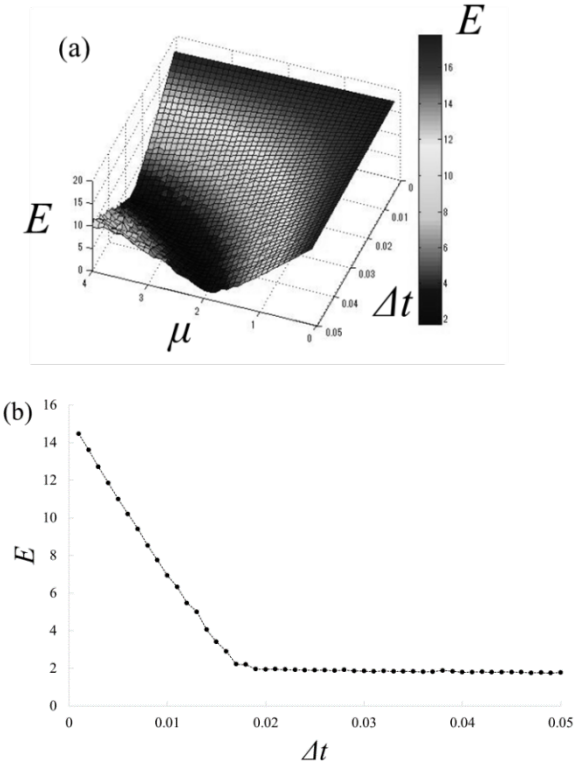


Figure 3: A typical evaluation, *E*, for numerical solutions to the mathematical model for Control with the following parameters: (μ, Δt) (a), Δt (b).

$$\frac{\partial y}{dt} = -(4a_y y^3 + 3b_y y^2 + 2c_y y + d_y) + \mu_y w_y(t) \quad (8)$$

Here, the constants μ_x and μ_y represent the noise amplitudes. We rewrote Eq. (7, 8) into a difference equation and obtained numerical solutions with the Runge-Kutta-Gill formula as the numerical calculus.

In Eq. (7, 8), the initial values of (x, y) were set to be (0, 0). Pseudorandom numbers (mean ± standard deviation: 1 ± 1) were generated to substitute for white Gaussian noise w. The noise amplitude μ and time step Δt were set to be 0.1, 0.2, …, 4.0 and 0.001, 0.002, …, 0.05, respectively. Numerical analysis was employed for 11,200 steps, and the first 10,000 steps of the numerical solutions were discarded due to dependence of the initial value. The area of sway (Xs) and the total locus length (Ys) were also calculated in these numerical solutions. Designating the measured the area of sway and the total locus length as (Xr) and (Yr), errors (E) between the numerical solutions of the mathematical model and measured data were defined as follows:

$$E = \sqrt{\frac{\sqrt{Yr}}{Xr}(Xr - Xs)^2 + (\sqrt{Yr} - \sqrt{Ys})^2}. \quad (9)$$

We then estimated the smallest E for each parameter μ and Δt (Figure 3a), and the variance of E on E-Δt plane as

$$var(\Delta t^*) = \frac{1}{\#data} \sum_{\Delta t^* \le \Delta t \le 0.05} \{E(\Delta t) - mean(E)\}^2, \quad (10)$$

was calculated to determine the plateau of the error E. The left bound Δt^* was defined as the optimum value for the plateau in the case that the variance (10) exceeded 0.3 in this study (Figure 3b). This parameter was calculated for each subject.

5. Result

Stabilograms measured during exposure to a 2D video clip were compared with those of a 3D video clip (Figures 4). We also calculated the area of sway and the total locus length for each stabilogram (Figures 5). These were compared by Wilcoxon signed-rank test with each index for the stabilometry. The significance level was set to be 0.10. The results of the Wilcoxon signed-rank test showed that the value of the total locus length while viewing the 3D video clip was significantly smaller than that while viewing the 2D video clip ($p<0.10$).

Histograms of each component were obtained from subjects' stabilograms in each condition. The TAPFs were constructed from the histograms using Eqs. (5, 6). The TAPFs were sufficiently regressed by polynomials of degree four. Eqs. (7, 8) were rewritten into difference equations, and the numerical solutions were obtained with the Runge-Kutta-Gill formula (Figures 6). The area of sway and total locus length were also calculated in these numerical solutions. The smaller the E, the better the description the numerical simulation could give (Figures 7). The value of errors E was evaluated with the control parameters Δt and μ in the abovementioned difference equations (Figures 8). The smallest Δt was obtained while viewing the 3D video clip ($p<0.10$) (Figure 8b). The largest μ was obtained while

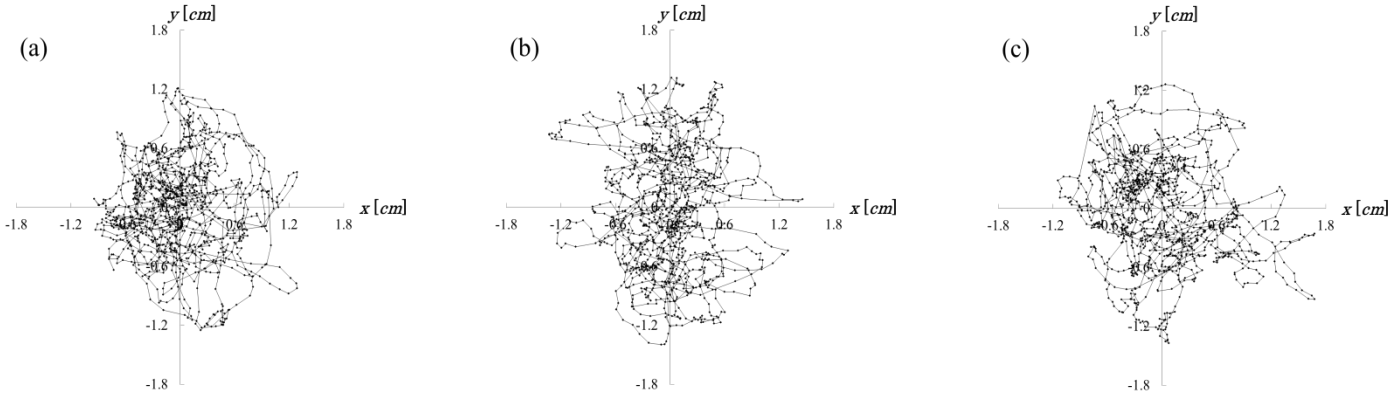


Figure 4: Typical stabilograms obtained from a subject: Control (a); the 2D (b); the 3D (c).

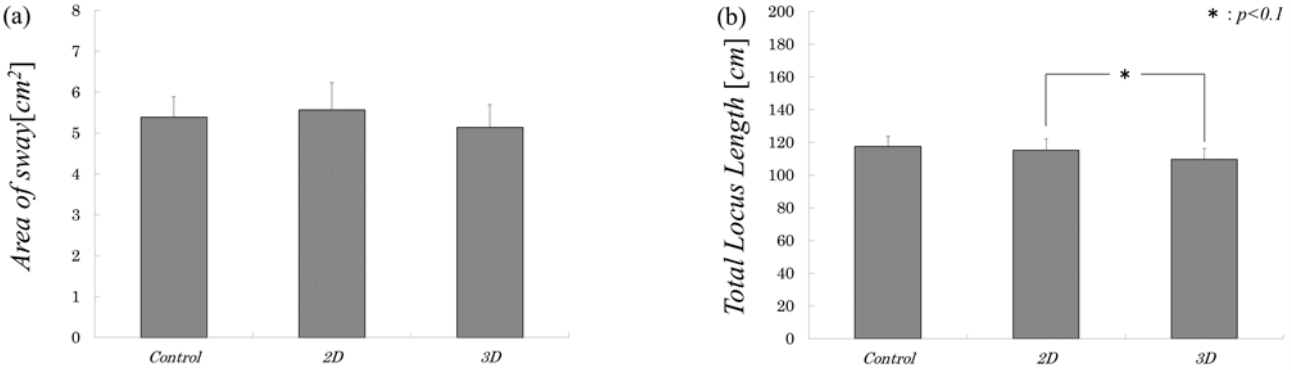


Figure 5: Results of sway values (average ± SE): the area of sway (a); the total locus length (b).

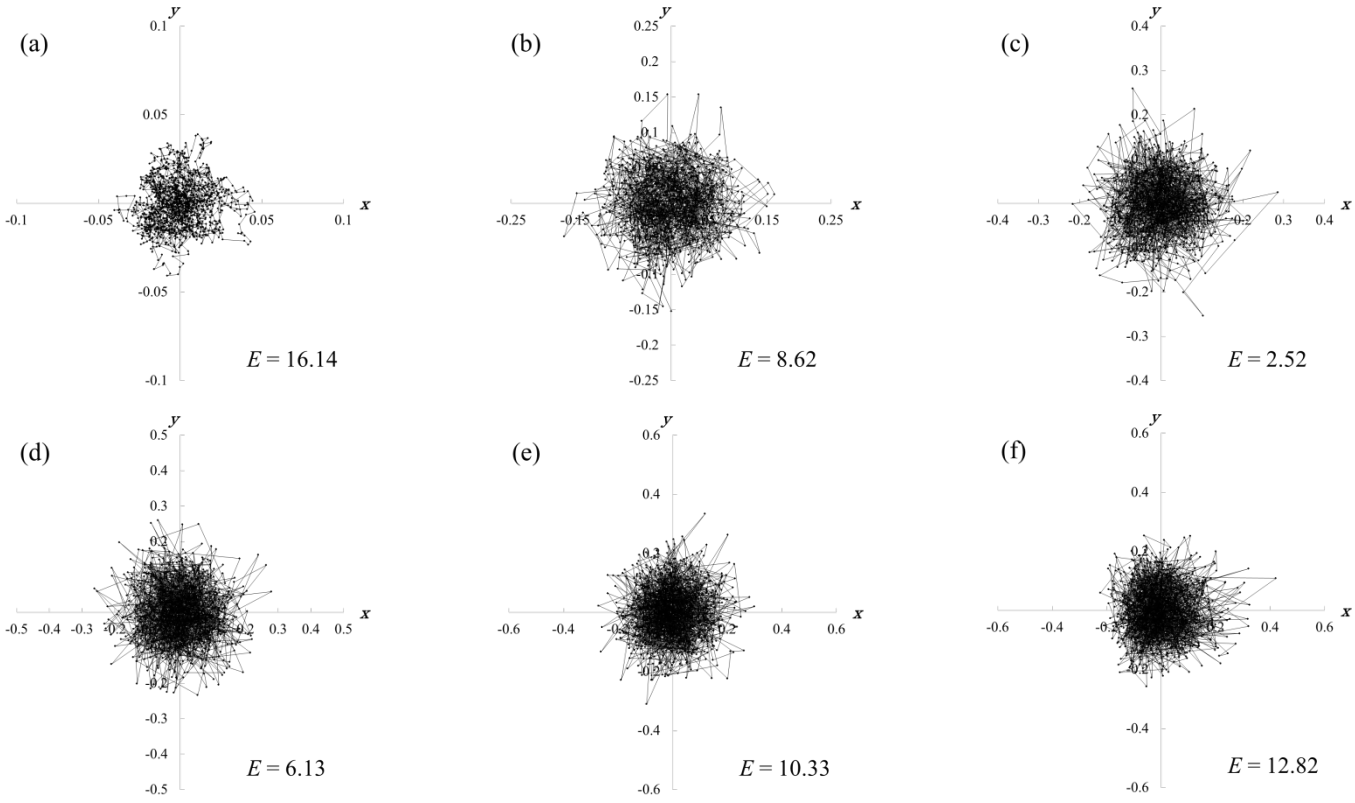


Figure 6: Typical examples of numerical solutions at $\mu = 3.9$ and the following time step:

$\Delta t = 0.001$ (a); $\Delta t = 0.01$ (b); $\Delta t = 0.02$ (c); $\Delta t = 0.03$ (d); $\Delta t = 0.04$ (e); $\Delta t = 0.05$ (f).

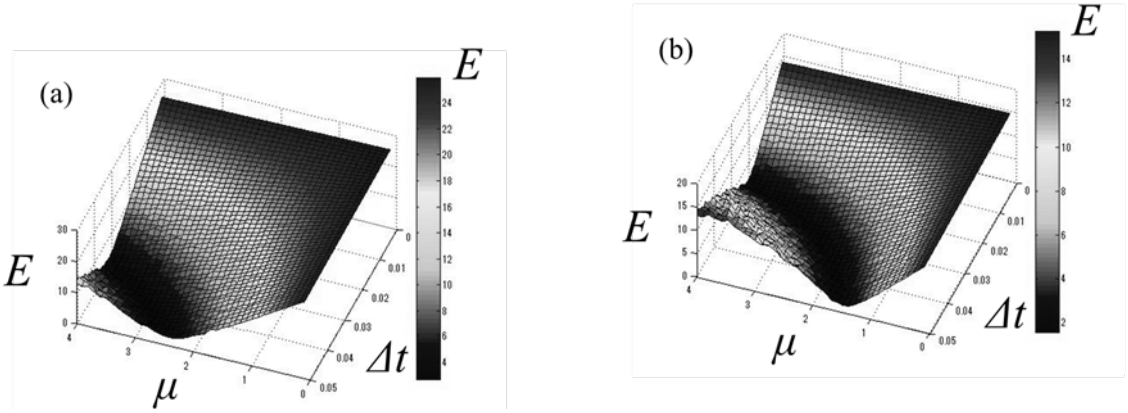


Figure 7: A typical evaluation for numerical solutions to the mathematical model while viewing the video clip: the 2D (a); the 3D (b).

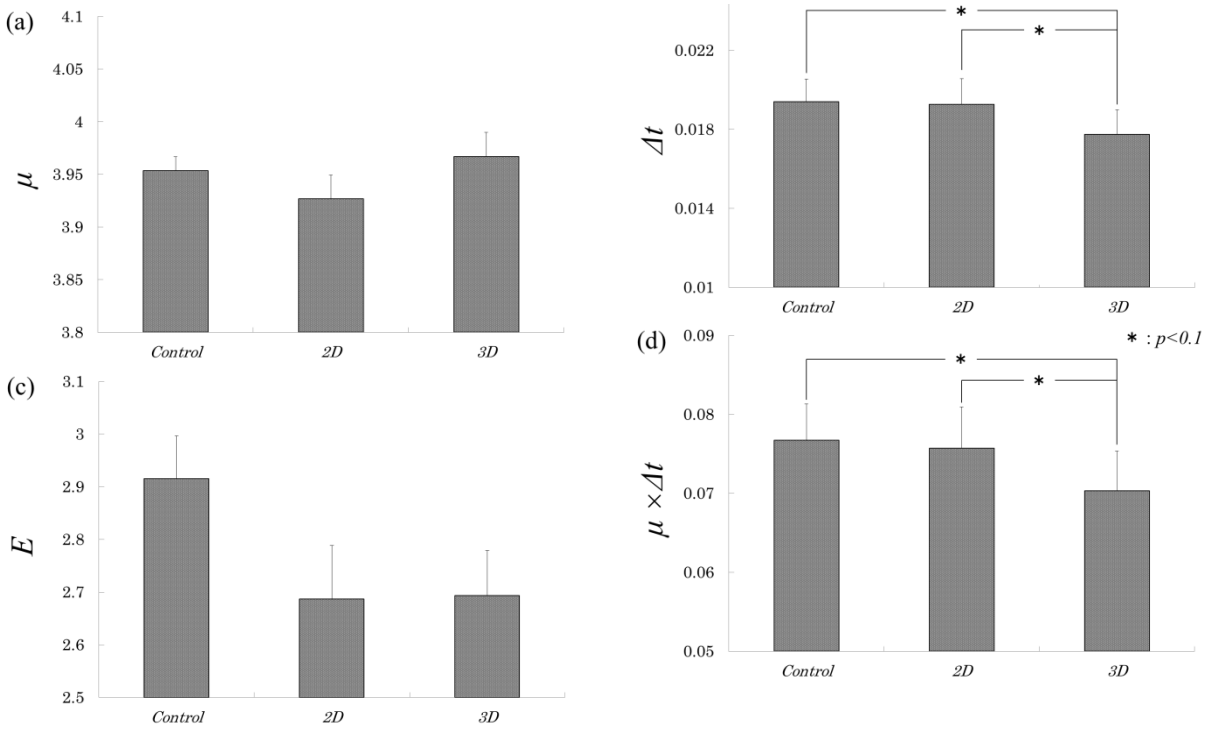


Figure 8: Results of optimum value (average ± SE): μ (a); Δt (b); E (c); $\mu \times \Delta t$ (d).

viewing the 3D video clip, and the smallest one was obtained while viewing the 2D video clip (Figure 8a). In this paper, $\mu \times \Delta t$ was evaluated for rigidity in the postural control. As shown in Figure 8d, the value of $\mu \times \Delta t$ while viewing the 3D video clip was smaller than in the other cases ($p<0.10$).

6. Discussion

Previously, we evaluated body sway by conducting stabilometry studies with simple use of the analytical indices for stabilograms. In this study, we further obtained TAPFs in the SDEs as mathematical models of body sway. We also examined whether there is any remarkable evolution in the potential functions to control standing posture during exposure to stereoscopic video clips.

The previous research showed that the sway values after viewing a 3D video clip are larger than those in other conditions, including after viewing the 2D. However, as a result of the Wilcoxon signed-rank test in this study, the value of the total locus length while viewing the 3D video clip was found to be significantly smaller than that while viewing the 2D video clip ($p<0.10$). Therefore, we proposed the mathematical model while viewing the video clips and evaluated the stabilograms by numerical analysis. In the numerical solution, the area of sway and the total locus length were proportional to the product $\mu \times \Delta t$. In this study, the results when Δt was around 0.019 and μ was around 3.9 showed values close to the sway values in the measurements, such as the area of sway and the total locus length.

A flexible system to control body sway reduces the product of $\mu \times \Delta t$. Fine postural control, which has been regarded as an anomalous system to control the upright posture, might be seen

before the outbreak of motion sickness. This anomalous process is considered to be a clue to elucidating the procedure of the motion sickness and predicting the time when the first symptom of motion sickness will appear. In the next step, we will extend the viewing time of the video and use a subjective questionnaire to examine the relationship between the VIMS and body sway.

We have also succeeded in showing that postural control can be evaluated by numerical analysis with our mathematical model in the case that differences cannot be seen in the stabilograms with use of previous indices such as the area of sway and total locus length. Using a mathematical model for the evaluation of the stabilograms may contribute to the elucidation of the postural control system.

7. Conclusion

Previously, we evaluated body sway by conducting stabilometry studies with simple use of the analytical indices for stabilograms. In this study, we obtained TAPFs in the SDEs as mathematical models of body sway. We also examined whether there is any remarkable evolution in the potential functions to control the standing posture during the exposure to stereoscopic video clips. As a result, we verified that 3D viewing effects on our equilibrium function can be seen with the use of a mathematical model.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was supported in part by the Japan Society for the Promotion of Science, Grant-in-Aid for Scientific Research (B) Number 24300046 and (C) Number 26350004.

References

[1] D. Gabor, “A New Microscopic Principle,” Nature, **161**, 777–779, 1948.

[2] International standard organization, “IWA3: 2005 Image Safety-Reducing Determinism in a Time Series,” Phys. Rev. Lett., **70**, 530–582, 1993.

[3] Y. Sumio, I. Shinji, “Visual Comfort and Fatigue Based on Accommodation Response for Stereoscopic Image,” The Institute of Image Information and Television Engineers, **55** (5), 711–717, 2001.

[4] N.H. Barmack, “Central Vestibular System: Vestibular Nuclei and Posterior Cerebellum,” Brain Res. Bull., 60, 511–541, 2003.

[5] Japan Society for Equilibrium Research, “Standard of Stabilometry,” Equilib. Res., **42**, 367–369, 1983.

[6] T.S. Kaptyen, W. Bles, C.J. Njiokiktjien, L. Kodde, C.H. Massen, J.M. Mol, “Standarization in Platform Stabilometry being a part of Posturography,” Agreessologie, **24**, 321–326, 1983.

[7] M. Hase, Y. Ohta, “Meaning of Barycentric Position and Measurement Method,” J. of Environ. Eng., **8**, 220–221, 2006.

[8] K. Yoshikawa, F. Kinoshita, K. Miyashita, A. Sugiura, T. Kojima, H. Takada, M. Miyao, “Effects of Two-Minute Stereoscopic Viewing on Human Balance Function,” In: Antona, M., Stephanidis C. (eds.) UAHCI/HCII 2015 Part II. LNCS, **9176**, 297–304, 2015.

[9] M. Takada, Y. Fukui, Y. Matsuura, M. Sato, H. Takada, “Peripheral Viewing during Exposure to a 2D/3D Video Clip: Effects on the Human Body,” Environ Health Prev Med., **20**, 79-89, 2015.

[10] P.A. Goldie, T.M. Bach, O.M. Evans, “Force Platform Measures for Evaluating Postural Control: Reliability and Validity,” Arch. Phys. Med. Rehabil., **70**, 510–517. 1986.

[11] R.E.A. Emmerrik, R.L.V. Sprague, K.M. Newell, “Assessment of Sway Dynamics in Tardive Dyskinesia and Developmental Disability,” Sway Profile Orientation and Stereotypy. Moving Dis., **8**, 305-314, 1993.

[12] J.J. Collins, C.J.D. Luca, “Open Loop and Closed-Loop Control of Posture: A Random-Walk Analysis of Center of Pressure Trajectories,” Exp. Brain Res., **95**, 308–318, 1993.

[13] K.M. Newell, S.M. Slobounov, E.S. Slobounova, P.C. Molenaar, “Stochastic Processes in Postural Center of Pressure Profiles,” Exp. Brain Res., **113**, 158–164, 1997.

[14] H. Takada, Y. Kitaoka, Y. Shimizu, “Mathematical Index and Model in Stabilometry,” Forma, **16**, 17–46, 2001.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 127-133 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

A Survey of Text Mining in Social Media: Facebook and Twitter Perspectives

Said A. Salloum[1,2*], Mostafa Al-Emran[3], Azza Abdel Monem[4], Khaled Shaalan[1]

[1]*Faculty of Engineering & IT, The British University in Dubai, UAE.*

[2]*University of Fujairah, UAE.*

[3]*Faculty of Computer Systems and Software Engineering, Universiti Malaysia Pahang, Malaysia.*

[4]*Faculty of Computer and Information Sciences, Ain Shams University, Egypt.*

ARTICLE INFO

Article history:
Received: 12 December, 2016
Accepted: 06 January, 2017
Online: 28 January, 2017

Keywords :
Text Mining
Social Media
Facebook
Twitter

ABSTRACT

Text mining has become one of the trendy fields that has been incorporated in several research fields such as computational linguistics, Information Retrieval (IR) and data mining. Natural Language Processing (NLP) techniques were used to extract knowledge from the textual text that is written by human beings. Text mining reads an unstructured form of data to provide meaningful information patterns in a shortest time period. Social networking sites are a great source of communication as most of the people in today's world use these sites in their daily lives to keep connected to each other. It becomes a common practice to not write a sentence with correct grammar and spelling. This practice may lead to different kinds of ambiguities like lexical, syntactic, and semantic and due to this type of unclear data, it is hard to find out the actual data order. Accordingly, we are conducting an investigation with the aim of looking for different text mining methods to get various textual orders on social media websites. This survey aims to describe how studies in social media have used text analytics and text mining techniques for the purpose of identifying the key themes in the data. This survey focused on analyzing the text mining studies related to Facebook and Twitter; the two dominant social media in the world. Results of this survey can serve as the baselines for future text mining research.

1. Introduction

As we know that there are various social networking sites available, Facebook and Twitter are considered as the most crowded ones [1], [2]. These networking sites have made it easy to communicate with friends and family members without making any much effort [3], [4]. People related to different values come closer to each other by sharing their ideas, interests, and knowledge [5]. These days, it becomes very easy for anyone to meet the people of their interests for learning and sharing precious information [6], [7].

The advancement in technology has shrunk the world. The distances look closer and sharing information looks easier [8]. Through these social networks, people can easily and confidently share their point of views [9], [10] regarding various global issues by uploading their posts, text comments and blogs [11]. A study by [12] claimed that social media including Google Apps facilitate the way people learn, collaborate, and share ideas with each other. Moreover, social media has been incorporated by several learning forms such as e-learning and m-learning [13], [14]. Whatever the scenario is, people don't like to use structured sentences, correct grammar and spellings [6]. Not matter, whether they are searching something on the site, posting any comment or connecting people through various discussion forums. People use irregular data patterns to convey their messages. It seems like they have a shortage of time but due to the use of this unstructured language, it is not an easy task to bring out the correct and regular data patterns. On different social networking sites, the most common method of interaction with each other is through text. People share their knowledge and information through blogs, posts, and chats by writing in their own languages. The basic use of the text mining methods is to make the text clear to make it easy for anyone to write or search in the most appropriate manner [15].

*Corresponding Author: Said A. Salloum, University of Fujairah, UAE.
Tel: +971507679647 Email: Salloum78@live.com
www.astesj.com
https://dx.doi.org/10.25046/aj020115

As people write words or sentences with errors, so in order to let them write or search with proper grammar and structured sentences, text mining approach [16] is used. Text mining means the extraction of the data which is not familiar to anyone. If we compare web searching with text mining then both the terms are vastly different from each other. If we talk about web searching, then you are fully aware of what you are going to search. But in the case of text mining, the main focus is to bring out the most appropriate data in accordance with the written text, no matter whether it is structured or not. This technique only requires a particular alphabet in order to dig out the data which is then further transformed into different suggestions and expectations. Text mining seems to grasp the entire automatic natural language processing. For instance, exploration of linkage structures, references in academic writing and hyperlinks in the Web writing are important sources of data that lie outside the conventional area of NLP. NLP is one of the hot topics that concerns about the interrelation among the huge amount of unstructured text on social media [17], besides the analysis and interpretation of human-being languages [18], [19].

Several research articles were collected from various databases in order to be analyzed and used in this survey. The search terms include "Text mining with social media", "Text mining with Facebook", and "Text mining with Twitter". This survey is categorized as follows: section 2 provides a complete background about the text mining field. Other related studies are addressed by section 3. Conclusion and future perspectives are presented in section 4.

2. Background

Businesses have identified data-driven approaches as the ideal blueprint for their growth. It is easier to understand this theory. After all, wouldn't it benefit a company to get an idea about the perception of its products in the market without having to consult individual reviews from everyone? Wouldn't it be better if they could gauge which political candidate is ideal for their public image without having to analyze them all individually? This is why market study and research are some of the most highly invested fields in the world right now. Social networking sites like Twitter and Facebook are ideal for this purpose. Posts or messages shared by people on these platforms with their friends remain freely accessible or are kept confidential. They give businesses the chance to scoop up public sentiments [20], [21] about topics that they are interested to share by a large group of people.

The processing of surveys and public impressions using specially designed computational systems is a shared objective of inter-connected fields like subjectivity analysis, opinion mining [22], and sentiment analysis. Creating problem-solving techniques or methods to define the structure and precedence or for summarizing opinionated messages for particular topics [23], occasions or products is another target of the survey. For example, these methods could be used for gauging support for particular occasions or items, or determining thumbs down or thumbs up votes for specific movies based on their reviews.

2.1. Text Mining

Text mining makes it easy to obtain a meaningful and structured data from the irregular data patterns [24], [25], and [26]. It is really not an easy task for the computers to understand the unstructured data [27], [28] and make it structured. Human beings can perform this task without any further efforts due to the availability of different linguistic techniques. However, human beings are limited in terms of speed and space as comparing to computers. That is, computers are much better than humans to do these tasks. Most of the existing data in any organization is represented in a text format, so if we compare data mining with text mining then text mining is more important [29]. But as text mining is used for structuring the unstructured text data then this task is more demanding as compared to data mining. In general, the data related to social media sites is not collected for the research purpose [30], it is mandatory to change the structure of the data coming from the social media. 80% of the available text on the web is unstructured while only 20% is structured [31].

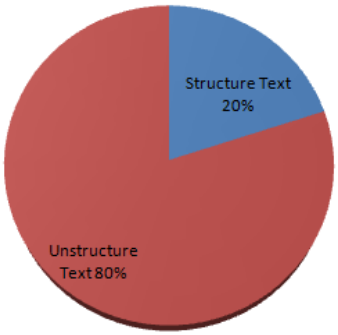


Figure 1: Text available on the web.

2.2. Text Mining vs. Data Mining

In the case of posting comments on any post on different social networking sites, there is not a single structured technique available which causes problems in the direct usage of the data. Data available in the text format has much more importance and that is why text mining is generating much business value [32]. A study by [33] stated that data mining represents the derivation of a meaningful pattern or principles from a spatial database for determining a particular issue or issues. Data mining is different from text mining [34]. A study by [35] pointed out that text mining is much more complex than data mining because it contains irregular and unstructured data patterns, whereas data mining is dealing with the structured sets of data. The tools that were used in data mining were only dealing with structured data [34]. Text mining is like an intelligence system which is extracting proper words or sentences from the improper words and then transforming those words into the particular suggestions. Text mining is basically a new field having the main purpose of data recovery, machine learning, information mining and computational linguistics [36].

2.3. Text mining in social networks

The importance of text mining has been increased due to the significant contributions in the field of technology. Data mining as reported by [37] is also important but due to the advancement,

text mining is taking its place. It is really a big effort to convey valuable information and knowledge [38] through powerful handling and mining processes from the irregular form of information. In this era, structured data has lost its importance and the unstructured data has gained the popularity. Most of the organizations are going towards text mining and forgetting the concept of data mining [39]. Scholars of [40] reported that all the social networking sites are providing a great space to individuals to facilitate interaction and share their views and opinions. The best thing which these sites are doing is that it has become easy for the individuals to understand a particular person depending upon his or her activities. Through all these activities, people related to different customs and values have come closer to each other because of having the better understanding of each other's emotions, perceptions and areas of interest. At this time, user interfaces are going to be equipped with personality based qualities [41]. Personalized designs were used in e-commerce [42], [43], e-learning, and information filtering for enhancing different styles and skills.

3. Text mining efforts in resolving various NLP issues

A study by [44] stated that text mining is responsible for structuring the irregular data patterns written in the human language. As most of the people interact with each other in the form of text so for those people who are not able to share structured form of data, text mining is the best technique to handle these situations. Among others, NLP is considered as the most amazing research field. The main goal of NLP is to seek information regarding how the computer systems are examining and getting information from the languages of human beings to create applications of high quality [17]. The art of sharing meaningful information with the help of uncommon and meaningless data is truly a good thing. Text mining technique as described by [45] examines the content for extracting the meaningful data which can be used for particular purposes. It looks like text mining that is going to include the overall NLP scheme [46] in its system in order to effectively examine the human language and to structure the unstructured data patterns accordingly. As the technology is advancing day by day, text mining system will get better and better and this is what all people are looking for.

3.1. Text mining in Facebook

The social networks are growing at a rapid rate without a break. Most importantly, the unstructured data is being stored on these networks as they act as a large pool and this data pertains to a host of domains containing governments, businesses, and health. Data mining techniques tend to transform the unstructured data for its placement within a systematic arrangement [47]. Nowadays, Facebook is one of the most popular social media. This media is used by a large number of people on earth for expressing their ideas, thoughts, sorrows, pleasures and poems [48]. Researchers had chosen a number of Facebook variables that were expected to develop the right situation for carrying out our investigations. The valuable statistics of user's personality is provided by the Facebook profiles and activities, which exposes the actual objects instead of projected or idealized character [49]. The digital data has currently witnessed an enormous growth. The key area of interest among professionals is now data mining and knowledge discovery. Moreover, a strong need has been felt to transform such data into useful knowledge and information. A number of applications like business management and market analysis have realized the benefits from the information and knowledge extracted out of large scale data. Information is stored in text form across various applications so one of the up-to-date areas for research is text mining. The hard issue is extracting the user required information. The knowledge discovery process has an important step which is believed to be the Text Mining. The hidden information is extracted from unstructured to semi-structured data in this process. Extracting information from a number of written resources and its automatic discovery is called as Text mining. Moreover, computers are also used for the needful and to meet this goal.

Scholars of [50] illustrated the text mining techniques, methods, and challenges. These successful techniques would be described to give usefulness over information acquisition during text mining. The study discussed the situations where each technology could be beneficial for a different number of users. A number of business organizations would be examined by mining data that has been exposed by their employees on LinkedIn, Facebook, and other openly available sources. A network of informal social connections among employees is extracted through web crawler developed for this purpose. According to the findings, leadership roles can be identified within the organization and this could be achieved absolutely by using machine learning techniques besides centrality analysis. Clustering the social network of an organization and collecting available information within each cluster can result in the valuable non-trivial perceptions. A key asset or a considerable threat to the primary organization can be the knowledge about the network of informal relationships. Besides analyzing social networks of the organizations, algorithms and methods used to gather data from freely available sources would be presented by this paper. A web crawler was developed to obtain profiles of employees from six targeted organizations and this was done by collecting the Facebook data. A social network topology was created for each organization, and machine-learning algorithms and centrality measures were implemented so that the hidden leadership positions within each company could be discovered. Moreover, the social community clusters inside these organizations were also revealed by the algorithms, which gave us understanding about the communication network of each company in addition to the structure of the organization.

According to a study by [51], it has become clear that social media data is simply susceptible to misuse. The scheme encompasses structured approach and its application. Furthermore, it entails performing a statistical cluster analysis in addition to the comprehensive analysis of social media comments so that researchers could determine the inter-relationships among key factors. The qualitative social media data can be quantified by these schemes and subsequently cluster them based on their similar features, and then they can be used as decision-making tools. The SAMSUNG Mobile Facebook page, where Samsung smartphones were introduced, was used for the data acquisition process. The comment published by Facebook users on the captioned Facebook page is referred to as the "Data". In a period of 3 months, almost 128371 comments were downloaded. The English comments only were undergone through the analysis process. Afterward, the conceptual analysis was used by the content analysis and

ultimately statistical cluster analysis was performed by carrying out relational analysis. Hence, social media data is integrated by applying the statistical cluster analysis and it is performed based on the output of the conceptual analysis. The researchers are consequently enabled to categorize a large dataset into many subsets, at times, referred to as objects. One of the disciplines of its application is marketing. Factors that can be manageable in some cases are also minimized by these types of techniques.

A study by [52] explored the social data as a systematic data mining architecture. Findings indicated that Facebook as a social networking site is the major source of data. Besides this approach, information on "my wall" post regarding myself, age and comments from the Facebook all are emphasized by the author. It has been taken as a raw data, which is applied later to study and monitor the analytical tactics. In addition, the study investigated images for the advertisement of their products and for the decision-making process. A number of data mining techniques precede the coercion of intellectual knowledge from social data. Mainly, it organizes the key information and other applied activities in which users are attributed regarding their colleagues on social networking sites (i.e. Facebook). For the recovery on Facebook user database, Facebook API performs Application Secret key and Facebook API Key are executed by Facebook API. As a result, WEKA files and data mining techniques are supported to collect certain data into the secondary database, while the text data is represented by the detached data.

Researchers of [41] explored the applicability of representing user's personality based on the extracted features from the Facebook data. The classification techniques and their utilities were completely analyzed with regard to the inspirational research outcomes. A sample of 250 user instances from Facebook formed the research study and this sample was from about 10,000 status updates, which was delivered by the My Personality project [53]. The study has the following two interconnected objectives: (1) having knowledge about the pertinent personality-correlated indicators that presents user data implicitly or explicitly in Facebook, and (2) identifying the feasibility of prognostic character demonstration so that upcoming intelligent systems could be supported. The study emphasized on the promotion of pertinent features in a model, through which the enhanced output of the classifiers under evaluation could be observed.

3.2. Text mining in Twitter

A significant size of research has been occupied by the Twitter data analysis over the last couple of years [54]. Large spectrums of domains are using this data, some of which are using it for academic research and others for applications [55]. New improvements regarding twitter data are presented by this section. The document collection from various resources triggers the "Text Mining" process. A particular document would be retrieved by Text mining tool and this document is pre-processed by checking the character sets and format [56]. Subsequently, a text analysis phase would monitor the document. Semantic analysis is used to derive high-quality information from text; this is referred to "Text analysis". The market has a lot of text analysis techniques. Professionals can use combinations of techniques subject to the goal of the organization. Researchers tend to repeat the text analysis techniques till the time information is acquired. A management information system is capable of incorporating the resulting information, and as a result, significant knowledge is produced for the user of that information system [57]. A key issue in text mining is intricacy of natural language. The ambiguity problem is much dense in the natural language. There are multiple meanings of a single word and multiple words can possess same meaning. Ambiguity is referred to as the understanding of a word which has more than one possible meaning. Noise has emerged in extracted information as a result of this ambiguity. Since usability and flexibility are the main parts of ambiguity, it cannot be removed from the natural language. One phrase or sentence can have multiple understandings, so there is a chance we can obtain a number of meanings. The work is still undeveloped and a particular domain is correlated with the suggested approach while the experts have attempted to resolve the ambiguity problem by performing a number of research studies. As there is uncertainty/vagueness in the semantic meanings of many discovered words, so it is very difficult to answer the requirements of the user.

Scholars of [58] developed and formulated an automatic classification technique through which potentially abuse-indicating user posts could be identified and evaluating the likelihood of social media usage as a source for automatic monitoring of drug medication abuse. In this regard, Twitter user posts (tweets) were collected and these were linked with three commonly abused medications (Oxycodone, Adderall, and Quetiapine). Besides interpreting a control medication (metformin), which is not the subject of abuse due to its process, nearly 6400 tweets were manually annotated, where these three medications were pointed out. The annotated data was qualitatively and quantitatively analyzed to determine as to whether or not signals of drug medication abuse are presented in Twitter posts. To sum up, Twitter's value was assessed in exploring the patterns of abuse over time and an automatic supervised classification technique was also designed, in which the purpose was to observe and separate the posts containing signals of medication abuse from those that do not. According to the findings of investigations, Twitter posts have yielded clear signals of medication abuse. As compared to the proportion for the control medication (i.e., metformin: 0.3 %), there is a very high ratio of tweets containing abuse signals for the three case medications (Adderall: 23 %, oxycodone: 12 %, quetiapine: 5.0 %). In addition, almost 82 % accuracy (medication abuse class recall: 0.51, precision: 0.41, F-measure: 0.46) has been achieved through the automatic classification approach. The Study demonstrated how the abuse patterns over time can be analyzed by using the classification data and its goal is to illustrate the effectiveness of automatic classification. As a result, it is found that abuse-related information for medications can be significantly acquired from social media, and the research indicates that natural language processing and supervised classification are the automatic approaches that have potentials for future monitoring and intervention assignments. With respect to supervised learning, the lack of sufficient training data is believed to be the largest shortcoming of the study. Both annotation and automatic classification are hindered by the lack of context and ambiguity in tweets. During the course of annotations, many ambiguous tweets were found and services of pharmacology expert were hired to address these issues. As a result of these ambiguities, the undefined situation is observed in the binary classification process and this inadequacy will continue until the time fine-tuned annotation rules could be specified by the future annotation rules.

A study by [59] applied the text mining approach on a large dataset of tweets. The complete Twitter timelines of 10 academic libraries were used to collect the dataset for this research. Nearly

23,707 tweets formed the total dataset, where there were 7625 hashtags, 17,848 mentions, and 5974 retweets. Inconsistency among academic libraries is found in the distribution of tweets. "Open" was the most repeated word that was used by the academic libraries in different perspectives. It was observed that "special collections" was the most frequent bigram (two-word sequence) in the aggregated tweets. While "save the date" was the most recurrent tri-gram (three-word sequence). In the semantic analysis, words such as "insight, knowledge, and information about cultural and personal relations" were the most frequent word categories. Moreover, "Resources" was the most widespread category of the tweets among all the selected academic libraries. The significance of data and text-mining approaches are reported within the study and their purpose is to gain an insight with the aggregate social data of academic libraries so that the process of decision-making and strategic planning could become facilitated for marketing of services and patron outreach. The 10 academic libraries from top global universities have undergone the text mining approach. The study aimed to illustrate their Twitter usage and to examine their tweet content.

As far as social media is concerned, decision-making is supported and user-generated text is analyzed through text mining and content analysis [60]. By employing an archiving service (twimemachine.com) in December 2014, the complete Twitter timelines of 10 academic libraries were taken into account to collect the dataset for this research. The libraries of 10 highest-ranking universities from the global Shanghai Ranking were chosen for that purpose. The language of the university must be English-based, which was the condition for selection and selection was restricted to only one library if there was more than one library in the university. Certain weaknesses were found in the study, for example, all of the libraries are English-language libraries in the sample and only 10 academic libraries were considered for the analysis. This gap must be filled in future by applying the analysis to a dataset from diversified academic libraries, including non-English language libraries. Consequently, a complete understanding of tweet patterns would be acknowledged. The future inquiry can also incorporate the international or cross-cultural comparisons. Any discrepancy among libraries in their tweets' content affected by the number and interaction of followers could be highlighted by the analysis and its findings. The accuracy of the tweet categorization tool has yielded the inadequate findings, and the said tool needs to be substantiated through other machine-learning models along with their applications.

Researchers of [55] demonstrated in a smoking cessation nicotine patch study an innovative Twitter recruitment system that is deployed by the group. The study aimed to describe the methodology and used to address the issue of digital recruitment. Furthermore, designing a rule-based system with the provision of system specification besides representing the data mining approaches and algorithms (classification and association analysis) using Twitter data. Twitter's streaming API captured two sets of streaming tweets, which were collected for the study. Ten search terms, (i.e. quitting, quit, nicotine, smoking, smoke, patches, cig, cigarette, ecig, cigs, marijuana) were used to gather the first set. The second set of tweets contains 30 terms, in which the terms from the first set were included. Moreover, the second set is a superset of the first one. A number of studies have been conducted to review the information gathering methods. As unstructured data sets are in the textual format, the use of various procedures of text mining has been tackled by many research studies. Nonetheless, the data sets on the social networking websites are not mainly discussed by these studies. A study by [50] applied various text mining techniques. The study would describe the application of these strategies in the social networking websites. In the field of intelligent text analysis, the latest improvements would also be examined in the survey. The study focused on two key techniques pertaining to the text mining field, namely classification and clustering. Usually, they are operated for the study of the unstructured text accessible on the extensive scale frameworks. Prior to the start of World Cup, a total of approximately 30,000 tweets were used by [61]. Moreover, an algorithm was used for integrating the consensus matrix and the DBSCAN algorithm. Consequently, the concerned tweets on those prevailing topics were available to him. Afterward, the clustering analysis was applied to seek the topics discussed by the tweets. The tweets were grouped utilizing the k-means [62], Non-Negative Matrix Factorization (NMF), and a popular clustering algorithm. After that, the results were compared. Similar results were delivered by both algorithms. However, NMF became faster and the researchers could easily interpret the outcomes.

A study by [1] initiated a workflow to gain an insight into both the large-scale data mining methods and qualitative analysis. Twitter posts of engineering students were the primary concern. The basic goal was to identify their issues in their academic experiences. The study conducted a qualitative analysis of samples obtained from around 25,000 tweets that were associated with the engineering students and their college life. The encounter troubles of engineering students were discovered during the study. For example, a large volume of study, sleep deprivation and lack of social engagement. Considering these outcomes, a multi-label classification algorithm was implemented to categorize tweets in lieu of students' queries. The algorithm was applied on approximately 35,000 tweets streamed at the geo-location of Purdue University. At the first instance, the concerned authorities have addressed the experiences and issues of the students and social media data was used to expose the issues. Moreover, a study by [1] also developed a multi-label classifier so that tweets founded within the content evaluation phase could be organized. A number of renowned classifiers are significantly consumed in machine learning domain and data mining process. With Comparison to other state-of-the-art multi-label classifiers, the Naïve Bayes classifiers were found proficient on the dataset.

A study by [63] discussed the clustering technique, the execution of correlation and association analyses to social media. The investigation of insurance Twitter posts was carried out to assess this matter. Consequently, recognizing theories and keywords in the social media data has become an easy task, due to which the information by insurers and its application would be facilitated. After having a detailed analysis, client queries and the potential market would be proactively addressed with usefulness and the findings of the analysis are to be effectively implemented in suitable fields. According to this evaluation, the overall 68,370 tweets were utilized. Two additional kinds of evaluation need to be applied to the data. The first is the clustering analysis, through which the tweets depending on their similarities or dissimilarities would be merged. An Association Analysis is the second one whereas the occurrences of particular composed words were discovered.

Authors of [64] stated that sentiment analysis through social media usage has witnessed a huge interest from scholars in the last few years. In that, the authors discussed the influence of tweets'

sentiment on elections and the impact of the elections' results on web sentiment.

4. Conclusion and Future work

The method of communication with each other has now completely changed due to the progress in the field of social media. Nowadays, modernization can be seen everywhere and based on that; the information production is touching the altitudes. Currently, the new companies are moving forward to take an active part in transforming the communication method [65]. The keywords and phrases' particularization can become helpful to different companies in order to shape their future. In the present study, we have highlighted the state-of-the-art research work regarding the implementation of text mining in the most dominant social media (Facebook and Twitter). From the point of view of several scholars, Text mining was explained through various models. Moreover, different authentic references are also provided to support the research work. As a result, text mining can be classified into text clustering, text categorization, association rule extraction and trend analysis according to applications. With the passage of time, text mining is going to be progressed well.

We can observe from the surveyed literature that Arabic text in social media is overlooked from the point of view of several text mining studies. As a result, this gap opens the door for many text mining scholars to bridge that gap through conducting various studies in the field of text mining in the Arabic language context. A study by [66] argued that researchers analyzing the Arabic post are seldom found, focusing on the text mining of English, albeit the Arabic post on social media is present in bulk amount. Scholars of [67] outlined its strange and peculiar characteristics as the reasons behind this attitude. From the surveyed literature, we have observed that researchers have paid less attention to sentiment analysis in the Arabic text. The sophisticated tasks of parsing and sense disambiguation fortify production of target lists of the most recurrent grammatical structures and senses of polysemous words, and the potential for syntactic and semantic ambiguity is found to be high [68]. As a future work, we are highly interested in examining the text mining techniques on Arabic textual data from Facebook and Twitter. In addition, future research should take sentiment analysis of Arabic text into consideration. The Arabic language is convoluted morphologically, possesses free word order, punctuation seldom found and short vowels are avoided in the written form of Standard Arabic. Hence, context is essential to eradicate prevailing ambiguity from apparently identical forms which is significant in recognizing opinions.

References

[1] Chen, X., Vorvoreanu, M., & Madhavan, K. (2014). Mining social media data for understanding students' learning experiences. *IEEE Transactions on Learning Technologies*, *7*(3), 246-259.

[2] Buettner, R. (2016). Predicting user behavior in electronic markets based on personality-mining in large online social networks. *Electronic Markets*, 1-19.

[3] Baumer, E. P., Sinclair, J., & Tomlinson, B. (2010, April). America is like Metamucil: fostering critical and creative thinking about metaphor in political blogs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1437-1446). ACM.

[4] Stieglitz, S., Dang-Xuan, L., Bruns, A., & Neuberger, C. (2014). Social media analytics. *Wirtschaftsinformatik*, *56*(2), 101-109.

[5] Kasture, N. R., & Bhilare, P. B. (2015, February). An Approach for Sentiment analysis on social networking sites. In *Computing Communication Control and Automation (ICCUBEA), 2015 International Conference on* (pp. 390-395). IEEE.

[6] Sorensen, L. (2009, May). User managed trust in social networking-Comparing Facebook, MySpace and Linkedin. In *Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology, 2009. Wireless VITAE 2009. 1st International Conference on* (pp. 427-431). IEEE.

[7] Naaman, M. (2012). Social multimedia: highlighting opportunities for search and mining of multimedia data in social media applications. *Multimedia Tools and Applications*, *56*(1), 9-34.

[8] Evans, B. M., Kairam, S., & Pirolli, P. (2010). Do your friends make you smarter?: An analysis of social strategies in online information seeking. *Information Processing & Management*, *46*(6), 679-692.

[9] Li, J., & Khan, S. U. (2009, November). MobiSN: Semantics-based mobile ad hoc social network framework. In *Global Telecommunications Conference, 2009. GLOBECOM 2009. IEEE* (pp. 1-6). IEEE.

[10] Liu, X., Wang, M., & Huet, B. (2016). Event analysis in social multimedia: a survey. *Frontiers of Computer Science*, 1-14.

[11] Tsytsarau, M., & Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, *24*(3), 478-514.

[12] Al-Emran, M., & Malik, S. I. (2016). The Impact of Google Apps at Work: Higher Educational Perspective. *International Journal of Interactive Mobile Technologies (iJIM)*, *10*(4), 85-88.

[13] Al-Emran, M. N. H. (2014). Investigating Students' and Faculty members' Attitudes Towards the Use of Mobile Learning in Higher Educational Environments at the Gulf Region.

[14] Al-Emran, M., & Shaalan, K. (2015, August). Learners and educators attitudes towards mobile learning in higher education: State of the art. In *Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on* (pp. 907-913). IEEE.

[15] Irfan, R., King, C. K., Grages, D., Ewen, S., Khan, S. U., Madani, S. A., ...& Tziritas, N. (2015). A survey on text mining in social networks. *The Knowledge Engineering Review*, *30*(02), 157-170.

[16] Berry Michael, W. (2004). Automatic Discovery of Similar Words. *Survey of Text Mining: Clustering, Classification and Retrieval", Springer Verlag, New York*, *200*, 24-43.

[17] Salloum, S. A., Al-Emran, M., & Shaalan, K. (2016). A Survey of Lexical Functional Grammar in the Arabic Context. *Int. J. Com. Net. Tech*, *4*(3).

[18] Al Emran, M., & Shaalan, K. (2014, September). A Survey of Intelligent Language Tutoring Systems. In *Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on* (pp. 393-399). IEEE.

[19] Al-Emran, M., Zaza, S., & Shaalan, K. (2015, May). Parsing modern standard Arabic using Treebank resources. In *Information and Communication Technology Research (ICTRC), 2015 International Conference on* (pp. 80-83). IEEE.

[20] Alfaro, C., Cano-Montero, J., Gómez, J., Moguerza, J. M., & Ortega, F. (2016). A multi-stage method for content classification and opinion mining on weblog comments. *Annals of Operations Research*, *236*(1), 197-213.

[21] Yang, L., Geng, X., & Liao, H. (2016). A web sentiment analysis method on fuzzy clustering for mobile social media users. *EURASIP Journal on Wireless Communications and Networking*, *2016*(1), 1.

[22] Robinson, R., Goh, T. T., & Zhang, R. (2012). Textual factors in online product reviews: a foundation for a more influential approach to opinion mining. *Electronic Commerce Research*, *12*(3), 301-330.

[23] Rahmani, A., Chen, A., Sarhan, A., Jida, J., Rifaie, M., & Alhajj, R. (2014). Social media analysis and summarization for opinion mining: a business case study. *Social Network Analysis and Mining*, *4*(1), 1-11.

[24] Grimes, S. (2008). Unstructured data and the 80 percent rule. *CarabridgeBridgepoints*.

[25] Hung, J. L., & Zhang, K. (2012). Examining mobile learning trends 2003–2008: A categorical meta-trend analysis using text mining techniques. *Journal of Computing in Higher education*, *24*(1), 1-17.

[26] Feldman, R., & Dagan, I. (1995, August). Knowledge Discovery in Textual Databases (KDT). In *KDD* (Vol. 95, pp. 112-117).

[27] Rajman, M., & Besançon, R. (1998). Text mining: natural language techniques and text mining applications. In *Data mining and reverse engineering* (pp. 50-64). Springer US.

[28] Gök, A., Waterworth, A., & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics*, *102*(1), 653-671.

[29] Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2005). Tapping into the power of text mining.

[30] SØRENSEN, H. T., Sabroe, S., & OLSEN, J. (1996). A framework for evaluation of secondary data sources for epidemiological research. *International journal of epidemiology*, *25*(2), 435-442.

[31] Zhang, J. Q., Craciun, G., & Shin, D. (2010). When does electronic word-of-mouth matter? A study of consumer product reviews. *Journal of Business Research*, *63*(12), 1336-1341.

[32] Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, *1*(1), 60-76.

[33] Zaza, S., & Al-Emran, M. (2015, October). Mining and Exploration of Credit Cards Data in UAE. In *2015 Fifth International Conference on e-Learning (econf)* (pp. 275-279). IEEE.

[34] Navathe, S. B., & Ramez, E. (2000). Data warehousing and data mining. *Fundamentals of Database Systems*, 841-872.

[35] Akilan, A. (2015, February). Text mining: Challenges and future directions. In*Electronics and Communication Systems (ICECS), 2015 2nd International Conference on* (pp. 1679-1684). IEEE.

[36] Sukanya, M., & Biruntha, S. (2012, August). Techniques on text mining. In*Advanced Communication Control and Computing Technologies (ICACCCT), 2012 IEEE International Conference on* (pp. 269-271). IEEE.

[37] Piatetsky-Shapiro, G. (2007). Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from "university" to "business" and "analytics". *Data Mining and Knowledge Discovery*, *15*(1), 99-105.

[38] Weninger, T. (2014). An exploration of submissions and discussions in social news: mining collective intelligence of reddit. *Social Network Analysis and Mining*, *4*(1), 1-19.

[39] Chakraborty, G., & Krishna, M. (2014). Analysis of unstructured data: Applications of text analytics and sentiment mining. In *SAS global forum* (pp. 1288-2014).

[40] Stieglitz, S., & Dang-Xuan, L. (2013). Social media and political communication: a social media analytics framework. *Social Network Analysis and Mining*, *3*(4), 1277-1291.

[41] Markovikj, D., Gievska, S., Kosinski, M., & Stillwell, D. (2013, June). Mining facebook data for predictive personality modeling. In *Proceedings of the 7th international AAAI conference on Weblogs and Social Media (ICWSM 2013), Boston, MA, USA*.

[42] Zhang, Y., & Yu, T. (2012). Mining trust relationships from online social networks. *Journal of Computer Science and Technology*, *27*(3), 492-505.

[43] Yan, B. N., Lee, T. S., & Lee, T. P. (2015). Analysis of research papers on E-commerce (2000–2013): based on a text mining approach. *Scientometrics*,*105*(1), 403-417.

[44] Witten, I. H. (2005). Text mining. *Practical handbook of Internet computing*, 14-1.

[45] Steinberger, R. (2012). A survey of methods to ease the development of highly multilingual text mining applications. *Language Resources and Evaluation*, *46*(2), 155-176.

[46] Schoder, D., Gloor, P. A., & Metaxas, P. T. (2013). Special Issue on Social Media. *KI*, *27*(1), 5-8.

[47] Injadat, M., Salo, F., &Nassif, A. B. (2016). Data mining techniques in social media: A survey. *Neurocomputing*.

[48] Kamal, S., & Arefin, M. S. (2016). Impact analysis of facebook in family bonding. *Social Network Analysis and Mining*, *6*(1), 1-14.

[49] Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., & Gosling, S. D. (2010). Facebook profiles reflect actual personality, not self-idealization. *Psychological science*.

[50] Fire, M., & Puzis, R. (2012). Organization mining using online social networks. *Networks and Spatial Economics*, 1-34.

[51] Chan, H. K., Lacka, E., Yee, R. W., & Lim, M. K. (2014, December). A case study on mining social media data. In *2014 IEEE International Conference on Industrial Engineering and Engineering Management* (pp. 593-596). IEEE.

[52] Rahman, M. M. (2012). Mining social data to extract intellectual knowledge. *arXiv preprint arXiv:1209.5345*.

[53] Celli, F., Pianesi, F., Stillwell, D., & Kosinski, M. (2013, June). Workshop on computational personality recognition (shared task). In *Proceedings of the Workshop on Computational Personality Recognition*.

[54] Reips, U. D., & Garaizar, P. (2011). Mining twitter: A source for psychological wisdom of the crowds. *Behavior research methods*, *43*(3), 635-642.

[55] Hamed, A. A., Wu, X., & Rubin, A. (2014). A twitter recruitment intelligent system: association rule mining for smoking cessation. *Social Network Analysis and Mining*, *4*(1), 1-19.

[56] Dey, L., & Haque, S. M. (2009). Opinion mining from noisy text data. *International Journal on Document Analysis and Recognition (IJDAR)*, *12*(3), 205-226.

[57] Gaikwad, S. V., Chaugule, A., & Patil, P. (2014). Text mining methods and techniques. International Journal of Computer Applications, 85(17).

[58] Sarker, A., O'Connor, K., Ginn, R., Scotch, M., Smith, K., Malone, D., & Gonzalez, G. (2016). Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from Twitter. *Drug safety*, *39*(3), 231-240.

[59] Al-Daihani, S. M., & Abrahams, A. (2016). A Text Mining Analysis of Academic Libraries' Tweets. *The Journal of Academic Librarianship*, *42*(2), 135-143.

[60] Abrahams, A. S., Fan, W., Wang, G. A., Zhang, Z. J., & Jiao, J. (2015). An integrated text analytic framework for product defect discovery. *Production and Operations Management*, *24*(6), 975-990.

[61] Godfrey, D., Johns, C., Meyer, C., Race, S., & Sadek, C. (2014). A case study in text mining: Interpreting twitter data from world cup tweets. *arXiv preprint arXiv:1408.5427*.

[62] Velez, D., Sueiras, J., Ortega, A., & Velez, J. F. (2015). A method for K-Means seeds generation applied to text mining. *Statistical Methods & Applications*, 1-23.

[63] Mosley Jr, R. C. (2012). Social media analytics: Data mining applied to insurance Twitter posts. In *Casualty Actuarial Society E-Forum, Winter 2012 Volume 2* (p. 1).

[64] Kermanidis, K. L., & Maragoudakis, M. (2013). Political sentiment analysis of tweets before and after the Greek elections of May 2012. *International Journal of Social Network Mining*, *1*(3-4), 298-317.

[65] Wu, J., Sun, H., & Tan, Y. (2013). Social media research: A review. *Journal of Systems Science and Systems Engineering*, *22*(3), 257-282.

[66] Habash, N. (2010). Introduction to Arabic natural language processing. Synthesis Lectures on Human Language Technologies, 3(1), 1-187, Morgan & Claypool Publishers.

[67] Shaalan, K., Abo Bakr, H., & Ziedan, I. (2007). Transferring Egyptian Colloquial into Modern Standard Arabic, International Conference on Recent Advances in Natural Language Processing, PP. 525-529, Bulgaria.

[68] Farghaly, A., Shaalan, K. (2009). Arabic Natural Language Processing: Challenges and Solutions. ACM Transactions on Asian Language Information Processing (TALIP), ACM, 8(4), 1-22.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 134-148 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

Privacy-by-Design(PbD) IoT Framework : A Case of Location Privacy Mitigation Strategies for Near Field Communication (NFC) Tag Sensor

V.Ragunatha Nadarajah, Manmeet Mahinderjit Singh*

School of Computer Sciences, University of Science Malaysia, 11800, Malaysia

ARTICLE INFO

Article history:
Received: 15 December, 2016
Accepted: 11 January, 2017
Online: 28 January, 2017

Keywords:
NFC- Near Field Communication
RFID - Radio Frequency Identification
MITM - Man-In-The-Middle
PbD – Privacy-by-Design
MIDAS - Multifactor Identifications Attendance System
NDEF - NFC Data Exchange Format

ABSTRACT

Near Field Communication (NFC) technology is a short range (range about 10cm) standard extended from the core standard Radio Frequency Identifier (RFID). These technologies are a portion of wireless communication technology. Even though NFC technologies benefit in various field, but it's still exposed to multiple type of privacy attacks and threat as well since the communication occur in an open environment. The filtering technique been perform on the tag in order to get access to the embedded information. As solution based on tag filtering techniques, existing NFC filtering, Intent filtering has merged together with Bloom filtering from RFID technology. This help in term of elimination the duplicate tag and verify the receiving tag. Meanwhile, as a content protection to NFC Data Exchange Format (NDEF) message been transmitted through the communication channel, Advance Encryption Standard (AES) 128bit has been implemented on the NDEF message. AES provide solution to encrypt the NDEF message which has been communicated. Bloom filtering performed the hashing operation using MD5 technique as a verification of registered user to the NFC system. While the default Intent filtering direct the user to the selected invocation as registered on the tag after the Bloom filtering verification. Besides that, implementation of AES cryptographic in NDEF message, took approximately about 80 trillion years++ to crack the key using brute force attack. Communication of two legitimate entities is secured with AES encryption. Hence, secured user validation or filtering with encrypted message, prevent the possibility for MITM attacker to retrieve sensitive or personal information. The overall framework provide a better security solution compare to the existing framework.

1. Introduction

Near Field Communication (NFC) technology is a short range (range about 10cm) standard extended from the core standard Radio Frequency Identifier (RFID). Extended standard of RFID which is ISO 14443 Type A and Type B were the standards that were incorporated to NFC [1, 3]. These standards are commonly used in NXP Mifare cards [2], followed by Vicinity Cards [5] that are used for item management which was incorporated to ISO 15693 standard and FeliCa [2], the famous NFC technology by Sony that has been standardized to ISO 18092 standard. Since NFC is the extended standard of RFID, most of NFC-embedded devices as well as applications began to be compatible with the RFID devices or infrastructures [2]. These advantages were from the combination of all those standards into NFC standard. Furthermore, NFC-embedded devices such as NFC enabled mobile phones to be switched easily between an active reader modes to passive tag. In other words, it merges the design of a reader and a contactless card on a single device.

Privacy information is about controlling or more generally, filtering access to personal information. The filtering is performed on the tag in order to get access to the particular information. As

*Corresponding Author: Manmeet Mahinderjit Singh, School of Computer Sciences, University of Science Malaysia, 11800, Malaysia
Email: manmeet@usm.my

per statement above, the major focus of this research is on the studying technique in tag filtering and the possible attacks toward NFC, as well as to propose a proper solution or enhancement that can reduce privacy-based attacks. NFC Intent filtering is one of the embedded technologies which is the focus of this research as well.

In reflecting the privacy, content protecting is one of the concern in this research. Encryption of the NDEF message or payload of the communication has been classified by the strength of the key. As per this research, the Advance Encryption Standard (AES) has been selected as the encryption algorithm for the content protection. AES is one of the stronger and faster encryption standard compared with Data Encryption Standard (DES) and Triple-DES.

Privacy by Design (PbD) is a public guideline or approach which concentrates or focuses on privacy spectrum to systems engineering which takes privacy into account throughout the whole engineering process. In this approach, there are seven major principles discussed. Embedded into design, is one of the principle which defines the privacy technique/algorithm which is embedded in the architecture of the system. This particular principle has been used in this research as a guideline in embedding privacy approach in a system.

2. Literature Review

2.1. Near Field Communication (NFC) Technology

Near Field Communication (NFC) technology is an extended from the core standard Radio Frequency Identifier (RFID) which support a short range communication about 10cm. ISO 14443 Type A and Type B were the standards incorporated to NFC [1, 3]. These standards are commonly used in NXP Mifare cards [2], followed by Vicinity Cards [30]that are used for item management which is incorporated to ISO 15693 standard and FeliCa, [2]which is the famous NFC technology by Sony that has been standardized to ISO 18092 standard. Since NFC is the extended standard of RFID, most of the NFC-embedded devices as well as applications have been compatible with the RFID devices or infrastructures [2]. These advantages are from the combination of all those standards into a NFC standard. Furthermore, NFC-embedded devices such as NFC has enabled mobile phones to be switched easily between an active reader modes to passive tag. In other words, it merges the design of a reader and a contactless card on a single device.

NFC operates in three different modes which are the read/write mode, followed by peer-to-peer mode and finally the tag emulation mode. NFC Devices are able to communicate at approximately 10cm and the communication mode indirectly reflects to the standard as it's a short range wireless communication extended from the RFID technology [2]. Furthermore, NFC-embedded devices would be able to perform in the active and passive mode. In reflecting to common RFID type communication where a device acts as the reader (initiator) and as a passive tag (target) explains the passive mode situation [2]. Meanwhile, if both devices are communicating and generating their own induction defines the active mode of NFC technology.

2.2. Radio Frequency Identification (RFID)

Radio Frequency Identification (RFID) networks exist in a broad range of environments and their rapid proliferation has been underway for quite some time. Commonly, RFID systems consist of tiny integrated circuits equipped with antennas or well known as RFID tags, that communicate in a few methods with their reading devices also known as RFID readers using electromagnetic fields at one of the several standard radio frequencies [2,4]. Additionally, there is usually a back-end database that collects information related to the physically tagged objects.

RFID systems are vulnerable to a broad range of malicious attacks ranging from passive eavesdropping to active interference. Unlike in wired networks, where computing systems typically have both centralized and host-based defenses such as firewalls, attacks against RFID networks can target decentralized parts of the system infrastructure, since RFID readers and RFID tags operate in an inherently unstable and potentially noisy environment [6]. In addition, the RFID technology is evolving quickly as the tags are multiplying and shrinking and so the threats they are susceptible to, are similarly evolving. Thus, it becomes increasingly difficult to have a global view of the problem as it is vulnerable to a wide range of attacks.

There are four (4) typical frequency ranges that the RFID systems commonly run which is Low Frequency (LF) ranging from 125kHz to 13.2 kHz, High Frequency (HF) operating at 13.56MHz, Ultra-high frequency (UHF) ranging from 860MHz to 960MHz, and microwave frequency starting from 3.1GHz up to about 10GHz [2][3][4].

2.3. NFC Operation Mode

NFC enabled devices can operate in three (3) different operations which are reader/writer mode, peer-to-peer (P2P) mode and card emulation mode as shown in Figure 1. The NFC Forum technical specifications unlock the full capabilities of NFC technology for the different operating modes and are based on the ISO/IEC 18092 NFC IP-1, JIS X 6319-4 and ISO/IEC 14443 contactless smart card standards, also referred to as NFC-A, NFC-B and NFC-F in NFC Forum specifications [4].

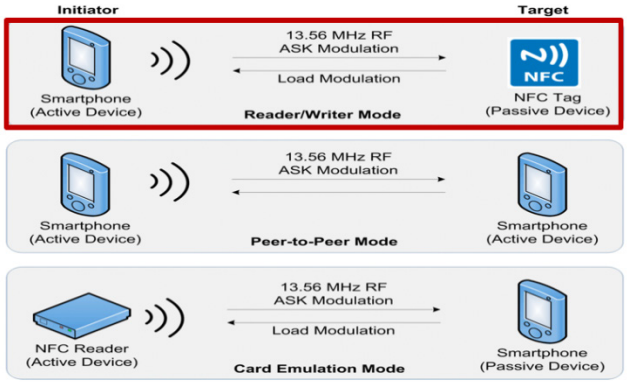


Figure 1: NFC Operation Mode [4]

- Reader/Writer Mode - NFC-enabled device which able to exchange data with NFC Forum-mandated tags, such as a tag embedded in a NFC smart poster [4]. It means that in

the reader/writer mode, when a NFC tag is put in a close range or coverage area to a NFC device, the device can read data from the tag as well as store data into the tag. The reader/writer mode on the RF interface conforms to the ISO 14443 and FeliCa schemes [4].

- Peer-to-Peer (P2P) - can be described as two NFC devices that are able to communicate with each other to exchange data and share files. This means that a NFC device user can exchange information promptly. In our daily life, we can realize that P2P operates as music download, share Bluetooth or WiFi set up parameters or exchange data such as digital photos, videos or phone address book. The Peer-to-Peer mode is standardized on the ISO/IEC 18092 standard [4].
- Card Emulation - threats NFC devices as smart cards, allowing users to perform transactions such as credit cards and smart cards [4]. With just a single touch, the function of purchases, ticketing, and transit access control can be fully achieved. An external reader is required when the NFC device acts like a traditional contactless smart card.

2.4. *Operating Principle of NFC Reader/Writer Mode*

When a NFC application starts to work, the NFC reader/phone generates a Radio Frequency (RF) sine wave to release energy to the tag in its coverage distance and retrieve data from the tag. Usually, the sine wave is transmitted at 13.56MHz frequency and it will form an area of magnetic flux or well known as RF signal transmission tunnel [4]. The tag which is close to the magnetic flux area, will receive energy from the reader and then generate a counter frequency, which can modify the frequency properties of the original sine wave created by the NFC reader/phone. After the NFC reader/phone detects the modification, it confirms that there is a NFC tag nearby. With the target NFC tag lock-in, data are be transferred between the NFC reader/phone and the NFC tag respectively by the radio wave. The reader/writer mode operation supports NFC Data Exchange Format (NDEF) and this mode has been chosen in this research project. NDEF message is the default messaging for NFC communication which support this operation mode are an advantage for us to choose this mode. Meanwhile, this mode is the current commonly used mode in NFC embedded devices as it could provide tag reader and writer operation. Somehow, most of the NFC device manufacturing is tend to use this mode as the default operation mode as it provide the basic operation to read and write on a tag. At the same time, man-in-the-middle attacks is frequently happened in this kind of NFC embedded devices as users able to retrieve data from a tag as well as modify the data. Since our focus is on man-in-the-middle attack, read and write NFC operation mode is the best suite the research project which provided the best platform to perform the research on the attacks and security approach.

2.5. *NFC Data Exchange Format (NDEF)*

NFC Data Exchange Format (NDEF) is a lightweight, binary message format that can be used to encapsulate one or more application-defined payloads of arbitrary type and size into a single message construct [4, 7]. The NDEF specification is well defined as a message encapsulation format to exchange information between a magnetic flux and RF signal transmission tunnel. NDEF message is composed of numerous records. The record amount in an NDEF message depending on the tag type and calling application [4]. Each NDEF record contains a header and a payload. The payload is described by type, length and an optional identifier encoded in an NDEF record header structure. Payload type refers to the data type that is being carried in the payload of a record and this is used to guide the processing of a payload [4, 7]. Usually the payload can be of one of a variety of different types such as text, URL, Multipurpose Internet Mail Extensions (MIME) media, including NFC-specific data type. The optional payload identifier allows user applications to identify the payload carried within an NDEF record. For NFC-specific data types, the payload contents must be defined in a NFC Record Type Definition file, RTD.

2.6. *Privacy by Design (PbD)*

Privacy by Design (PbD) advances the view that the future of privacy cannot be assured solely by compliance with regulatory frameworks. The objectives of PbD is to ensure privacy and gain personal control over one's information [8]. While for organizations, is to gain a sustainable competitive advantage [8]. There are seven foundational principles of PbD discussed below:

Table 1: PbD Foundational Principles

PbD Foundational Principles	Privacy	Security
1. Proactive not Reactive; Preventative not Remedial	Anticipate and prevent privacy. Prevent from privacy risks to materialize.	Begin with the end in mind. Proactive implementation of security.
2. Default Setting	Build privacy measures directly into any IT system.	Implement "Secure by Default" policies.
3. Embedded into Design	Embed privacy into the design and architecture of It system.	Apply Software Security Assurance practices. Use of Trusted Platform Module.
4. Positive-Sum	Accommodate all legitimate interests and objectives in a positive-sum. "Win-win manner.	Resolve conflicts to seek win-win.
5. End-to-End Security	Secure life-cycle management. Ensure cradle-to-grave.	Ensure confidentiality, integrity and availability.
6. Visibility & Transparency	Component parts of IT systems practices visible and transparent.	Strengthen security through open standard.

7.	**Respect for the User**	Respect and protect interests of the individual. Keep it user-centric.	Respect and protect the interests of all information owners.

2.7. NFC Security Challenges

Security in a communication system is well defined as the prevention of unauthorized access and manipulation of data. Security approach is a basic necessary item that need to be added in every system. Security in a system can be classified into two major parts which are Confidentiality, Integrity, Availability (CIA) and X.800.

- CIA - exposed in three main principles in security which are confidentiality, integrity and availability. Confidentiality defines the authorization principle whereby only those with sufficient privileges and a demonstrated need may be able to get access to certain information. While integrity is the principle of ensuring information is maintained in a complete and uncorrupted state. The integrity of information is threatened when it is exposed to corruption, damage, destruction, or other disruption of its authentic state. Availability is the principle that allows entities either a person or other peripheral devices to access information in a usable format without interference or obstruction. Availability does not imply that the information is accessible to any user; rather, it means availability to authorized users [9]. We could classify those security threats and attacks that are related to CIA in 3 major layers of the NFC system. Most of the attacked happened on the communication as initially NFC architecture has lack of security approach been implemented. At the same time poor communication channel in NFC interactions is also a good reason most of the attack occur at the communication layer. Security approach at hardware and back-end devices can be manage with an authentication process or access control approach. But different at communication layer which will required encryption process to ensure the message communicated is secure.
- X.800 - service provided by a protocol layer of communicating open systems, which ensures adequate security of the systems or of data transfers or communication. X.800 considers on authentication, access control, data confidentiality, data integrity, non-repudiation and availability. Unauthorized tag reading can be considered as one of the X.800 security attacks. Unauthorized tags including hidden tags involves third party attempts to access a secure NFC communication. Key compromise can also be a concern of X.800 treat where multiple type of crypto attacks are tend to open communication of two entities.

2.8. NFC Privacy Challenges

Privacy in NFC threats refers to the sensitive information or data that have been retrieved by an attacker. This can be classified into four major categories including location, user, time and data.

- Location - privacy in terms of location points to the attacks in which the information retrieved is the location of an NFC tag user or the tag location itself. Cloning, reprogramming or spoofing and swapping a tag can be considered as impersonation of a NFC tag which benefic an attacker to retrieve NFC user location details. Relay attack is the intention of Man-in-the-Middle attack [10]. Here, the both legitimate NFC entities are fooled that they are communicating directly with each other but they didn't realize the third party access in between the communication. Tracking and hot listing are the most significant privacy threats. Silent involvement of an unauthorized NFC tags or party without giving any alert or sign of activity describes the tracking [10]. Tracking here refers to collecting of personal information to track a particular user. Meanwhile, hot listing refers to the collection of information or object that could be used in future by an attacker to perform more direct attacks [10, 11].
- Time - privacy in time refers to those attacks or threats which the information gained is the time of an NFC user and the time of the NFC communication that have taken place. Relay attack allows the attacker to retrieve the data or more specifically referring to time [10]. Tag and reader are fooled by the third party access in between the communication where the sender and receiver believe that they are communicating directly with each other. Tracking and hot listing are the most significant privacy threats. Silent involvement of an unauthorized NFC tags or party without giving any alert or sign of activity describes the tracking [10].
- User - privacy in terms of user defines the attacks or threats in which the information retrieved is the user's sensitive or private information via a NFC tag user and the system itself. In impersonation attack, the attacker might imitate the identity of tags or readers and the system to retrieve a particular user identity [12]. Modification and retrieval of the data refers to the sender and receiver details [12] is most commonly referring to relay attack.
- Data - Privacy that points to the data is those attacks of which the information gained is the sensitive information or data of an NFC tag, system and the user's details of the NFC communication that have taken place. Data modification in privacy reflects to impersonation attacks of NFC tags which includes involvement of cloning and reprogramming a tag as well as swapping a tag [12, 13, 14]. This data is modified for their personal purpose as well as for self-satisfaction. While relay attack happened when an attacker played the role as man-in-middle attacker where they usually place their own or personal devices linking the legitimate NFC entities in intention to intercept and modify or retrieve data communicated in between tags and readers. While information injection reflects to buffer overflow attacks. The overflow data will flow over stacks that caused it to be executed as well when the system attempts to process the buffer [1, 16].

2.9. Existing Filtering and Duplication Elimination Technique

Here we have identified a few techniques of duplication elimination which are currently being used in the RFID technology.

- Intersection Algorithm - compares collected/read data between two readers. If the same data exists in both readers on the similar network reader, this data is moved to a specific shelf/database/array [17].

```
Algorithm: Intersection Algorithm
Input: Reader A, Reader B
Output: Shelf S
begin
for (Every member of Reader A) do
for (Every member of Reader B) do
if (Reader A = Reader B) then
Put member of Reader A into Shelf S
else
if (Reader B = Reader A) then
Put member of Reader B into Shelf S
end if
end for
end for
end
```

- Relative Complement Algorithms - also compares collected/read data between two readers. But if there are duplications between both readers and similar network reader, these data is ignored [17].

```
Algorithm: Relative Complement Algorithm
Input: Reader C, Reader D
Output: Shelf S
begin
for (Every member of Reader C) do
for (Every member of Reader D) do
if (Reader C != Reader D) then
Put member of Reader C into Shelf S
else
if (Reader D != Reader C) then
Put member of Reader D into Shelf S
end if
end for
end for
end
```

- Randomization Algorithm - this algorithm will randomly assign values between "0" and "1" to every read tag. If the outcome is equals to "0", the specific collected/read data is allocated to the shelf/database/array one (1). Otherwise, the collected/read data is moved to the shelf/database/array two (2) as the outcome is "1" [17, 18]. However, in the real randomization, there may not be equal number of tags allocated between two shelves/databases/arrays.

```
Algorithm: Randomization Algorithm
Input: Reader E
Output: Shelf SA, Shelf SB
begin
for (Every member of Reader E) do
Randomize number between 0 AND 1
if (Reader E = 0) then
Put member of Reader E into Shelf S1
else
if (Reader E = 1) then
Put member of Reader E into Shelf S2
end if
end if
end for
end
```

- Bloom Filtering - data filtering process that occurs in the RFID middleware can be classified into two types of low level data filtering and semantics data filtering [19]. Raw RFID data stream is cleaned at the low level data filtering while data has been filtered according to the demands from the system at semantic data filtering. Bloom filtering has the capability to identify a data whether the data is in the set or not. This is why Bloom filtering is considered as a space-efficient-probabilistic data structure [19]. Referring to figure 2, k defines the number of hashing function and data which represents in bit array of size m.

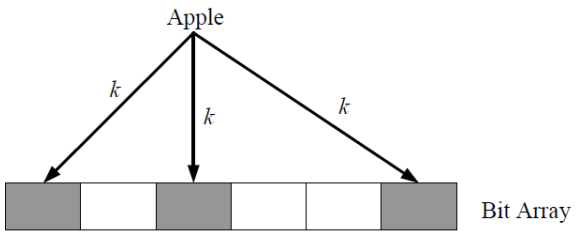


Figure 2: Structure of the Bloom filter [19]

Bloom filtering can perform two basic operations which are test and add. Test operation is used to verify the appearance of an element in the set. Return value of false shows that the element is not in the set while value of true definitely means the element is in the set. Alternatively, add operation provides the functionality to add an element into a set. K different hashing function is feed in order to add an element into bloom filtering. To verify whether the element has been stored in the set, the k value hashing function is feed. Bloom filtering is a best space-efficient probabilistic algorithm for tag filtering technique. While Bloom filtering also, specify the filtering technique according to the tag technologies.

2.10. NFC Intent Filtering Technique

To start a NFC application when an NFC tag is scanned, the NFC application can filter for one, two, or all three of the NFC intents in the Android manifest. However, usually the first filtering option is the ACTION_NDEF_DISCOVERED intent for the most control of when the application starts [33]. The

ACTION_TECH_DISCOVERED intent is a fallback for ACTION_NDEF_DISCOVERED when no applications filter for ACTION_NDEF_DISCOVERED or for when the payload is not NDEF [20,21]. Filtering for ACTION_TAG_DISCOVERED is usually too general of a category to filter on. Most of the time, a NFC applications will filter for ACTION_NDEF_DISCOVERED or ACTION_TECH_DISCOVERED before ACTION_TAG_DISCOVERED, so the application has a low probability of starting. ACTION_TAG_DISCOVERED is only available as a last resort for applications to filter for in the cases where no other applications are installed to handle the ACTION_NDEF_DISCOVERED or ACTION_TECH_DISCOVERED intent [21]. Figure 3 show the processes of intent filtering.

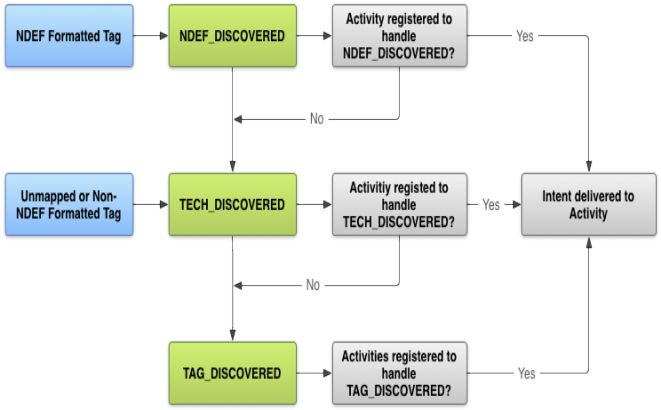


Figure 3: Intent processes [20].

2.11. *Content protection*

Content protection is the security approach on the NDEF message which that began transmitted through the NFC communication channel. Encryption is one of the technique used to protect the NDEF message or content. Below are a few encryption standards that probably could be used as content protection for NDEF message.

Advanced Encryption Standard (AES) has been used as the encryption standard or content protection for the NDEF message transmitted over the NFC communication channel. AES has been selected as it provides better key size compared to DES and 3DES. Table 2 shows the comparison between AES, DES and 3DES.

Table 2: Comparison between AES, DES and 3DES

Algorithms	Strength	Weakness
Data Encryption Standard (DES)	• Brute force search looks hard (56-bit keys)	• Able to encrypt using DES Cracker.
Triple-DES (3DES)	• Able to support up to 128-bit keys.	• Efficiency/security: Bigger block size desirable.
Advance Encryption Standard (AES)	• Stronger & faster than Triple-DES (128-bit block size, 128/192/256-bit keys)	• cost – computational due to implementation characteristics

2.12. *Multifactor Identification Attendance System (MIDAS)*

Multifactor Identification Attendance System (MIDAS), a NFC-based system to handle student attendance [22]. MIDAS is an integrated multi-factor system which means identification knowingly by usage of NFC and biometric via face for a university based attendance system. Figure 4 shows the flow of the MIDAS operation.

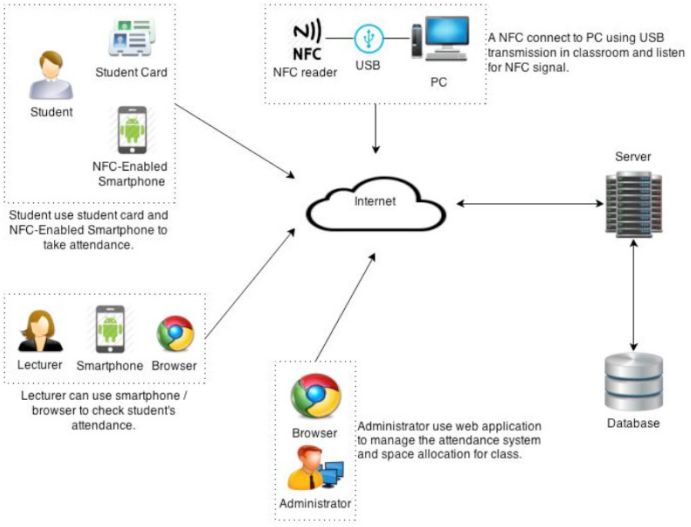


Figure 4: MIDAS operation [22]

The system is implemented with combinations of sensing technologies such as NFC and biometrics including face detection and recognition. NFC reader is used for reading the signal from the NFC-embedded smartphone or card. A dedicated server is deployed to handle requests from the receiving NFC signal and manage the student's attendance information and space allocation for class information. This give a fast and convenient way for administrators, lecturers, and student access system. Students' attendance is presented on the web page as well as android application. This process makes the system look more efficient and faster in terms of data processing. MIDAS has been selected as the real-implementation for the proposed framework as provide a standard operation of NFC system. This application read and write data from a tag as the attendance. Which mean its support read and write mode. Beside that MIDAS is using NDEF message format to transmit the attendance from a tag to reader. These condition suite our research project as in this research project, read/write mode NFC operation tag and also NDEF mapped tag is used. At the same time, since MIDAS using the NDEF format for message transmission, by default intent filtering method is used as the filtering technique.

2.13. *Summary*

In this research project, the NFC reader/writer mode of NFC technology has been chosen. When a NFC tag is put in a close range or coverage area to a NFC device, the device can read data

from the tag as well as store data into the tag. The main reason is, this type of mode has been chosen as MITM attacks is more frequent here. On the other hand, privacy-based has been chosen as it provides personal information of a particular or targeted person as well as the tag information. In most of bank cards or account hack cases, the hacker will look for the last transaction history. Tag and personal information are the main focus of an attacker who perform the MITM attacks.

Intent filtering is the existing filtering technique in NFC technology. In this research, we have merged the NFC intent filtering technique with RFID Bloom filtering technique as this ensure a tight filtering technique. Bloom filtering provides access control process while intent filtering provides tag filtering and process embedded on the NFC tag. In order to provide a preferred privacy protection in NFC communication, the payload should be protected as well. Advance Encryption Standard (AES) has been chosen as the content protection to safeguard the content of payload which will contain the personal and sensitive data. AES could provide big key sizes which makes the encryption more secured when compared to DES and 3DES.

While PbD, discussed the principle which we have chosen to follow in this project. PbD has been chosen as the guideline to this research project as it was the latest or current principle or guideline in ensuring privacy as an aspect of information technology development. Privacy is embedded into the design and architecture of the NFC system based tag filtering technique with content protection. This shows that privacy has become an essential component of the core functionality being delivered. Privacy is integral to the system, without diminishing functionality.

3. Tag filtering

Since NFC technology is new in this information technology era, only few algorithms have been proposed for extending the tag filtering process to perform the filtration on the valid NFC tag as well as the reader. There are a variety of NFC tags that can be read with a handheld device and also desktop applications. The spectrum ranges from simple NFC stickers and NFC key rings to complex NFC cards including bank cards with integrated cryptographic hardware. The tags implemented also differ in their chip technology. The major part is the NDEF, which is supported by most tags and readers. In addition, we could say that NDEF's tags are the most used contactless chip technology worldwide. Some tags can be read and written, while others are read-only or encrypted.

We has proposed an improved algorithm for NFC tag filtering, which combines benefits from the RFID Bloom filtering techniques. The proposed framework is focus on improving the tag filtering techniques which point to privacy-based attacks. The performance or evaluation of the proposed algorithm and previously presented algorithms as the benchmark are compared under a variety of conditions. Here the tag filtering technique has been selected as this process play the major role in NFC communication as tag filtering process react as access control for a tag to communicate with another NFC embedded device. Figure 5 shows how the enhanced tag filtering process.

In order to test the filtering technique, the Multifactor Identifications Attendance System (MIDAS) has been used which was developed by David Ong, Dr. Manmeet. The enhancement has been updated on the filtering techniques where Bloom filtering technique has been merged with the NFC Intent filtering technique.

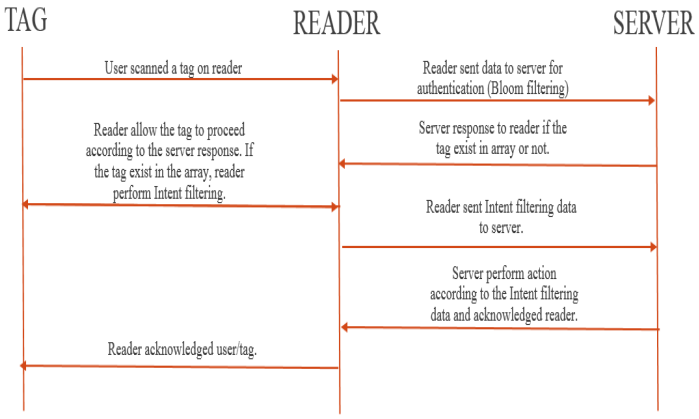


Figure 5: Process of tag filtering enhancement.

3.1. NFC Intent Filtering

In this research we used the MIFARE tag as the NFC tag/card used to store the student details according to the MIDAS system. MIFARE tag uses the standard NFC frequency which is 13.56Mhz and is compatible with most of the latest NFC-embedded devices. At the same time, MIFARE tag is NDEF formatted tag. MIDAS used NDEF technique to discover the tag attached to the reader. Figure 6 shows the NDEF process flow.

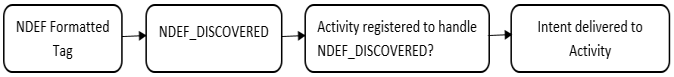


Figure 6: NDEF activity

When the reader detects a tag (MIFARE tag) with NDEF technology and registered with MIDAS, an Intent filtering is triggered which means the tag will direct the user straight to the MIDAS as the tag's activity has been registered to the MIDAS system. This process will not perform any access control verification on the tag. There is a big possibility for the reader to discover a clone tag which has been copied exactly as the registered tag. The table 3 shows the result of what Intent filtering does.

Table 3: Result of Intent filtering

Characters	Responses	
	Yes	No
Authentication Function		x
Responses to un-registered tag		x
Responses to clone / duplicated tag	x	
Access control		x
Perform registered activity	x	

In reflecting to existing NFC intent filtering it only perform those activities registered to the tag. Unfortunately, intent filtering has does not have the capability to authenticate the read tag. Which means a reader will only respond to a tag which has been registered to the specific NFC application. Otherwise it will discard the process. Besides that, intent filtering also not able to identify a clone or duplicate tag. Since intent filtering only direct a user as per the registered tag, it could not validate the read tag. Even though intent filter is the first stage of NFC operation, it does not help much in access control. Intent filtering is not designed for access control instead it only direct user according to the registered activities.

3.2. NFC Intent Filtering Merge with Bloom Filtering

This part discussed the proposed framework that focus on improving the tag filtering techniques which point to privacy-based attacks. Man-In-The-Middle (MITM) attack has been chosen as the major threat to the personal and sensitive information. Connection between two NFC parties or devices can be interrupted by the third party which is called the Man-In-Middle attack. The third party tricks the two legitimate parties to be the other legitimate party thus, routing the communication between the two parties to go through the third party [23]. Particularly, MITM attack has been selected in this project as this kind of attack has high possibility for modification of message contents and replay attacks. In terms of Confidentiality, Integrity and Availability (CIA), MITM attack involved in all three of it. Figure 7 shows the details on the MITM attack.

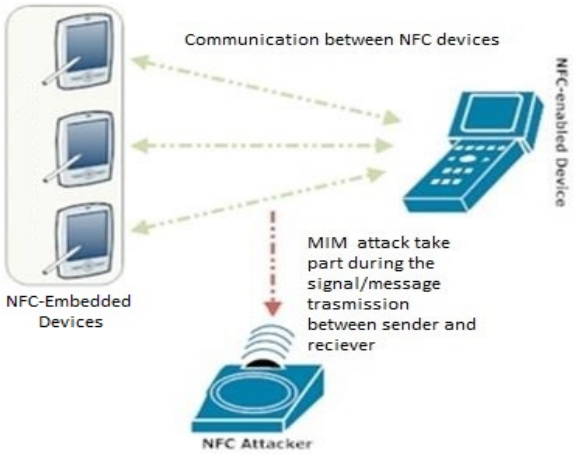


Figure 7: MITM attack performed during NFC communication

The MITM attack is evaluated with the enhanced filtering techniques where the NFC Intent filtering has been merged with Bloom filtering, while the attack is evaluated according to privacy-based. NDEF mapped or embedded tag technology (MIFARE) is used as the tag in order to test the existing Intent filtering and the enhanced filtering technique. Figure 8 shows the proposed filtering techniques.

The enhanced filtering technique performed an additional process of access control and duplication elimination of tag. This process helps to reduce third party or unauthorized access to the system. This means, indirectly the bloom filtering technique will eliminate the clone tag that was discovered and can be considered as the MITM attack. Table 4 is the result of the enhanced tag filtering techniques.

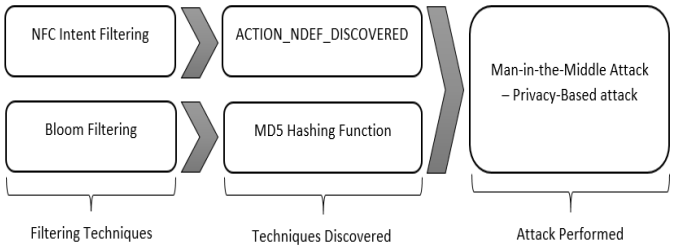


Figure 8: proposed filtering technique against MITM attacks

Table 4: Result of enhanced tag filtering

Characters	Responses	
	Yes	**No**
Authentication Function	x	
Responses to un-registered tag		x
Responses to clone / duplicated tag		x
Access control	x	
Perform registered activity	x	

The proposed framework perform Bloom filtering as an initial process of NFC operation which will indirectly authenticate the tag or user before proceed with intent filtering, perform activities as per registered. Here the intent filtering will only respond the tag or user who registered to the specific NFC application. Since Bloom filter is implemented on the proposed framework, clone or duplicated tag is automatically validated and eliminated. Bloom filter provided a solution as access control to NFC system or application before it perform the activities as per registered. Valid tag or user is carried forward to intent filter that direct the user the registered application.

3.3. Discussion

In reflection to the result of Intent filtering (Table 3), the NFC activities are performed according to what has been registered on the tag. The filtering part tries to discover the registered activities on the tag instead of the duplication or unauthorized tag. The reader detects a tag with NDEF technology and registered with MIDAS, the Intent filtering started to direct the user straight to the MIDAS as the tag's activity has been registered to the MIDAS system and at the same time it reads the information on the tag. First of all, this process does not allow any authentication process on the tag attached to the reader. Indirectly, the tag is not performing any access control verification for unauthorized access. In a case where a clone or duplicate tag has been used, the system still respond to the tag as the registered activities are copied exactly from the original tag. There is a high possibility where MITM attack has been performed on the tag in order to retrieve the personal information on the tag.

While reflecting to the result of the enhanced filtering technique (Table 4), Bloom filtering helps to perform the authentication process in order to verify the authorized access and at the same time, it indirectly eliminate the duplicate or clone tag. As the enhancement has been done on top of Intent filtering technique, the system will discover the registered activities once the bloom filtering process is complete. This means, the tag will direct the user to the registered activities as usual but before that bloom filtering will take place to eliminate the duplicate or unauthorized tag. In this case, signal from MITM attack is eliminated as the duplicate of original tag. At the same time, Intent and Bloom filtering will not response to the un-registered tag, which means, those tags that are not registered to the MIDAS will be automatically eliminated by the system.

4. Content Protection

Content protection is the encryption technique forced on the content of any particular information or message carried on an open network. In this process, we do not use the trusted third party and thus can reduce the communication cost and obtain reliability [24, 25]. Most of the RFID or NFC devices are authenticated to readers and to the backend system using strong cryptography like DES, 3DES, AES and RSA. All modern cards support symmetric cryptography such as 3DES or AES, while some higher-grade cards already support asymmetric cryptography such as RSA [26]. When asymmetric encryption is used, no valuable master keys need to be stored in the door controller, which makes the resulting design and maintenance less complex. Depending on the card capabilities, symmetric AES encryption is used actively by most of the MIFARE card since AES could provide high capability in terms of encryption and decryption. Table 5 shows the strength and weakness of the AES encryption standard.

Table 5: Strengths and Weaknesses of AES

Algorithms	Strength	Weakness
Advance Encryption Standard (AES) [60,61]	• Stronger and faster than Triple-DES. • 128-bit block size. • 128/192/256-bit keys. • Support hardware and Software • Less susceptible to cryptanalysis	• Cost – computational due to implementation characteristics. • Complex encryption and encryption.

4.1. Proposed Algorithm

In this research project, the Advance Encryption Standard (AES) has been chosen as the algorithm for content protection. AES algorithm could provide a stronger and faster encryption when compared to DES and Triple-DES since AES has a maximum of 256-bit keys size as well as 128-bits block size. At the same time, AES is less susceptible to cryptanalysis which can be concluded that AES is more secure than 3DES. Hence, this is the reason why AES begin used in this research project instead of 3DES even though 3DES could support up to 128bit key size. Block size up to 128bit compare to 3DES with 64bit block size, it is less open to security attacks or threats. The number of AES encryption rounds increases with the number of key length. With a big block size and key size, AES provide a better encryption of NDEF message begin communicated between two entities. Which means AES increase the security level against cryptanalysis since it consist a large number of encryption round with a big block and key size. AES could operate in hardware and software which benefit us on this research project as NFC system required NFC devices and NFC applications. Figure 9 shows the flow of AES encryption and decryption as the content protection in the NFC system.

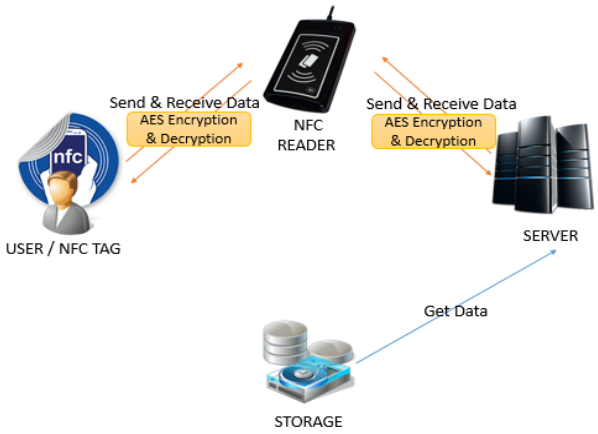


Figure 9: AES implementation in the NFC system

The key length used in the encryption and decryption determines the practical feasibility of performing a brute-force attack. These can be concluded that with big key sizes, it is more difficult to crack or decrypt as compared to small key sizes. Table 6 shows the possible number of key combinations with respect to key size.

Table 6: Key combinations versus key size

Key Size	Possible combinations (2^k)
1-bit	2
2-bit	4
4-bit	16
8-bit	256
16-bit	65536
32-bit	4.2×10^9
56-bit (DES)	7.2×10^{16}
64-bit	1.8×10^{19}
128-bit (AES)	3.4×10^{38}
192-bit (AES)	6.2×10^{57}
256-bit (AES)	1.1×10^{77}

The exponential increase in possible combinations as the key size increases. "DES" is part of a symmetric cryptographic algorithm with a key size of 56 bits that has been cracked in the past using brute force attack. There is also a physical argument that a 128-bit symmetric key is computationally secure against brute-force attack. Just consider the following:

Faster supercomputer: 33.86 Petaflops = 33.86 x 1015 Flops [28] (Flops = Floating point operations per second)

Number of Flops required per combination check: Assume for now 1000

Number of combination checks per second = (33.86 x 1015) / 1000 = 33.86 x 1012

Number of seconds in one year = 365 x 24 x 60 x 60 = 31536000

Number of years to crack AES with 128-bit Key = (3.4 x 1038) / [(33.86 x 1012) x 31536000]

= (0.103 x 1026)/31536000

= 3.2 x 1016

= 80 trillion years++

Even with a supercomputer, it will take about 80 trillion years++ to crack the 128-bit AES key using brute force attack. Table 13 shows the time taken to crack AES using brute force attack.

Table 7: Time to crack cryptographic

Key Size	Time to Crack
128-bit	3.2×10^{16}
192-bit	5.8×10^{42}
256-bit	1.01×10^{71}

The key size used for encryption should always be large enough so that it could not be cracked by modern computers despite considering advancements in processor speeds.

4.2. Discussion

This research project, AES 128 has been used as the content protection for the payload of NFC message. Referring to table 11 above, even with a supercomputer, it would take 80 trillion years++ to crack 128-bit AES key using brute force attack. This is more than the age of the universe (estimation of 13.75 billion years). By default, NFC message or NDEF record consist of a Type Name Format (TNF), payload, payload type, and payload identifier. Payload is the most important part of an NDEF record where it is the content of the message we are transmitting. Here the TNF is defined on how to interpret the payload type. This means, the default architecture of NFC does not provide protection on the content it carries through an open network. By implementation of AES cryptographic, NDEF message or record that are transmitted through the NFC communication is well protected from unauthorized access and modification. Even though the tag filtering technique has been implemented, as a prevention to privacy attacks, content protection will provide the better solution. Reflecting the PbD approach, as "Embedded into Design" principle is selected, here privacy prevention has been embedded into the NFC system and architecture.

5. Analysis and evaluation

This part discussed further on the analysis and evaluation conducted on the proposed work. Basically, analysis and evaluation defines the research goals into defined functions and operation of the proposed work. Security analysis is a method which helps to calculate the value of various attacks against the proposed work. Below are the few parts where the security analysis has been conducted:

5.1. Man-In-The-Middle (MITM) attack.

The Man-in-the-middle attack (MITM) intercept a communication between two legitimate entities. An adversary placed a hidden device between the legitimate NFC entities in the intention to modify or retrieve data communicated between the tags and readers. This allows the attacker to retrieve the data or more specifically referring to personal information. Tag and reader are fooled by the third party access in between the communication where the sender and receiver believe that they are communicating directly with each other.

Reflecting to default Intent filtering technique in NFC architecture, the filtering technique will discover the registered activities on the tag and allow the user to perform further action accordingly as activities were registered. The reader detects a tag with NDEF technology and registered with particular activities, the Intent filtering will take place and direct the user as the activities registered on it. This process could not detect or eliminate unauthorized access or duplicate tag attempt to access the system. Furthermore, the tag does not perform any access control or verification on the tag read. This situation allows an attacker to use a clone or duplicate tag to access the NFC system and by right the reader will still respond to the tag as the registered activities including tag information are copied exactly as the original tag. There is a high possibility where MITM attack been performed on the communication channel during NFC communication. The attack is performed on the NDEF message intention to retrieve the tag information as well as personal information on the tag. Indirectly, this helped the attacker to prepare a duplicate tag from the information gathered through the attack. In future, the attacker could access the NFC system without any hitch.

While reflecting to the proposed work, Bloom filtering helped to perform the user verification or authentication process in order to verify the authorized access to the NFC system. Bloom filtering has the capability to eliminate the duplicated tag in which Intent filtering do not have. Since the proposed work has been done on top Intent filtering, the system performed Bloom filter first and followed by Intent filter which discovers the registered activities on the tag. Furthermore, the NDEF message which a reader retrieve from a tag is discovered with AES encryption in order to protect the content of NDEF message from MITM attack. In this case, the duplicate tag retrieved from MITM attack is eliminated as the duplicate tag and at the same time, the content of a NDEF message are well protected with AES encryption.

5.2. Brute Force Attack

A brute force attack is an attempt of trial-and-error method used to obtain information or encrypt a key of some cryptographic. Brute force attacks may be used by criminals to crack encrypted data, or by security analysts for testing purpose. An attack of this nature can be time and resource-consuming. Hence the name "brute force attack", it only success in some situation where usually based on computing power and the number of combinations tried rather than an ingenious algorithm. NDEF message with an encryption of AES will require an attacker of MITM attack to encrypt the key in order to get the original message of NDEF.

As content protection of NDEF message, AES 128-bit has been used in the proposed work. Referring to table 6 in the previous chapter, even with a supercomputer with 33.86 Petaflops [28], it would take approximately about 80 trillion years++ to crack 128-bit AES key using brute force attack. This is more than the age of the universe (estimate 13.75 billion years) [28]. A different goes to brute force attack which is performed using a "Brute Force Attack Tool" which is available online nowadays.

By default, NDEF message is implemented with Type Name Format (TNF), payload, payload type, and payload identifier. Payload is the most important part of an NDEF record where it's the content of the message (personal and tag information is located) we are transmitting. Here the TNF define us on how to interpret the payload type. In this case, by default, there is no such content protection technique implemented on the content it is carrying through on the NFC communication. This will give an opportunity to a MITM attacker to simply retrieve NDEF message and interpret the tag and personal information communicated through the network. Implementation of AES cryptographic in NDEF message will cause the attacker in encrypting the key and interpret the original message. Even though the tag filtering technique has been implemented, as a prevention to privacy attack, content protection provided the better solution.

5.3. Privacy Attack

Privacy in NFC threats refers to the sensitive information or data that have been retrieved by an attacker via MITM attack. This can be classified into four major categories which are location, user, time and data. Location based privacy refers to the attacks in which the information retrieved is the geographical location of an NFC tag user. Meanwhile, time based privacy points threats in which the information gained is the time of the NFC communication that have taken place. Usually, location and time come together in a same package or session. Privacy based data, define those attacks where the information gained is the sensitive information including tag information. Finally, privacy based user is the threats in which the information retrieved is the user's sensitive or personal information.

Privacy attack happens when an attacker tries to get the sensitive information on the NDEF message and interpret the message for their personal or some other cybercrime. Usually, those privacy information is located into NDEF message that carried through the network. In reflecting to proposed work, as privacy prevention, the NDEF message has been encrypted with AES 128-bit. Implementation of encryption standard on the message transmitted through an open network, provided better prevention toward privacy attack and leakage of sensitive information on the NFC tag. Besides that, tag validation performed by Bloom filtering also avoided the duplication or clone tag registered to the system.

5.4. Proposed NFC Filtering Framework

The proposed NFC filtering framework offers hashing operations that merges Intent filtering with Bloom filtering to authenticate access, and have AES cryptographic as the prevention to content protection. Figure 10 shows the proposed NFC filtering framework.

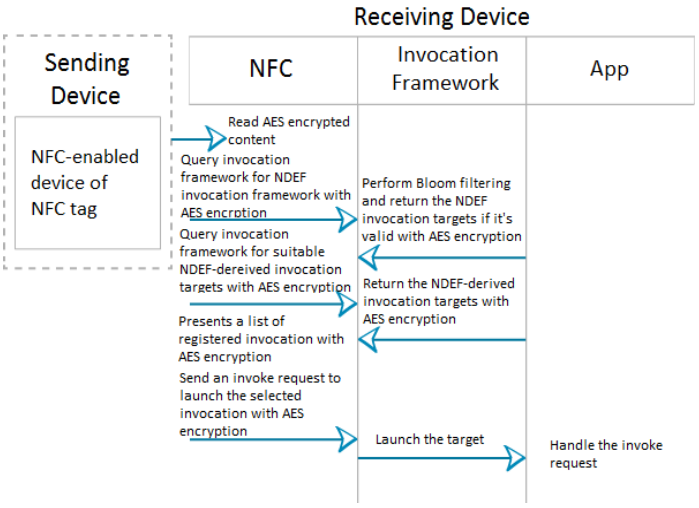


Figure 10: Proposed NFC filtering framework

AES 128bit cryptographic has been implemented into NDEF message and it's indirectly help in encrypting the original NDEF message with a key size of 128bit. The encryption standard performed the encryption on the NDEF message and sent it over the communication channels. MITM attack might be able to retrieve the NDEF message but unfortunately, the message is encrypted with AES encryption. By using brute force attack, it will take about 80 trillion years++ to crack the key and this nature can be timely and resource-consuming. On top of this, Bloom filtering help to perform hashing operation using MD5 technique to verify or authenticate access level to the NFC system. This operation help to prevent from MITM attack by eliminating the duplicate tag which act as the access control to the NFC system. Communication between two legitimate entities are more secure against MITM attack who attempt to retrieve tag or personal information over the network.

5.5. Proposed NFC Filtering with PbD

Privacy by Design (PbD) advances the view that the future of privacy cannot be assured solely by compliance with regulatory guidelines. In this research project, we have referred to PbD's third foundational principle as a guideline which is embedded into the design. In other words, privacy prevention should be implemented into the architecture of IT systems [8]. The result is that privacy becomes an essential component of the core functionality being delivered. Privacy is integral to the system, without diminishing functionality.

In the proposed NFC filtering, AES encryption standard has been implemented on the NDEF message been transmitted. Bloom filtering has been merged together with the Intent filtering technique. Indirectly, this shows that privacy prevention has been implemented into the system or architecture of the NFC system. We could conclude that the proposed NFC filtering is succeeding the third foundational principle of PbD principles as the privacy approach is embedded on the NFC system itself.

5.6. Possible threats in MIDAS

It is a common issue in most of the IT system, where there are some possible threats that might take place on the system. Below are the possible threats that might occur in MIDAS system.

- Card Cloning - card cloning is the major issue or threat of the MIDAS system. Although each of the card has its own ID and it's only readable. The student ID card can be easily duplicated with the same card information using some online model tools.
- Encryption - the server does not own any security certification or encryption standard to ensure the security of the communication channel. This is very important for the data being transmitted over the communication channel in order to prevent man-in-the-middle attack.

5.7. Proposed Filtering Framework in MIDAS

Students would attempt to tag NFC-enabled smartphone or card to represent their attendance. A NDEF message is sent over through the channel to the reader. Here, the system performed the Bloom filtering to verify the received tag ID. Once the tag has been verified, the system proceed with the Intent filtering by reading the registered activities on the tag. This means the tag directed the user to the selected invocation target that has been registered on the tag. This process carried the operation of ACTION_NDEF_DISCOVERED. According to the MIDAS operation, the activities registered will login the session to the attendance system and followed by submitting of the attendance. The login session is the authentication process to the system while Bloom filtering operates to verify the tag ID and eliminate duplicate tags. In this situation, the system will indirectly prevent towards MITM attack where duplicate tags or ID are automatically eliminated by the Bloom filtering. Meanwhile, the message been transmitted across the communication between the tag and reader has been secured with the AES 128bit encryption standard. The NDEF message communicated is encrypted before it is transmitted through the NFC communication channel. In some cases, if an attacker is able to break the Bloom filtering barriers, he or she is able to retrieve the NDEF message. But unfortunately, this message has been encrypted with AES encryption standard. This cause time and resource-consuming for the attacker to interpret the original NDEF message.

5.8. Comparison of Proposed Framework vs Existing Filtering

NFC tags can range in complexity and standard of the tag selected. Simple tags offer just read and write operations, sometimes with one-time-programmable areas to make the tag read-only. The most common MIFARE tags contain operating environments, allowing NDEF message interactions with code execution on the tag. Figure 11 shows the common NFC operation.

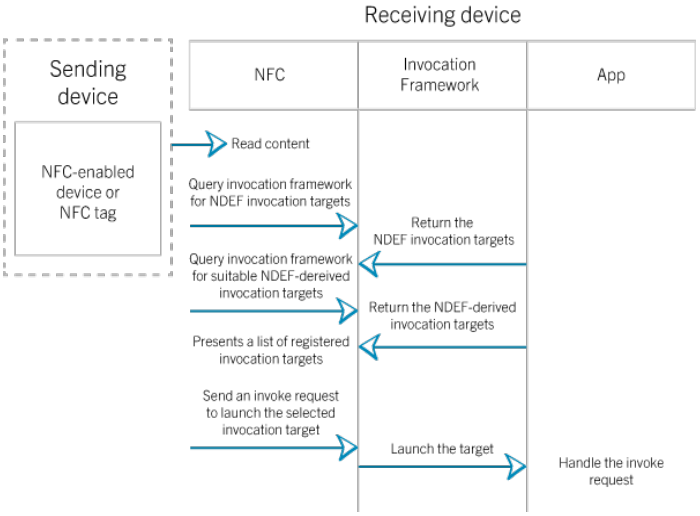


Figure 11: Common NFC Operation

NFC Data Exchange Format (NDEF) is a binary message format that can be used to encapsulate one or more application-defined payloads. NDEF is well defined as a message encapsulation format to exchange information between a RF signal transmission tunnel as well as NFC communication. A reader detects a tag with NDEF technology, then proceed with the Intent filtering in order for the user to gain access to the NFC system. Next, the tag began direct the user to the selected invocation target that has been registered on the tag. This process carried the operation of ACTION_NDEF_DISCOVERED.

Alternatively, proposed NFC filtering framework is more complex which offer hashing operations that merged Intent filtering with Bloom filtering to authenticate access, and have AES cryptographic as the prevention to content protection. Implementation of AES cryptographic in NDEF message, help to encrypt the original NDEF message with a key size of 128bit which will cause approximately about 80 trillion years++ to crack the key using brute force attack.

On top of this, Bloom filtering performed the hashing operation using MD5 technique to verify or authenticate access level to the NFC system. This operation help to eliminate the duplicate tag which Intent filtering does not handle. The Bloom filtering is followed by Intent filtering where the user is directed to the selected invocation as registered on the tag after the Bloom filtering verification. Communication of two legitimate entities is secured with AES encryption. Therefore, the possibility for a MITM attacker to retrieve tag or personal information over the network is very low since the attacker will require a long process to decrypt the NDEF message.

6. Discussion and conclusion

This chapter discussed the overall completed research project including the limitation of the research project. This chapter also discuss the successful achievements and the future enhancement of the research project as the weaknesses of the completed project.

6.1. *Security Enhancement Using Proposed Technique for Filtering*

Bloom filter is a technique used in RFID technology which is space effective and probabilistic data structure. The major benefit of the Bloom filters is its strong space advantage over other data structures for representing sets, such as self-balancing binary search trees, tries, hash tables, or simple arrays or linked lists of the entries. A bloom filter does not store the elements themselves, and this is the crucial point. You do not use a bloom filter to test if an element is present, you use it to test whether it is certainly not present. This lets you to not do extra work for elements that do not exist in a set.

The use of Bloom filtering in NFC technology prevents from MITM attacks in terms of eliminating the duplicate tag by avoiding the third entities who attempt to gain access communication between two legitimate entities. Implementation of Bloom filtering in NFC architecture provided better security solution in which the current Intent filtering does not provide. Besides that, all in significantly less space than something like a hash table which is likely going to be partially on disk for large data sets make Bloom filtering much more capable to be implemented in NFC technologies. You may use a bloom filter in conjunction with a structure like a hash table, once you are certain the element has a chance of being present.

6.2. *Security Enhancement Using Content Protection*

The Advanced Encryption Standard or AES is a symmetric block cipher used to protect classified information and is implemented in software and hardware throughout the world to encrypt sensitive data. In this research project, AES 128-bit has been implemented on the NDEF message as the content protection of the message being communicated. The selection process to find this new encryption algorithm was fully open to ensure a thorough, transparent analysis of the designs. The use of AES 192bit or 256bit could provide a more complex encryption standard. AES is more secure than its predecessors such as DES and 3DES as the algorithm is stronger and uses longer key size. It also enables faster encryption than DES and 3DES, making it ideal for NDEF message as well as NFC technology.

Security enhancement in terms of content protection is highly recommended as encryption standard could provide a better and secure NDEF transmitted message. Indirectly, a MITM attacker would require high time and resource in interpreting the NDEF message in order to get the original message. Since NFC commination occur in an open environment, there is high possibility for an attacker to simply gain access to the channel and retrieve the sensitive and personal information on the tag including the tag information. Encryption of NDEF message avoided this as the NDEF message has been encrypted with complex encryption algorithm and it time and resource-consuming.

6.3. *Framework: PbD with Proposed NFC Filtering*

Privacy by Design (PbD) provides a set of seven major foundational principles that we could follow to prevent privacy threats or attacks. The result is that privacy becomes an essential component of the core functionality being delivered. This approach is rooted in a belief that reliable protection in a complex ecosystem can only be achieved through an integrated design approach. Practice of using privacy guidelines is indirectly guide us to concerns on privacy as an essential component of a core functionality of a system. Since NFC technologies is a new era of wireless communication or sensor networking, there is pretty much possible threats and privacy attacks. This is with the consideration that privacy prevention provide a better protection on securing the sensitive and personal information. In reflecting this research project, privacy prevention would help to prevent leakage of tag and personal information during the commination which might be caused by MITM attack. Table 15 shows the achievement of proposed work on using PbD as the privacy guideline for this research project. We could conclude that the proposed NFC filtering is succeeding the third foundational principle of PbD principles as the privacy approach is embedded on the NFC system itself.

Table 8: PbD with proposed framework

PbD Foundational Principles	**Privacy**	**Achievement**
1. **Proactive not Reactive; Preventative not Remedial**	Anticipate and prevent privacy. Prevent from privacy risks to materialize.	Implementation AES encryption provide prevention of privacy.
2. **Default Setting**	Build privacy measures directly into any IT system.	Filtering technique as the build-in privacy approach.
3. **Embedded into Design**	Embed privacy into the design and architecture of IT system.	Bloom filtering and AES encryption is embedded into NFC architecture.
4. **Positive-Sum**	Accommodate all legitimate interests and objectives in a positive-sum. “Win-win manner.	Privacy approach in proposed framework does not against the security approach in existing NFC architecture.
5. **End-to-End Security**	Secure life-cycle management. Ensure cradle-to-grave.	The proposed framework provide end-to-end security with an encryption technique.
6. **Visibility & Transparency**	Component parts of IT systems practices visible and transparent.	The security and privacy approach implemented is transparent to end-user.
7. **Respect for the User**	Respect and protect interests of the individual. Keep it user-centric.	Interest of end-user is much appreciate and respect without any limitation to utilize the NFC system.

6.4. *Limitation of the Proposed Framework*

Limitations are influences or weaknesses of the proposed framework which can be considered as the future work. They are the shortcomings, conditions or influences that cannot be controlled or can be considered for the future enhancement work. The limitations of the proposed framework are described below.

- Content Protection - AES 128bit has been used as the content protection in this framework. Referring to table 11 above, even with a supercomputer, it would take about 80 trillion years++ to crack 128-bit AES key using brute force attack. But unfortunately, this is not adequate nowadays since there are some tolls available for brute force attacks. This might help an attacker in terms of time and recourse-consumption. For tighter security, we might consider on using AES 192bit or 256bit key size which would require a long time to crack. For public-key cryptosystems, we could go for Rivest-Shamir-Adleman (RSA) cryptographic for secure data transmission.

- Filtering technique - Bloom filtering technique has been merged together with Intent filtering in this research project. Bloom filter is a space with effective and probabilistic data structure. The major benefit of the bloom filter is its size, the constant number of bits and is set upon initialization. Meanwhile, given the fact that the advantages provided by the bloom filter outweigh the small limitation. In this research project, we have used MD5 hashing function instead of MD5. We could consider on using Secure Hash Algorithm 1 (SHA1) which could produce a 160-bit (20-byte) hash value known as a message digest.

6.5. *Research Contribution*

Near Field Communication (NFC) technology play an important role nowadays and depends on contactless equipped devices. NFC technologies or contactless equipment devices minimize human man power or work load on handling varied comprehensive tasks in their daily routine. For example, the use of Smart Card/Tag on highways toll, enables a person to just pass by the sensors and the amount is deducted exactly as how we pay at the toll counter. NFC technology is a part of the RFID technology, core from wireless communication technologies that tent to be exposed to multiple security and privacy threats or issues as the communication occur in an open environment.

The most common threats in NFC communication is Man-in-Middle attacks that attempt to corrupt tag data and the most significant issue is to gain access to personal information and sensitive area. The contribution in this research is towards the issue or problem that most users come across when using NFC devices. This research has focused on the privacy threats caused by Man-in-the-Middle attack. In other words, the sensitive or personal information or data that could be retrieved from a NFC-embedded device which is active in an open environment. For example, an attacker might have the intention to retrieve the last transaction history of an NFC smart card or bank card. Those can be counter measured with access control which could pretend as an unauthorized access or communication as the first level security during when the communication is established.

This research has been conducted on the main purpose to enhance the tag filtering technique in NFC architecture. At the moment, Intent filtering is the build-in technique in NFC architecture which perform filtering according to what has been registered to the read tag. Bloom filter technique has been merged together with Intent filtering, so the bloom filtering could filter the unauthorized access as well as eliminate the duplicate tag readings. Since the existing Intent filtering only be able to filter activities registered to the tag, bloom filter cut off the duplicate tag and pretend from unauthorized access to the NFC system. Additional to this privacy prevention, content protection modules have been implemented on the payload of the NFC message. The AES approach has been used in order to encrypt and decrypt the NFC message before it is sent over to the open network. Indirectly, this encryption technique help in terms of content protection of privacy information.

Both tag filtering and content protecting enhancement on the NFC architecture have followed the PbD foundational principle which consist of 7 major principles. This principles defined that privacy approach should be an essential component of the core functionality being delivered of an IT system.

7. Conclusion

Referring to the research in NFC, near field technology provide a useful impact on the usability of hand held devices in various contexts where it could facilitate a user's experience by making it possible to perform multiple services with a single NFC-enabled device. However, every technology has their own cons where NFC has a potentially dramatic impact on users' privacy since the communication occur in an open environment.

User's privacy refers to personal information or sensitive data and credit card details that are stored in NFC-embedded devices which would become the target for hackers and cyber-criminals such as Man-in-the-Middle attack that could cause multiple threats. Development in the NFC technology, together with some aspects such as usability, and level of security need to account for the reasons explained above.

Poor architecture or system implementation, caused multiple security and privacy attacks. In a case of poor communication channel or lack of implementation of security approach can cause leakage of sensitive and personal information. Implementation of Bloom filtering performed the verification of registered user to the NFC system and eliminate duplicate or third party access. Besides that, implementation of AES cryptographic in NDEF message, ensure communication between two legitimate entities is secured. Hence, secured user validation or filtering with encrypted message, prevent the possibility for MITM attacker to retrieve sensitive or personal information.

Acknowledgment

With grace of God, I managed to complete the research project. Even though I has faced a lot of difficulties upon completing of this research. Here by, I would like to express my gratitude towards several individuals those who directly and indirectly guide, help and support on completing this research. First of all, I would like to express my gratitude to my supervisor, Dr. Manmeet Kaur Mahinderjit Singh for her guidance, encouragement and patience

for conducting this research. Without her supervision, I might not be able to complete this research project.

I would also like to thanks each and every one who support me in term of financial and materials, especially to my mother, Mrs. N.Krishnaveni and also my future wife, Ms. M.Krishnapriyah. I really appreciate their effort in help me toward this research completion. Hopefully, all the experiences that I has gained to complete this research project will eventually come to help in the future. May all praises go to mighty God.

References

[1] Aikaterini Mitrokotsa and Michael Beye and Pedro Peris-Lopez, Classification of RFID Threats based on Security Principles, Delft PhD Thesis, University of Technology, 2012, pg. 1-22.

[2] Bing Dai, The Product Authentication Application Design Based on NFC, Master Degree, Vaasa University of Applied Sciences, 8-17 March 16, 2015.

[3] Aikaterini Mitrokotsa, Melanie R. Rieback and Andrew S. Tanenbaum, Classification of RFID Attacks, PhD, Department of Computer Science, Vrije Universiteit, 5-12, 2012.

[4] Liam Church and Maria Moloney, State of the Art for Near Field Communication: security and privacy within the field, Escher Group Ltd, Ireland, 5-6, May 10th 2012.

[5] Siti Salwani Yaacob, Hairulnizam Mahdin, An Overview on Various RFID Data Filtering Techniques Based on Bloom Filter Approach, Master Degree, Universiti Tun Hussein Onn Malaysia, 8-41, 2013.

[6] David Ong, Dr Manmeet Kaur, MIDAS - Multifactor Identification Attendance System, Bachelor Degree, University of Science, Malaysia, 32-68, 2015.

[7] Ann Cavoukian, Mobile Near Field Communications: Keep It Secure and Private, ISSA Journal, 1-2, August 2012.

[8] Ann Cavoukian, Privacy by design – The 7 Foundational Principles, Information & Privacy Commissioner Ontario, Canada, 1-2, January 2011.

[9] Anurag Kumar Jain and Devendra Shanbhag, Addressing Security and Privacy Risks in Mobile Applications, IEEE Computer Society, 25-29, Sept/Oct 2012.

[10] Prapassara Pupunwiwat, Bela Stantic, Location Filtering and Duplication Elimination for RFID Data Stream, International Journal of Principles and Applications of Information Science and Technology, 13-22, Dec 2007.

[11] EMCA International, NFC-SEC - NFCIP-1 Security Services and Protocol - Cryptography Standard using ECDH and AES, EMCA International, 56-59, Dec 9, 2008.

[12] Andreas Rohr, Karsten Nohl, Henryk Plotz, Establishing Security Best Practices in Access Control, Security Research Labs, Publication version, v.1.0, 78-83, 2007.

[13] BIC, Near Field Communication (NFC) and the use of Radio Frequency Identification (RFID) in Libraries, BIC, 1-2, 2012.

[14] International Organisation for Standardisation. ISO/IEC 15693-1. Identification cards - Contactless integrated circuit(s) cards - Vicinity cards, 2-4, 2000.

[15] [6] NFC Forum, NFC Data Exchange Format (NDEF), NFCForum-TS-NDEF_1.0, 3-8, 2006.

[16] BIC, Near Field Communication (NFC) and the use of Radio Frequency Identification (RFID) in Libraries, BIC, 4-7, 2012.

[17] Uwe Trottmann, NFC – Possibilities and Risks, Seminar Future Internet WS2012, Network Architectures and Services, 13-26, Feb 2013.

[18] Ernst Haselsteiner and Klemens Breitfub, Security in Near Field Communication (NFC) - Strengths and Weaknesses, Philips Semiconductors, 11-13, 2012.

[19] Y. Bai, F.Wang, and P. Liu. Efficiently Filtering RFID Data Streams. In CleanDB, 9-34, 2006.

[20] Min Kyung Jeon, Kee Hyun Choi, Bong Jae Kim, Dong Ryeol Shin, Efficient and Secure Copy Protection System for Digital Content in IoT Environments, Research Notes in Information Science (RNIS), 44-48, June 14th 2013.

[21] Wiem Tounsi, Security and Privacy Controls in RFID Systems Applied to EPCglobal Networks, Telecom Bretagne, 33-48, Jan 16, 2015.

[22] Cheng-Hao Chen, Iuon-Chang Lin, Chou-Chen Yang, NFC Attacks Analysis and Survey, 2014 Eighth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, 4-6, 2014.

[23] Reading NFC Tags with Android, 18 Feb 2016, retrieved from: http://code.tutsplus.com/tutorials/reading-nfc-tags-with-android--mobile-17278

[24] Android, NFC Basics, 20 May 2016, retrieved form: https://developer.android.com/guide/topics/connectivity/nfc/nfc.html

[25] Advantages and Disadvantages of NFC, 2 March 2016, retrieved from: http://near-field.blogspot.my/p/pros-cons.html

[26] US to Challenge China for World's Fastest Supercomputer, 16 May 2016, retrieved from: http://thediplomat.com/2015/08/us-to-challenge-china-for-worlds-fastest-supercomputer/

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 149-156 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

Semantic modeling of portfolio assessment in e-learning environment

Lucila Romero[1*], Milagros Gutierrez[2], Laura Caliusco[3]

[1] *lucila.rb@gmail.com –GIDIS Research Group –Facultad de Ingeniería y Ciencias Hídricas Universidad Nacional del Litoral, 3000, Santa Fe, Argentina*

[2] *mmgutier@frsf.utn.edu.ar - CIDISI Research Center–UTN–Facultad Regional Santa Fe, 3000, Santa Fe, Argentina*

[3] *mcaliusc@frsf.utn.edu.ar - CIDISI Research Center–UTN–Facultad Regional Santa Fe, 3000, Santa Fe, Argentina*

ARTICLE INFO

Article history:
Received: 16 December, 2016
Accepted: 19 January, 2017
Online: 28 January, 2017

Keywords:
Pofrtfolio
elearning
semanticmodel
ontology network
assessment
higher education

ABSTRACT

In learning environment, portfolio is used as a tool to keep track of learner's progress. Particularly, when it comes to e-learning, continuous assessment allows greater customization and efficiency in learning process and prevents students lost interest in their study. Also, each student has his own characteristics and learning skills that must be taken into account in order to keep learner`s interest. So, personalized monitoring is the key to guarantee the success of technology-based education. In this context, portfolio assessment emerge as the solution because is an easy way to allow teacher organize and personalize assessment according to students characteristic and need. A portfolio assessment can contain various types of assessment like formative assessment, summative assessment, hetero or self-assessment and use different instruments like multiple choice questions, conceptual maps, and essay among others. So, a portfolio assessment represents a compilation of all assessments must be solved by a student in a course, it documents progress and set targets.

In previous work, it has been proposed a conceptual framework that consist of an ontology network named AOnet which is a semantic tool conceptualizing different types of assessments. Continuing that work, this paper presents a proposal to implement portfolios assessment in e-learning environments. The proposal consists of a semantic model that describes key components and relations of this domain to set the bases to develop a tool to generate, manage and perform portfolios assessment.

1. Introduction

One of the most important task in e-learning is to assess student, because constant assessment avoid students lost interest in their studies. The use of different evaluation methods make students feel that connection is established between them and instructors, and their learning efforts are properly assessed [1](Sun, Tsai, Finger, Chaen & Yeh, 2008). As Bolivar says [2], assessment is a systematic procedure used by teachers to determine the level of student knowledge in a particular discipline, before, during and at the end of an academic subject. Note that assessment should not just be about grading but about assisting the process of learning itself. Accordingly, assessment should be developed and refined in order to assist student to learn effectively and efficiently [3]. From a pedagogical point of view, e-learning demands greater customization and efficiency to leading learners. Then, it requires identify how and when to assess and considering which outfit to be used. Even though portfolios are used in e-learning as tool to keep track of student progress, when its main objective is to assess, it is known as portfolio assessment [4]. In previous work it has been presented a conceptual framework consist of an ontology network, called AONet, which is a semantic tool that implements different kind of assessment used in e-learning [5]. Following with this activity, this work presents an e-portfolio assessment

*Corresponding Author: Lucila Romero, GIDIS Research Group –Facultad de Ingeniería y Ciencias Hídricas Universidad Nacional del Litoral, 3000, Santa Fe, Argentina, lucila.rb@gmail.com

specification. As Lorenzo & Ittelson says [6], an e-portfolio is a collection of digital artifacts like demonstrations, comments, resources and results that characterize a person, group or institution. E-portfolios are widely used to assess students with the advantage of including different type of assessment like essay, conceptual map, objective test that allow teacher evaluate different knowledge level.

In e-learning context, there is a lot of work that address technological aspect but a few work about pedagogical one. In this way, this work contributes to address pedagogical issues because it proposes a conceptual model of portfolio based on ontology. This technology allows coding expert knowledge through logical rules.

This work is organized as follows. Next section explains the background knowledge: the AONet ontology network and the use of e-portfolio in education. Next a literature revision is presented. Then the Portfolio ontology is introduced and a case study is developed. Finally conclusions and future works are presented.

2. Background

This work is part of a mayor task done by author in the e-learning area. In this section, it is presented the ontology network called AOnet [5]. An ontology network is a set of ontologies related together via a variety of different relationships such as mapping, modularization, version, and dependency. The elements of this set are called networked ontologies [7].

2.1. Extending the AOnet ontology network

AONet was developed as the conceptual description of a tool for semi-automatic e-assessment development in the e-learning context. Then, AONet conceptualizes in a modular way, all aspect of assessment domain. In first stage, it has considered three different domains: course topic, educational resources, and assessment. Figure 1 shows AOnet, its domain and ontologies in each of these, which are related one to another through meta relationships.

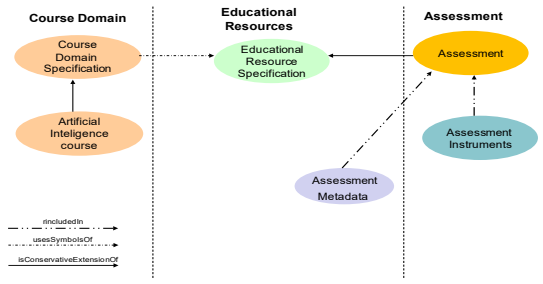


Figure 1: AONet ontology network

In this case, an assessment is an educational resource and as such, it has metadata that describe it. In order to develop an assessment, different instruments can be used. An assessment evaluates topics or subjects that must be learned, which are conceptualizing in Course Domain ontology.

A new stage of this network is presented. In this stage, a new domain was added to the AONet: the Agent domain that represents the stakeholders involved in an educational process since the stakeholders could play different roles that in some cases is necesary to identify. In [8] it is presented a discussion about stakeholders involved in e-learning process considering highe education. In addition, a new ontology in the EducationalResources domain was added: the Portfolio ontology that represents portfolio educational resource. Figure 2 shows the new version of the AONet ontology network.

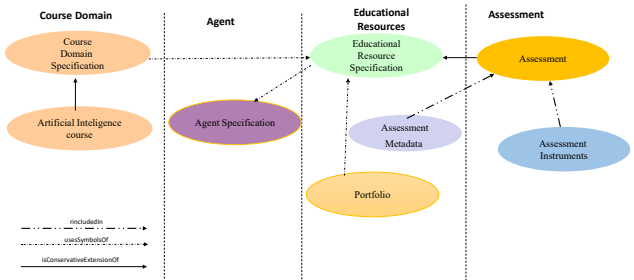


Figure 2: The extended AONet ontology network

2.2. ePortfolios in learning

The electronic portfolio (e-portfolio), in education, is a collection of a students' work that can advance learning by providing a way for them to organize, archive, and display work. The uses of e-portfolios are most common in the courses with departments of education. Most preservice teachers are asked to compile an e-portfolio to demonstrate competencies needed to gain teaching certification or licensure. Student e-portfolios are increasingly being used in other disciplines such as communications, math, business, nursing, engineering and architecture. In education e-portfolios have five major functions [9]:

(1) Document skills and learning; (2) Record and track development within a program; (3) Plan educational programs; (4) Evaluate and monitor performance; (5) Evaluate a course.

The primary aim of an e-portfolio may be to collect evidence for summative assessment, to demonstrate achievement, record progress and set targets – as in records of achievement and individual learning plans (ILPs) – or to nurture a continuing process of personal development and reflective learning, more commonly experienced in higher and continuing education contexts, but now also occurring in further education and schools [10].

E-Portfolios have become a popular alternative to paper-based portfolios because they provide the opportunity to review, communicate and give feedback in an asynchronous manner. Therefore, during the last years, a great amount of e-portfolios system have been developed [11]. These systems use different technologies that students and teachers could use to create multiple e-portfolios from the same repository or set of repositories. Most of the e-portfolio systems are integrated to Learning Management Systems (LMS), giving lot of functionalities.

When e-portfolios is part of an assessment framework, there are advantages and disadvantages for practitioners. On the one hand, digital portfolios offer more efficient working practices, enabling marking and verification to take place incrementally. On the other hand, the diversity of evidence contained in portfolios can make them harder and more time consuming to assess [12]. So, in this context, it is necessary a tool to facilitate the assessment.

3. Literature review

EPorfolios are widely used in the field of academic institutions as a valuable tool for continuous learning, not only for careers offered in distance learning but also for traditional careers. In

recent years, there have been a lot of e-portfolios systems as shown in [11]. Using different technologies, they offer learners and teachers the ability to create their own portfolios that customize their career advancement. Many of these tools are integrated into Learning Management Systems (known by its acronym LMS), offering different features. Some allow the institution to have control over these systems in order to maintain the image, others fail to provide social networking functionality needed in this age. The difficulties in implementing these technologies persist due mainly to technical and teaching skills that teachers are expected to have.

That is, the ePortfolio that was originally static web pages, have evolved demanding the availability of commercial and open source tools in the form of web applications for management of ePortfolios. These tools are developed on formats that do not reflect open standards, and do not facilitate the exchange of the information contained among other limitations. This makes difficult to share the artifacts contained in an ePortfolio between different tools or applications and, even more, sharing an ePortfolio of a student between different educational platforms or LMSs in the context of different educational institutions.

Current limitations on the use of portfolios in the learning process could be grouped as follows: First, the use of ePortfolio to estimate learning should be accepted by teachers. To do this, a tool to support the generation and management of portfolios whose content is valid and reliable, from a pedagogical perspective is needed [13]. This means it must be defined a mechanism to validate if the portfolio covers all learning objectives of a course and meets certain pedagogical principles and allows assessment as proposed by [2]. Secondly, systems must support e-learning portfolios with different types of devices with different levels of complexity and have to adapt to different stages of learning. Thus, the e-learning systems should allow teachers to manage ePortfolios containing diagnostic, summative and formative assessments, while must be capable to manage self-assessments, co-evaluation and hetero-assessments. Third, ePortfolio management should be solid against the collaborative work to resolve differences arising from the administration of multicultural information and the advent of complex environments semantically incompatible.

In order to optimize the benefits of using ePortfolios the integration of the components that comprise it and determine the level of learning should be favored. As part of its components, which are called artifacts, teacher can find work performed by learners, evaluations and results or achievements. These artifacts represent means to facilitate educational continuity between programs within an educational institution. and became into evidence that can be shared and integrated between institutions and organizations throughout their academic performance. To do this requires specific technologies to present a solution for interoperability.

The drawbacks associated with interoperability occur when trying to share and reuse heterogeneous information resources [14]. With the use of technological solutions for the interoperability of information, it is not required to have prior agreement between the institutions managing education on LMS platforms for sharing and reuse of components or devices covered in ePortfolio.

As part of technologies that provide solutions for interoperability and the need for integration of heterogeneous and diverse information, the use of Semantic Web technologies in e-learning in recent years is evident. The Semantic technologies can be exploited as a platform for the implementation of a system of e-learning as it provides all the tools that this type of education requires [15]: ontology based conceptualizations of learning materials, components standardization to share information and educational courses in composition with proactive delivery of learning materials through LMS [16].

The use of semantic technologies for the management of ePortfolios is also evident. In [17] the authors use an ontology to formalize ePortfolios, representing taxonomically different types of portfolios in learning. Although defined ontology allows the description of the portfolios by properties, work does not contemplate the use of standards for it. The work does not address the management of ePortfolios and does not include the learning assessment process.

In their paper, Taibi et al [18] use ontologies to model ePortfolios and social relations in informal learning environments. To do this, they extend the FOAF ontology, modeling interactions and collaborative work based on social networks. While the authors mention standards, they are only based on IMS (IMS Learning Model Portfolio) and do not address in depth the aspect of assessment ePortfolios.

In [19] authors propose a model of ePortfolio based on an ontology but is aimed at promoting self-regulated learning (SRL). In SRL students are aware of their knowledge, skills and, therefore, they consider learning as a controllable process and assume greater responsibility in the outcome of this process. In this work, authors do not consider rules or restrictions that allow specific recommendations of pedagogical nature.

4. Portfolio assessment ontology

In order to conceptualize the portfolio domain, it has been defined the Portfolio ontology, which is shown in Figure 3. The main concept is Portfolio, which has two subtypes: LearnerPortfolio and TeacherPortfolio, representing portfolios owning by students and teachers respectively. The LearnerPortfolio concept has the PortfolioAssessment subtype. Also in figure 3, it can perceive that Portfolio is composed by artifact, a concept representing whatever learning object, for instance Assessment is a subtype of artifact.

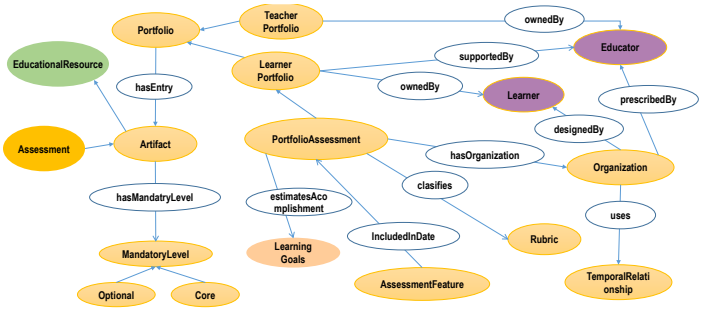


Figure 3: Portfolio ontology

As is declared in [20, 21] a portfolio contains artifacts created by student and teacher such as problems solutions, reflections, descriptions, teacher and classmate feedback, among others. In the ontology, the Portfolio concept is related to the concept Artifact through hasEntry relationship. An artifact is an educational resource, a situation expressed through the inheritance relationship between the Artifact and EducationalResource. The latter concept

is shown in green color in order to highlight it is belonging to the EducationalResourceSpecification ontology, which models all educational resources used in the teaching-learning process such as books, notes chair, presentations, videos, assessment among others. This ontology is not shown because it is beyond the scope of this work. The elements that conform a portfolio have different mandatory levels, represented by the MandatoryLevel concept, which is subclassified into Core and Optional. An artifact has an Optional level associated can or cannot be selected to integrate the student's portfolio, but one that has the Core level associated must be included in the students' portfolio.

A PortfolioAssessment must have organization and content [21]. As regards organization, it can be prescribed by teacher or designed by student also [21, 22]. The hasOrganization relationship relates Portfolio with Organization concept. Then, there are two relations that relate Organization with the responsible agent for setting it: designedBy and prescribedBy. Note that in figure, the agents Educator and Learner appear in different colors because are belonging to another ontology in the network.

Organizing a portfolio primarily establishes the relationships that must exist between the elements belonging to portfolio. For example, there will be temporal relationships to identify in which order elements should be accessed. For this type of relationship, it is proposed to use the temporal relationship defined by Allen [23]: Before, after, during, meets, overlaps, start and finish.

As regards content, it inherits the structure from Portfolio, which has been described previously. In this case, a PorfolioAssessment limits its content to Assessment only. This restriction is set in the logic rule shown in equation (1).

The AssessmentFeature concept labels both, time when Assessment is added to the student portfolio and time when it is resolved or accessed by student.

A LearnerPortfolio belongs to a student and is supervised by teacher. This is expressed in the ontology through ownedBy relationship, that relates it with Learner, and supervisedBy relationship that relates it with the Educator concept. In the case of TeacherPortfolio occurs only ownedBy relationship relating Educator with TeacherPorfolio concepts.

The PortfolioAssessment allows teacher, responsible for the teaching-learning process, to determine whether the learning objectives are achieved [24]. Through LearningGoal concept, the teacher sets goals that students must achieve. Each of these goals will be associated with elements of the student portfolio.

$\forall partOfPortfolio.PortfolioAssessment \sqsubseteq Assessment$	(1)

Moreover, PortfolioAssessment has rubric, which describes criteria or standards associated with learning objective. In the Portfolio ontology, the hasRubric relationship relates PortfolioAssessment and Rubric concepts.

As regards the level of knowledge that is expected to reach the student, revised Bloom's taxonomy was used [25- 26].

The KnowlegeLevel concept conceptualizes the levels of knowledge, which will be associated with the goals that must be achieved.

The ontology that describes assessment are shown in Figure 4. According with the actors involve in the assessment it has identified three types of assessments [27]: (a) Self Assessment is an assessment taking and solved by an student, in this case, the student play the role of teacher and student. He/she evaluates his/her own progress and suggest his/her own qualifications, (b) PeerAssessment is the assessment perform among peers, it allows students to reflect critically and possibly suggest scores on learning of their peers, (c) HeteroAssessment allows teacher assess a group of students.

The hasEvaluator and hasAssessed relations restrict the agents involved in each type of assessment. According with that, hasEvaluator and hasAssessed relations link SelfAssessment with Learner classes. In the same way, the hasEvaluator relation relates HeteroAssessment with Educator and hasAssessed relates HeteroAssessment with Learner classes. Also, PeerAssessmen class has four relations instead of two. One pair of hasEvaluator and hasAssessed relations relates PeerAssessment class with Educator classes meaning that the assessment takes effect between peer educators. The other pair of relation relate PeerAssessment with Learner meaning that the assessment takes effect between peer learners.

Table 1 shows the logical rules to restrict the type of assessment according with agent involved.

Because a portfolio assessment is used for assessing student, it is expected it has at least same selfAssessments associated. In the model presented, this logic restriction is established by equation (2).

$Portfolio \sqcap \exists hasEntry.SelfAssessment \sqsubseteq Assessment$	(2)

The link of a PortfolioAssessment with the Assessment class also allows of including in a portfolio different moments in the evaluation process [28- 30]: (a) formative assessment (Formative concept), when you want to find out if the learning objectives are being achieved and what needs to be done to improve student performance. It performs throughout the entire course and its objective is to help students in their learning process and point out shortcomings and mistakes. (b) Summative assessments (Summative concept) is designed with the aim to measure and judge learning in order to certify, assign ratings or determining promotions. In general, this type of assessment is taken when the course is ending. It is a way for teachers, to validate if the goals set at the beginning were met.

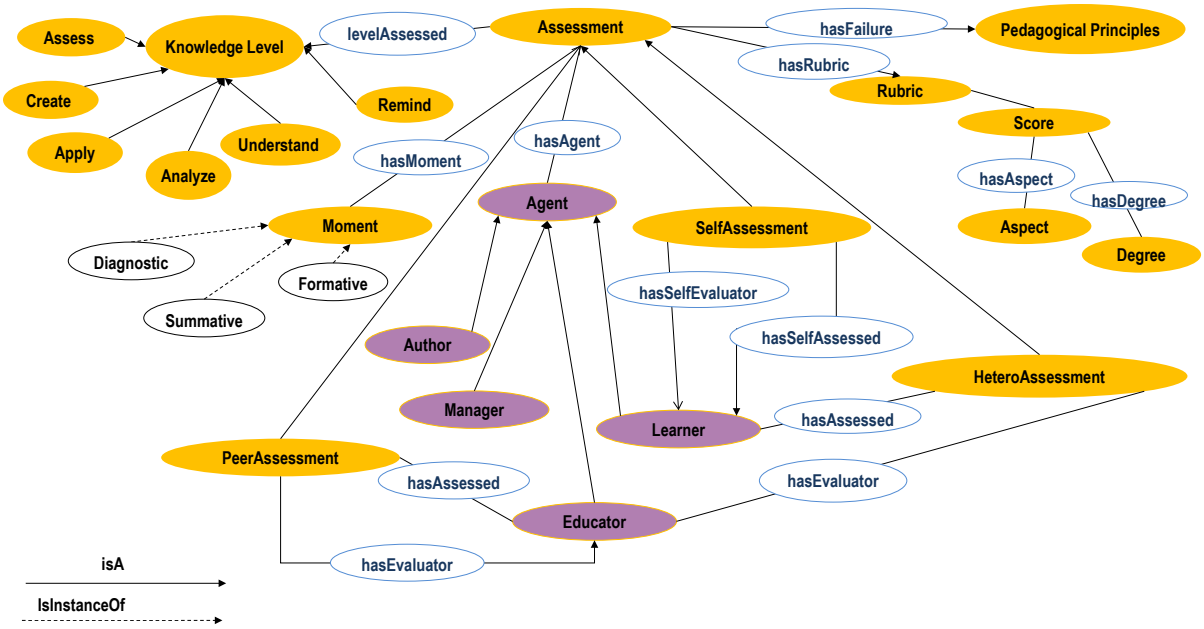


Figure 4: Assessment ontology

Table 1: Assessment ontology formalization

Description	First-order Logic
A peer assessment takes effect between peer educators or between peer learners.	$\Im \models (\forall x,y,z(PeerAssessment(x) \wedge hasEvaluator(x,y) \wedge hasAssessed(x,z)) \Rightarrow ((Educator(y) \wedge Educator(z)) \vee (Learner(y) \wedge Learner(z))))$
A hetero assessment has an educator as the evaluator and a learner assessed.	$\Im \models (\forall x,y,z(HeteroAssessment(x) \wedge hasEvaluator(x,y) \wedge hasAssessed(x,z)) \Rightarrow (Educator(y) \wedge Learner(z)))$
A self assessment has a learner, which is both evaluator and assessed.	$\Im \models (\forall x,y(SelfAssessment(x) \wedge hasEvaluator(x,y) \wedge hasAssessed(x,y)) \Rightarrow Learner(y))$
An assessment always has an author.	$\Im \models (\forall x\ Assessment(x) \Rightarrow \exists y\ (hasAgent(x,y) \wedge Author(y)))$

(c) Diagnostic assessments (Diagnostic concept), take effect early in the teaching process with the aim to determine the level of knowledge of students before starting the learning process. It is a way for teacher, to adjust and adapt the course content according to the results [2]. These concepts are linked to the concept Assessment through hasMoment relationship.

Summarizing, AOnet defines a set of tools that can be used in the definition of assessments and thus achieve different levels of knowledge. This work has been presented in [5].

5. Case study

For a better understanding of the ontology network, a case study is presented. The case study shows instances that correspond to real cases of subjects from Ingeniería en Sistemas de Información study of Universidad Tecnológica Nacional, Facultad Regional Santa Fe.

In order to continue with the implementation of the AOnet considering Portfolios assessment, new relationships linking concepts from different ontologies of the network were incorporated.

5.1. AOnet Meta relation implementation

Figure 5 shows established and implemented relationships for the integration of different ontologies of the network. As stated above, the Portfolio concept is related to Learner concept through ownedBy relationship, and is related to the concept Educator through the relationship supervisedBy (Learner and Educator are concepts of Ontology Agent). Portfolio concept is also associated

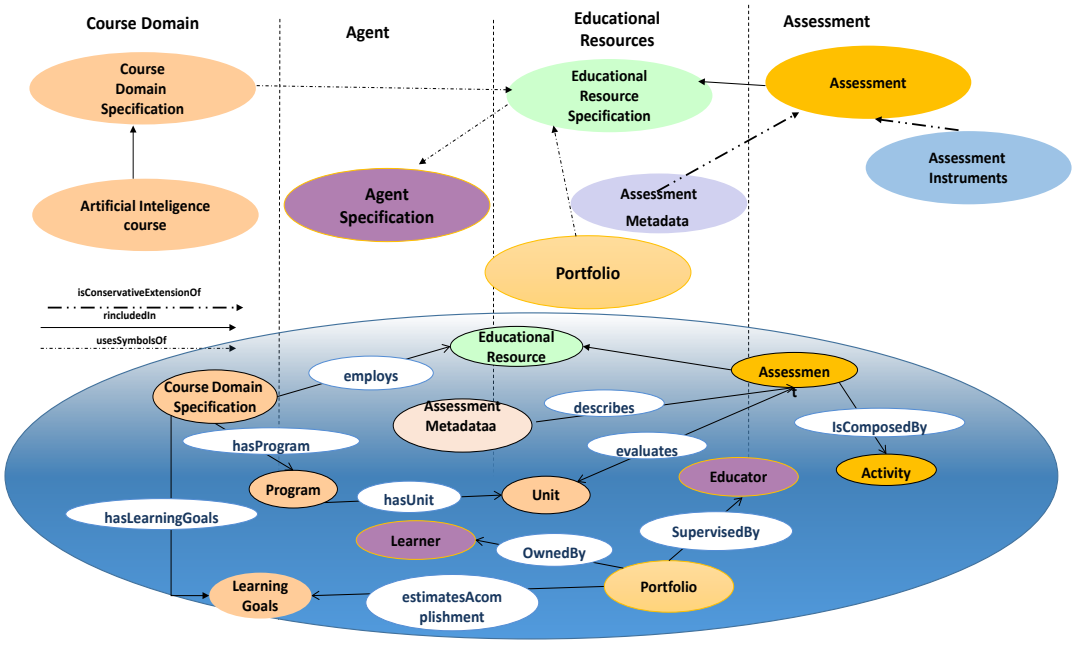


Figure 5. AOnet implementation

with the concept LearningGoal as a portfolio of assessments to estimate the achievement of the learning objectives of the subject.

Assessment concept is an EducationalResource specialization concept. The Assessment concept is related to AssessmentMetadata concept through isDescribedBy relationship, stating that an assessment is described by the corresponding metadata. In practice the ontology also implements the inverse relationship named describes, which states that metadata modeled in AssessmentMetadata concept describes assessments belonging to Assessment class.

5.2. AOnet instantiation

Figure 6 shows some AONet instances used to exemplify the implementation of portfolios for assessment corresponding to the course mentioned above.

Figure 6 shows Artificial Intelligence course (AICourse concept). Artificial Intelligence course has two units: Machine Learning unit and Ingtelligent Agent unit (machineLearningUnit and intelligentAgentUnit instances of AIUnit concept, links expressed through the relationship hasUnit.

MachineLearningAssessment and IntelligentAgentAssessment are instances of Assessment concept. The first one evaluates the unit machineLearningUnit of AICourse course and the second assessment evaluates the intelligentAgentUnit.

AIPortfolioAssessment is a portfolio of assessments (instantiated in Assessment class) belonging to the student María Gómez (instance mariaGomez of Learner class) and is supervised by the teacher responsible for the subject Milagros Gutiérrez (instance milagrosGutierrez from Educator concept), situation expressed through relationships ownedBy and supervisedBy).

AIPortfolioAssessment contains assessments of intelligentAgentUnit and machineLearningUnit. This situation is expressed through the relationship hasEntry that links instances of AIPortfolioAssessment with instances of Assessment concept.

AIPortfolioAssessment has an organization (represented by aIPortfolioOrganization instance and hasOrganization relationship) that uses the before instance of TemporalRelationship concept to indicate that in the portfolio assessments mentioned, assessments corresponding to machineLearningUnit must be resolved before assessments of intelligentAgentUnit.

AIPortfolioAssessment allows educators to set learning objectives, for example to understand the concept of Intelligent Agent. This situation is expressed by the relation estimateAccomplishment that links iAPortfolioAssessment (instance of PortfolioAssessment) with understandIntelligentAgentConcept i(nstance of LearningGoal concept).

The case study presented validates the ontology network from the point of view of their coverage, ie, as an appropriate instrument to express the necessary concepts and relations between them to define portfolios. This ontology is the conceptual framework to build a tool with the necessary functionalities for portfolio management.

One of the advantages that immediately arise when considering ontologies based tools is the representation of knowledge. A fundamental difference between generating knowledge and generating data, is that knowledge can be used by agents to learn, infer new knowledge and communicate with other agents.

Thus, this tool can guide learners in the order that is necessary to incorporate the concepts of a particular material, or noting

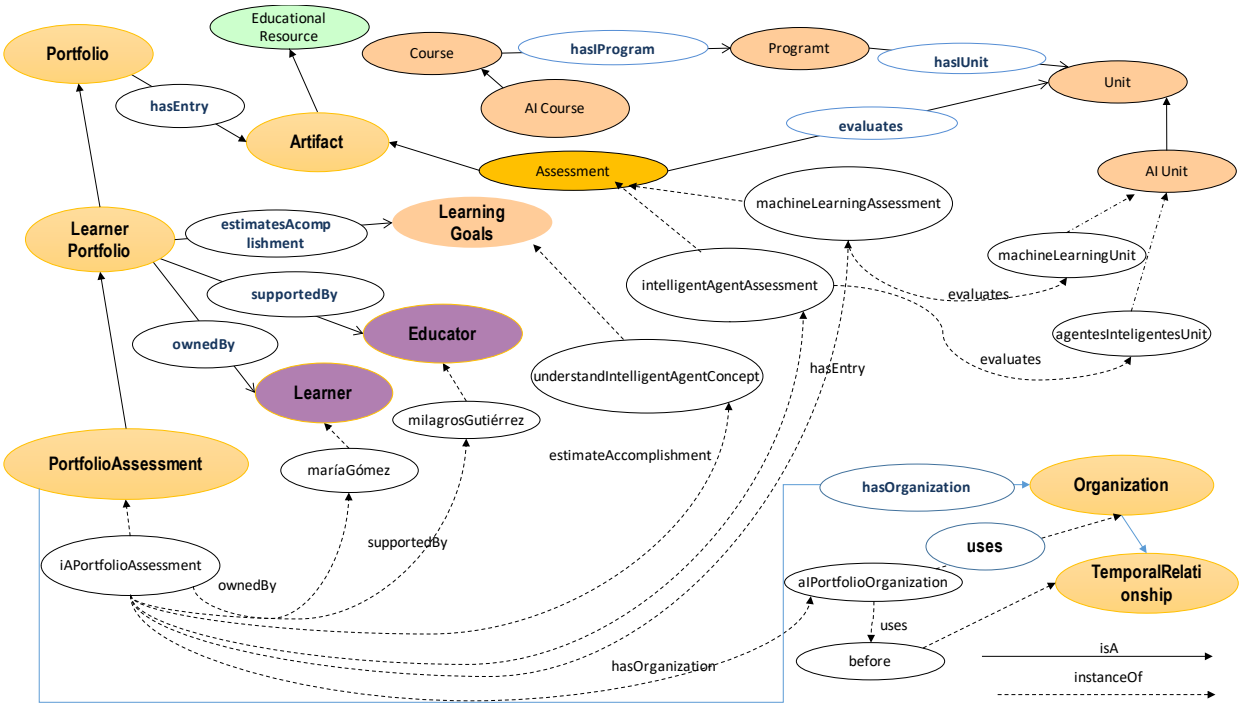


Figure 6. Artificial Intelligence Case Study

teachers on the lack of assessments for a given unit or in a particular moment. The tool can also provide means of search and retrieval of learning materials in an efficient way. In short, the use of ontologies to describe a domain can treat the information from a semantic point of view resulting in intelligent tools for knowledge.

6. Conclusions

The pedagogical objectives of e-portfolios are various: they allow students to describe their learning path, increase awareness of their strengths and weaknesses, take responsibility and increase their autonomy and have a unified way of presenting their competences. However, to ensure the success of a learning activity, teachers must consider different variables that make the e-portfolio assessment harder and more time consuming.

The importance of a system based on formal rules that assist the assessment of an e-portfolio is undeniable. In this work an ontology that formalize portfolios for the learning process assessment in an e-learning context was presented. This ontology is a part of the AONet ontology network that model different areas of interest to be considered in the assessment domain assisted by technologies. The advantages in the use of this ontology network is the domain modular organization that facilitate the collaborative work.

As a continuation of previous work, this contribution can go a step further by incorporating the concept of portfolio as a valuable tool to assess learning considering key aspects in the execution of these techniques, such as intervening actors, levels of knowledge to evaluate, instruments to use, goals to achieve, and so on.

Through logical rules restrictions on modeling these concepts were defined. Rules are also used to represent pedagogical aspects that although not restrict the model, can be used as a guide for teachers who use these tools to assess their students and improve teaching process in general. These pedagogical rules are part of the proposed formalizing the target domain. That is, not only seeks to present a new ontology but is intended with this work contribute to building tools that support teachers in their daily work by providing recommendations based on expert knowledge and improving, in this way, quality teaching process in its entirety.

The use of the ontologies and semantic web tools may provide advanced features for evaluating competency and performance, sharing resources, and seeking help. As a future work we will conduct experiments and analysis methods to continue to evaluate and improve the AONet Ontology Network and to incorporate this new ontology in a tool that is being developed for the semi-automatic generation of assessments as presented in [5].

Acknowledgements

The authors are grateful to Universidad Tecnológica Nacional (UTN), Consejo Nacional de Investigaciones Cientícas y Técnicas (CONICET) and Universidad Nacional del Litoral (UNL) for their financial support.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Sun, P.C.; Tsai, R.J.; Finger, G.; Chen, Y.Y. & Yeh, D. (2008). What drives a successful e-learning? An empirical investigation of the critical factors influencing learnir satisfaction, in Computers and education, Elsevier, 1183 – 1202

[2] Bolivar, C. (2011). Pruebas de rendimiento académico. Technical report. Programa interinstitucional doctorado en educación

[3] Roberts, T. S. (2006). Self, peer and group assessment in e-learning. IGI Global

[4] Chang,C.; Tseng, K. & Lou, S. (2012). A comparative analysis of the consistency and difference among teacher-assessment, student self-assessment and peer-assessment in a web-based portfolio assessment

environment for high school students. In Computers & Education 58, 303-320.

[5] Romero, L., North, M., Gutiérrez, M., Caliusco, L. (2015). Pedagogically-driven ontology network for conceptualizing the e-learning assessment domain. In Journal of educational technology and society, 18(4), 312-330. http://www.ifets.info/issues.php?show=current

[6] Lorenzo, G & Ittelson, J. (2005). An overview of ePortfolios. In Educause Learning Initiative. From https://net.educause.edu/ir/library/pdf/ELI3001.pdf

[7] Allocca, C., D'aquin, M., & Motta, E. (2009). DOOR-towards a formalization of ontology relations. International Conference on Knowledge Engineering and Ontology Development (KEOD), Madera, Portugal. Retrieved from http://oro.open.ac.uk/24326/1/keod09.pdf

[8] Romero, L.J., Ballejos, L.C., Gutiérrez, M.M. & Caliusco, M.L. (2014). Stakeholder's analysis in e-learning software process development, in European alliance for innovation. 15(2), ISSN 2032-9253, 2015 http://eudl.eu/doi/10.4108/el.2.5.e4

[9] Mendoza-Calderón, M.A. & Ramirez-Buentello, J. (2006). Handbook of Research on ePortfolios. Facilitating Reflection Through ePortfolio. In Tecnológico de Monterrey. Hershey, USA. Ali Jafari (Ed). 484-493 ISBN 1-59140-890-3.

[10] Gray, Lisa (2008). Effective practice with e-Portfolios: Supporting 21st century learning. From: http://www.jisc.ac.uk/media/documents/publications/effectivepracticeeportfolios.pdf

[11] Chen, H. & Ittelson, J. (2002). EPAC ePortfolio-related Tools and Technologies. From http://epac.pbworks.com/w/page/12559687/FrontPage [accedido, 16/09/2016].

[12] Strivens, J. (2006). Efficient assessment of portfolios. Fom: http://www.open.ac.uk/opencetl/sites/www.open.ac.uk.opencetl/files/files/ecms/web-content/Strivens-(2006)-Efficient-assessment-of-portfolios.pdf

[13] Olfos, R. & Zulantay, H. (2007). Reliability and Validity of Authentic Assessment in a Web Based Course. Educational Technology & Society, 10 (4), 156-173.

[14] Sheth A. (1998). Changing Focus on Interoperability in Information Systems, Syntax, Structure to Semantics. Interoperating Geographic Information Systems. M F Goodchild, M J Egenhofer, R Fgeas and C AKottman (eds).

[15] Stojanovic, L., Staab, S. & Stuber, R. (2001). E-Learning based on the Semantic Web. In WebNet2001 - World Conference on the WWW and Internet, Orlando, Florida, USA.

[16] Chung, G., Niami, D. & BewLey, W. (2003). Assessment Applications of Ontologies. In Annual Meeting of the American Educational Research Association.

[17] Lougheed, P., Bogyo, B., Brokenshire, D., & Kumar, V. (2005). Towards formalizing electronic portfolios. In Workshop on Applications of Semantic Web Technologies for e-Learning at the Knowledge Capture. From http://www.brokenshire.ca/david/artifacts/TowardsFormalizingElectronicPortfolios.pdf

[18] Taibi, D., Gentile, M., Fulantelli, G., & Allegra, M. (2013). An Ontology to Model e-portfolio and Social Relationship in Web 2.0 Informal Learning Environments. In International Journal of Computers Communications & Control, 9836(5).

[19] Nguyen, L. T., & Ikeda, M. (2014). ePortfolio System Design Based on Ontological Model of Self-Regulated Learning. In 3rd International Conference on Advanced Applied Informatics (IIAIAAI), 301-306, IEEE ed.

[20] Chang, C., Liang, C. & Chen, Y. H. (2013). Is Learner Self-assessment Reliable and Valid in a Web-based Portfolio Environment for High School Students?. In Computers & Education, 60(1), 325-334.

[21] Van der Schaaf, M., Baartman, L. & Prins, F. (2012). Exploring the Role of Assessment Criteria during Teachers' Collaborative Judgement Processes of Students' Portfolios. In Assessment & Evaluation in Higher Education, 37(7), 847-860.

[22] Vance, G., Williamson, A., Frearson, R., O'Connor, N., Davison, J., Steele, C. & Burford, B. (2013). Evaluation of an Established Learning Portfolio. In Clinical Teacher, 10(1), 21-26.

[23] Allen, J. (1983). Maintaining knowledge about temporal intervals. In Comunication of the ACM, 26, 832-843.

[24] Scherba de Valenzuela, J. (2002). Defining portfolio assessment. From http://www.unm.edu/~devalenz/handouts/portfolio.html

[25] Bloom, B. & Krathwohl, D. (1956). Taxonomy of educational objectives. The classification of educational goals by a committee of college and university examiners. Handbook 1. Cognitive domain. New york, Addison-Wesley.

[26] Anderson, L.W. & Krathwohl, D. (2001). A Taxonomy for Learning, Teaching and Assessing: a Revision of Bloom's Taxonomy of Educational Objectives. Longman, New York.

[27] Roberts, T. S. (2006). Self, peer and group assessment in e-learning. IGI Global.

[28] Quesada Castillo, R. (2006). Evaluación del aprendizaje en la educación a distancia "en línea". In RED: Revista de Educación a Distancia, 5(6).

[29] Monteiro, M. & Lobato Miranda, G. (2012). Validation of the Electronic Portfolio Student Perspective Instrument. Sistemas y Tecnologias de Informacion. In 7ma Conferencia Ibérica de Información. Madrid, España. 1(1) Alvaro Rocha, José A. Calvo-Manzano, Luis Paulo Reis, Manuel Pérez Cota Ed.

[30] Ramírez Vega A., Fallas Hidalgo, M. & Chacón Rivas, M. (2012). Motor de Juegos para la creación de evaluaciones en e'learning. In 7ma Conferencia Ibérica de Información. Madrid, España.1(2), Alvaro Rocha, José A. Calvo-Manzano, Luis Paulo Reis, Manuel Pérez Cota Ed.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 157-166 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

Management of Health Information in Malawi: Role of Technology

Patrick Albert Chikumba*

PhD Student, Department of Informatics, University of Oslo, 0316, Oslo, Norway

Applied Sciences Faculty, Department of Computing and Information Technology, University of Malawi, Malawi

ARTICLE INFO

Article history:
Received: 17 December, 2016
Accepted: 20 January, 2017
Online: 28 January, 2017

Keywords :
DHIS2
Health information
Health information management
Health management
Health management information system

ABSTRACT

This paper is an extended version of the conference paper presented at IST Africa Week Conference 2016 and it discusses in detail the existing technology gaps using DHIS2 (District Health Information System 2.0) as an example, and how Geographic Information System (GIS) and mobile application, as specific examples of technology, can enhance health management information system (HMIS) in Malawi. The paper focuses on management of health information. When organisation information is made available, it is expected that the decision-makers use it objectively making rational decisions. This can be achieved by how the information is organized, integrated and presented probably through technology. Along with the increase in strengthening HMIS, questions of how to support the management of information at various organizational levels arise. Research on technologies in health management in developing countries has been on single technologies. Therefore, in this paper, the interest is on multiple technologies and how they support each other to enhance health information management. It has been observed that when it comes to health information management, HMIS employs a mix of paper-based and technology-based practices. Taking into account the infrastructure in Malawi, as in many developing countries, this is probably the most feasible approach. Hence, discussions of existing technology gaps include both paper-based and technology-based practices and how to better support health information management practices through this mixed use of media. The case study confirms that technology plays a role in strengthening HMIS. However, this should be supported by enhancing a culture of information management. It has been noted that DHIS2 is the main information system but it requires the enhancement through inclusions of other technologies. The DHIS2 alone cannot do everything.

1. Introduction

This paper is an extended version of the conference paper presented at IST Africa Week Conference 2016 [1]; excluding the case of Burkina Faso. This paper discusses in detail the existing technology gaps using DHIS2 (District Health Information System 2.0) as an example, and how Geographic Information System (GIS) and mobile application, as specific examples of technologies, can enhance health management information system (HMIS) in Malawi. HMIS is an information system for health management at district, state, regional, and national levels [2]. Malawi HMIS supports community, facility, district, zonal and national levels. When we talk of HMIS in Malawi, DHIS2, as the central data repository, is the main focus. Hence, in this paper, discussions are around DHIS2. DHIS2 (www.dhis2.org) is a tool for collection, validation, analysis, and presentation of aggregate and patient-based statistical data, tailored (but not limited) to integrated health information management activities.

The availability of operational, effective and efficient HMIS is an essential component of the health management capacity [3] because it aims to ensure the appropriate and effective use of resources [4]. In Malawi, since 2002, there have been various

*Corresponding Author: Patrick Albert Chikumba, Department of Informatics, University of Oslo, 0316, Oslo, Norway
Phone: +265885123533, +4795754367
Email: pchikumba@poly.ac.mw, patricc@ifi.uio.no

initiatives in order to strengthen its national HMIS; different technologies have been implemented including DHIS2.

Demand for health information exists [4]. Health managers always need health information when they are carrying out their activities. The health managers need necessary and sufficient information for monitoring, evaluating, and planning their activities [5]. In developing countries, it is important to strengthen health management teams by providing them with resources [6] such as information. When information is made available, it is expected that the decision-makers (e.g. in this study, health managers) use it objectively making rational decisions [3]. This can be achieved by how the information is organized, integrated and presented probably through technology.

Along with the increase in strengthening HMIS in Malawi, as in other developing countries, questions of how to support the management of information at various organizational levels arise. Research has been done on technologies in health management in developing countries. However, the focus has been on single technologies; for instance, recently in Malawi, there has been research on mobile health [7, 8], league table [9], GIS [10, 11], among others. In this paper, the interest is on multiple technologies and how they support each other to enhance health information management. From this perspective, two research questions have been drawn: How do technologies support each other in enhancing the management of health information? How can a new technology be integrated with existing ones in order to strengthen health management information systems? To answer these questions, the empirical data has been analysed and discussed from the perspectives of information cycle [12] and lines of development in health information systems (HIS) [13]. These perspectives are explained in the next section.

2. Related Literature

Health management is broad and includes, for example: care of health facilities, provision and handling of resources, coordination of health programmes, and reporting on health of communities and performance of health teams. According to Mintzberg framework, managerial roles are categorized as interpersonal, informational and decisional roles, which are further re-grouped in Shapiro framework as information generation and transmission, and formulation and execution of decisions [14]. In this paper, the focus is on the informational role, in which a manager is the monitor, disseminator, and spokesperson [14]. In order to play all these informational roles, health managers require managing properly information at their disposal, which include data collection, analysis, interpretation, utilization and dissemination of such information.

The important role of information in managerial work has been always recognized [15]. Information is a precondition for action [16] and it is gathered and used because it helps make a choice [17]. For instance, an organisation can use information to make sense of change in its environment; create new knowledge for innovation; and make decisions about courses in action. It is important for the information to have value in order to be useful. Information needs be in formats to support multiple users; otherwise it is of little value. The value of information depends in a well-defined way on the information's relevance to the decision to be made [17]. Schulte et al. [16] argue that the value is added to information when being transferred and distributed, for example, by translating from technical to general language, repackaging or organizing.

Hence, dissemination and communication are essential attributes of a health information system (HIS). HIS is any system that captures, stores, manages or transmits information related to the health of individuals or the activities of organisations that work within the health sector. As Almunawar and Anshari [18] point out, HIS is the interaction between people, process and technology to support the provision of essential information aiming at improving the quality of health care services. Health managers can use multiple kinds of information in combination and select information to suit their purposes. In this context, the technology plays a role in the sense that an organisation needs to use effective tools to generate and disseminate the required information. It has been observed that there are significant improvements in healthcare due to information and communication technologies [14]. Even in India, it has been observed that the computerization has led towards implementation of a good information system for service delivery, planning, monitoring, and supervision [19].

To achieve what have been discussed above, in summary, it is necessary to consider [20-22] (a) free information flow which is enabled by provision of technological infrastructure and organizational culture that secures support, enthusiasm and cooperation of staff and management; (b) information to be sent in a way that is useful to others without compromising confidentiality and packaged for all partners with considerable integration across organizational barriers; (c) working cooperatively and sharing information within and outside; and (d) information channels that guide workers to the pertinent data.

These issues concerning the management of information, particularly in health management, can be understood by considering theories presented by Braa and Sahay [12]. They point out that active and engaged management of information at all levels are key objectives for all efforts to strengthen HIS by considering inter-connected practices concerning analysis, interpretation, presentation and dissemination of information. Information management practices should be strengthened along various dimensions of the entire information cycle [12]. The information cycle is a set of stages of collection, collation and storage, processing, analysis, presentation, interpretation, dissemination and use of information. National HMIS in developing countries, like in Malawi, follow these stages; and at different stages, different technologies are used.

Research on the application of ICT in HIS is pursued in various areas including health management [23]. HIS can be composed of automated and non-automated agents, including computers and humans, for managing health information. To generate such information, the relevance of systematically processing data, information and knowledge is recognized [13]. A good HIS should provide access to its users on reliable, authoritative, useable, understandable, and comparative data.

Although HIS has evolved through several different technologies [18], primarily, it is not about technology; technologies are to support processes. Since 2002, in Malawi HMIS, various technologies have been introduced with the aim of strengthening the national HMIS, which has been also happening in other developing countries such as Tanzania [5], India [19], Pakistan [24], and Nigeria [2]. Therefore, in order to understand the nature of these technologies in HMIS (particularly DHIS2, GIS and mobile), with reference to what has been discussed above, different lines (or trends) of development in HIS from Haux [13] have been applied, which are briefly described below:

1. *Towards computer-based information processing tools* – moving from paper-based processing and storage to computer-based processing and storage; although the adopting of technologies in HIS is increasing, paper-based is still favoured due to ease of use and legal reasons, among others.
2. *From local to global information system architectures* – initially small and functionally limited applications have been developed for specialized departments or functions, but now development is focusing on national, regional or global systems.
3. *From health care professionals to patients and consumers* - at the beginning, computerised HIS were primarily intended to support health care professionals and administrative staff in hospitals; now it is recognized that HIS have to directly support all people with health questions and problems.
4. *From using data only for patient care to research* - until the last decade, there was an almost exclusive use of HIS data for patient care and administrative purposes, with some use for quality management and controlling; now there is an ability to extend the possibility of using data, primarily for patient care, health care planning and, above all, for clinical research.
5. *From technical to strategic information management priorities* – in computerised information systems, the main focus has been shifted from technical problems to organizational problems, social issues and change management aspects and as a result, the strategic long-term information management is regarded as a serious and necessary task.
6. *Inclusion of new types of data* - with a higher degree of use of computer-supported information processing tools, not only the functionality of HIS is extended, the types of data to be considered continuously increased.
7. *Inclusion of new technologies* – there has been the increase of functionality in computer-supported HIS with new extensions, e.g. enabling technologies for health monitoring, GPS-related technologies, and mobile-based applications.

In this paper, the assumption is that developments in HIS are related to technology improvements which can be well understood by considering stages in the information cycle; determining at which organizational levels and activities, expected technology can be introduced and support respectively. The author takes the perspectives of information cycle and lines of development in HIS as frameworks which can help him to understand various roles that technology can play in the health information management practices.

3. Research Context

This case study was conducted in Malawi health sector in 2015 at both national and district levels. Malawi is a landlocked country in southeast Africa and it borders with Tanzania to the northeast, Zambia to the northwest, and Mozambique to the east, south and west. The government is the main provider of health care in Malawi through Ministry of Health (MoH). About 80% of population lives in rural areas and the population is affected by various diseases such as malaria, tuberculosis (TB), HIV/AIDS and nutrition-related diseases.

Malawi HMIS was introduced in 2002 as part of its health reforms in order to house all health data for decision making. It recognizes the evidence-based decision making which has been demonstrated by revising its national HIS policy with the main principle of "*information for action, action for improving efficiency, quality, and equitable coverage*". The primary objective is to provide up-to-date information for policy makers, planners, researchers and health programme managers that would allow guidance in the development, monitoring and evaluation of health. This HMIS has five levels: community, facility, district, zone and nation, which reflects the national organization of the health system. The health zone is a catchment area consisting of health districts within an administrative region. The health district is divided further into smaller catchment areas (i.e. health facility catchment areas) within which there are communities being composed of villages.

MoH is working with various stakeholders, both local and international, in setting-up and supporting various electronic health information systems (eHIS). These initiatives aim at strengthening the health systems by enhancing its capacity to use information for local action in supporting evidence-based decision making. They are also supported by various strategies and policies which emphasize the utilization of ICT in health sector with the aim of improving access and use of information. For instance, eHealth Strategy focuses on some areas of improving access to health services; improving healthcare; and strengthening monitoring and evaluation. It is expected that these issues can be implemented and managed by using sustainable national priority ehealth solutions that provide easy access to needed information in order to improve efficiency and effectiveness of health delivery services. National HIS strategy emphasizes the adoption and promotion of ICT use in the health sector. Even national ICT policy highlights the importance of integrating ICT in the health delivery system at all levels in order to improve management of information dissemination, among others.

In Malawi, there are various health programmes and in this study, malaria, nutrition, and family planning have been chosen as cases. Malaria is a major public health and economic problem in Malawi. Young children and pregnant women are the most at-risk population for malaria-related morbidity and mortality. Undernutrition particularly in women and children affects public health and development in Malawi; for example, the children suffer from chronic undernutrition (stunting) and micronutrient deficiencies. Malawi has made tremendous progress to increase contraceptive use and reduce fertility; and family planning services are provided in the health facilities. In addition, Central Monitoring and Evaluation Division (CMED) has also been chosen as a fourth case. To strengthen HMIS, MoH established CMED which is responsible for, among others, managing data and supporting technologies.

4. Research Methodology

This is the case study research in which qualitative interpretive methods were adopted with the aim of understanding social and natural settings of health managers, health coordinators, and HMIS officers (i.e. data managers) and particularly their health information management practices. The case study is an empirical inquiry that investigates a contemporary phenomenon within its real life context and multiple choices of evidence are used [25, 26]. The qualitative research helped the author to understand different ways of how health managers, health coordinators, and HMIS officers look at reality [27] and also to produce factual descriptions based on face-to-face knowledge of these participants in their every-day settings [28]. The aim of interpretive methods of research in information system (IS) is to produce an understanding

of the context of IS, and the process whereby IS influences and is influenced by the context [29].

Purposive sampling was used to select Blantyre health district, health programmes (malaria, nutrition, and family planning) and CMED, in which the sample was selected based on a particular characteristic [30]; i.e. the basis of researcher's own knowledge of population, its elements, and the nature of research aims [31]. Blantyre health district was chosen because it is one of the largest districts with the population of slightly over 1 million and it has both city (having a central/referral hospital) and rural settings. In addition, it was close to where the author lived; as a result it was relatively cheap to visit in terms of logistics. CMED was chosen because it is only the division within MoH which is responsible for managing data, implementing systems and providing technical support at all levels. All three chosen health programmes involve both children and women who are the most at-risk population as far as diseases are concerned in Malawi and they are fully implemented in DHIS2.

In this case study, data was collected through interviews as the main method and supplemented with document analysis and participant observation. At national level, semi-structured interviews were conducted with one statistician from CMED and one member of DHIS2 team while at district level, two HMIS officers (one from the central hospital and another from District Health Office) and three health programme coordinators (of chosen health programmes) were interviewed. The in-depth interview allowed the author to delve deeply into social and personal matters [32] and to ask follow-up questions and use non-verbal communication to his advantage [33, 34]. Data collected through these interviews include information flow from community level to national level, challenges being faced, sources of information, presentation of information, technical support provision, technologies and information systems being used, data capturing and processing, and information availability.

Document analysis involved written data sources such as meeting minutes, reports, policies, strategies, manuals, job descriptions, and forms. The main objectives were to (a) understand policies, procedures and strategies that govern the participants in their work in respect to management of health information; (b) understand various roles of the participants; and (c) understand various ways of presenting data and information. Participant observations were done at the national level in order to understand institutional contexts in which the participants work. The author was part of DHIS2 team; implementing GIS. This method allowed the author to interact with participants and become part of their community [33] aiming at gaining an understanding of physical, social, and cultural contexts in which the participants live. This permitted the author evaluating whether there are unintended and unanticipated consequences which need to be addressed and whether there are areas of experience not touched by practice to which attention should be given.

This case study employed thematic analysis approach which is a qualitative analytic method for identifying, analysing and reporting patterns (themes) within data [35]. Data was analysed by classifying according to the major themes that were defined in the conceptual model which was derived from information cycle [12] and the lines of development in HIS [13].

5. Health Information Management Practices

As earlier mentioned, the health system in Malawi consists of five administrative levels: nation, zone, district, health facility and community. In this section, health information management practices are presented per these administrative levels. Data is collected and aggregated at facility level, and then integrated and analysed at district level. As in developing countries, e.g. Tanzania HMIS [5], the district is identified as a focal geographic unit in Malawi for integrating multiple health programmes and their information systems. The integration and analysis are done through the computerized system (DHIS2). Officers at zone and national levels rely heavily on districts in terms of data availability. It has been noted that majority of health programmes follow this hierarchy systematically from one level to another (e.g. malaria, nutrition, and family planning). Figure 1 shows the administrative levels and what happens at each level including technologies being used. All health programmes are supported at all levels by CMED which is the custodian of data and technologies.

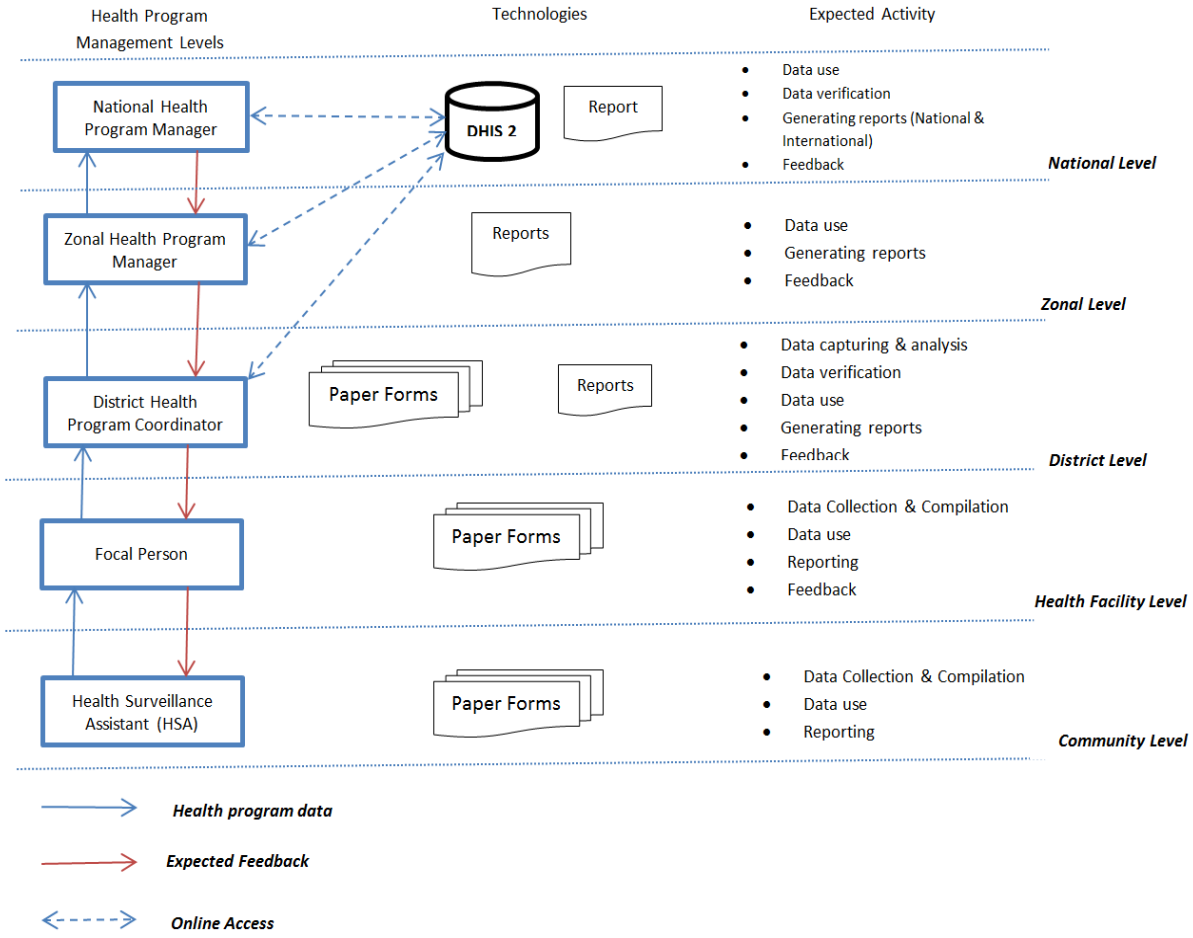
In Malawi, there are health programmes with well-established data flows from the community up to national level. Each health programme has a manager at the national level and a coordinator at district level. At the health facility level, the responsibility is given to the health facility in-charge (or manager) as a focal person. CMED, as a HIS unit, has its own structure and well-established data flow. It has officers at all levels except at the community level.

The Ministry of Health (MoH) established CMED in the Planning Department as one way of strengthening its HMIS. The overall objective of CMED is to continuously collect, analyse and use data to monitor and evaluate progress towards achieving goals and objectives of the health sector including the national development agenda. CMED involves in coordination, data management, advocate and facilitation of the information use in various activities such as policy formulation, planning and program implementation at all levels.

5.1. Facility and Community Levels

Every health facility (e.g. hospitals and health centres) has communities that it serves. At the community level, there are village and outreach clinics that are managed by community health workers (CHWs) who are referred as health surveillance assistants (HSAs) in Malawi. CHWs collect data using paper forms and send them to their respective health facilities. Similarly, at the facility level, health programmes have individual registers in which patient level data is recorded. Key players at health facilities are facility managers (mainly clinical officers or nurses) and CHWs. The health facility managers, as focal persons for health programmes, are responsible for compiling and reporting data to the district level with the support from CHWs. Some health facilities have statistical/data clerks, who aggregate, compile and report data on behalf of focal persons.

In each health programme, paper forms are used to report aggregate program level data to the respective district health programme coordinator. For monitoring and evaluation, there is another paper form (referred as HMIS form) which is used to record summary data from various health programmes at the health facility. This HMIS form is sent to respective HMIS officers at the district level. Apart from data recording, compilation and reporting, it is expected at the facility level to have some kind of data analysis and use for local actions. *Participant R* commented: "Some health facility managers call for stakeholder meetings in their respective catchment areas to discuss health issues in these meetings collected data is used as evidence …" Since there are no computerised information systems at health facilities, data is prepared manually.



For central hospitals, as referral hospitals, data is collected from wards (taken as the 'community level') using paper forms and then sent to departments (taken as the 'health facility level'). In each department, there is a focal person who is responsible for compiling and reporting aggregate departmental level data to the HMIS officer at the central hospital. At the departmental level, it is also expected to analyse and use data for local actions and this analysis can be done using DHIS2 and other software such as spreadsheets and word processors. *Participant S* said: "At a central hospital, technology for information processing and management is mainly at the departmental level where health personnel use computers to prepare reports."

5.2. District Level

Health programme coordinators receive aggregate program level data from health facilities and then verify the data before being captured into DHIS2 with the support from HMIS officer and data clerks. Similarly, HMIS officers get HMIS forms from the same health facilities; verify them and capture the data into DHIS2. At a central hospital, HMIS officer gets paper forms from department focal persons from which the summary data is compiled in HMIS forms and then captures the data into DHIS2. Specific health programme reports from central hospitals are sent directly to the respective health managers at the national level.

DHIS2 is used in data analysis, reporting, presentation, interpretation and even sharing. Since DHIS2 is the web-based system, it can be accessed at all levels by only registered users and who have the Internet access. At the district level, all health programme coordinators, HMIS officers and data clerks have access to DHIS2. However, it has been observed that HMIS officers have more access time than the health programme coordinators as *Participant T* commented "... in most of the time I ask our HMIS officer to extract data for me from DHIS2 and even capturing into DHS2 … I am always busy with coordination of my health programme activities and supervisory visits to health facilities …" All three health programme coordinators extract data from DHIS2 by themselves or through the HMIS officer when preparing their reports. It is expected that HMIS officers produce district HMIS bulletins for their respective districts at least once a year and also send them to the zonal and national levels. Apart from DHIS2, some software tools, such as Excel, Word and PowerPoint, are also used.

5.3. National Level

At the national level, MoH through CMED is promoting the use of DHIS2 as a central data repository and some health programmes extract their data from DHIS2 provided the programmes and their data are available in DHIS2. HIS policy states that all programmes and partners with interest to collect

health data on basic health services shall utilize HMIS, and it further emphasizes that MoH shall establish separate national repositories for patient level data and aggregate data (currently DHIS2). However, in some cases, health programmes need data which is not available in DHIS2 and hence, they request it from other sources. Key players at the national level are health programme managers, economists and statisticians from CMED, and DHIS2 team, who all have access to DHIS2. MoH has inadequate technical capacity at the national level and as a result, majority of members of DHIS2 team are from development partners.

Although DHIS2 has a capability of data analysis and presentation, sometimes data is analyzed using other technologies, such as Excel, due to operational limitations of DHIS2 or required data is not available in DHIS2. For instance, *Participant U* said: "DHIS2 cannot do everything … we have some indicators that require spatial analysis and presentation …"

5.4. Extensions to DHIS2

Since its introduction in 2012, there have been inclusions of technologies to support DHIS2. In this paper, two technologies have been identified: DHIS2 mobile and DHIS2 GIS which enhance HMIS at different stages. DHIS2 mobile was introduced in 2012 with the aim of improving timeliness of data collection and reporting. Paper-based system has resulted in many challenges including delays in reporting to districts. However, this mobile technology is only available in fifteen (15) out of twenty-nine (29) health districts, and also for few health programmes. DHIS2 mobile is a module within DHIS2 which allows data capture into DHIS2 through mobile phones at the data collection points (e.g. health facilities). Data is transferred from registers to paper forms and then captured into mobile application which transfers the data directly into DHIS2. Due to instability of the technology, paper forms are still used for backup purposes.

According to Moyo et al. [36], DHIS2 mobile has brought improvements to the timeliness of data reporting to districts although it has increased the workload of data clerks due to double reporting, i.e. same data is sent through mobile application and paper forms. They observed that for the successful use of DHIS2 mobile, some technical, organizational and behavioral issues need to be addressed; such as instability of, attitudes of users towards, and management support to the system, among others.

From 2015, MoH is introducing DHIS2 GIS in order to enhance data analysis, integration and presentation. It has been observed that many health programmes have recognized the power of GIS in data integration and visualization. From the interviews in this study, every participant, said "Yes" to GIS and below are some comments from the participants:

- *Participant V:* "Sometimes people come to my office asking for thematic maps particularly on distribution of facilities, catchment areas, road networks and disease surveillance which need GIS."
- *Participant X:* "GIS analysis should include utilization of health facility by its catchment area population."
- *Participant Y:* "We are interested using GIS to establish distances e.g. CHWs travel for resupply for their village clinics; mapping also activities covered by CHWs"
- *Participant Z:* "Our interest is to create maps of all established health services to conduct a gap analysis which will reveal the populations living in deprivation."

Demand for GIS in MoH was there as early as 2002 when a booklet of maps of the health facilities in each district across the country was produced and distributed in CD. Since there have been various GIS initiatives such as purchasing of GPS in 2005, training HMIS officers from 2009, collecting coordinates of health facilities across the country in 2013, and mapping of village and outreach clinics in 2015-2016. In 2016, DHIS GIS was deployed with the expectation of being operational in 2017. DHIS GIS includes all health facilities down to village and outreach clinics; and all health programmes are to use the system.

6. Existing Technology Gaps

6.1. Mix of paper-based and technology-based practices

The case study has shown information management practices at different administrative levels (see Table 1). This section discusses some interesting points when it comes to technological support of information management. Malawi has put up national HMIS that covers the majority of the health sector. As shown in Table 1, when it comes to health information management, HMIS employs a mix of paper-based and technology-based practices. Taking into account the infrastructure in Malawi, as in many developing countries, this is probably the most feasible approach. Therefore, discussions of existing technology gaps include both paper-based and technology-based practices and how to better support health information management practices through this mixed use of media.

With reference to the information cycle [12], the case study shows that paper-based practices are mainly at community and facility levels while technology-based practices are at upper levels. The use of technology is rather concentrated around analysis, process and presentation of information. These activities are essential because they may add value to the information. Health managers need the information that makes sense and which knowledge can be generated from [37]. However, data collection and dissemination practices are also important that need to be enhanced through technology. Data collection is a fundamental attribute of every information system; without data HMIS is useless. Since transfer and distribution can add value to information as well [16], good dissemination and communication practices are essential. It has been observed that data collection is only supported to a very limited extent by technology, while dissemination processes are almost not supported by the technology. This can be due to both organisation- and technology-related factors. Sometimes, emails are used at individual level to distribute data or information; the email system is not institutionalized.

Data collection is done at community and facility levels in rural areas where access to technology is still limited and this is a well-known and -documented obstacle in other studies. Some well-known challenges are lack of electricity and little telecommunication coverage. It is difficult to extend DHIS2 to the facility level, since it requires the Internet connectivity and computers which are limited in health facilities. This forces health facilities to use paper forms for managing data. However, it has been noted that in some cases data is not reported to the district on time due to challenges brought by the use of paper forms, e.g. an increase in workload to health personnel, as also mentioned in Moyo et al. [36]. To ease workloads, MoH deployed statistical/data clerks to manage data and introduced DHIS2 mobile to some health facilities.

Table 1: Use of Technologies in Malawi HMIS

Information Cycle Steps	Administrative Level	Technology Used
Collect	Community and facility levels	• Paper forms in all 29 health districts • DHIS2 Mobile in only 15 health districts
Collate and store	Facility level	• Paper forms in all health districts
	District level	• DHIS2 and paper forms for backup in all health districts
Process and analyse	Facility level	• Manually done at a small scale
	District and national levels	• Using DHIS2 at a large scale • Other software tools e.g. Excel, are used
Present	Facility level	• Manually done at a small scale – paper-based
	District and national levels	• DHIS2 – text, graphs, tables (maps, yet to be used from 2017) • Other software tools e.g. Excel, are used
Disseminate	All levels	• Paper-based e.g. forms, bulletins, reports • DHIS2 – only for registered users • Emails and Internet
Use	Facility level at a small scale	• Not available
	District and national level at a large scale	• Not available

Although mobiles are used for reporting at health facilities in order to improve the information flow by enhancing the technological infrastructure [20-22], paper forms are still in use which results in increasing workload for statistical clerks. According to Moyo et al. [36], some challenges of using the DHIS2 mobile are technical while others are organizational in nature: for example, mobiles not functioning, poor screens for data entry, lack of management commitment, and even individual attitudes towards the technology. It seems these issues are contributing to a failure of scaling the technology to other health facilities and health programmes; as a result the whole project will probably fail; making health facilities to go back to the paper forms.

Malawi has adopted a system use practice that builds on quite few data managers (e.g. HMIS officers and DHIS2 team) who are key personnel in health information management practices. Although health programme coordinators and CMED officials, for example, have access to DHIS2 they usually depend on HMIS officers and DHIS2 team respectively to extract data from and even enter data into DHIS2. This can be due to lack of time, lack of access and even lack of skills [38]. It has been noted that HMIS officers and DHIS2 team have received more training and less attention have been given to health managers and coordinators who are actual information users.

On the other hand, HMIS officers and DHIS2 team can become a bottleneck and there is a risk that this might inhibit data management and use. Sometimes, HMIS officers are not physically available or busy with other issues which result in not providing assistance required. In this context, there is a possibility that the health programme coordinators use other sources of data apart from DHIS2. Although developing countries have implemented their respective national HMIS aiming to use its information, what should be understood is that health managers rely on both formal and informal sources of obtaining information to make rational decisions; in some cases they use information other than that is provided by formal organizational information systems [3]. This issue can be dealt with by providing more people in the health sector with access to the system, but it would furthermore require that this access is followed up by training. This can also promote a free information flow and practices of information sharing among information users [20-22].

However, some studies, such as that of Williamson and Kaasbøll [14], have argued that health managers depend much of the information from informal sources because, for example, formal health information systems are not comprehensive or readily available. This has also been evidenced in this case study in the way how DHIS2 is utilised. The case study points to an issue regarding the adaptation of DHIS2 and its appropriateness. Current DHIS2 in Malawi is mainly for data analysis, process, integration, presentation, and small extent of dissemination. CMED is promoting DHIS2 as the formal health information system. However, the main challenge is that no many health managers are using it due to various reasons.

For instance, DHIS2 can only be accessed by registered users, which leads to a limited dissemination of information. It means every time when unregistered user want data or information from

DHIS2 he/she needs to consult a registered user particularly HMIS officers and DHIS2 team. As observed in this case study (see Table 1), MoH depends much on paper-based system to distribute reports to information users which makes it difficult to access the information by all users. The dissemination can be improved by publishing health information on the Internet as pointed out in the HIS policy: *"MoH shall maintain a web portal to make ready-to-use data available to all stakeholders"*. Another issue is the availability of data in DHIS2. No all health programmes have data in DHIS2.

Although DHIS2 is the formal information systems for MoH, some software tools such as Excel are used to large extent. This can be because Excel is commonly available and does not need an access to the Internet. It is a package which is easy to learn and use and having all necessary basic tools for simple data analysis. As mentioned in the previous section, the current DHIS2 has some limitations. In terms of analysis and integration, the DHIS2 is a good system. It requires enhancing its presentation and dissemination capabilities. For instance, MoH has recognized the power of maps in data presentation and visualization; as a result it is in a process of implementing GIS as part of DHIS2 (i.e. DHIS2 GIS). Presenting data in maps can provide more insight than a table of the same data [39] and be effective means for communicating messages clearly even to those who are not familiar with technology.

Considering the characteristics of a good HMIS – such as free flow of information, presentation of information, cooperative working, and information channels – the case study points to the following findings. The free information flow in inhibited both by technological and organizational factors. Poor Internet and paper forms make it hard for information to flow, while availability of data managers creates a bottleneck for the information flow. This is also true for the creation of channels to the pertinent data. Data managers hold the expertise on this, but if they are not available the channel is blocked. Presentation and packaging of information is taken well care of by the data managers, but could be further enhanced by the technology in terms of dissemination.

6.2. Role of DHIS2

With reference to theories of Haux [13] of different lines of development in HIS, the study has shown that DHIS2 plays various roles in Malawi HMIS. For instance, DHIS2 is the information processing tool; provides national information system architectures; supports strategic information management; among others. For more details see Table 2. As observed in this study, DHIS2 cannot do everything. Hence, it needs to link or integrate with other systems. Therefore, further discussions are on how to manage this integration.

Table 2: Lines of Development in Malawi HMIS

Line of Development	**Status in Malawi HMIS**
Towards computer-based information processing tools	DHIS2 was implemented with the aim of moving from paper-based to computer-based analysis, processing, storage, and presentation. Although DHIS2 is the main formal information system, some activities are still done on paper. Even the introductions of DHIS2 GIS and DHIS2 Mobile have the purpose of enhancing processing and reporting respectively.
From local to global information system architectures	The introduction of DHIS2 has helped to move towards the national health information system through which different health programmes are being integrated. CMED is working tirelessly on the implementation of integrated national health information system. Although some health programmes have individual applications, DHIS2 plays the role of data integration of such applications. However, the main issue is interoperability between DHIS2 and other systems which requires much research and development on how best it can be handled.
From health care professionals to patients and consumers	DHIS2 has not yet reach a level of supporting patients and consumers. Originally, DHIS2 has been designed for aggregate data for health administrators and managers.
From using data only for patient care to research	DHIS2 has enhanced the availability and accessibility of data. It has been observed that, to some extent, players in health management, such as health programme managers and coordinators, health M&E officers, and health statisticians, are now requesting data from DHIS2 for planning and controlling activities. Even some people outside MoH come and ask for data which CMED get from DHIS2.
From technical to strategic information management priorities	It has been observed that most of time of implementing DHIS2 is for data management. DHIS2 team spends most its time on organizing data, customizing data entry forms and checking quality, among others. For instance, the setting up of DHIS2 GIS took only two days while cleaning and uploading spatial data (coordinates of all health facilities and districts) into DHIS2 GIS took almost eight months.
Inclusion of new types of data	When DHIS2 was being introduced in 2012, few health programmes were included. Now, majority of health programmes have been and more will be included in DHIS2. This leads towards the inclusion of new types of data.
Inclusion of new technologies	Since the introduction of DHIS2 in 2012, there have been a lot of extensions such as DHIS2 GIS and DHIS2 Mobile. In addition, CMED and development partners are in progress of integrating DHIS2 with other health information systems in order to improve availability and accessibility of data.

The DHIS2 was introduced as a computer-based information processing tool with the aim of moving from paper-based processing and storage [13] at district and national levels; focusing on specific information processing capabilities. From the perspective of information cycle, DHIS2 supports, may be, just half of activities in HMIS. Paper-based processing and storage is still available at facility level and dissemination and communication practices are still supported by the paper-based systems. This translates to the existence of technology gaps that need to be filled by introducing new technologies to support DHIS2, such as GIS and mobile included in this paper.

Introduction of new technologies or information systems to enhance the DHIS2 needs good plans and management commitments. As illustrated in Figure 2, it is essential to define how a new technology will fit in the existing context (e.g. DHIS2 environment). For example, the implementation of GIS was completely within the architecture of DHIS2. It depends fully on existing components of DHIS2 in terms of database and user interfaces; GIS is inbuilt in DHIS2. While for the mobile system, one part is within DHIS2 and another part is supported by other systems e.g. national mobile networks. In some cases (not included in this study) DHIS2 can either feed or get data from other independent systems (illustrated as System A in Figure 2). In this context, interoperability issues are considered.

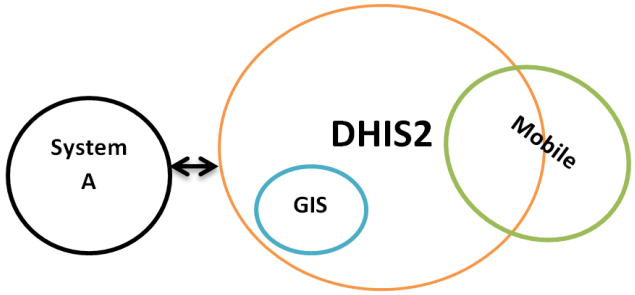


Figure 2: Ways of Linking DHIS2 to other systems/technologies

The author suggests the following when introducing a new technology in HMIS:

- identifying and prioritizing existing technology gaps that need much attention and this should be done with relation to activities in information cycle and administrative levels (e.g. DHIS2 mobile for data collection at the facility level and DHIS2 GIS for analysis and presentation at district and national levels);
- involving many health programmes that will benefit from the new technology and accommodating as much specific program needs as possible to increase a chance of acceptance;
- identifying interfaces for linking new technologies with the DHIS2 as a central system; and
- considering organizational factors including data issues (e.g. standards, sharing, etc.), people (roles, and required skills and knowledge), and procedures, policies and strategies, among others.

7. Conclusion

The case study confirms that technology plays various roles in strengthening HMIS. This has been also observed in some studies in developing countries such as Tanzania [5], India [19], Pakistan [24], and Nigeria [2]. For example, in India case, the computerization has enabled implementation of a good system in health care. In Nigeria case, the issue was strengthening HMIS through introduction of new technologies.

In recent years, in developing countries, there have been reforms in their health sectors by introducing HMIS and one component is to build formal information systems for providing necessary and sufficient information to facilitate evidence-based decision making and effective management of health at all levels of the organisation. In some developing countries (e.g. Malawi, Tanzania and Nigeria) DHIS2 has been implemented as their formal information systems for the health management. It seems that the technology needs to be supported by paper-based systems as observed in this case study and cases of Tanzania [5], India [19], Pakistan [24], and Nigeria [2]. Activities in some levels (particularly, lower levels) are supported by paper forms while other levels (e.g. upper levels) use computerised information systems. This means that HMIS cannot exclude paper-based systems.

Another observation is that when a new technology is being introduced, it is usually an extension to the formal information system (e.g. DHIS2). This case study has shown that the introduction of mobile application and GIS are the extensions to DHIS2. This has also happened in Nigeria where DHIS2 is the formal information system and mobile technology was introduced as an extension to DHIS2.

Therefore, to elaborate the mix of paper-based and technology-based practices, and extensions to the formal information systems using the case of Malawi, this case study focused on two: how different technologies support each other in enhancing the management of health information; and how a new technology can be integrated with existing formal information systems in order to strengthen health management information systems. From the analysis and discussions in this paper, it is concluded that in developing countries, we need both paper-based and technology-based systems for the health information management and these systems support each other. We should gradually move from the paper-based to technology-based environment by introducing new technologies as extensions to what we have as our formal information systems. Due to our infrastructural challenges, it is difficult to implement computerised information systems that can support all activities of information management and at all organizational levels.

However, due to well-known challenges of paper-based systems in developing, such as delay in reporting, it is obvious that technology-based systems will be affected regardless how good they are. For example, in Malawi [36] and Nigeria [2], mobile systems were introduced to minimize challenges of manual data collection, e.g. data timeliness and accuracy. The author suggests that further investigations should be taken to establish how best paper-based and technology-based systems can be interfaced in order to improve the efficiency.

As earlier pointed out, developments in HIS are not only on technologies, organizational, social, economic, political and

cultural issues should be considered. A culture of information management should be enhanced. It has been noted that in Malawi, DHIS2 is the main information system which leads towards moving from local to national information system architectures [13] in the sense that it is taken as a national platform for health programmes. It also addresses organizational problems rather than technical ones; e.g. integration of health programmes to improve data availability and accessibility.

References

[1] P. A. Chikumba, and S. L. Ramussen, "Management and use of health information in Malawi and Burkina Faso: the tole of technology" in IST-Africa Week Conference 2016, Durban South Africa, 2016.

[2] I. Asangansi, "Understanding HMIS implementation in a developing country Ministry of Health context - an institutional logics perspective" Online Journal of Public Health Informatics, **4**(3), 2012.

[3] R. I. Mutemwa, "HMIS and decision-making in Zambia: re-thinking information solutions for district health management in decentralized health systems" Oxford University Press in association with The London School of Hygiene and Tropical Medicine, 2005.

[4] A. Garrib, N. Stoops, A. McKenzie, L. Dlamini, T. Govender, J. Rohde, and K. Herbst, "An evaluation of the District Health Information System in rural South Africa" South African Medical Journal, **98**(7), 549-552, 2008.

[5] M. Smith, S. Madon, A. Anifalaje, M. Lazarro-Malecela, and E. Michael, "Integrated Health Information Systems in Tanzania: Experience and challenges" The Electronic Journal on Information Systems in Developing Countries, **33**(1), 1-21, 2008.

[6] C. P. Conn, P. Jenkins, and S. O. Touray, "Strengthening health management: experience of district teams in The Gambia" Health Policy and Planning, **11**(1), 64-71, 1996.

[7] T. D. Manda, and T. A. Sanner, "The mobile is part of a whole: implementing and evaluating mHealth from an Information Infrastructure perspective" International Journal of User-Driven Healthcare, **4**(1), 1-16, 2014.

[8] M. S. Chawani, and C. Ngoma, "Use of mobile technology to support provision of cimmunity-based maternal and neonatal care in developing countries" in International Conference on Health Informatics, Rome Italy, 2011.

[9] C. Moyo, M. H. Frøyen, J. I. Sæbø, and J. J. Kaasbøll, "Using performance league tables to provide feedback in health information systems in Malawi" in ifip9.4 Conference, Negombo Sri Lanka, 265-276, 2015.

[10] B. Msiska, and P. Chikumba, "Challenges and opportunities in using GIS for monitoring and management of HIV/AIDS: A case study from Malawi" Malawi Journal of Applied Science and Innovation, 1(2), 27–36, 2015.

[11] P. A. Chikumba, "Application of Geographic Information System (GIS) in drug logistics management information system (LMIS) at district level in Malawi: opportunities and challenges" AFRICOMM 2009. LNICST, 105–115: Springer, Heidelberg, 2010.

[12] J. Braa, and S. Sahay, Integrated Health Information architecture - power to the users: design, development and use, Sanjay Sethi for MATRIX Publishers, 2012.

[13] R. Haux, "Health information systems — past, present, future" International Journal of Medical Informatics, **75**, 268—281, 2006.

[14] L. Williamson, and J. Kaasbøll, "Health information and managerial work: exploring the link" in International Conference on Social Implications of Computers in Developing Countries, Dubai UAE, 2009.

[15] C. W. Choo, "Environmental scanning: acquisition and use of information by Chief Executive Officers in the Canadian Telecommunications Industry" PhD Thesis, University of Toronto, 1993.

[16] P. A. Schulte, A. Okun, C. M. Stephenson, M. Colligan, H. Ahlers, C. Gjessing, G. Loos, R. W. Niemeier, and M. H. Sweeney, "Information dissemination and use: critical components in occupational safety and health" American Journal of Industrial Medicine, **44**, 515–531, 2003.

[17] M. S. Feldman, and J. G. March, "Information in organization as signal and symbol" Administrative Science Quarterly, **26**(2), 171-186, 1981.

[18] M. N. Almunawar, and M. Anshari, "Health Information Systems (HIS): concept and technology" in International Conference Informatics Development, Cornell, 2011.

[19] A. Krishnan, B. Nongkynrih, K. Yadav, S. Singh, and V. Gupta, "Evaluation of computerized health management information system for primary health care in rural India" BMC Health Services Research, **10**(310), 2010.

[20] K. Bystrom, and K. Jarvelin, "Task complexity affects information seeking and use" Information Processing & Management, **31**(2), 191-213, 1995.

[21] A. Curry, and C. Moore, "Assessing information culture — an exploratory model" International Journal of Information Management, **23**, 91–110, 2003.

[22] D. O. Case, "lnformation behavior," Annual Review of Information Science and Technology, **40**(1), 293–327, 2006.

[23] M. S. Chawani, "Development of electronic medical record systems for maternal health services in rural settings: An action research study from Malawi" PhD Thesis, University of Oslo, 2014.

[24] R. Kumar, B. T. Shaikh, A. K. Chandio, and J. Ahmed, "Role of Health Management Information System in disease reporting at a rural district of Sindh" Pakistan Journal of Public Health, **2**(2), 10-12, 2012.

[25] C. Yen, M. Woolley, and K. Hsieh, "Action case research: a method for the accumulation of design theory/practice knowledge in practice" Working papers in Art & Design, **01**/2002, 2002.

[26] L. O. Johansson, B. Cronquist, and H. Kjellin, "Visualization as a tool in action case research" in 6th European Conference on Research Methodology for Business and Management Studies, UK, 2007.

[27] B. Hancock, E. Ockleford, and K. Windridge, An Introduction to Qualitative Research: The NIHR RDS EM/YH, 2007.

[28] G. J. Ebrahim, and K. R. Sullivan, Mother and Child Health - Research Methods, Book Aid, 1995.

[29] K. E. Berntsen, J. Sampson, and T. Østerlie, "Interpretive research methods in computer science" DIF8916 Method Essay, Norwegian University of Science and Technology, 2004.

[30] L. R. Frey, C. H. Botan, and G. L. Kreps, Investigating communication: an introduction to research methods, Allyn and Bacon, 2000.

[31] E. Babbie, Survey research methods, Wadsworth Publishing Company, 1990.

[32] B. DiCicco-Bloom, and B. F. Crabtree, "Making sense of qualitative research: The qualitative research interview" Medical Education, **40**, 314-321, 2006.

[33] D. L. Driscoll, "Introduction to primary research: observations, surveys, and interviews" Writing Spaces: Readings on Writing, **2**, 2011.

[34] N. Mathers, N. Fox, and A. Hunn, "Trent Focus for research and development in primary health care: using interviews in a research project" Trent Focus, 1998.

[35] V. Braun, and V. Clarke, "Using thematic analysis in Psychology" Qualitative Research in Psychology, **3**, 77-101, 2002.

[36] C. Moyo, T. Nkhonjera, and J. Kaasbøll, "Using mobile technology to improve timeliness of data reporting: A case study from Malawi" in IST-Africa Conference Week 2015, Lilongwe Malawi, 2015.

[37] C. W. Choo, "The knowing organization: how organizations use information to construct meaning, create knowledge and make decisions" International Journal of Information Management, **16**(5), 329-340, 1996.

[38] R. Bertulis, "Barriers to accessing evidence-based information" Nursing Standard, **22**(36), 35-39, 2008.

[39] R. P. Fisher, and B. A. Myers, "Free and simple GIS as appropriate for health mapping in a low resource setting: a case study in eastern Indonesia" International Journal of Health Geographics 2011, **10**(15), 2011.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 167-171 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

Development of a new lines of sight analyzer while playing sport

Shinya Mochiduki[1], Miyuki Suganuma[1], Gaito Shoji[2], Mitsuho Yamada[1]

[1]*Graduate of school of Information and Telecommunication Engineering, Tokai University, 108-8619, Japan*

[2]*School of Information and Telecommunication Engineering, Tokai University, 108-8619, Japan*

ARTICLE INFO

Article history:
Received: 20 December, 2016
Accepted: 20 January, 2017
Online: 28 January, 2017

Keywords :
Eye movement
head movement
athlete
baseball
lines of sight

ABSTRACT

The Olympics will be held in Tokyo in 2020, and the training of the athlete using technology has been gaining attention. In an effort to refine the competitive ability of top athletes by evaluating their performance objectively, we have focused on eye movement and head movement. Since the field of view moves according to the athlete's head movement, which is a problem for the conventional method of measuring eye movement, we proposed a new method of analysis of lines of sight which can record head movement during a competition and make it easier to analyze by superimposing the lines of sight on an externally recorded fixed image. With the goal of measuring the lines of sight of an athlete during an actual competition, we made a video during a competition and had an athlete observe the video in a laboratory. First we compared the video in which only the eye movement was measured and the field-of-view image moved according to the head movement with another video in which the head movement and eye movement were measured and the image did not move in spite of the occurrence of head movement. The results of the experiment, which involved baseball as the competitive sport, showed the effectiveness of our proposed system. Furthermore, we showed the difference between the lines of sight of an experienced and an inexperienced catcher.

1. Introduction

Various efforts to improve the movement ability and the play of top athletes by objectively estimating their competitive ability will be undertaken before the Tokyo Olympic Games in 2020.

Sports research for the skilled players and the unskilled players have been researched a lot [1,2]. When working to improve an athlete's performance, the movement of the athlete's eyes is a particular area of focus. Skilled athletes obtain important information through peculiar visual search patterns, not disorderly eye movement [3]. The measurements in the actual athletic sites are also conducted [4]. For example, it is often said that a player should not take his or her eyes off the ball in ball games, but the speed of the ball is too fast for anyone to follow with smooth pursuit eye movement, especially in tennis. Tennis players predict the landing point of a tennis ball and move his/her eyes to that point [5]. On the other hand, in an experiment that focused on the accuracy of smooth pursuit eye movement, the position error in pursuing a target moving at 70 deg/sec was found to be 6~8 degrees for a general university student, although this error is less than 2 degrees for university students who are table tennis players [6].

The convenonal measurement of the eye movement of athletes has been performed using a system in which the athlete wears eye-movement-measuring equipment while playing [5] or a system in which the athlete's head is held stationary on a head-chin rest while he or she views a video showing action from his or her sport [5]. Movement of the line of sight is not only determined by eye movement; head movement is also important. When eye-movement-measuring equipment is worn by an athlete and the movement of the eyes is measured, the camera attached to the athlete's forehead moves with the player's intense movement, and analysis of the video after the experiment becomes difficult. Though head movement is indispensable for acquiring information from the wide viewing angle of human vision, head movement itself has not often been measured. Equipment with which eye movement and head movement can be measured at the same time

*Corresponding Author: Shinya Mochiduki, Graduate of school of Information and Telecommunication Engineering Tokai University, +810334411171, s.mochiduki@star.tokai-u.jp

was developed [7], and a new method was proposed for the analysis of the coordination of head and eye movement in sport. One objective of this research is to actually measure the lines of sight of an athlete using equipment worn by the athlete, but in the present research, an athlete in a laboratory observes video of a game which was recorded using a 4K camera in order to assess the method.

2. Experimental Method

An experiment was performed involving baseball as the sport to be studied. We fixed a wearable camera, a Panasonic HX-A500 [8] on a tripod behind the catcher and behind the runner and recorded a video. A 55-inch screen (Regza 55X3, Toshiba, Tokyo, Japan) on which a 4K picture can be displayed was used, and an experimental picture was output. The distance between the display and the subject was set to 51 cm, which was 0.75 times the screen height, to reproduce the view of the catcher, and the viewing angle was 102 degrees. For the eye-movement measurements, the system used a corneal reflex eye-tracking recorder (EMR-8B, Nac Image Technology, Tokyo, Japan) [9]. To measure head movement, a Patriot magnetic position detecting system (Polhemus, Colchester, VT, USA) was used, which can measure motion at 6 degrees of freedom (DOF) [10]. The sensors of the magnetic position-detecting system and the eye-tracking recorder were attached to the subject's cap. The source coil was fixed on a wooden pillar installed at the subject's right-hand side. We decided not to use ironwork around the subject and during the measurement to avoid any obstruction of magnetism. The data from the eye-movement and head-movement measuring systems were unified and changed into gaze movement on a PC [7]. The equipment used in the experiment is shown in Table 1.

Table 1. Equipment used in the experiment

Video	4K Wearable camera (Panasonic HX-A500)[8]
Eye movement measurement	EMR-8B, Nac Image Technology, Tokyo, Japan [9]
Head movement measurement	Patriot magnetic position detecting system (Polhemus, Colchester, VT, USA)[10]

The system configuration is shown in Figure 1. The 4K video was played on a personal computer and displayed on a 4K monitor (REGZA 55X3 55, Toshiba, Tokyo). For the composition of the 4K video that was viewed by the subject and the gaze movement, the system uses ATEM Production Studio 4K Switchers [11] (Blackmagic Design, Fremont, CA, USA). The 4K video is connected to the ATEM Production Switchers through an HDMI output, and a superimposed image representing the movement of the subject's gaze is displayed on the experimenter's display and recorded in the solid-state drive of a non-compression video recorder (HyperDeck Shuttle; Blackmagic Design, Fremont, CA, USA). In this way, it is possible to superimpose a player's line of sight, which is a combination of head movement and eye movement, on the wide-view picture, which corresponds to the view of the player and is recorded from behind the player by a fixed camera on a tripod. The subjects were three baseball players at our university; all were males ranging from 21~22 years old; two usually played as catchers and one usually played as an outfielder.

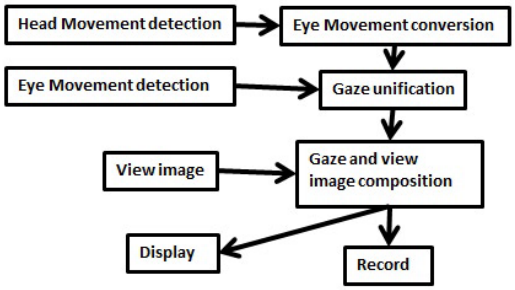


Figure 1. System configuration

3. Experimental Procedure

We took video of an intra-squad game of a club that plays "a Japanese version of baseball played with a hard rubber ball", Highrings, which was made up of students from our university. The video was recorded from behind the catcher, from beside the first baseman, and from beside the third baseman using a 4K camera fixed on a tripod so as not to interfere with the game. The video was edited to a length of around five minutes and separated into three scenes, a scene shot from behind the catcher, a scene shot from the first baseman's side, and a scene shot from the third baseman's side, and a still image of a fixation point on the center of the screen was inserted between scenes to confirm the eye movement accuracy. Before the experiment, the subjects were instructed to watch the video as if actually playing in the game, for example, as a catcher for the scene shot from behind the catcher, and as a runner for the scenes shot from beside the first baseman and third baseman. We had the subjects watch the video wearing eye-movement and head-movement sensors without fixing the head.

4. Experimental Result

Here we propose a new method of analysis of eye and head coordination during a sports competition involving the simultaneous measurement of eye movement and head movement. An image of the player's view moves according to the head movement by means of a conventional system in which eye movement is superimposed on a view image at the forehead part of the player. Figure 2 shows an example of the view image moving according to the head movement. On the other hand, the line of sight (head movement + eye movement) is superimposed on the fixed image taken from behind the player using the newly developed system. Since the lines of sight of the player are superimposed on the fixed camera image taken from behind the player, the lines of sight are reproduced on the fixed image instead of the view image, which moves according to the head movement as shown in Figure 3.

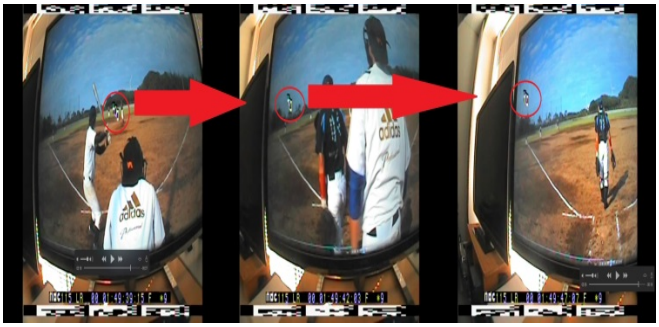


Figure 2. Example of eye movement by conventional experimental method

Figure 3. Example of line of sight (head movement + eye movement) determined by our newly proposed method

Here, we discuss the results obtained when the subjects viewed the video recorded from behind the catcher. This video assumed that a catcher watches a first-base runner cautiously because the catcher is responsible for catching any pitch thrown by the pitcher that bounces near home plate to keep the runner from advancing. The runner's focus is on the next base because there is a possibility that the ball could get past the catcher when the pitch bounces near home plate. Therefore, when a pitcher throws a pitch that bounces near home plate, the catcher catches it or blocks it with his or her body, then confirms whether or not the runner is running to the next base.

Figure 4 indicates the result when the lines of sight of subjects A, B and C are all included. First, the lines of sight of subjects A and C, who have played the position of catcher, were on home plate where a ball thrown by the pitcher bounced, and they then moved their gaze to the first-base runner. The line of sight of subject B, who has not played the catcher position, always pursued the ball. The part enclosed in a blue circle in Figure 4 shows that the lines of sight of subjects A and C both went toward a place where the line of sight of subject B did not go.

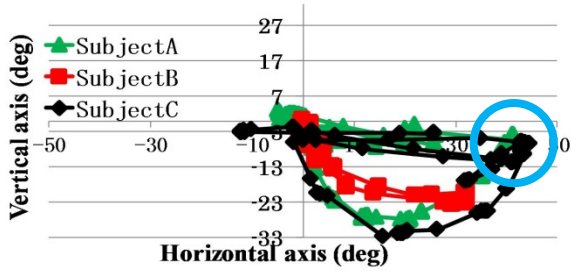


Figure 4. Three subjects' lines of sight (eye movement + head movement) as a catcher

Two runners on first base and second base, two outs, and the fourth player is the batter in the situation in the second experiment. The first-base runner and second-base runner start as soon as the batter hits the ball. As for this video, the batter hits an extra-base hit past the outfielders to the left, and the first-base runner and second-base runner reach home plate. The batter who hit the ball is advancing to second base. Figures 5 and 6 show the Eye movement of subjects A and B, respectively. Figure 7 indicates the amount of head movement of both subjects, and Figure 8 shows the result when the lines of sight of subjects A and B are combined. Regarding the lines of sight of subject A, who had experience as a catcher, movement to follow the ball accounted for most of his movement, but lines of sight to watch the movement of the batter who ran were also seen. The result for subject B, who had no experience as a catcher, showed eye movement that mostly followed the ball. In subject A, there was less head movement than in subject B when we compared the amounts of head movement from Figure 7. From Figure 8, in which the movement of the lines of sight of subjects A and B is shown simultaneously, only subject A is looking around second base (a light blue circle in the figure), which the batter eventually reached.

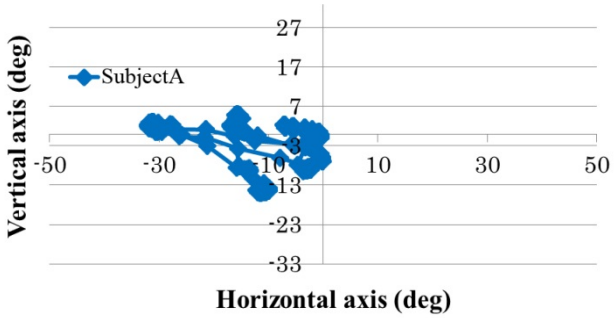


Figure 5. Subject A's Eye movement as a catcher

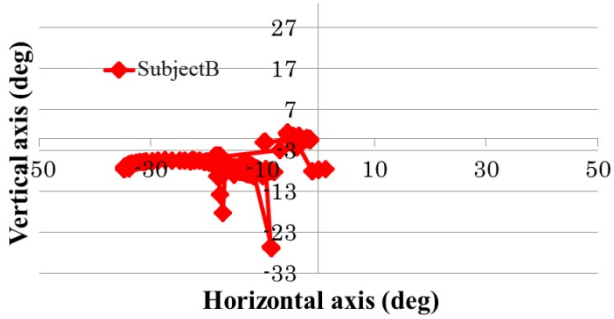


Figure 6. Subject B's eye movement as a catcher

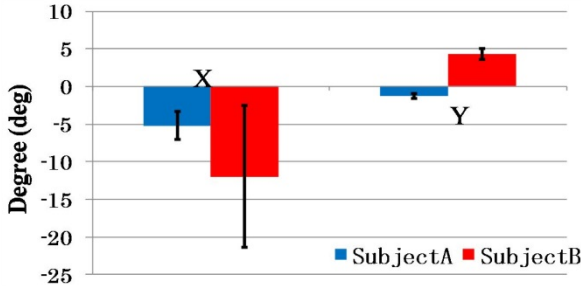


Figure 7. Amount of movement of the head

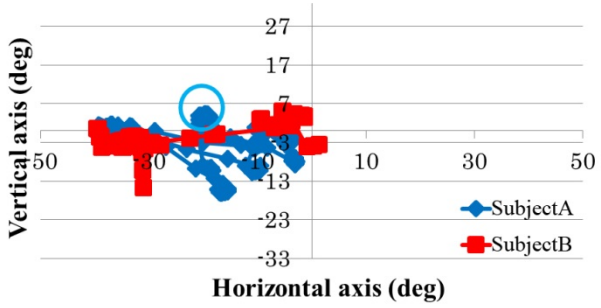


Figure 8. Two subjects' lines of sight (eye movement + head movement) as a catcher

5. Discussion

From the first finding, a difference was seen in the movement of lines of sight in a subject who had no experience as a catcher compared with a subject who was an experienced catcher. The subject who was an experienced catcher showed a tendency to mind the runner while following the ball while the subject without experience as a catcher always followed the ball. It was suggested that the viewpoint of the game was different for a player in a defensive position based on the movement of the lines of sight.

The second finding suggested that a subject who was an experienced catcher monitored a wide field of view using only eye movement and with little head movement, because the moving area of the lines of sight was not very much different between the subject who was an experienced catcher and the subject who was not, even though the subject who was an experienced catcher showed very little head movement. Although it is said that high-level athletes move their heads well and monitor a wide field of vision, it is also said that athletes running marathons who become tired show slight head movement. In other words, runners who are running stably run without their heads moving, even slightly. A tendency to monitor a wide field of vision closely by eye movement was seen in the subjects who were experienced catchers and who did not move their heads significantly in this experiment. While the head movement was small in these subjects, they showed more eye movement by saccade than an inexperienced person.

Here we compare the result in the past research and the present experiment. Gorman et al. and Land et al. have reported that experienced players predictively look ahead their lines of sight to the next destination [12, 13]. In our experimental results, the experienced players predict the runners' movement and they moved their lines of sight to the first base. The result that the no experienced player focuses his attention exclusively on the ball is also supports the result of Williams et al. [14]. From these results, we considered that there are no differences between the past researches result and the present experiment by proposed method.

Urgesi et al. has reported that actual player predict the next movement better than the watching learner in the volleyball learning experiment between two groups: learn by playing and learn by watching others' play [15]. This result also supports our result because we could see the differences between the experienced catcher and the unexperienced catcher in our present experiment results.

6. Summary

Given that the Olympics will be held in Tokyo in 2020, the training of athletes using technology has begun to attract a great deal of attention. In Japan, an effort to improve athletes' performance is being led by a national scientific center. In order to improve the athletic capabilities of elite athletes by evaluating their skills objectively, we have studied their eye movement and head movement. Since the field of view moves according to the head movement of the athlete, which was a problem for the conventional method of measuring eye movement, we proposed a new method of analysis of lines of sight which could record the head movement during competition and make it easier to analyze by superimposing the lines of sight on an externally recorded fixed image. The ultimate objective is to attach the equipment to the athlete and measure the lines of sight during competition, but in this study we recorded a video during competition before the experiment was conducted to confirm the effectiveness of the method, and we had athletes view the video in a laboratory.

We recorded our video during a baseball game. First, we compared a video in which only eye movement was measured and the field-of-view image moved according to the head movement of the player with another video in which the head movement and eye movement were both measured and the image did not move in spite of the player's head movement. The experiment showed differences between the eye movement of an experienced person and that of an inexperienced person, particularly between those who had experience playing the position of catcher and those who did not. We would like to measure the line of sight of an athlete actually playing a game by attaching the head and eye movement sensors to an athlete in the future.

Acknowledgment

We would like to express our sincere gratitude to the "High rings" student baseball team at our university for kindly allowing us to take movies of them playing.

References

[1] S.Mori, T.Shimada, "Expert anticipation from deceptive action,Attention" Perception, & Psychophysics,May 2013, 75(4), 751–770, 2013.

[2] S.Mori, "Spatiotemporal occlusion of biological motion reveals anticipatory processes of expert players" JSPS Grant Kickoff Workshop "Experimental study on perception-action systems in real complex environments", 2009.

[3] T. Kato, "Skill of the expert who looked from visual science, the latest sports psychology trace and prospects" Sport psychology,163-174, 2004.

[4] A.M.Williams, K.Davids, J.G.Williams. Visual perception and action in sport, E & FN Spon. 1999.

[5] M. Yamada, T. Fukuda, "A New Sight-line Displacements Analyser and Its Application to TV Program Production SMPTE Journal" J SMPTE., **99**(1), 16-26, 1990.

[6] M. Saito, Y. Ogata, M. Yamada, "Development of eye-movement analysis equipment for clarifying the visual characteristics of moving images" The First International Workshop on Image Media Quality and its Application,135-140,2007.

[7] H. Takahira, K. Kikuchi, M. Yamada, "A System for Measuring Gaze Movement and Hand Movement Simultaneously for Hand-Held Devices" IEICE Trans. Commun, **98**(1), 51-61, 2015.

[8] http://panasonic.jp/wearable/a500/in Japanese

[9] http://www.nacinc.com/datasheets/archive/EMR9-Data-Sheet-June-09.pdf ,latest version of EMR8-B

[10] http://polhemus.com/motion-tracking/all-trackers/patriot/

[11] https://www.blackmagicdesign.com/jp/products/atem/techspecs/W-APS-04

[12] A. D. Gorman, B. Abernethy, D. Farrow, ",Investigating the anticipatory nature of pattern perception in sport.", Mem Cognit, 39(5), 894-901, 2011.

[13] M. F. Land, P. McLeod, "From eye movements to actions: how batsmen hit the ball" Nature Neuroscience 3, 1340 – 1345, 2000.

[14] A. M. Williams, K. Davids, L. Burwitz, J. G. Williams, "Visual search strategies in experienced and inexperienced soccer players" Research Quarterly for Exercise and Sport, 65(2), 127−135, 1994.

[15] C. Urgesi, M. M. Savonitto, F. Fabbro, S. M. Aglioti, "Long- and short-term plastic modeling of action prediction abilities in volleyball" Psychol Res., 76(4), 542-560, 2012.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 172-176 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

Implementation a Secure Electronic Medical Records Exchange System Based on S/MIME.

Chien Hua Wu[*,1], Ruey Kei Chiu[2]

[1]Graduate Institute of Business Administration, Fu Jen Catholic University, 242, Taiwan

[2]Department of Information Management, Fu Jen Catholic University, 242, Taiwan

ARTICLE INFO

Article history:
Received: 21 December, 2016
Accepted: 19 January, 2017
Online: 28 January, 2017

Keywords :
Electronic Medical Record
Message Security
RESTFul
S/MIME

ABSTRACT

The exchange of electronic medical records can reduce the preservation and the use of papers of medical records for management issues. The sharing of electronic medical records has been effective in Taiwan. Now days, enterprises are sharing their electronic medical records through the Exchange Center of EMR under the Virtual Private Network but slightly less secure. This study aims to propose a security mechanism for the sharing of electronic medical records. The combination of security mechanism of S/MIME message level and RESTFul Service were adopted to build a secure mechanism for the sharing of electronic medical records. Two scenarios were simulated and implemented to verify the feasibility of this mechanism. From the results of the simulation presented, it has been conclude that the use of RESTful and S/MIME can enhance the security exchange of the electronic medical records.

1. Introduction

The Ministry of Health and Welfare [1] planned to conduct an electronic medical record exchange center(EEC) from 2009 to share electronic medical records for hospitals in Taiwan. There are already five categories of electronic medical records that can be exchanged through the EEC between hospitals under Virtual Private Network [2]. The development of electronic medical records in the hospital now has a considerable effect. After having legislation [3] of electronic medical records, hospitals can use electronic medical records, do not need to create and save papers of medical records. Hwang et al. [4] indicated that information quality of EMR exchange was the key factor which influencing users. EMR allows physicians rapid access to medical treatment in different hospitals, save the use of medical resources. This paper references to the prevailing exchange EHR architecture. Aim to explore a core technology and the security approach among health care information systems. Two scenarios were simulated and implemented to evaluate and verify the feasibility of such a mechanism. The purpose of this study was to propose a security mechanism, and to achieve the information exchange security:(1) Authentication of EMR;(2) Confidentiality storage of EMR;(3) Integrity of EMR;(4) Non-repudiation of EMR exchange with each stakeholder..

*Corresponding Author: Chien Hua Wu
Graduate Institute of Business Administration, Fu Jen Catholic University, 242, Taiwan
Email: kevin930202@gmail.com

2. Background

2.1. Medical Information share Standards

The Health Level Seven International [5] was founded in 1987. American National Standards Institute (ANSI) [6] and the International Organization for Standardization (ISO) [7] accredited international standards for the EMR exchange and sharing that support clinical practice. Due to the wide range of medical services covered by the industry, such as medical care, medicines, medical equipment, medical information, health care, etc. The main objective of HL7 is to develop a commonality and interoperability system. The Level 7 layer also supports the secure authentication and identification of data exchange. HL7 standards can be quickly applied in hospital and can be easily integrated with several of the other systems. Clinical Document Architecture, Release Two (CDA R2), became an ANSI-approved [8] HL7 Standard in May 2005. CDA documents are encoded in Extensible Markup Language (XML) that specifies the structure and semantics of a clinical document. A CDA document can include text, images, sounds, and other multimedia content. Digital Imaging and Communications in Medicine (DICOM) is currently the standard format widely used in hospitals for medical imaging message [9]. This standard was announced by the committee (ACR-NEMA) which established by American College of Radiology (ACR) and the National Electrical Manufacturers Association (NEMA), published in 1993, and officially named DICOM 3.0 to help the image storage, content distribution and

viewing of medical images, such as Computerized Tomography (CT), Magnetic Resonance Imaging (MRI) and ultrasound.

2.2. Representational State Transfer

Representational State Transfer (REST) [10] is a design concept. This concept comes from the Roy published PhD thesis. He proposed REST software architecture style as an abstract model of network applications, but it is not a standard. REST systems interface with external systems as web resources, each resource will have a URI (Uniform Resource Identifier). It relies on a stateless, client-server, cacheable communications protocol. Because Web applications in HTML only defined to the GET POST, cause little use to other methods such as PUT and DELETE on Web-based applications as well as HEAD, STATUS and other methods. REST is simple interface often used to describe any use of XML (or YAML, JSON, plain text), without having to rely on other mechanisms (such as SOAP). Compared to other commonly used Web Service standards, such as SOAP and XML-RPC, it is more simple and easy to use. It has the following characteristics:(1) All of the API is Resource form;(2) This service can accept and return MIME-TYPE, also can return XML / JPG / TXT and other formats; (3)Supporting the operation of the various HTTP methods (such as GET, POST, PUT, DELETE).

2.3. Multipurpose Internet Mail Extensions

Multipurpose Internet Mail Extensions (MIME) is a network messaging applied to flexible message format standard [11]. It extends the standard of E-mail, MIME standard can support transmission such as images, audio, video, and other binary file. MIME message format consists of Header and Body. Header is a set of Header Fields, Body contains a single Party or more Parties. MIME Header provides the information structure and encoding. MIME Body is the actual message content, supports a variety of data formats, sometimes also referred to as "Payload". Secure MIME (S/MIME) is a standard message format [12]. S/MIME provide MIME message format standard encryption and digital signatures to send and receive secure messages in MIME format on the web. It provides digital signature and encryption; these two security mechanisms are based on RSA public key infrastructure (PKI).

2.4. Comparison to Web Services

Web Service is based on the Simple Object Access Protocol (SOAP) agreement, WS-Security [13] is the core of Web services security standards. Gabriel et al. [14], respectively, used the GET and POST methods to compare different security mechanisms among them. In the conditions of plain text, encryption or signature, the results show that RESTful services were processed more efficiently than Web Services. Cesare et al. [15] describes the differences between REST services and WS- * services. He used a variety of architectural decision models to determine which type of service were more appropriate. Their result shows that REST was more suitable for basic and ad-hoc integration scenarios. When business requirements demand a higher quality of service WS- * was more flexible.

3. Materials and Methods

3.1. System Architecture

A Secure Electronic Medical Record Services system (SEMRS) conducted in this paper references an existing electronic medical records exchange to enhance the message security. Figure 1 presents the whole architecture of SEMRS. Step 1 through Step3 presents the EMR upload and EMR index registry process of Hospital A, step 4 through Step 5presents the EMR index store query process of Hospital B, step 6 through step 7 presents EMR retrieve process of Hospital B.

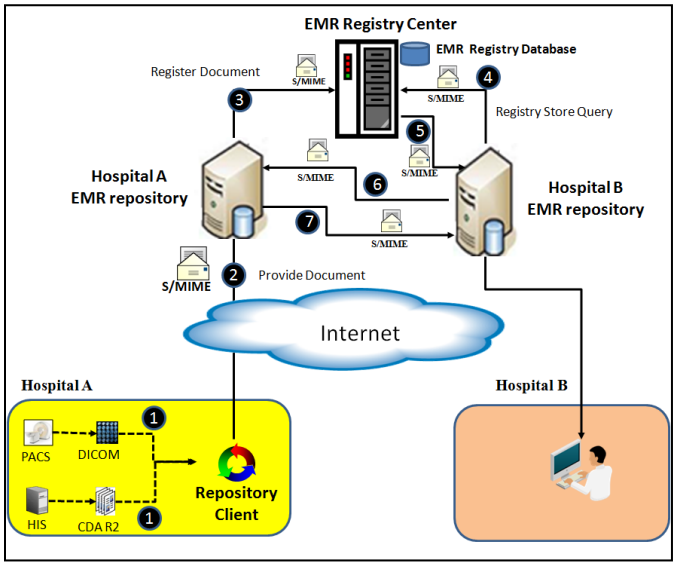


Figure 1. System Architecture of SEMRS.

- Step 1 through Step 3 are described as follows: (1)Hospital A puts CDA R2 documents and DICOM documents which will be upload to repository to the Repository Client of Hospital.(2)Repository Client of Hospital A packaged the CDA R2 documents and DICOM documents into a S/MIME envelope then upload it to EMR repository of Hospital A through RESTFul service.(3)EMR repository of Hospital A used RESTFul service to register received EMR index to EMR Registry Center via RESTFul services using S/MIME.
- Step 4 through Step 5 are described as follows:(4)EMR Repository of Hospital B sent EMR registry store query request to EMR Registry Center via RESTFul services using S/MIME.(5)EMR Registry Center response it to the EMR repository of Hospital B through RESTFul service using S/MIME.
- Step 6 through Step 7 are described as follows:(6)EMR Repository of Hospital B sent a retrieve request to EMR Repository of Hospital A via RESTFul services using S/MIME.(7)EMR Repository of Hospital A packaged the documents into a S/MIME envelope then response it to the EMR repository of Hospital B through RESTFul service.

3.2. Message Structure

Figure 2 presents the message structure of SEMRS. Table 1 lists the usage of MIME header tags which were used in SEMRS.MIME body presents in figure 2 which encrypted CDA document is in body part 1, encrypted message digest is in body part 2, encrypted digital signature is in body part 3, encrypted one-time password is in body part 4,encrypted DICOM documents are stored from body part 5 to the end part.

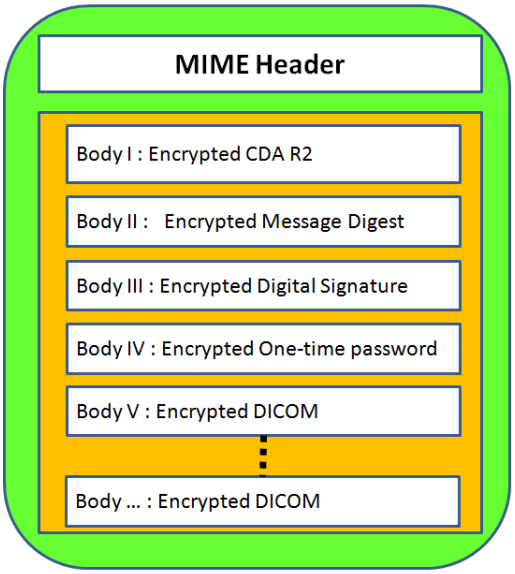


Figure 2 Message Structure of SEMRS

3.3. Algorith of SEMRS

Figure 3 represents the processes of Hospital A from step 1 through step 9, corresponding to the steps of algorithm in Table 2 respectively. The processes involved in this principle are:(1) Dynamically generated one-time password;(2)Using one-time password to encrypt all MIME Header tag values;(3)Using one-time password to encrypt CDA R2 document then puts it into MIME body part 1;(4)Using CDA R2 document to generate message digest;(5)Using one-time password to encrypt message digest then put it into MIME body part 2;(6)Using sender's private key and message digest to create digital signature;(7)Using one-time password to encrypt digital signature then put it into MIME body part 3;(8) Using receiver public key to encrypt one-time password then put it into MIME body part 4;(9) Using one-time password to encrypt DICOM document then put it into MIME body part 5 and so on.

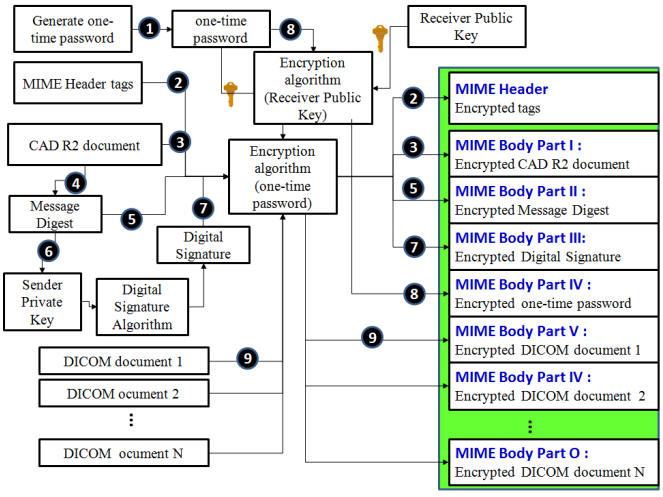


Figure 3 The scenario of Hospital A

Figure 4 represents the processes of Hospital B from step 1 through step 7 correspond to the steps of algorithm in Table 3 respectively. The processes involved in this principle are :(1)Extract encrypted one-time password from MIME body;(2) Decrypted it using receiver's private key then get one-time password;(3) Decrypted MIME header tags using one-time password;(4) Extract MIME body part 1 of encrypted CDA R2 document then decrypted it using one-time password;(5) Extract MIME body part 2 of message digest then decrypted it using one-time password;(6) Extract MIME body part 3 of digital signature then decrypted it using one-time password;(7) Extract MIME body part 4 of DICOM document then decrypted it using one-time password and so on. To verify the digital signature after successfully decrypted.

Table 1 Header Tags

MIME Header Tag	Tag Description
X-EEC-Sender	Sender
X-EEC-Receiver	Receiver
X-EEC-SymmetricAlg	One-time password Algorithm
X-EEC-AsymmetricAlg	PKI Algorithm
X-EEC-DistAlg	Message Digest Algorithm
X-EEC-SignAlg	Digital Signature Algorithm
X-EEC-BodypartCnt	MIME Body Part count
X-EEC-HeaderCnt	MIME Header tag count
X-EEC-MessageType	EMR category

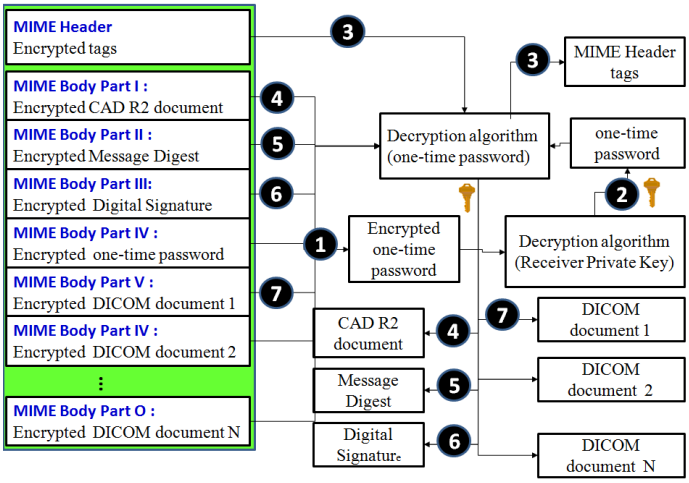


Figure 4: The scenario of Hospital B

4. Experimental Results

4.1. Provide and Register EMR of Hospital A

Hospital A packages CDA R2 and DICOM document into MIME envelope then uploads to Repository of Hospital A and registers to Registry EMR center. Part A of figure 5 shows the encrypted MIME header and Figure 6 shows the encrypted MIME body. After successfully uploaded to the Registry EMR center and signature verification success. It indicates that this transaction has non-repudiation.

4.2. Hospital B Retrieve EMR from Hospital A

Hospital B inquires patient's EMR records from Registry EMR center then retrieves desired EMR records from Repository of Hospital A to Repository of Hospital B. Figure 6 presents the patient's EMR record which successfully retrieved from Hospital A to Hospital B using RESTful service.

Table 2: Algorithm of Hospital A (Provider)

STEP	ALGIRITHM
Step 1	OTPKey = KeyGenerator(KeyAlg)
Step 2	For all header.name Mhd[name].value=encryptData(header.value, OTPKey, OTPKeyAlg) End for

Step 3	EncEMR = encryptData (hisEMR, OTPKey, OTPKeyAlg)
	If EncEMR is not null then
	Mpart.AddBodyPart(EncEMR)
	End if

Step 4	mDist = digest(hisEMR, DistAlg)

Step 5	EncDist = encryptData (mDist , OTPKey, OTPKeyAlg)
	If EncDist is not null then
	Mpart .AddBodyPart(EncDist)
	End if

Step 6	SignEMR = sign(mDist, SenderPriKey, SignAlg) byteSign = concateToByte(mDist, FinPrt)

Step 7	EncSign = encryptData (SignEMR, OTPKey, OTPKeyAlg)
	If EncSign is not null then
	Mpart .AddBodyPart(EncSign)
	EndIf

Step 8	EncOTPKey = encryptData (OTPKey, ReceiverPubKey, PKIAlg)
	If EncOTPKey is not null then
	Mpart.AddBodyPart(EncOTPKey)
	End if

Step 9	EncDICOM = encryptData (hisDICOM, OTPKey, OTPKeyAlg)
	If EncEMR is not null then
	Mpart.AddBodyPart(EncDICOM)
	End if

Table 3: Algorithm of Hospital B (Consumer)

STEP	ALGIRITHM
Step 1	EncOTPKey = Mpart .GetBodyPart(4)
Step 2	OTPKey = decryptData (EncOTPKey, ReceiverPriKey, PKIAlg)
Step 3	For all header.name Mhd[name].value=decryptData(heder.value,OTPKey, OTPKeyAlg) End for
Step 4	EncEMR =Mpart .GetBodyPart(1) hisEMR = decryptData (hisEMR, OTPKey, OTPKeyAlg)
Step 5	EncDist =Mpart .GetBodyPart(2) mDist = decryptData (EncDist , OTPKey, OTPKeyAlg)
Step 6	EncSign =Mpart .GetBodyPart(3) SignEMR = encryptData (EncSign, OTPKey, OTPKeyAlg)
Step 7	EncDICOM =Mpart .GetBodyPart(5) hisDICOM = decryptData (EncDICOM, OTPKey, OTPKeyAlg)

5. Conclusion and Discussion

This paper, was focused on how to use S/MIME and RESTful Web service frameworks to develop a secure mechanism of EMR documents exchange. With the flexibility of REST and MIME envelope, the RESTful Web service frameworks have become more acceptable. Is has been presented a solution to provide security for S/MIME equivalent to VPN. The proposed approach respects the REST philosophy by implementing the message security with MIME envelope. This approach enforces the message to be encrypted and protected during transmission. It was also applied, message digest and digital signature to verification. Thus, the proposed approach can achieve the exchange of confidentiality, integrity, authentication and non-repudiation. When needed patient's electronic medical records can be easily accessed by any hospital. It can be integrated in-house system of hospitals and provide a way to exchange EMR documents securely between different enterprises. Enterprises can exchange patient's information when it is convenient to access the patient's treatment. This avoids repetitive inspections to save medical resources. In order to know if the mechanism can improve or not the security, it must be tested by EEC.

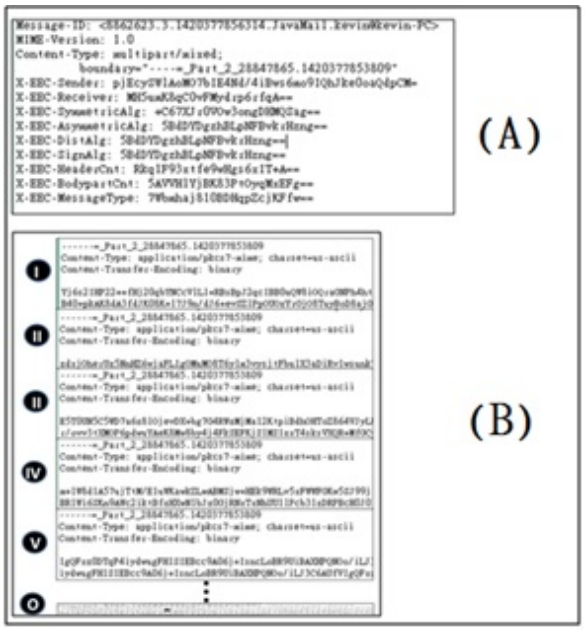


Figure 5 Secure MIME of SEMRS

Figure 6 Retrieved EMR document from Hospital A

References

[1] Ministry of Health and Welfare , Health Care Platinum program. Available:

http://www.mohw.gov.tw/news/448238669, 2009.
[2] National Helth Insurance Administration , Healthcare Information Network Service System (VPN) User Manual, 2012.
[3] Ministry of Justice , Law & Regulations Database of The Republic of China., Available: http://law.moj.gov.tw/LawClass/LawAll.aspx?PCode=L0020021, 2014.
[4] H. G., Hwang, C. H. Lu, J. L. Hsiao and R. F. Chen , "Factors Influencing Benefits of Electronic Medical Records Exchange : Physician Perspectives"., Journal of e-business., 11(1), pp. 95-118, 2009.
[5] Health Level Seven International , About HL7., Available: http://www.hl7.org/about/index.cfm?ref=nav, 2017.
[6] American National Standards Institute , HL7 V3 Normative ., Available: http://webstore.ansi.org/RecordDetail.aspx?sku=HL7+V3+Normative-2011, 2011.
[7] ISO , Data Exchange Standards -- HL7 Clinical Document Architecture, Release 2., Available: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=44429, 2015.
[8] American National Standards Institute (2005), "HL7 Version 3 Standard: Clinical Document Architecture (CDA), Release 2"., Available: http://webstore.ansi.org/RecordDetail.aspx?sku=ANSI%2FHL7+CDA+R2-2005+(R2010), 2015.
[9] NEMA ,. "DICOM PS3.1 2014c - Introduction and Overview"., Available: http://medical.nema.org/medical/dicom/current/output/pdf/part01.pdf, 2013.
[10] T. F. Roy , "Architectural Styles and the Design of Network-based Software Architectures"., Ph.D. dissertation, University of California, Irvine, USA, 2000.
[11] SoftwareAG , "MIME-S/MIME Developer's Guide Version 8.2 "., 2011.
[12] Internet Engineering Task Force , "Secure/Multipurpose Internet Mail Extensions (S/MIME) Version 3.2 Message specification" , RFC 5751, 2010.
[13] OASIS , "Web Services Security: SOAP Message Security 1.1 (WS-Security 2004)"., OASIS Standard Specification, Available: http://docs.oasis-open.org/wss/v1.1/wss-v1.1-spec-os-SOAPMessageSecurity.pdf, 2006
[14] S. Gabriel, S. D. O. Anderson, M. Julien and R. Yves , "Enabling Message Security for RESTful Services"., IEEE 19th International Conference on Web Services (ICWS 2012) , Hawaii, USA, 2012.
[15] P. Gesare, Z. Olaf and L. Frank , "RESTful Web Services vs. "Big" Web Services: Making the Right Architectural Decision"., 17th International World Wide Web Conference (WWW 2008) , Beijing, China, 2008.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 177-185 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

Development of a Motorized Afifia Mowing Machine Design for Controlling Environmental Conservation and Menace for Home Use

Gbasouzor Austin Ikechukwu[*,1], Mbunwe Josephine Muncho[2]

[1]*Department of Mechanical Engineering, Chukwuemeka Odumegwu Ojukwu University, Formerly Anambra State University, P. M. B. 02 Uli. Nigeria.*

[2]*Department of Electrical Engineering, University of Nigeria Nsukka Nigeria.*

ARTICLE INFO

Article history:
Received: 21 December, 2016
Accepted: 18 January, 2017
Online: 28 January, 2017

Keywords:
Convenience,
Conventional manual method,
Environmental Conservation,
Lawn Grass, Mowing, Reel,
Rotary Cutters, Rotating Blades

ABSTRACT

Technology has become more affordable and penetrates every aspect of daily life, even in developing country like Nigeria. However many of the users in developing countries are still finding difficulty in using the technologies due to lack of experience as they undergo a technology leap. The aim of this research work explores the approach in designing, development of a motorized Afifia (grass) mowing machine. This research was considered because of the unhygienic environmental conservation and its menace. An estimate of 20N was adopted as the required force to cut lawns and based on this design force of 70N was chosen. This design force was the basis of characterizing the selection of materials use, as a result it was found that the machine is 85% efficient based on the area mowed per hour which is 390.6m^2.

1. Introduction

A mower is a person or machine that cuts (mows) grass or other plants that grow on the ground. We may want to keep a cleaning environment around our houses. Then we need to remove unnecessary grass on the compound and also if somebody need to make a new sport ground or maintain a current ports ground, we need to put a certain grass level. To do that, we use a very tool which known as lawn mower. Thus a lawn mower is a machine that uses a revolving blade or blades to cut a lawn at an even length. Actually there are plenty of various types of lawn mowers, but we need to choose the best one according to our work. There are 3 types of lawn mowers which are,

- Push type: a human needs to push it
- Ride on type: a human rides it
- Multi-gang type: pulled behind a tractor

Firstly, we need to identify which kind of lawn mower needed for our work

- For or home purposes, we need to use push type.
- For sports ground purposes, we need ride on type.
- For even bigger grounds such as golf grounds are multi-gang type.

A smaller mower used for lawns and sports grounds (playing fields) is called a lawn mower or grounds mower, which is often self-powered, or may also be small enough to be pushed by the operator. Grounds mowers have reel or rotary cutters. Larger mowers or mower-conditioners are mainly used to cut grass (or other crops) for hay or silage and often place the cut material into rows,which are referred to as windrows, Swathers (or windrowers) are also used to cut grass (and grain crops). Prior to the invention and adoption of mechanized mowers, (and today in places where use of a mower is impractical or uneconomical), grass and grain crops were cut by hand using scythes or sickles. Lawn mower is simply described as a machine used for cutting or mowing grass. Petrol driven rotary lawn mower is a piece of equipment that provides the convenience of keeping the lawn grass short. It serves as a substitute for the conventional manual method and pushed gasoline lawn mower; which has been found not only inefficient but so labour intensive.

A lawn mower is a machine utilizing one or more revolving blades to cut (afifia/grass) surface to an even height. The height of the cut grass may be fixed by the design of the mower, but generally is adjustable by the operator, typically by a single master lever, or by a lever or nut and bolt on each of the machine's wheels. The blades may be powered by muscle, with wheels mechanically connected to the cutting lades so that when the mower is pushed forward, the blades spin, or the machine

[*]Corresponding Author: Gbasouzor Austin Ikechukwu is a Ph.D researcher and Lecturer in the Department of Mechanical Engineering, Chukwuemeka Odumegwu Ojukwu University, Formerly Anambra State University, P.M.B. 02 Uli. Nigeria,

E-mail: unconditionaldivineventure@yahoo.com

may have a battery-powered or plug-in electric motor. The most common power source for lawn mowers is a small (typically one cylinder) internal combustion engine, particularly for larger, self-propelled mowers. Smaller mowers often lack any form of propulsion, requiring human power to move over a surface; "walk-behind" mowers are self- propelled, requiring a human only to walk behind and guide them. Larger lawn mowers are usually either self-propelled "walk-behind" types, or more often, are "ride-on" mowers, equipped so the 'operator can ride on the mower and control it.

1.1. The Application of Lawn Mower

The application of the lawn mower is extensive. The rise of sporting activities like football, cricket and golf has necessitated the use of these machines more extensively.

Moreover, the use of motorized lawn mower has helped in saving time and energy when compared with manual cutting of grasses.

But even with the above advantages of motorized lawn mower, it is still found that manual labour is still largely applied in Nigeria. This may be as a result of the cost of producing a lawn mower. It is therefore our intention to construct and fabricate motorized petrol driven lawn mower at a greatly reduced cost so that the common man can afford it.

1.2. Aims and Objective

This machine was constructed with, the design of a tricycle unlike the convectional four wheel drive. A snuffer was constructed as part of the machine to keep Mown grasses on a straight line along the path of the Lawn mower. This serves to protect the operation from propelled broken pebbles. Also, it is our inspiration to design and construct a functional motorized tricycle gasoline powered Lawn mower at a reduced cost, compared to the conventional four wheel driven, in the market.

- This method is injurious, laborious, and unhygienic and produces non-uniform cutting and thereby giving poor end finishing.
- It uses human energy, slow and labour intensive
- The design and development of motorized afifia mowing machine will provide a relief of human labour speed up cutting and produces uniform cutting or clearing for a better end finishing

1.3. Scope of Research

The design and fabrication is considered with the constraints and consideration discussed. The study aim to produce a cost effective lawn mower that is driven by a petrol engine with the capability to mow lawns to an optimal height of 20mm (2cm) Using simple belt and chain drive mechanism in order to ease operation and maintenance of the machine.

2. Enviromental Importance of Fefea/Grass Moving Machine

Grass cutting is a tedious work for man became it requires a lot of energy especially using local hand-tools. This difficulty led to the development of devices that could ease the cutting process (Drake and Hermick, 1983). Some of these devices include the Reel mower, electric trimmer and the Hand sheare. Although these' devices still depended on the use of hands, it was relatively easier than using cutlasses. In the process of developing these simple devices, the LAWN MOWER was invented (Greamer, 1964). The first Lawn mower was invented in 1830 by EDWIN BREAD BUDDING, an engineer from Stroud, Gloucester- shire, England. He developed the idea from observing a machine in a local cloth mill which used a cutting cylinder (or blade reel) mounted on a bench to trim cloth to make a smooth finish. After weaving, Budding realized that a similar concept could be used in cutting grass if the mechanism could be mounted on a wheeled frame to make the blade rotate close to the Lawn's surface. He then collaborated with a local Engineer, John Ferrabee and together they developed the idea and produced the first Lawn mower in a factory at Stroud. The machine was made of art iron with geared wheels transmitting power from the rear roller to the cutting cylinder. Budding and Ferrabee were shrewd enough to allow other companies build copies of their mower under license. The most successful of these was Ransom's of Ipswich which began producing mowers as early as 1932.

The company has continued making Lawn mowers since then and it is now the world's largest manufacturer of Lawn care equipment (www.kellerstudio.com.4th April, 2008). Other companies apart from Ransoms of Ipswich picked up Budding's idea. Thomas Green was one of these companies and he produced the first chain driven mower in 1859, named the “Silens Messor” which means the silent cutter. It used a chain to transmit power from the rear roller to the cutting cylinder. These machines were lighter and quietly than the gear driven mowers that preceded them although they were, slightly more expensive. At about the same time, Alexander shanks of Abroath introduced it's range of Caledonian mower. All were available with either gear or chain drive and Grass collection boxes were optional. Around 1990, one of the best known English machines was the Ransom's automation which was available in chain or gear driven models. J. P engineering of Leicester, founded after world war one, produced a range of very popular chain driven mowers. About this time an operator could ride behind animals that pulled large machines. These were the first riding mowers (www.wikipedia.com.4th April 2008).

The development of these mowers was aided by the rise in popularity of sports like lawn tennis, football, cricket and rugby, helped prompt spread of these inventions. James Sunner of Lancashire patented the first steam- powered lawn mower in 1893 his machine burn petrol and/or paraffin oil (kerosene) as a fuel.

After numerous advances, the machines were sold by the stilt fertilizer and Insecticide Company of Manchester. The company they controlled was called the Ley Land steam motor company. Numerous manufacturers entered the field with gasoline driven mowers after the turn of the century. The first grass boxes were flat trays but they took their present shape in the 1860's the roller driven Lawn mower has changed very little since 1930. Gang lawn mowers, those with multiple set of blades were built in the USA in 1919 by a minister, Worthington. His "company was taken over by the Jacob seen corporation but his name is still cast on the frames of the gang units. Rotary mowers were not developed until engines were small and powerful enough to run the blades at a high speed. In 19301 power specialities Itd introduced a gas power rotary mower. However, the machines continue to attract the interest of collectors and enthusiast art throughout the world, which was why the Old Lawn mower club was formed in 1990. (www.reference.com.4th April 2008). EgbeJule Blessing designed an electric Lawn mower at the University of

Portharcourt which had the advantages of quick start, less maintenance cost and produced little or no noise Egbejule, 2006). His design however, was limited in use nee it depended on electricity which is never constant in Nigeria. Based on these findings we intend to review Egbejule's design producing a motorized lawn mower that uses a petrol engine as it's power source to eliminate the dependence on electricity. It is our utmost desire to produce this machine at affordable rate.

3. Design of Afifia/Grass Mowing Machine

3.1. Material Selection

For an intelligent and resourceful design, the designer must clearly know the materials, which are available and the properties they posses. Selection of materials depends on factors such as the intensities and types of stresses to which the component is subjected to, whether it is flexible or rigid or it is to experience high temperature or corrosive actions and how it lends itself to processes of manufacture i.e forging, machining e.t.c.

Therefore the designer's selection will be influenced by the following factors;

1) Strength
2) Weight
3) Appearance
4) Manufacture
5) Cost of production

These also will determine the variation between success and failure of the machine. We can further classify the above factors into four classes;

a. Service requirement
b. Fabrication requirement
c. Maintenance requirement
d. Economic requirement

3.2. Service Requirement

In deciding the choice of a material, the material has to have a certain properties that can fulfill the desired role. These properties called service requirement include strength, stiffness, hardness, corrosion resistance, heat resistance e.t.c.

3.3. Fabrication Requirement

A material must possess some properties that enable it to be worked on. These properties include forgeability, malleability, ductility and ability to be welded. Material undergoing forgeability are heated to a high temperature close to its melting point and then shaped to a desired structure. Malleability entails that the material should be able to take sheet like forms while ductility requires that the material is able to be drawn into the form of a wire. The material must be able to be joined by welding.

3.4. Economic Requirement

This is about the most important factor for the selection of materials because it determines the cost of production, which in turn determines the price of the product and its use by consumers. If the total cost of production is high, consumers will look for cheaper alternatives which will eventually defeat the purpose of the production.

Bearing In mind that the two main aims of production are satisfying human needs and also to make profit, a producer must judiciously select relatively cheaper, but reliable and appropriate materials for production.

This will reduce the overall cost of production. This in turn ensures hat human needs are satisfied and ultimately justifies this project work.

3.5. Maintenance Requirement

This enables the interchangeability of failed parts. Failed parts are expected after the machine has served for some good number of years. But, it does not always go this way if not properly maintained, machine parts can get bad before the expected time or life time. It therefore means that maintenance involves far more than changing parts. It includes day to day activities like clearing, lubricating, tightening of loosed bolts and nuts as all of these have a bearing on the life span "of the product.

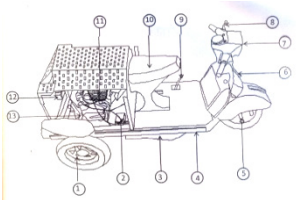
3.6. Choice of Material

Based ·on the above considerations, the materials for all parts of the machine were selected. These materials are tabulated in table 3.1 below.

Table 1: Materials selected

Components	Material	Reasons
Chasis	Mild steel	Tough, hard and malleable
Blade	Steel	Light weight and resistance to corrosion
Bolts and nuts	Mild steel	Tough, hard and malleable
Handle	Mild steel	Tough, hard and malleable
Shaft	Medium carbon steel	High tensile strength, heat and corrosion resistance
Belt	Leather	Tough

Fig 1: The Designed Afifia Mowing Machine



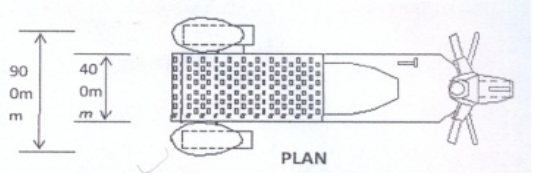


Fig 2: Assembly view of the Afifia mowing Machine

Table 2: Items

Items	Descriptions	Items	Description
1	Tire	7	Head lamp
2	Battery	8	Clutch lever
3	Snuffer	9	Gear pedal
4	Chassis	10	Seat
5	Shock Absorber	11	Engine
6	Brake lever	12	Rear lamp
13	Shaft		

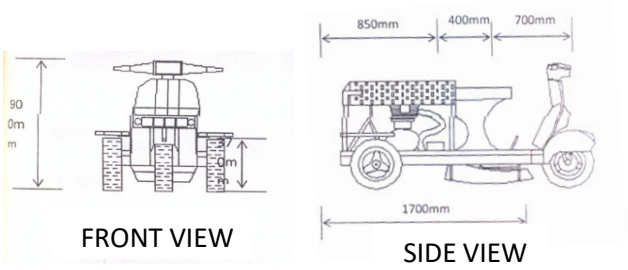


Fig 3: Different Technical Views of the Machine

Fig 4: Orthographic Projection of the Machine

3.7. Manufacturing Process of the Chasis

The chasis was constructed from mild steel with thickness 7mm. the sheet was cut with the aid of chisel, hammer, and hacksaw and was developed into the shape of the frame. The mild steel was joined together by metal arc welding. This utilizes heat of the arc between a continuously fed consumable electrode and the work to be welded. The heat of the arc melts the surface of the base metal and the end of the electrode. The metal melted off, the electrode is transferred across the arc to the molten pool. Shielding of the molten pool, the arc and the surrounding area is provided by an envelope of gas fed through a nozzle. The shielding gas surrounds the area to protect it from contamination from the atmosphere.

3.8. Heat treatment of the Blade

The blade was subjected to heat until it became red hot. This is done to distort the grain structures of the metal, after which cooling was controlled in a bow containing oil. This is an attempt to bring the grain structure to a coarse form. This process is known as hardening. The essence of the aforementioned process is to increase the toughness, hardness and strength of the blade, thereby preventing wear to the blade as it makes contact with stones, pebbles, broken bottles etc

3.9. Shaft

The shaft was machined with lathe to a diameter of 70mm and coupled directly to the wheels by means of a lock nut and a key.

3.10. Coupling

The blade was coupled directly to the gear by means of a lock nut and a key.

3.11. Lawn Mower Assembly

Lawn mower components are listed in the table below

Table 3: List of Lawn Component

S/n	Element	Quantity
1	1 C engine	1
2	1 C engine seating	1
3	Blade	1
4	Front tyres	1
5	Rear tyres	2
6	steering	1
7	operators seat	1
8	Steel chassis	4

The assembling of the component to form the lawn mower is as follows:

An 1C engine positioned at the rear of the chassis is bolted to the chassis. The rear axles with two tyres of 37cm diameter are bolted to the wheels while the blade is directly coupled to the gear in the mower housing. A front tyre of 37cm diameter is also fixed. All tyres are detachable and adjustable to enable changes when bad and to adjust the height of the lawn mower respectively.

3.12. Surface Finish

This process entails remover of spatter formation arising from the intense heat during welding process. This was achieved with the help of grinding disc. Thereafter, energy paper was used to smoothen the affected parts. This process was followed by painting the frame with a black oxide. This was done to prevent corrosion. The machine was then finished by painting the body red.

The design procedure for this study will be discussed under the major heading.

- Problem formulation.
- Design specification.
- Feasibility study.
- Preliminary design
- Detailed design.

3.13. Problem Formulation

There is a, general problem of keeping the environment clean. This often requires keeping lawns at an acceptable height to avoid the problem of harboring snakes and other dangerous animals. A lawn mower will be necessary in achieving this. But, while lawn mower can satisfy this needs, it may not be affordable. A cost effective motorized town mower will therefore be necessary to fully satisfy needs.

A market survey was conducted to ascertain the present cost of a motorized lawn mower of 20.00Hp (14.914kw) engine capacity sells for #400,000; a 15.00 Hp (11.1855kw) engine capacity sells for #300,000. Literature review of past work on this research work was also examined in order to understand the areas that need improvement.

Considering the following constraints:
1. Minimum cost
2. Limitation of side effect.
3. Technological limitation.
4. Minimum performance requirement.

3.14. Design Specification

This project is aimed at producing a cost effective one sealer tricycle motorized lawn mower to carry a maximum lead of 80kg.

The motorized lawn mower is going to be mechanically powered with a Qlink 125 engine capacity: - one cylinder, 4 stroke petrol engine

Horse power: .- maximum power, 18hp· (13.1kw)
Max, Output: - 14p.s/ 75000rpm
Max. Torque: - 14N.m /6000rpm
Compression ration: - 9: 5: 1

The design will be typified by the following features;
- A steel chassis for strength and durability.
- Shaft and bearing for the rear wheel.
- Steel metal cutting blade of height 350mm (35cm)
- The handle bar.
- The chain drives transmission for the rear wheel.
- The belt drive transmission for the blade
- Intermediate pulley for tensioning the belt.
- The gear system on the blade.
- The adjuster for the blade.
- The gear lever
- The breaking system.
- The throttle and clutch on the handle
- Position of the seat.
- Force required to move the machine and to cut grass

3.15. Feasibility Study

Different system could be employed in satisfying the design of the motorized lawn mower in order to achieve the aim of this project. A proper feasibility study was carried out from this study. It was observed that different mechanisms can be used to achieve the aim of this work e.g, an electric motor can be used as a source of power. A mechanical system that involves the use of direct gears to transmit the torque from engine to blade and wheels, a robotic system can also be used. From above analysis, we decided to use a petrol engine as the source of power. This was chosen because;

1. An electric motor will require the use of long wire during operation. This is inconvenient and may really not appeal to everyone, moreover, in country like Nigeria, where electric power supply is not really available, the use of electric motor will be in appropriate.
2. The robotic system is far too expensive to be considered adopting, it defeat the aim of this project.
3. The use of direct gear transmission will require a lot of machining; also, it follows with a large noise due to vibrations from the gear.

In the system chosen, the rear axle should be operated on a manual transmission through gears liked with chain and sprocket supported on the shaft of the axle (4 speed, constant mesh gear box).

The blade will be coupled directly to the shaft from the bevel gear box. The bevel gear box will be derived by the engine direct shaft which will be transmitted by belt drive from the engine shaft to the bevel gear box. The cooling system will be forced air cooled. The break will be on internal expending drum consisting of two shoes.

The gears and bearing life will be 712, 00 infinite shaft life. The engine will be located at the rear market. Segment will be of low cost market safety features. Low centre of gravity (C.G) to ensure firm grip of road wheels on the ground cost limitations: < #250,000.

3.16. Preliminary Design of the Chassis

The chassis is an arrangement of angle bars joined together by welding, the chassis is required to carry the engine weight, the mower house, the operations weight and acts as a mounting support for the handle and wheels.

The chassis should be able to keep all the component on. It in there corrects relative positions in spite of all varying stresses which they may be subjected to. As the base on which all other components are attached, it is expected that the weight of the engine, the operators weight, accelerating torques as well as centrifugal forces will all act on the chassis.

3.16.1. Shaft and Bearings for The Rear

An axle shaft is used for transmission of bearing moment. The shaft is coupled with three radial ball bearing which acts as a support to the wheels and also able to withstand the corresponding torque.

3.16.2. Cutting Blade

This is the cutting element of the lawn mower. It is coupled to the bevel gean shaft by a nut and a key for proper fastening. It has a cutting edge on either side along its length. It is made from hardened steel of length 35cm and breath 5cm.

The blade will be acted upon by three forces

(1) Shear force that is impacted' on the blade as it cuts lawns.

(2) The weight of the blade that tends to drag it out of its attachment paint.

(3) Centrifugal force that tends to pull it out of its path of rotation.

3.16.3. The Handle (Steering)

The steering id through a linkage, with the handle bar directly connected to the front wheel or via a form of Ackerman steering geometry. The handle (steering) carries the throttle control lever, clutch control lever and brake control lever.

The handle will be acted upon by a torque requires to turn the handle by the operator.

3.16.4. Chain-Drive

This is the mechanism that will be introduced in transmitting the output motion of the gear shaft from the engine to the rear wheels (axle).

In designing the chain drive, we ·are to determine the velocity ration of the chain drive, select the minimum number of teeth on the pinion, design power by using the service factor, and choose the type of chain and chain length.

3.16.5. Belt - Drive

The belt drive mechanism will be introduced to transmit the motion from the direct shaft of the engine to the bevel gear box which will drive the blade.

3.16.6. Intermediate Pulley

This will be introduced for constant tensioning of the belt. This is because of different level of the mower housing. The idler pulley will be design in such a way that it will be moving to and fro so that it aims will be achieve.

3.16.7. The Bevel Gear Box

Here, a bevel gear will be used to change the axis of rotational motion of the engine shaft. A miter gears connect two shaft whose axes intersect at right angle. In designing the bevel, we consider the number of teeth of the bevel gear that correspond to the speed required by the blade to mow a particular square size of a lawn at a particular time.

3.17.5. The Blade Adjuster

Here, a mechanism will be chosen to perform this operation. The arrangement is as shown below

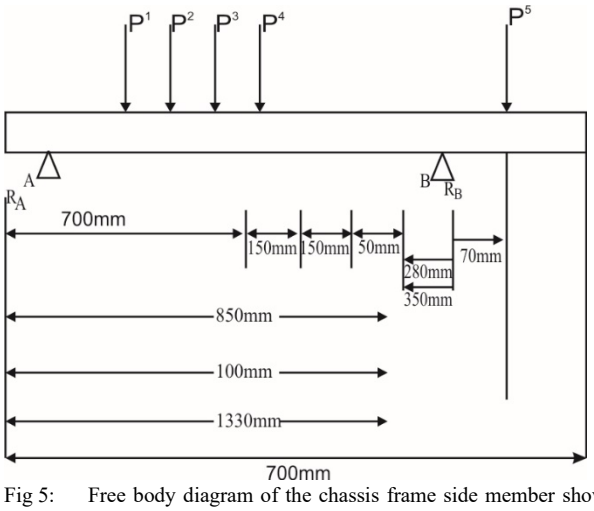


Fig 5: Free body diagram of the chassis frame side member showing the various load distributions.

3.16.8. The Gear Lever

A mechanism will be chosen to serve as gear lever close to the position of the operator. Here, we consider it to be a four bar mechanism as shown below.

3.16.9. The Braking System

The brakes will be an internal expanding drum consisting of two shoes. The effect of this brake is control at the handle of the machine. This is linked by a flexible cord which is operated by pushing the engine brake lever when brake is necessary.

3.16.10. The Throttle and Clutch On The Handle

This is used to control the speed and engagement of gear for operation of the engine. It is formed from a coupling of iron and plastic anchored to the handle. It is linked to the engine by a flexible cord which is operated by pushing the engine throttle when in tension and releases it when free.

3.16.11. Position Of Seat

The position of the seat and height of the seat will be put into consideration for efficient comfort during operation. Here, we shall consider the maximum weight of the operator and expectable height of the operator.

3.16.12. Force Required To Move The Machine And To Mow Grass.

The amount of force needed to mow a lawn is less than 20N. For greater efficiency, a machine that will deliver a Larger force more than that of the cutting force and moving force plus friction will be adopted. Required force = Static friction + Dynamic friction + Cutting force + Maximum force for movement.

4. Fundamental Distributions Analysis of the Machine

Having completed the preliminary design, the various part of the lawn mower will be designed.

4.1. Chassis Frame Design

Considering the side member of the frame, the several load reactions can be distributed and presented as shown below

Where;
P_1 = weight of the mower house with blade - 117.72 and gear box
P_2= weight of body and frame = 441.5N
P_3 = weight of seat 8 operator = 833. 85N
P_4 = weight of front engine support = 294.3N
P_5=weight of rear engine support = 343.35N
P_A = Reaction at support A,
P_B= reaction at support B.

4.2. Determination of the Magnitude of the Support Reaction at A and B

For equilibrium of the system,
The resultant moment must be zero
i.e., $\sum M_A=0$
$\sum M_B = 0$
The resultant vertical force must be zero
i.e. $\sum fy = 0$,
Taking moment about B,
i.e. $\sum M_B=0$,

Table 4: Internal stresses

Section	Shear force
A-1	F_{A1} F_{1A}= RA= 465.882N
1.2	F_{12}= F_{21} = R_A= - P_1 = 348.162N
2-3	F_{23}= F_{32}= R_A= P_1- P_2- 93.338N
3-4	F_{34}= F_{43}= R_A-P_1- P_2- P_3 P_4=-927.188N
3-B	F_{4B}= F_{BA}= R_A- P_1- P_2= -1221.488N
B-5	F_{B5}= F_{5B}= R_A- P_1- P_2-R_B=-2786.328N

4.3. Design Procedure Of Chain-Drive

The chain drive is design below velocity Ratio of the chain drive,

$$V.R = \frac{T_2}{T_1} \qquad N1=1800rpm$$
$$N2 = 710 \text{ rpm}$$

Where T_1 = Number of teeth on the smaller sprocket = 15
T2 = Number of on the large sprocket = 38

$$\therefore V.R = \frac{T_2}{T_1} = \frac{38}{15} = 2.5333AY\ 3$$

4.4. Rear Axle Shaft Design

The shaft with a sprocket or gear mounted on the bearings is shown in fig. it is subjected to combined twisting moment and bending moment.

Power, P = 13.1kw= 13,100w, N = 710rpm, D = 0.193m
L =1.2m, Normal load (w) = 1832N
Torque transmitted by the shaft,

$$T = \frac{PX60}{2\pi N} = \frac{13100X60}{2\pi X710} = 176.2Nm$$

Tangential force on the gear,

$$T_t = \frac{2T}{D} = \frac{2X176.2}{0.1193} = 1825.91N$$

Since the sprocket is mounted at the middle of the shaft, therefore maximum bending moment at the centre of the gear

$$M = \frac{W.L}{4} = \frac{1832X1.2}{4} = 549.6Nm$$

According to maximum normal stress theory, equivalent bending moment,

$$M_e = {}^1/_2 (M + \sqrt{M^2 + T^2}) = (M + T^2)$$
$$= {}^1/_2 (549.6 + 577.2) = 563.4Nm$$
$$= 563.4 \text{ X } 10^3 \text{N mm}$$

Also equivalent bending moment (M_e),

$$563.4 \text{ X } 10^3 = \frac{\pi}{32} \text{ X } \delta_b \text{ X } d3 = \frac{\pi X17.5Xd^3}{32}$$

$$d^3 = 327928.3015$$
$$d = 3\sqrt{327928.3015} = 68.96mm \text{ say } 70mm$$

4.5. Design of Belt-Drive

N_1 =7500rpm = N_2, d_1, = d_2 = 100mm = 0.1m
p= 13.1kw, $r_1 = r_2$ = 0.05m, μ= 0.3

speed of the belt (v)

$$v = \frac{\pi d_1 N_1}{60} = \pi \text{ x } \frac{0.1 \text{ x } 7500}{60} = 39.27m/s$$

Power transmitted (p),
$13.1 \text{ X } 10^3 = (T_1 - T_2)\ 39.27$
$\therefore T_1 - T_2 = 13.1x\ 10^3/39.27 = 333.59N$ …….. (i)
We know that
$2.3 \log (T_1/T_2) = \mu\theta 1 = 0.3 \text{ X } \pi$ ($\therefore \theta = 180^0\ \pi$rad)

$$\therefore \text{ Log } (T_1/T_2) = \frac{0.942}{2.3} = 0.4096$$

Or $T_1/T_2 = 2.57$ ..... (ii)
From equations (i) and (ii),
$T_1 - T_2 = 333.59N$ and $T_1/T_2 = 2.57$
$T_1 + 2.57\ T_2$

$\therefore\ 2.57\ T_2 - T_2 = 333.59$
$1.57\ T_2 = 333.59$
$T_2 = 333.59/1.57 = 2.1.48N$
And $T_1 = 2.57 \text{ X } 212.48 = 546N$

Let b = Width of the belt in meters
t = thickness of the belt = 10mm

Since the velocity of the belt is more than 10m/s, therefore centrifugal tension must- be taken into consideration. The density of the leather belt used is 1000kg/m^3

4.6. Design of Bevel Gear Box

A 90° bevel gearing arrangement is used to drive the mower blade. It is designed thus,

Module and face width for the pinion
Let m = Module in mm
b = face width in mm = L/4

Dp = Pitch circle diameter of pinion
Velocity ration,

$$V.R. = \frac{T_G}{T_p} = \frac{44}{12} = 3.667$$

$\therefore$ Speed of the gear,

$$N_G = \frac{N_p}{V.R} = \frac{7500}{3.667} = 2045rpm$$

4.7. Pressure on Wheels

The wheel will be acted upon by the three forces.
1. The weight of the engine (W_e)
2. The weight of the chassis (W_c)
3. The weight of the blade (W_b)
We = 20 x 9.81 = 196.2N
We = density of chassis materials
X volume of chassis (V_c) x 9.81

V_c= Area of chassis x thickness = 1,020m^3
= 1.7 x 0.6 x 0.007 = 0.00714m^3
= 7.14 x 10-3m^3
Density of chasis material
= 7861kg/m^3
W_c = 7861 x 7.14 x 10^{-3}
= 56.12754 x 9.81
= 550.6N
Total weight on the wheels
= 196.2 + 550.6 + 11.7
= 758.5N

$$\text{Force acting on the wheels} = \frac{758.5}{4}$$

= 189.6N

4.8. Determination of the Area Mowed by the Blade During one Revolution of the Wheel

The linear distance covered in one revolution of the wheel, Id
$L_d = IT_{dw}$ = IT x 0.195 = 0.61m
Diameter of the blade, db = 0.35m
Area covered in one revolution Ar

Ar = Ld x db

= 0.61 x 0.35

= 0.21m^2

4.9. Determination of Area Mowed in one Hour

Time taken to mow 0.61m
(1 rev of wheel) = 2sec

$$V = \frac{S}{} = \frac{0.61}{}$$

t 2
= 0.31m/s
Assuming this speed to be constant, the distance covered in one hour (3600s) will be:
S = Vt

= 0.31 x 3600 = 1116m/hr
1116m will be covered in one hour
Area mowed in one hour = distance covered in one hour x diameter of blade
= 1116 x 0.35
=390.6m^2

5. Testing and Results

5.1. Test and Results

The machine on complexion was tested on lawn and found working normally with little or no vibration. It was tested on many occasions without giving any sort of problems showing a very satisfactory performance.

The machine was operated for an hour and it was found that the area mowed amounted to 211m. This was as opposed to the expected performance based on calculations.

As a result of the foregoing, we were able to estimate the efficiency based on the performance criteria earlier stated.

$$\text{Efficiency} = \frac{\text{expected area mowed based on calculation - actual area mowed}}{\text{actual area mowed}} \times 100$$

$$\text{Efficiency} = \frac{390.6 - 211}{211} \times 100 = 85\%$$

5.2. Discussion

The efficiency of the machine was found to be satisfactory from the results obtained above. This shows that even if the materials were not the best, satisfactory equipment could be produced in other for the machine to work properly; the following procedures must be followed.

5.3. Operational Procedures

The mower uses a dry cell battery to switch on the engine. The starting procedures are as follows:

1. Switch on the ignition.
2. Press the start button until the engine starts.
3. Push the clutch lever and engage the gear.
4. Then drive the mower to the area to be mowed ensuring that there are no loose stones, twigs, branches, logs e.t.c.
5. Mow along the lawn to give a uniform cut.
6. After mowing, disengage the gear and switch off the engine.

5.4. Precautions to be taken

The necessary precautions are as follows:

- Make sure you understand the operation of your mower, especially how to stop it.
- Make sure all protective devices are in place on the mower.
- Wear proper eye protection like plastic eye glasses or safety glasses to protect your eye from flying debris.
- Use earmuff to reduce noise level to the ear.
- Keep curios kids and pets at a safe distance away from the mowing area.
- Mow high sloppy areas from side to side.
- Before operation is made, make sure that there are no stones, debris, bottles, tins, pebbles or any hard object on the grass that could damage the blade.

5.5. Maintenance and Storage

- Before carrying any maintenance or cleaning operation, stop the engine.
- Periodically check the blade to see if it requires re-sharpening or replacement.
- clean and lubricate dry component of necessary
- clean above and below deck to remove dirt, leaves and debris, pay close attention to the blade housing.

5.6. Blade Maintenance

- First jack up the machine and place it on axle stand.
- check to see if the blade is bent
- locate a reference point on one side and known the height of the blade, tip at that location, then rotate the blade and check the height of the opposite blade tip. There should be no significant difference. A difference of say 0.003175m or greater shows that the blade is bent.
- Slight makes and bends can be cleaned up with a file while the blade is still in position. If damage is severe, a replacement will be needed.

5.7. Change of Blade

Changing blade is carried out by a qualified personnel, is carried out as follows,

- Obtain a new blade from the market with the same dimensions as the one in the mower.
- Remove fasteners, i.e bolts and nuts.
- Replace the old blade with the new one.
- Replace fasteners on the new blade

Conclusion

A motorized lawn maker which helps to push away obstacles has been built with the list cost possible. The cost of a motorized lawn mower of the same capacity in the market is about N400, 000 while the one produced in this research cost about N250, 000. Comparing the two cost shows that their design achieved its target tremendously. If given the opportunity to produce a mass production of our mower, the cost of production could be as low as N200, 000. It is our belief that our innovation can put more mowers in the hands of consumers and eliminate the dependence of manual labour in our country Nigeria. Finally, it is believed hope that this study will symbolically represent our efforts which are aimed to serve our great country as engineers eliminating problems of the common man.

Recommendation

In our industrialized world, advances in technology are on the increase. It is therefore necessary that we improve on our indigenous technology. While the motorized lawn mower is very useful in eliminating manual labour used in cutting and trimming lawns, the operator is still required to drive the lawn mower during -operation. This too, requires expanding time energy. The advancement in robotics has given rise to robotic lawn mowers. In this type of mower, the operator is not require to push or drive the lawn mower as he cuts the laws. He is just expected to mark out the area and program it on the machine.

The machine then moves within the marked areas and cut the lawns on it to the desired height. This machine is quite expensive. It is our recommendation that a research be carried out on this type of lawn mower with a view of producing it on affordable price to consumers who need it.

References

[1] Akele S. m. G and sulaimon A.O, (2003) Automotive for beginners Nigeria. Pagasms publishers Auchi page 36.
[2] Drake C.W and Helmick C. G (1983), industrial motor application, Halsted Inc. New York. Page 62
[3] Egbejule B. (2006), design and fabrication of lawn mower, project report, university of port-Harcourt
[4] Emmanuel Alvin O. (2011), design and fabrication of two seater sports car, project report, Anambra state university, Uli.
[5] Greamer R.H. (1964), machine design, Delhi Rajcudra Revindra publishers ltd. Ram Nager, New Delhi.
Page 423
[6] Khurmi R.S and Gupta J.K. (2004) theory of machines.
[7] Oko C. O. C and Abam D. P. 5 (2006) Engineering Professional practice and procedures, university of port- Harcourt press. Page 63.
[8] Kinnander, Ola (October 25,, 2012). “Rise of the Lawn-Cuting Machines”. Bloomberg Businessweek. “Mower History”. Oldlawnmowerclub.co.uk.
[9] Mower History”. The Old Lawnmower Club Collection, Preservation and Disay of Old lawn Mower. N.P., n.d. Web. 29 Feb 2012

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 186-191 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

Theoretical Expression of the Balance Function during Galvanic Vestibular Stimulation in Dual Space

Hiroki Takada*

Department of Human & Artificial Intelligence Systems, Graduate School of Engineering, University of Fukui, 910-8507, Japan

ARTICLE INFO

Article history:
Received: 17 December, 2016
Accepted: 20 January, 2017
Online: 28 January, 2017

Keywords:
Body sway,
Stabilogram, Dual space
Stochastic differential equation (SDE),
Temporally averaged potential function,

ABSTRACT

A method to construct a stochastic differential equation describing a non-Gaussian process in a stationary state and to obtain a static potential function in terms of the temporal average is proposed. However, it has been suggested that the potential function changes temporally through some analysis of climatic change. In this paper, a mathematical model of body sway is theoretically constructed during galvanic vestibular stimulation. The mathematical model is not regarded as stationary due to perturbation. We discuss a new expression for the temporal variations of the model with the use of a motion process in a dual space composed of coefficients of the temporally averaged potential function and the possibility to estimate the number of unobservable variables.

1. Introduction

There are several methods to derive a mathematical model for a motion process from time series data. In mathematical models directly derived from observed data, the relation among only time series data describes a state value Z; the other state variables are not essential to describe the motion process. The auto-regression (AR) model is well known as a mathematical model to describe Gaussian time series with distributions that are regarded as normal. This mathematical model has succeeded in describing various stationary processes. However, there are currently no established methods to derive a mathematical model directly from non-Gaussian time series data for the sway of the center of gravity of a human body [1], meteorological elements [2], or exchange rates of JP¥ to US$ [3]. Moreover, it is difficult to extract hidden variations in the state variables that are not directly observable. We propose a method to construct stochastic differential equations (SDEs) describing the non-Gaussian processes in stationary states and to obtain a static potential function in the meaning of the temporal average.

In recent years, it has been suggested that the potential function changes temporally through some analysis of climatic change. Air temperature has shown an upward tendency in statistical tests in previous research [4]. These tests were reasonable to show a general tendency; however, these were insufficient to analyze details of the temporal variations. Shimizu and Takada [2] found temporal variations in the form of histograms on which a new peak appears at higher temperature and variations in form of the potential function in the meaning of the temporal average, which is derived from the form of the histogram in agreement with Takada et al. [1]. The authors have proposed that variations in the form of the potential function are assumed to follow a delay convention of one-dimensional cut off of the wedge catastrophe [5]. If the same type of conventions were introduced into the temporally averaged potential, we could build up the following hypothesis involved in the physical system describing a climatic change as seen in air temperature.

Hypothesis A stationary point of the temporally averaged potential moves in agreement with a delay convention (Figure 1a) in catastrophe theory.

Assuming that these assumptions are satisfied, we can give a physical explanation of time series; for example, a downward arrow in Figure 1b indicates the appearance of a stable equilibrium point of high temperature in the 1990s. Moreover, this hypothesis gave us an opportunity to find a transition of the temporally averaged potential function. It was suggested that there are transition and temporal variations in the temporally averaged potential function caused by breaches of stationary states. There may be dynamic potentials depending on time in contrast to the static potential functions that we have constructed. The author believes that it is often the complexity seen in these systems that controls the biomaterial and living bodies. In this study, we focus on the system that controls the upright posture in the human body.

*Corresponding Author: Hiroki Takada, Department of Human & Artificial Intelligence Systems, Graduate School of Engineering, University of Fukui, 910-8507, Japan Email: takada@u-fukui.ac.jp

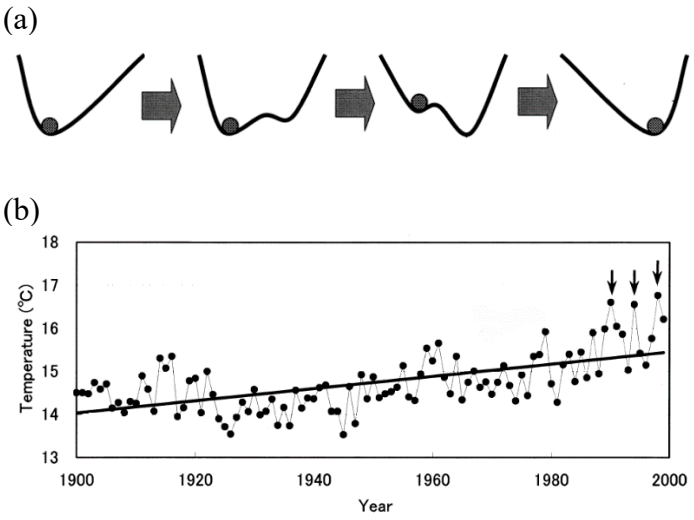


Figure 1: A description of the climate change, using the potential function (a) for year-averaged temperature at Nagoya in 20th century (b)

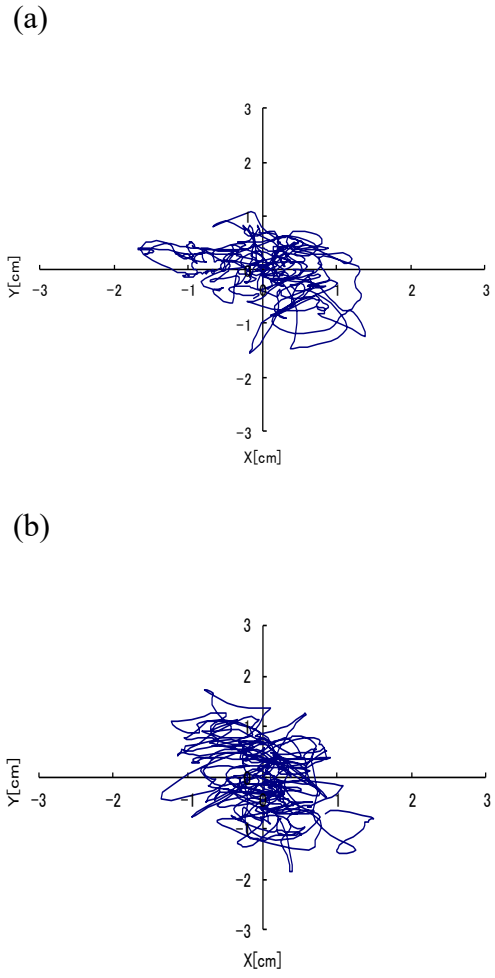


Figure 2: Typical stabilograms recorded during the resting state, with a young subject's eyes open (a) and closed (b)

The center of pressure (COP) on the Euclid space $\mathbf{E}^2$ □ (x, y) is measured as a time series that comprises the stabilograms (Figures 2 and 3). The time series of sways in stabilograms in the lateral (x) and anterior–posterior (y) directions can be assessed independently [6, 7]. The SDEs

$$\frac{\partial x}{\partial t} = -\frac{\partial}{\partial x}U_x(x) + w_x(t)$$

$$\frac{\partial y}{\partial t} = -\frac{\partial}{\partial y}U_y(y) + w_y(t)$$

are used as mathematical models to describe sways of the COP [1, 8-10]. Here, $w_x(t)$, $w_y(t)$ are white noise terms and U_x and U_y express their temporally averaged potential functions. In a previous study, the body sway was described by the Brownian motion in which U_x and U_y are expressed by parabolic functions [8-10]. In the last two decades, the limitations of this stochastic process have been realized by [1, 11, 12]. In the following section, we state a methodology to construct the temporally averaged potential functions based on the time series data.

2. Theory

In general, various stochastic processes are expressed by nonlinear SDEs for a random variable $z(t)$:

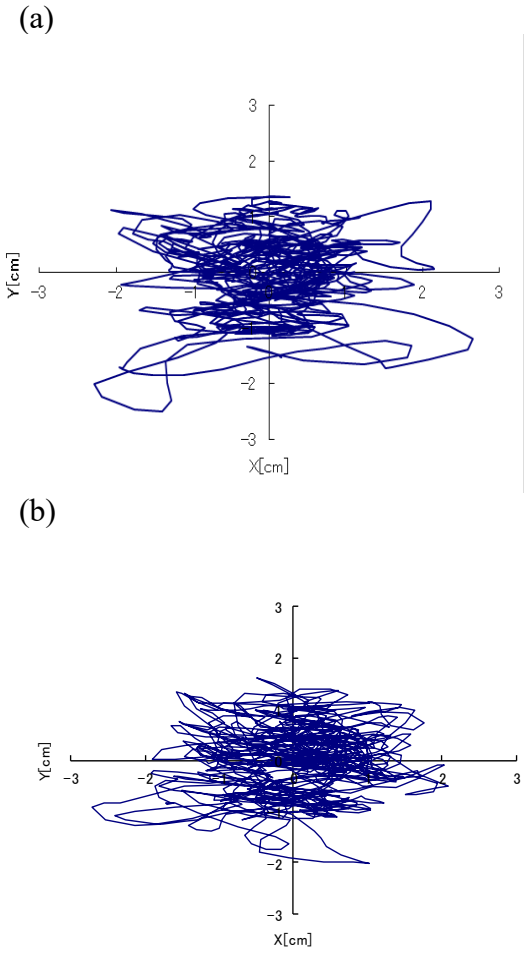


Figure 3: Typical stabilograms recorded during the galvanic vestibular stimulation, with a young subject's eyes open (a) and closed (b)

$$\frac{dz}{dt} = -f(z) + w(t) \quad (1)$$

where $w(t)$ indicates white noise and $f(z)$ is a nonlinear function. This SDE has already been applied to describe some time series data mentioned above. In particular, we suggest a method to construct an SDE as a mathematical model for these non-Gaussian time series data as follows.

Assumption 1. We assume that $z(t)$ is one of the Markov processes that is determined by $t(>t_0)$ and $z(t_0)$.

Assumption 2. We assume that $z(t)$ is not an anomalous process that extends far rapidly in a short time.

Based on these assumptions, a stochastic process can be described by a Fokker–Planck equation (FPE). The FPE is rewritten as follows for the distribution of the random variable $g(z \mid z_0, t)$:

$$\frac{\partial g(z \mid z_0, t)}{\partial t} = \frac{\partial}{\partial z}\{f(z)g(z \mid z_0, t)\} + \frac{1}{2}\frac{\partial^2}{\partial z^2} g(z \mid z_0, t) \qquad (2)$$

by a change of variables to normalize the second term of the FPE [13], called FPE normalization. Here, $g(z \mid z_0, t)$ indicates a conditional probability at time t in the initial condition of z_0. As Eq. (2) is a FPE, it does not have a coefficient function depending on the random variable $z(t)$ in the second term. This FPE uniquely corresponds to the SDE by calculating the moment of the transition probability for degree n (n = 1,2,3,…). Taking the average of both sides of Eq. (1) for any stationary interval K, one can obtain the following integro-differential equation:

$$\int_K dt\left[\frac{dz}{dt} + f(z)\right] = 0 \qquad (3)$$

In the meaning of the time average, the ordinary differential equation (ODE) $dz/dt = f(z)$ is satisfied. Thus, there is a good possibility that the mean values are controlled by the ODE. The curved surface $f(z) = 0$ is regarded as an equilibrium space for the SDE in the meaning of the temporal average (a temporally averaged equilibrium space), and the space integral of the function $f(z)$

$$U(z) = -\int_K f(z)dz \qquad (4)$$

is defined mathematically as a temporally averaged potential function. In fact, the potential function in the SDE seems to fluctuate; the potential function cannot strictly be obtained in a certain time. The authors thus considered the potential function to be a fluctuating potential.

A stationary solution $g(z)$ to the FPE (2) has been found under a natural boundary condition [14]. Under Assumptions 1 and 2, the author pointed out that it was important to analyze the forms of the histograms corresponding to the temporally averaged potential functions as

$$U(z) = -\frac{1}{2}\log\frac{g(z)}{C} \qquad (5)$$

because the stationary solution could be regarded as a density distribution through a long observation. We then regarded the stationary solution (a probability density function) as a density distribution, which is the normalized histogram of the time series data. That is, one can estimate the SDE as a mathematical model for time series data using Eq. (5). However, it is necessary to estimate a formula as a temporally averaged potential function from the standpoint of the numerical analysis. To approximate the temporally averaged potential function, we fit polynomials of degree n to the logarithmic density distribution under the following demands.

- **Demand on statistics:** The coefficients of determination R^2 for the optimal polynomial fitting to the logarithmic density distribution must be greater than 0.9.
- **Demand on geometry:** The potential function should be structurally stable taking into consideration the perturbation exerted on the control system.

3. Materials and Methods

Thirty-two healthy subjects voluntarily participated in the study; all of them were Japanese and lived in Nagoya and its environs. They were divided into two groups: a group of young people aged less than 22 years (20 ± 1 year) and a group of elderly people aged more than 65 years (70 ± 4 years). Each group included the same number of subjects. The following were the exclusion criteria for subjects: subjects working in a night shift, subjects with dependence on alcohol, subjects who consumed alcohol and caffeine-containing beverages after waking up and meals within two hours, subjects who may have had any otorhinolaryngologic or neurological disease in the past, except for conductive hearing impairment, which is commonly found in the elderly. The subjects were not prescribed drugs for any disease by doctors. The local ethics committee of the Nagoya University School of Health Science approved this study (approval number 7-129), and the subjects gave their informed consent prior to participation.

In the subsequent stabilometric analysis, we recorded the COP at rest and during GVS. We ensured that the body sway was not affected by environmental conditions; using an air conditioner, we adjusted the temperature to 25 °C in the exercise room, which was large, quiet, and bright. All subjects were tested from 10 am to 5 pm in the room of Nagoya University; they were positioned facing a wall on which a visual target was placed; the distance between the wall and subjects was 2 m.

Before the sway was recorded, the subjects stood still for 1 min in the Romberg posture with their feet together on the detection stand of a stabilometer (G5500, Anima Co., Ltd.). The COP sway was recorded (sampling frequency: 20 Hz) when the subjects stood with their eyes open (1 min) and looked at a visual target placed at a distance of 2 m or when their eyes were closed (subsequent 1 min). The stabilograms were simultaneously recorded using the stabilometer.

Every second, rectangular current impulses were output from an electronic stimulator (SEN-3301, Nihon Kohden Co., Ltd.). The duration of the current was set to 0.5 s. A small electric current (0.6–2.0 mA) was percutaneously applied to both sides of the mastoid processes through Ag/AgCl electrodes of an isolator (SS-104J, Nihon Kohden Co., Ltd.). We set the amplitude of the electric current to the maximum value obtained in the following anti-GVS test [15]; this value varied among subjects.

4. Results

Stabilometry was performed with subjects standing on stabilometer in the case of the resting state (Figures 2). Then, the

anti-GVS test was employed to the young and elderly subjects, and stabilograms were recorded during the GVS (Figures 3). In these figures, the vertical axis shows the anterior and posterior movements of the COP, and the horizontal axis shows the right and left movements of the COP. The amplitudes of the sway that was observed in the elderly subjects tended to be larger than those of the sway that was observed in the young subjects (Figures 2-3). Furthermore, the lateral movement of the COP was often observed during the GVS shown in Figures 3.

The amplitudes of the sway are affected by electric current during periodic GVS. It is considered that a periodic function $s(t)$ is added to the SDE as a forcing term, and the form of the potential function $U(z)$ changes [15]. The effective potential is expressed as

$$U_{eff}(z) = U(z) + s(t)z$$

where z is a space variable in the lateral direction x, that is, the direction of the GVS-evoked body sway. Using the sparse density [7] in the analysis of stabilograms, we investigate the evolution of the potential function $U(z)$ in this study. By comparing the stabilograms, we study the effects of aging and GVS on the sway of the COP.

5. Discussion

The nonlinear property of SDEs is important although the concept of simple muscle stiffness control during quiet standing has been accepted for a long time [16, 17]. However, the linear model [8–10] has been rejected by the experimental results of the relationship between the body posture angle and the ankle torque [18, 19], and the postural instability, which is called "microfall" [20]. The nonlinear property has also been found in the distribution of the COP [1].

A certain type of standardization is required to provide a mathematical model in accordance with the theory stated in Section 2 because the noise amplitude is set to be 1 in Eq. (1). To prevent arbitrariness in the standardization for each component, we discuss and propose a new scheme using

$$\frac{dx}{dt} = a(x) + \beta(x)w(t) \tag{6}$$

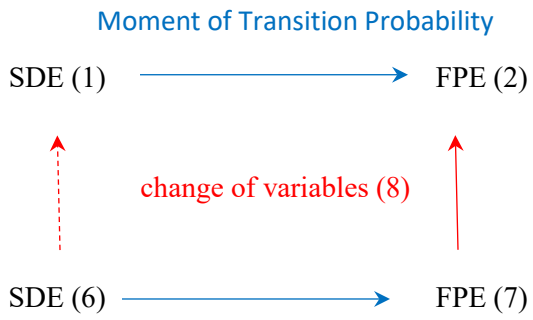


Figure 4: Correspondences between SDEs and FPEs

as a mathematical model of the time series in this section. By calculating the moment of the transition probability for degree n, this SDE corresponds to

$$\frac{\partial \psi(x|x_0,t)}{\partial t} = \frac{\partial}{\partial x}\left\{a(x)\psi(x|x_0,t)\right\} + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left\{b(x)\psi(x|x_0,t)\right\} \tag{7}$$

where $b(x) = \beta(x)^2$ $(n = 1,2,3,\ldots)$. The FPE (7) is rewritten as Eq. (2) by the change of variables $x \mapsto z$ as

$$dx = \beta dz \tag{8.1}$$

$$g = \beta\psi \tag{8.2}$$

$$f = -\frac{\alpha}{\beta} \tag{8.3}$$

$$s.t. \quad \alpha(x) = a(x) - \frac{1}{4}\frac{\partial b(x)}{\partial x}$$

for FPE normalization (Figure 4). Actually, Eq. (2) does not have a coefficient function depending on the random variable $z(t)$ in the second term. By this change of variables, the SDE (1) does not, however, correspond to Eq. (6) but corresponds to the following SDE:

$$\frac{dz}{dt} = -f + \frac{1}{2}\beta' + w(t),$$

where the prime $'$ denotes differentiation with respect to the random variable x. In the case of $\beta = const$, the SDE (1) obviously corresponds to Eq. (6), and a stationary solution to Eq. (7) can be found as

$$U(x) = -\frac{b}{2}\log\frac{\psi(x)}{C} \tag{9}$$

under a natural boundary condition. Otherwise, we should solve the following differential equation:

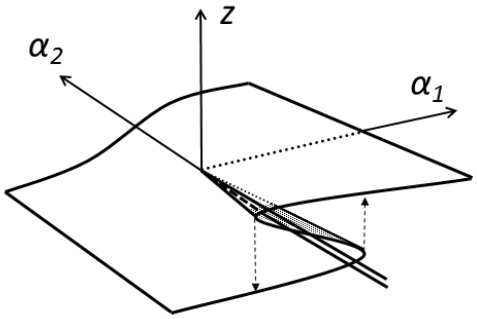


Figure 5: Riemann–Hugoniot manifold $f(z) = \alpha_1 - \alpha_2 z - z^3 = 0$. The potential function $U_z(z) = \frac{1}{4}z^4 + \frac{1}{2}\alpha_2 z^2 - \alpha_1 z$ is derived from Eq. (4). The delay convention dictates that catastrophe occurs when the local minimum of the potential disappears completely and only one global minimum exists (Figure 1a). In practical terms, according to the delay convention, a catastrophe occurs when the system reaches the edge of the bifurcation set (fold) after crossing the bisector.

$$bf' = 2\left[a - \frac{1}{2}b'\right]f$$

To construct an optimal description (9) for each generator of time series, we fit graphs of polynomials to the logarithmic histogram of the time series, $\log\psi$. For example, we herein consider the Riemann–Hugoniot manifold as an equilibrium space $f(z) = 0$ (Figure 5). By using Eq. (4), the double-well potential

can be extracted from the half line in Figure 5. In general, a polynomial of degree four

$$U(z) = \alpha_1 z + \alpha_2 z^2 + \alpha_3 z^3 + \alpha_4 z^4 \qquad (10)$$

corresponds to a point $(\alpha_1 \quad \alpha_2 \quad \alpha_3 \quad \alpha_4) \in U^{\times}$, which is called a dual space of a vector space spanned by $(z^1 \quad z^2 \quad z^3 \quad z^4)$.

A coefficient vector $(\alpha_1 \quad \alpha_2 \quad \alpha_3 \quad \alpha_4)$ might depend on time t although it was considered as a constant under Assumption 1 in this paper. Temporal variations in the coefficient vector can describe the variations in temporally averaged potential functions. There is a good possibility that the temporal variations in the coefficient vector extract the effects of not only the breaches of stationary states, but also unobservable state values. Moreover, we may find the number of the unobservable state values for a motion process by using the dimensional analysis for a linear space $U^{\times} \square (\alpha_1 \quad \alpha_2 \quad \alpha_3 \quad \alpha_4)$ on a field $\mathbf{R}$.

Let the basis of a vector space U and its dual space $U^{\times}$ be $\{e_i\}$ and $\{e^{*j}\}$, respectively. The linear space $U^{\times}$ is a dual space of the vector space U under the dual basis $\{e^{*j}\}$ as $e^{*j}(e_i) = \delta_i^j$, where δ_i^j expresses the Kronecker delta [21]. The motion process in the dual space help us find a temporal relation of the coefficients. Eliminating the temporal variable t, we can look for a mathematical expression of the unobservable state values as an algebraic curve, which is a four-dimensional curve, in the dual space. It is important to study the four-dimensional curve mathematically. If $\alpha_i(t)$ generated random numbers, the construction method could be applied to sequences $\{\alpha_i(t)\}_{t \in K}$ again. We can propose novel potentials in the mathematical models for coefficients of the temporally averaged potential functions or the unobservable state values. In this paper, these novel potential functions are called grand potentials. One can apply the grand potentials to time series analyses for the non-Gaussian and non-stationary processes or forecasts for those motion processes. We may classify strategy for changes of those motion processes on the basis of the grand potentials.

The following translation is useful to simplify the expression of Eq. (10) because the third term on the right-hand side in Eq. (10) goes to zero:

$$z \mapsto z - \frac{a_3}{4a_4}$$

The dual space is also degenerated as $(A_1 \quad A_2 \quad 0 \quad \alpha_4)$ and can be visualized as a three-dimensional space (Figure 6). The catastrophe map χ

$$\chi : (\alpha^1 \quad \alpha^2 \quad \alpha^3 \quad \alpha^4) \mapsto (\alpha^1 \quad \alpha^2)$$

is considered useful for the visualization of the motion process in the dual space. Using the boundary in the dual space (Figure 6), we examine the nonlinearity of the potential during the GVS and the other loads in the next step.

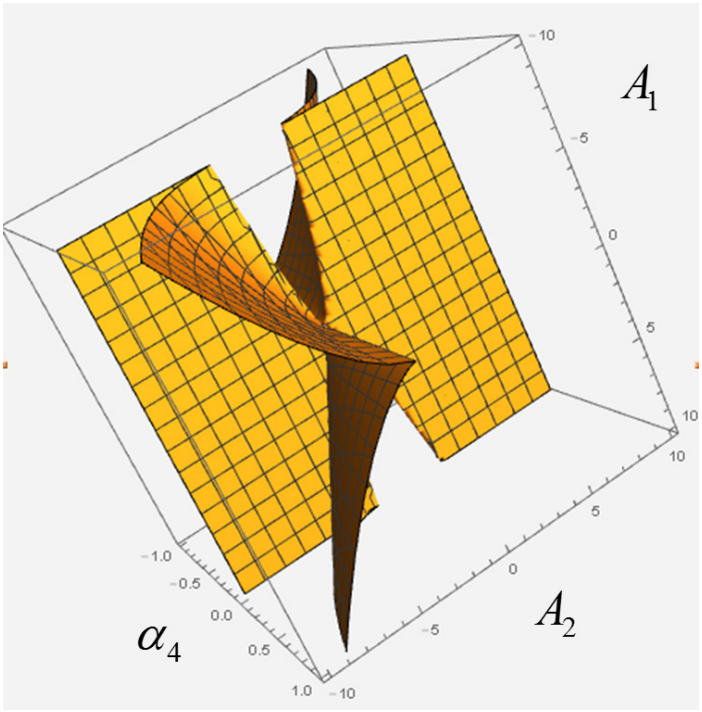


Figure 6: Boundary of the nonlinear potential that has multiple minimal points

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was supported in part by the Japan Society for the Promotion of Science, Grant-in-Aid for Scientific Research (C) Number 26350004.

References

[1] H. Takada, Y. Shimizu, Y. Kitaoka, "Mathematical index and model in Stabirometly," Forma, **16** (1), 17-46, 2001.

[2] Y. Shimizu, H. Takada, "Verification of air temperature variation with form of potential," Forma, **16** (4), 339-356, 2001.

[3] H. Takada,"Phase Transition with noise for time series of JP¥/US$ exchange rate," Forma, **31**(Special Issue), S41–S53, 2016.

[4] IPCC, Greenhouse Gas Inventory, Reporting Instructions, Inter -governmental Panel on Climate Change, 1995.

[5] T. Poston, I. N. Stewart, Catastrophe Theory and its Application, Pitman publishing limited, 1978.

[6] P. A. Goldie, T. M. Bach, O. M. Evans, "Force platform measures for evaluating postural control: Reliability and validity," Arch. Phys. Med. Rehabil., **70**, 510–517, 1986.

[7] H. Takada, "A Construction Method of the Mathematical Model on Time Series Data with Markov Property and Verification," Ph.D. Thesis, Meijyo Univ., 2004. (in Japanese)

[8] J. J. Collins, C. J. D. Luca, "Open loop and closed-loop control of posture: A random-walk analysis of center of pressure trajectories," Exp. Brain Res., **95**, 308–318, 1993.

[9] R. E. A. Emmerrik, R. L. V. Sprague, K. M. Newell, "Assessment of sway dynamics in tardive dyskinesia and developmental disability," Sway Profile Orientation and Stereotypy. Moving Dis., **8**, 305-314, 1993.

[10] K. M. Newell, S. M. Slobounov, E. S. Slobounova, P. C. Molenaar, "Stochastic processes in postural center of pressure profiles," Exp. Brain Res., **113**, 158–164, 1997.

[11] H. Takada, M. Miyao, "Visual fatigue and motion sickness induced by 3D video," Forma , **27** (Special Issue), S67-S76, 2012.

[12] Y. Asai, Y. Tasaka, K. Nomura, T. Nomura, M. Casadio, and P. Morasso, "A model of postural control in quiet standing: Robust compensation of delay-induced instability using intermittent acti-vation of feedback control," PLoS ONE, **4** (7), art. no.e6169, 2009.

[13] N. S. Goel, and N. Richter-Dyn, Theory of Stochastic Process in Biology, Industrial books, 33-75, 282-287, 1978.

[14] H. Harken, "Cooperative phenomena in systems far from thermal equilibrium and in nonphysical systems," Rev. Mod. Phys., **47**, 67-121, 1975.

[15] H. Takada, M. Takada, K. Tanaka, T. Shiozawa, M. Furuta, M. Miyao, "A simulated study of the deterioration in the equilibrium function with advancing age," Bulletin of Gifu University of Medical Science **3**, 109-117, 2009.

[16] D. A. Winter, A. E. Patla, F. Prince, M. Ishac, K. Gielo-Perczak, "Stiffness control of balance in quiet standing," J Neurophysio, **80**, 1211-1221, 1998.

[17] D. A. Winter, A. E. Patla, S. Rietdyk, M. G. Ishac, "Ankle muscle stiffness in the control of balance during quiet standing," J Neurophysio, **85**, 2630-2633, 2001.

[18] I. D. Loram, M. Lakie, "Human balancing of an inverted pendulum- position control by small, ballistic-like, throw and catch movments," J Physioligy, **540**, 1111-1124, 2002.

[19] I. D. Loram, C. N. Maganaris, M. Lakie, "Human postural sway results from frequent, ballistic bias impulses by soleus and gastrocnemius," J Physioligy, **564**(1), 295-311, 2005.

[20] A. Bottaro, Y. Yasutake, T. Nomura T., M. Casadio, P. Morasso, "Bounded stability of the quiet standing posture: an intermittent control model," Human Moving Science , **27**(3), 473-495, 2008.

[21] M. Wadachi, Differential Geometry and Topology, Iwanami, 1996. (in Japanese)

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 192-196 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

MIMO Performance and Decoupling Network: Analysis of Uniform Rectangular array Using Correlated-Based Stochastic Models

Obour Agyekum Kwame O-B*, Maxwell Oppong Afriyie[1], Paul Oswald Kwasi Anane[1], Affum Emmanuel Ampoma[1], Matthew Seddoh Akatey[2]

[1]*University of Electronic Science and Technology of China, School of Communication and Information Engineering, 611731, China*

[2]*University of Electronic Science and Technology of China, School of Energy Science, 611731, China*

ARTICLE INFO

Article history:
Received: 18 December, 2016
Accepted: 21 January, 2017
Online: 28 January, 2017

Keywords:
Azimuthal Spread (AS)
Decoupling Network
Elevation Spread (ES)
Mutual coupling (MC)
Spatial Correlation Function
Uniform Rectangular Array (URA)

ABSTRACT

We explore the dependency of MIMO performance on azimuthal spread (AS) and elevation spread (ES) using correlated-based stochastic models (CBSMs). We represent the transmitter as uniform rectangular array (URA), and derive an analytical function for spatial correlation, in terms of maximum power when phase gradient of the incident wave follows a Student's t-distribution. We model the correlated-based stochastic MIMO system to investigate the usefulness of the analytical function, under the condition that the magnitude and phase of mutual coupling and consequently the coupling matrix of the coupled receiving monopole array differs much from that of the decoupled array. Verification is achieved with the help of simulation results, which support the existing fact that the azimuthal spread is the fundamental determinant of system capacity. However, we observed that lower values of AS and ES increase the rate of deterioration in the symbol error rate (SER) and not antenna spacing after decoupling process.

1. Introduction

The correlation-based stochastic models (CBSMs) are mainly used for theoretical analysis to assess the performance of massive MIMO systems. However, these models could be applied to cases where the user equipment (UE) with multiple antennas works at millimeter wave [1]. Although the fifth generation (5G) wireless communication technology will offer alluring features such as higher transmission rate, there are still several challenges ahead in the deployment of this technology [2-3]. Very recently, antenna arrays have attracted attention in the wireless communication industry to enhance signal quality, coverage and capacity. This is because antenna configuration directly influences the channel characteristics and performance, therefore, it is a critical factor in deciding the channel characteristics in MIMO technology.

In view of this, all attempts to improve the performance of the existing MIMO technology must be investigated especially in regards to realistic channel models.

All endeavours to improve and enhance the performance of the existing MIMO technology need to be explored, specifically with regards to analyzing the MIMO characteristics of wireless channels.

*Corresponding Author: Obour Agyekum Kwame O-B, University of Electronic Science and Technology of China, 611731, China
Email: obour539@yahoo.com

The authors in [4] demonstrated that having the MIMO system in three dimensional view is the efficient method for managing spatial resource. Several array configurations have been the subject of investigations on MIMO technology, and most of the prior investigations considered simple system configurations using only AS [5–9].

Many refer to the work in [8] and [9] where researchers presented spatial correlation function for four-element receiving circular arrays, when the angle-of-arrival (AoA) followed Laplacian and truncated Gaussian distributions respectively, because the uniform circular array (UCA) configuration has been the subject of many investigations on MIMO technology. Although the authors derived the correlation function for UCA in terms of AS by considering that the ES of the incoming plane wave in the propagation geometry is 90°, their contribution provides fundamental basis for evaluating the spatial correlation function of URA in advanced MIMO technologies. Therefore, our objective in this paper is to experiment the CBSMs at 2.4 GHz. We modelled the transmitter as a uniform rectangular array, and examined the effects of spatial fading correlation and mutual coupling on channel capacity and symbol error rate (SER). To prove our concept, we quantify the mutual coupling of the receiving array by the receiving mutual impedance method (RMIM) [10], whereas, we examined the URA in the CST

Microwave Studio to determine the coupling matrix. Recently, authors in [11] proposed a spatial correlation function for MIMO channel model on the latest 3GPP standards and WINNER+ [12]. We are of the view that our concept will add to cover new aspects in MIMO modelling studies.

For purposes of clarity, the contributions in this paper are as follows: 1) We fabricate a decoupling network to study the dependence of system performance on antenna spacing, AS and ES of MIMO system, using CBSMs on the assumption that the magnitude and phase of the mutual coupling and consequently the coupling matrix of the coupled array varies from that of the decoupled array at the receiving end. 2) We derive an analytical expression for spatial correlation function of a URA, based on maximum power in the direction of maximum arrival using the Student's t-distribution. 3) Our outcomes show that azimuthal spread (AS) is the essential determinant of system performance. Also, lower values of AS and ES increase the rate of deterioration in the SER, and not antenna spacing.

This letter is organized as follows: Section II presents the design of the decoupling network. In Section III, we derive the closed-form expression for the spatial correlation function in terms of maximum power for URA using Student's t-distribution. Section IV presents analytical and simulation results. Finally, we give concluding remarks in Section V.

2. Operating Matrix of the Decoupling Network

The compensation network is designed using a power-divider and two rat-race couplers. The power divider has unequal power-dividing ratio with no active circuit elements to minimize extra circuit noise. The operating matrix for two-element receiving array is given by [13].

$$\begin{bmatrix} U_1 \\ U_2 \end{bmatrix} = \begin{bmatrix} 1 & -\frac{Z_{12}}{Z_L} \\ -\frac{Z_{21}}{Z_L} & 1 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} V_1 - \frac{Z_{12}}{Z_L} V_2 \\ V_2 - \frac{Z_{21}}{Z_L} V_1 \end{bmatrix} \quad (1)$$

where the inputs to network are coupled voltages V_1 and V_2 from the two monopole antennas, and the outputs of network are compensation voltages U_1 and U_2. The mutual impedance between the two monopole antennas is represented by Z_{12} and Z_L is the matched load impedance. In order to achieve the tri-band operation, we utilized resonators and stubs.

By sharing the stub with the resonator (parallel LC circuit) of two adjacent π -shaped structure circuits within the rat-race couplers, a decoupling network with working frequencies 2.4/3.5/5.5 GHz was designed. The length and impedance of stubs at different frequencies are calculated using the procedure in [14]. The decoupling network has been designed using ADS and fabricated using FR4 substrate with dielectric constant 4.8 as shown in Fig. 1. Simulated and measured responses are shown in Fig. 2. Measured results at 2.4/3.5/5.5 GHz demonstrate that isolation and insertion losses between input and output ports of more than 10 dB can be accomplished. To achieve best performance, the initial values of the impedance and the parallel resonator had to be increased to accomplish the tri-band operation. This accounted for the large differences between simulated and measured values in Fig. 2(a) and Fig. 2(b) at 3.5 GHz.

3. System Model

3.1. Proposed Spatial Correlation of URA

The spatial correlation between antennas at positions (n, p) and (m, q) following the procedure in [14] is expressed as

$$\rho[(n,p),(m,q)] = E[a_{np}(\varphi,\theta)a^*_{mq}(\varphi,\theta)] \quad (2)$$

$$\rho = \int_{\varphi}\int_{\theta} \left\{ e^{jZ\sin((\theta)\sin(\gamma+\varphi))}\, \mathrm{p}(\theta,\varphi)\sin(\theta) \right\} d\theta d\varphi \quad (3)$$

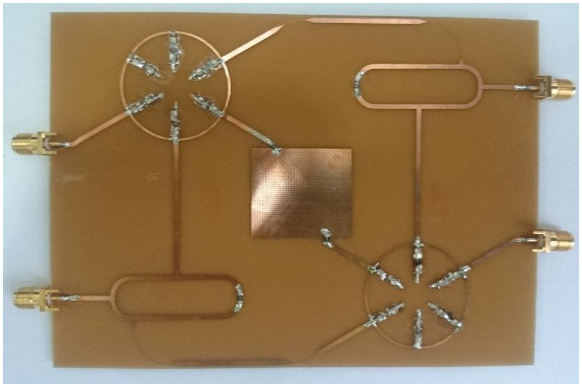


Fig. 1. Photograph of the fabricated compensation network

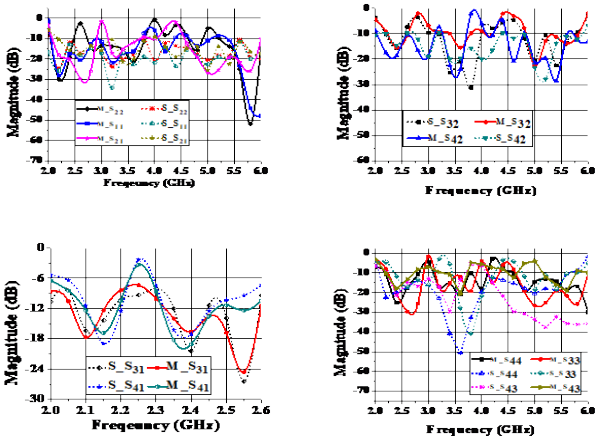


Fig. 2. Simulated and measured scattering parameters of the decoupling network for (a) input ports, (b), (c) input and outputs ports, and (d) output ports.

where $p(\theta,\varphi)$ is the joint spatial power spectrum. For our theoretical analysis, we have used the fact that the maximum power in the direction of maximum arrival is proportional to $p(\theta,\varphi)$. Therefore, we represent the joint spatial power spectrum $p(\theta,\varphi)$ in (3) by the maximum power $P_t(\alpha)$ of the incident plane wave and proportional to AS and ES. If the phase gradient has a Student's t-distribution as in [15], compared to the choice of joint spatial power spectrum of [14], then the phase in (3) can be expressed as $\beta = Zk\sin(\alpha)$, where $\alpha = \theta\sin(\gamma+\varphi)$. Following the methodology in [15], the probability density function of α according to Student's t-distribution is given by

$$p(\alpha)=(1/2)\left(s^2\cos(\alpha)\Big/\left(s^2+\left(\sin\alpha-\left(\overline{\sin(\alpha)}\right)^2\right)\right)^{3/2}\right) \quad (4)$$

where s is the measure of angular spread. In our analysis we assumed that $\left(\sin\alpha-\left(\overline{\sin\alpha}\right)\right)^2 \approx \sin^2\alpha$. The maximum power $P_t(\alpha)$ in the direction of maximum arrival is $P_t(\alpha)=\int p(\alpha)d\alpha$. With the above assumptions, we can rewrite (3) as

$$\rho=P_t(\alpha)\int_{\varphi}\int_{\theta}\left\{e^{jZ\sin(\theta)\sin(\psi)]}\sin(\theta)\right\}d\theta d\psi \quad (5)$$

$$\rho=(4\pi/Z)P_t(\alpha)\sin(Z) \quad (6)$$

A. Theoretical and S-parameter based mutual coupling

Incorporating the effect of mutual coupling, the channel vector can be written as follows

$$H=Z_m\rho^{1/2}H_o \quad (7)$$

Z_m and ρ denote the coupling and spatial correlation matrices at the receiver and transmitter respectively. The theoretical approximation of coupling matrix is expressed as

$$Z_M=(Z_A+Z_L)(Z_{12}+Z_LI)^{-1} \quad (8)$$

where Z_A, Z_L, Z_{12} are the antenna, load and mutual impedances respectively. With the above assumption that the magnitude and phase of the mutual coupling are different, we used RMIM measurement procedure shown in Fig. 3 to evaluate the mutual coupling between small receiving monopole arrays at the receiving end under two conditions [10] describe in Section IV. The mutual impedance Z_{12} is expressed as [10]

$$Z_{12}=\left(S_{11}-S_{21}^{(1)}/S_{21}^{(2)}\right)Z_L \quad (9)$$

where S_{11} is the measured scattering parameter at monopole 1's terminal with monopole 2 removed (taken away from the array), $S_{21}^{(1)}$ is the measured scattering parameter at monopole 1's terminal with monopole 2's terminal connected to Z_L and $S_{21}^{(2)}$ defines the measured scattering parameter at monopole 2's terminal with monopole 1's terminal connected to Z_L Moreover, in terms of scattering parameters [16] coupling matrix in (7) is given as

$$Z_M=(I+S_t)(I-S_t)^{-1}Z_L \quad (10)$$

where matched load impedance Z_L has a value of 50Ω and $S_t \in \mathsf{M}^{N\times N}$ is the S-parameter matrix of the antenna array. To determine the S-parameters at the transmitting end, the URA is modelled in CST Microwave Studio as a 4×4 monopole array, located in the $x-y$ plane.

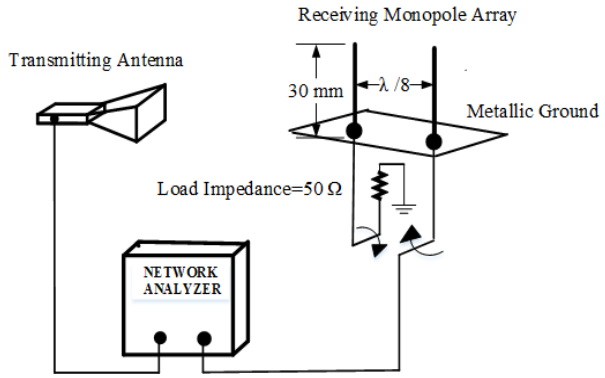


Fig. 3. Measuremwnt of receiving mutual impedance in anechoic chamber at 2.4 GHz

4. Numerical Results

In this section, we present the analytical performance and the simulation results of MIMO system using CBSMs. We modelled and analyzed the correlated-based stochastic MIMO channel, based on the two conditions that the magnitude and the phase of the mutual coupling and the coupling matrix of the coupled receiving array vary much from that of the decoupled array. The simulated channel is according to (6)-(10) and it is assumed that fading is correlated at both transmitter and receiver sides. The receiving array is represented by two-monopole element on a metallic ground and operating at 2.4 GHz. The monopoles have length of 30 mm, radius of 0.5 mm and element separation of $\lambda/8$. As indicated already, the formulation of coupling matrix of the receiving array is determined in anechoic chamber under two conditions using RMIM.

Condition I: the S-parameters of the coupled array are determined to estimate for the coupled voltages (V_1 and V_2), and also to formulate the coupling matrix using (9)-(10).

Condition II: the monopole array is connected to the decoupling network through equivalent length coaxial links. The scattering parameters of the output ports of the decoupling network are determined to formulate the coupling matrix and estimate for the compensated voltages (U_1 and U_2).

With the end goal of justifying the assumption that the magnitude and phase of mutual coupling of coupled array differs much from the decoupled array at the receiving end, there are three different types of voltages recorded in Table I. The last row in Table I is a ratio of the voltage obtained with monopole B to the voltage obtained with monopole A. It can be seen that the ratio of the compensated voltage is very close to the uncoupled voltages. This also demonstrates that the compensated voltages have successfully been removed off the coupling effect. The value of γ is determined for each antenna spacing of the URA and incorporated into the channel model.

4.1. Effects of AS and ES- Performance Tradeoff

We explore the dependency of narrowband MIMO channel capacity using uniform power allocation on AS and ES at different antenna spacing using CBSMs. Our results according to the

modelling of coupling capacity matrix at the receiving side in Fig. 4 and Fig. 5 support the existing fact of the effects of higher values of azimuthal spread (AS) on system performance. Our outcomes demonstrate that elevation spread (ES) does not affect system capacity as much as the azimuthal spread (AS) does. Moreover, it is therefore not surprising that spacing between antenna elements affects system performance. However, larger number of transmit antennas improve system performance in a fixed physical space.

4.2. Symbol error rate (SER) vs SNR

In Fig. 6, there is a close match between symbol error rate (SER) for $AS = ES = 30^{\circ}$ at 0.2λ and 0.5λ respectively. We observed that the SER after an increase of 10dB for $AS = 45^{\circ}$, $ES = 20^{\circ}$ at 0.5λ remains nearly the same as that for $AS = 20^{\circ}$, $ES = 45^{\circ}$ at 0.2λ. A close inspection of Fig. 6 reveals that the impacts of higher values of AS and antenna spacing on the channel capacity are more significant than SER. We note that lower values of AS and ES increase rate of deterioration in the SER and not antenna spacing after decoupling process.

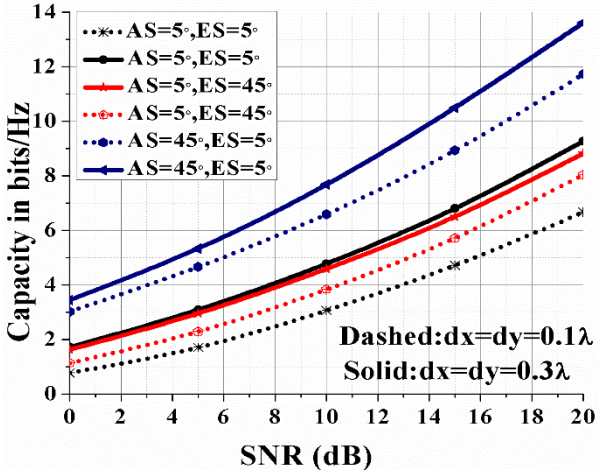


Fig. 4. Capacity vs SNR for antenna spacing 0.1λ and 0.3λ

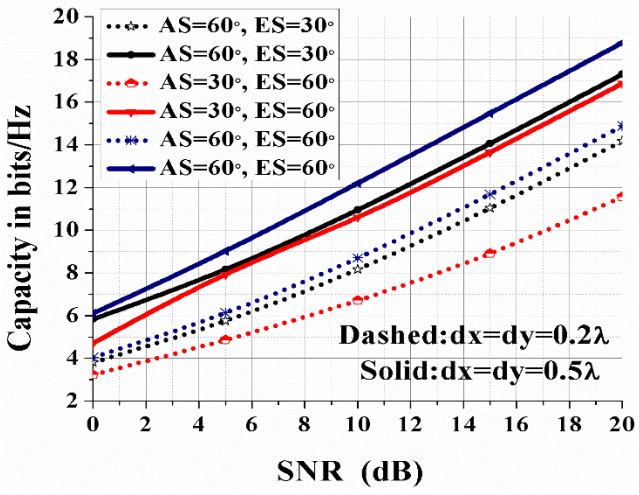


Fig. 5. Capacity vs SNR for antenna spacing 0.2λ and 0.5λ

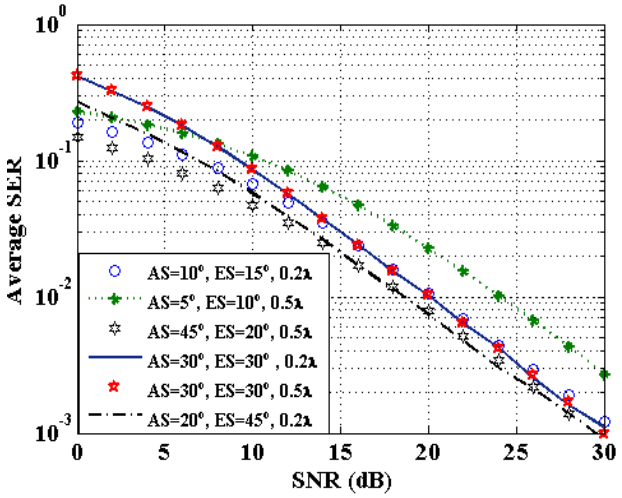


Fig. 6. SER vs SNR for QPSK with 4×4 URA

Table 1: Different Measured Voltages

		Uncoupled Voltages (reference)	Coupled voltages	Compensated voltages
Monopole A	mag (mV)	16.64	12.4	11.55
	angle (°)	-160.64	-166.67	34.967
Monopole B	mag (mV)	16.54	15.42	12.30
	angle (°)	-139.56	-141.46	55.16
B/A	mag (mV)	0.9939	1.2199	1.065
	angle (°)	21.08	25.208	20.193

5. Conclusion

We derived an expression for the spatial fading correlation of URA when the phase gradient follows Student's t-distribution. We investigated dependency of MIMO performance on azimuthal spread (AS) and elevation spread (ES) after decoupling process using CBSMs. Our results support the existing fact that the azimuthal spread (AS) is the essential determinant of system performance. Additionally, results confirm that lower estimations of AS and ES significantly expand the rate of deterioration in the SER, and not antenna element separation after decoupling process.

Appendix

The integral in (5) is evaluated in this appendix. Using the properties of Bessel function in the case of $n = 0$ as [17]

$$J_n(x) = \frac{1}{2}\int_{-\pi}^{\pi} e^{-j(n\tau - x\sin(\tau))} d\tau \overset{n=0}{=} \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{jx\sin(\tau)} d\tau \overset{2\pi}{=} \frac{1}{2\pi}\int_{0}^{2\pi} e^{jx\cos(u)} du \quad (11)$$

we can rewrite (5) as

$$\rho = P_t(\alpha)\left\{\int_0^{\pi}\sin(\theta)\left\{\int_0^{2\pi} e^{jx\cos(\psi)}d\psi\right\}d\theta\right\} \tag{12}$$

where $x = Z\sin(\theta)$. Using specific solutions in (11) such as

$$\int_{\Theta=0}^{\pi} J_o(x\sin(\Theta))\sin(\Theta)d\Theta \overset{x\geq 0}{=} \sqrt{(2\pi/x)}J_{\frac{1}{2}}(x) \tag{13}$$

and $J_{\frac{1}{2}}(x) \overset{x\geq 0}{=} \sqrt{(2/\pi x)}\sin(x)$, we obtain (6). To evaluate $P_t(\alpha) = \int p(\alpha)d\alpha$, we use the series [18]

$$\int(\cos\alpha)\sqrt{(1+\eta^2\sin^2\alpha)}d\alpha = (\sin\alpha/2)\sqrt{(1+\eta^2\sin^2\alpha)}$$
$$+(1/2\eta)\,\ell n\left\{\eta\sin\alpha+\sqrt{(1+\eta^2\sin^2\alpha)}\right\}$$

Using the series $\ln(\alpha) = \sum_{n=1}^{\infty}\frac{\left((\alpha-1)/(\alpha+1)\right)^{2n+1}}{2n+1}$ and

$$Sin^{2n}\alpha = \frac{1}{2^{2n}}\left\{\sum_{k=0}^{n-1}(-1)^{n-k}\binom{2n}{k}\cos(n-k)\alpha+\binom{2n}{n}\right\} \tag{14}$$

with well-known Taylor series of $\cos\alpha$ and $\sin\alpha$, (6) converges rapidly for small values of Z .

References

[1] Kan Zheng, Suling Ou and Xuefeng Yin "Massive MIMO Channel Models: A Survey *"International Journal of Antennas and Propagation",* vol. 2014, Article ID: 848071, Jun. 2014.

[2] Asvin Gohil, Hardik Modi and Shobhit K Patel "5G Technology of Mobile Communication: A Survey, IEEE Conference on Intelligent Systems and Signal Processing (ISSP), pp. 288 – 292, March 2013.

[3] Kan Zheng, Suling Ou and Xuefeng Yin "Massive MIMO Channel Models: A Survey", International Journal of Antennas and Propagation, vol. 2014, Article ID 848071, June 2014.

[4] Jianhua Zhang, Chun Pan, Feng Pei, Guangyi Liu, and Xiang Cheng "Three-Dimensional Fading Channel Models: A Survey of Elevation Angle Research," IEEE Communications Magazine, June 2014, pp. 218-226

[5] J. Zhou, S. Sasaki, S. Muramatsu, H. Kikuchi, and Y. Onozato, "Spatial correlation for a circular antenna array and its applications in wireless communication," in IEEE Global Telecommunications Conference, GLOBECOM '03., vol. 2, Dec 2003, pp. 1108–1113.

[6] A. Forenza, D. J. Love, and R. W. Heath, "Simplified spatial correlation models for clustered MIMO channels with different array configurations," IEEE Transactions on Vehicular Technology, vol. 56, no. 4, July. 2007, pp.1924–1934.

[7] J.-A. Tsai and B. Woerner, "The fading correlation function of a circular antenna array in mobile radio environment," in IEEE Global Telecommunications Conference, vol. 5, May 2001, pp. 3232–3236.

[8] Jiann-An Tsai, R. Michael Buehrer, and Brian D. Woerner "BER Performance of a Uniform Circular Array versus a Uniform Linear Array in a Mobile Radio Environment", IEEE Transactions on Wireless Communications, vol. 3, No. 3, May 2004, Page: 732.

[9] Jiann-An Tsai, R. Michael Buehrer "Spatial Fading Correlation Function of Circular Antenna Arrays With Laplacian Energy Distribution", IEEE Communications Letters, vol. 8, No. 5, May 2002, Page: 295

[10] H. T. Hui, "A new definition of mutual impedance for application in dipole receiving antenna arrays," IEEE Antennas Wireless Propag. Lett. vol. 3, pp. 364–367, 2004.

[11] Qurrat-Ul-Ain Nadeem, Abla Kammoun, M'erouane Debbah†, and Mohamed-Slim Alouini "Spatial Correlation Characterization of a Uniform Circular Array in 3D MIMO Systems," International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), July 2016, pp: 1 - 6

[12] 3GPP TR 36.873 V12 .0.0, "Study on 3D channel model for LTE," Sep.2014

[13] Y. Yu and H.T. Hui "Design of a Mutual Coupling Compensation Network for a Small Receiving Monopole Array", IEEE Trans. On Micro. Theory and Techniques, vol. 59, no. 9, pp. 2241-2245, Sept. 2011.

[14] S. K. Yong and J. S. Thompson, "A three-dimensional spatial fading correlation model for uniform rectangular arrays", IEEE Antennas and Wireless Propag. Lett, vol. 2, no. 1, pp. 182–185, 2003.

[15] J. B. Andersen and K. I. Pedersen, "Angle-of-arrival statistics for low resolution antenna", IEEE Trans. Antennas Propagat. vol. 50, No. 3, pp.391-395, Mar. 2002.

[16] D. Masouros, J. Chen, K. Tong, M. Sellathurai, and T. Ratnarajah, "Towards massive-MIMO transmitters: on the effects of deploying increasing antennas in fixed physical space," in Proc. of the Future Network and Mobile Summit, pp. 1–10, 2013.

[17] Abramowitz Milton, Stegun Irene "Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables", New York: Dover, ISBN 0-486-61272-4.

[18] Herbert Bristol Dwight, "Tables of Integrals and other Mathematical Data", 3rd ed, Macmillan Company, New York, 1957.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 197-205 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

Performance of Surge Arrester Installation to Enhance Protection

Mbunwe Muncho Josephine[1,*], Gbasouzor Austin Ikechukwu[2]

[1]Department of Electrical Engineering, University of Nigeria, Nsukka, 410101, Nigeria, muncho.mbunwe@unn.edu.ng

[2]Department of Mechanical Engineering, Chukwuemeka Odumegwu Ojukwu University, Uli, 431121, Nigeria, ai.gbasouzor@coou.edu.ng

ARTICLE INFO

Article history:
Received: 22 December, 2016
Accepted: 20 January, 2017
Online: 28 January, 2017

Keywords:
Surge
Arrester
Protection
Reliability
Technology

ABSTRACT

The effects of abnormal voltages on power system equipment and appliances in the home have raise concern as most of the equipments are very expensive. Each piece of electrical equipment in an electrical system needs to be protected from surges. To prevent damage to electrical equipment, surge protection considerations are paramount to a well designed electrical system. Lightning discharges are able to damage electric and electronic devices that usually have a low protection level and these are influenced by current or voltage pulses with a relatively low energy, which are induced by lightning currents. This calls for proper designed and configuration of surge arresters for protection on the particular appliances. A more efficient non-linear surge arrester, metal oxide varistor (MOV), should be introduced to handle these surges. This paper shows the selection of arresters laying more emphasis on the arresters for residential areas. In addition, application and installation of the arrester will be determined by the selected arrester. This paper selects the lowest rated surge arrester as it provides insulation when the system is under stress. It also selected station class and distribution class of arresters as they act as an open circuit under normal system operation and to bring the system back to its normal operation mode as the transient voltage is suppressed. Thus, reduces the risk of damage, which the protection measures can be characterized, by the reduction value of the economic loss to an acceptable level.

1. Introduction

A lightning protector also referred to as a lightning arrestor is a device used on electrical power systems and telecommunications systems to protect the insulation and conductors of the system from the damaging effects of lightning. The typical lightning protector, depending on the sizes, has a high voltage terminal and a ground terminal. When a lightning surge travels along the power line to the protector, the current from the surge is diverted through the arrestor, in most cases to earth [1].

In telegraphy and telephony, a lightning protector is placed where wires enter a structure, preventing damage to electronic instruments within and ensuring the safety of individuals nearby. Smaller versions of lightning protectors, also called surge protectors, are devices that are connected between each electrical conductor in power and communications systems and the earth [2]. These prevent the flow of the normal power or signal current to ground, but provide a path over which high voltage lightning current flows, bypassing the connected equipment. The purpose is to limit the rise voltage when a communications or power line is struck by lightning or being close to a lightning strike. If protection fails or is absent, lightning that strikes the electrical system introduces thousands of kilovolts that may damage the transmission lines, and can also cause damage to transformers and other electrical and electronic devices including home appliances [2-4].

Also exceeding the capability of an arrester will crack or puncture the metal-oxide disk(s), in effect reduce the arrester internal electrical resistance, thus, the condition will limit the arrester's ability to survive future system conditions. This

*Corresponding Author: Mbunwe Muncho Josephine, Department of Electrical Engineering, University of Nigeria Nsukka, +2348036675952;
E-mail: muncho.mbunwe@unn.edu.ng

condition does not jeopardize the insulation protection provided by the arrester, but in an unlikely case of complete failure of an arrester, a line-ground arc will develop and pressure will build up inside the housing of the arrester. This pressure will be safely vented to the outside and an external arc will be established provided the fault current is within the pressure relief fault current capability of the arrester. This low-voltage arc maintains equipment protection. Once an arrester has safely vented, it no longer possesses its pressure relief/fault current capability and should be replaced immediately. For a given application, the arrester selected should have a pressure/fault current capability greater than maximum short-circuit current available at the intended arrester location. This rating of arrester capability should include appropriate allowances for future growth in the system [3-7].

Surge arresters when installed are exposed to many physical factors on the earth. These factors that such arresters are exposed to when they are in operation are expected to affect them in one way or the other. These factors are temperature, spacing of the arresters and earth resistance. The villages around the Donga Matung Division of North West Province of Cameroon have the highest annual average number of lightning storms in the Cameroon. On average, between 40 and 90 thunderstorms hit the areas along the Binshua each year, while Nkambe and some villages along the Donga border have an average of less than 10 per year. So, it is easy to understand that surge protection of electrical equipment is a very important part of the electrical system design. Lightning strikes are not the only sources of voltage surges in the electrical system. The following are a few of the more frequently encountered causes of transient voltage surges:

- Surge voltages associated with switching capacitors;
- Surge voltages due to a failure in equipment insulation resulting in a short circuit on the distribution system;
- Surge voltages associated with the discharge of lightning arresters at other locations within the facility.

When capacitors are switched in and out of the circuit, it is possible to get a re-strike when interrupting the capacitor circuit current. A steep-front voltage excursion may be created from each re-strike. These voltage excursions may be high enough to damage rotating machines applied at the same voltage. A surge capacitor applied at the motor terminals can change the steepness of the wave front enough to protect the motor. A short circuit can cause a voltage surge in excess of 3 times the normal line to neutral crest value. The magnitude and steepness of the wave front is not as severe as that of a lightning strike, but can cause damage or weaken motor windings that do not have the higher Basic Impulse Insulation Level (BIL) ratings of other equipment.

When lightning, discharges through an arrester, surge currents are discharged into the grounding terminal. It is very important that substations and overhead lines be protected with well-grounded shield wires. It is also equally important that the ground system between pieces of equipment be bonded together with interconnected ground wires dedicated to the grounding system. When a surge is released on a line by direct strokes or induced strokes, the stroke travels in both directions from the point where the stroke originated. Wave velocity is an inverse function of the surge impedance. Waves travel on an overhead line at approximately 1000 ft. per microsecond, in cables about 300 – 600 ft. per microsecond and in a buried conductor about 300 ft. per microsecond. The velocity internal to a rotating machine may be only 25ft. per microsecond [4-8].

The current resulting from a traveling wave is equal to the voltage divided by the impedance, E/Z. Wave current is approximately two to four amps per kilovolt of surge voltage. Lightning waves on overhead lines gradually attenuate with travel. When the wave runs into a change in impedance (transformer, another line, etc.), the wave continues in the same direction at a different magnitude. It will also reflect back in the direction from which it came. When a wave E_1 traveling on surge impedance Z_1 encounters surge impedanceZ_2, the voltage on the new wave Z_2 becomes:

$$E_2 = E_1 \frac{2Z_2}{E_1 + E_2} \quad (1)$$

Note: as the new surge impedance Z_2 approaches infinity, representative of an open line, $E_2 = 2E_1$.

The reflected wave will actually double in magnitude in its return in the opposite direction. Unless the wave is discharged to ground (lightning arrester connected to ground), the reflected wave can severely damage electrical equipment. Surges produced by lightning have high magnitudes, but their durations are very short. The lightning discharge may reach its crest value in approximately 1 to 20 microseconds and produce conduction flashover voltages of 5 to 20 times normal in 1 microsecond or less. The wave shape is customarily expressed by two intervals associated with the wave geometry. The first time interval is between a virtual zero and crest; the second time interval is between the virtual zero and the half crest value on the wave tail. The wave is defined if the crest value is added to the two time interval designations. For example, a 20000 amp 10 x 20 microsecond current wave rises to a crest of 20000 amperes in10 microseconds after virtual zero and decays to 10000 amperes in 20microseconds after virtual zero.

In addition to component failures, it can cause system upset, lost data, erroneous signals and false system operations. Thus, surge arresters should be designed for all bus systems and system configuration to maintain system reliability regardless of the cause or magnitude of these transients. Surge arrester products should be invaluable to any business when planning for protection from unforeseen occurrences [3, 5, 9]. There are different series of surge arrester and where they can be applied in electrical engineering. Below is the list:

- Low-voltage surge arrestor
- Distribution arrestor
- The station type of common valve arrestor
- Magnetic blow valve station arrestor

- Protection of rotating machine using magnetic blow valve arrester:
- Line Magnetic blow valve arrester:
- DC or blowing valve-type arrester:
- Neutral protection arrester:
- Fiber-tube arrester:
- Plug-in Signal Arrester:
- High-frequency feeder arrester:
- Receptacle-type surge arrester:
- Signal Arrester:
- Network arrester:
- Coaxial cable lightning arrester [5-7].

2. Related Work

2.1. Principles of Surge Arresters (Heading 2)

Though there are different types and classes of surge arresters, they all work on the same general principle. Surge arrester works by conducting excess voltages from a signal or power-carrying conductor to ground.

2.1. Surge Arrester Specification

Most electrical equipment is rated for traveling wave voltage surge capability by the Impulse Test. The Impulse Test is most common and consists of applying a full-wave voltage surge of a specified crest value to the insulation of the equipment involved [6]. The crest value of the wave is called the Basic Impulse Insulation Level (BIL) of the equipment. Each type of electrical equipment has a standard BIL rating. Lightning arresters are coordinated with standard electrical equipment insulation levels so that they will protect the insulation against lightning over voltages [7]. This coordination is obtained by having an arrester that will discharge at a lower voltage level than the voltage required to break down the electrical equipment insulation. Equipment has certain applicable impulse levels or BIL as defined in industry standards [8]. Follow Current, Discharge current and voltage - The follow current is the current which flows from connected supply sources through lightning arrester following the passage of discharge current [11]. From the discharged current, the surge current flows through the arrester after the spark over, while the discharge voltage is the peak value of the voltage appearing between the line terminals and ground, during the passage of the discharge current [5, 9-11].

2.2. Surge Arresters Characteristics

The following types of electrical or electronic devices can be used to reduce or limit voltage surges. Some surge suppression systems use multiple technologies, since each method has its strong and weak points. The first six types listed operate primarily by diverting unwanted surge energy away from the protected load, through a protective component connected in a parallel (or shunted) topology. The last two methods also block unwanted energy by using a protective component connected in series with the power fed to the protected load, and additionally may shunt the unwanted energy like the earlier systems. Among the numerous types are [6]:

- Metal Oxide Varistor (MOV)

A metal oxide varistor consists of a bulk semiconductor material (typically sintered granular zinc oxide) that can conduct large currents (effectively short-circuits) when presented with a voltage above its rated voltage. MOVs typically limit voltages to about 3 to 4 times the normal circuit voltage by diverting surge current elsewhere than the protected load. MOVs may be connected in parallel to increase current capability and life expectancy, providing they are matched sets (unmatched MOVs have a tolerance of approximately ±20% on voltage ratings, which is not sufficient). MOVs have finite life expectancy and "degrade" when exposed to a few large transients, or many smaller transients. As a MOV degrades, its triggering voltage falls lower and lower. If the MOV is being used to protect a low-power signal line, the ultimate failure mode typically is a partial or complete short circuit of the line, terminating normal circuit operation.

When used in power applications, MOVs usually are thermal fused or otherwise protected to avoid persistent short circuits and other fire hazards. In a typical power strip, the visible circuit breaker is distinct from the internal thermal fuse, which is not normally visible to the end user. The circuit breaker has no function related to disconnecting an MOV. A thermal fuse or some equivalent solution protects from MOV generated hazards.

- Transient voltage suppression (TVS) diode

A TVS diode is a type of Zener diode, also called an avalanche diode or silicon avalanche diode (SAD), which can limit voltage spikes. These components provide the fastest limiting action of protective components (theoretically in picoseconds), but have a relatively low energy absorbing capability. Voltages can be clamped to less than twice the normal operation voltage. If current impulses remain within the device ratings, life expectancy is exceptionally long. If component ratings are exceeded, the diode may fail as a permanent short circuit; in such cases, protection may remain but normal circuit operation is terminated in the case of low-power signal lines. Due to their relatively-limited current capacity, TVS diodes are often restricted to circuits with smaller current spikes. TVS diodes are also used where spikes occur significantly more often than once a year, since this component will not degrade when used within its ratings. A unique type of TVS diode (trade names Transzorb or Transil) contains reversed paired series avalanche diodes for bi-polar operation.TVS diodes are often used in high-speed but low power circuits, such as occur in data communications. These devices can be paired in series with another diode to provide low capacitance as required in communication circuits.

- Thyristor Surge Protection Device (TSPD)

A Trisil is a type of thyristor surge protection device (TSPD), a specialized solid-state electronic device used in crowbar circuits to protect against overvoltage conditions. A SIDACtor is another thyristor type of device used for similar protective purposes. These thyristor-family devices can be viewed as having characteristics much like a spark gap or a gas discharged tube (GDT), but can operate much faster. They are related to TVS diodes, but can "breakover" to a low clamping voltage analogous to an ionized and conducting spark gap. After triggering, the low

clamping voltage allows large current surges to flow while limiting heat dissipation in the device.

- Gas discharge tube (GDT)

A gas discharge tube (GDT) is a sealed glass-enclosed device containing a special gas mixture trapped between two electrodes, which conduct electric current after becoming ionized by a high voltage spike. GDTs can conduct more current for their size than other components. Like MOVs, GDTs have a finite life expectancy, and can handle a few very large transients or a greater number of smaller transients. The typical failure mode occurs when the triggering voltage rises so high that the device becomes ineffective, although lightning surges can occasionally cause dead.

- Selenium Voltage Suppressor (SVS)

It usually has a longer life than a MOV. It is used mostly in high-energy DC circuits, like the exciter field of an alternator. It can dissipate power continuously, and it retains its clamping characteristics throughout the surge event, if properly sized.

- Quarter-Wave Coaxial Surge Arrestor

It is used in RF signal transmission paths, this technology features a tuned quarter-wavelength short-circuit stub that allows it to pass a bandwidth of frequencies, but presents a short to any other signals, especially down towards DC. The pass bands can be narrowband (about ±5% to ±10% bandwidth) or wideband (above ±25% to ±50% bandwidth). Quarter-wave coax surge arrestors have coaxial terminals, compatible with common coax cable connectors (especially N or 7-16 types). They provide the most rugged available protection for RF signals above 400 MHz; at these frequencies they can perform much better than the gas discharge cells typically used in the universal/broadband coax surge arrestors. Quarter-wave arrestors are useful for telecommunications applications, such as Wi-Fi at 2.4 or 5 GHz but less useful for TV/CATV frequencies. Since a quarter-wave arrestor shorts out the line for low frequencies, it is not compatible with systems which send DC power for a LNB up the coaxial downlink.

- Series Mode (SM) Surge Suppressors

These devices are not rated in joules because they operate differently from the other suppressors, and they do not depend on materials that inherently wear out during repeated surges. SM suppressors are primarily used to control transient voltage spikes on electrical power feeds to protected devices. They are essentially heavy-duty low-pass filters connected so that they allow 50/60 Hz line voltages through to the load, while blocking and diverting higher frequencies. This type of suppressor differs from others by using banks of inductors, capacitors and resistors that shunt voltage spikes to the neutral wire, whereas other designs shunt to the ground wire.

2.3. Factors that Affects Surge Arrester

Surge arresters when installed are exposed to many physical factors on the earth. These factors are temperature, spacing of the arresters and earth resistance. Figure 1 below shows the effect of temperature.

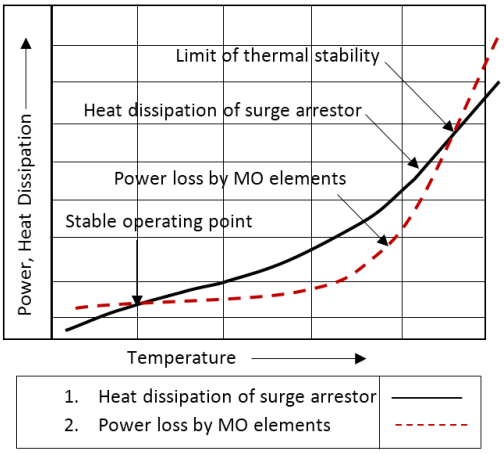


Figure 1.Effect of Temperature

2.3.1. Effect of Temperature

Arresters installed today are all metal-oxide (MO) arresters without gaps. ZnO material has negative thermal coefficient, the heat generation by the MO elements at a constant voltage will increase to a higher degree as the dissipation of this heat through the housing of the arrester. As a consequence, there are two intersections between heat generation and heat dissipation characteristic. After heating the MO elements of the arrester, by single or multiple current impulses, below the limit of thermal stability the arrester will always return to the stable operating point. However, after heating above the limit of thermal stability the arrester will become thermally unstable and be destroyed. The effect of thermal stability is strongly dependent on the thermal properties of the arrester.

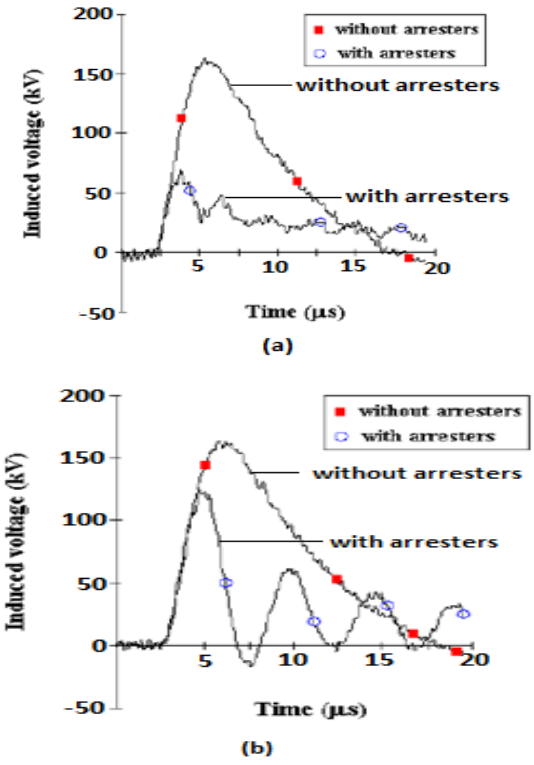


Figure 2. Effect of spacing of arresters

2.3.2. Effect of Spacing of the Arresters

The distance between adjacent surge arrester has influence on the induced voltage particularly if lightning strike point (in the case lightning) is nearly equidistant from the sets of surge arrester. The closer the arresters are, the lower the voltage magnitude will be. The figure 2(a) and 2(b) below compares the induced voltage on line with corresponding arrester spacing of 300m and 600m. The stroke current front time and time to half-value are about 3.2μs and 5.8μs, respectively. Figure 2 shows the effect of spacing of the arresters.

Figure 3 comprises the voltage induced on a line without arresters and a line with arrester but with earth resistance of 0Ω and 200Ω [9].

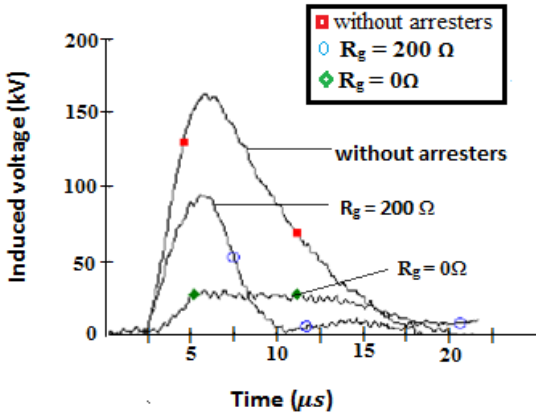


Figure 3.Voltage induced on the lines with and without surge arrester with earth resistance of 0Ω and 200Ω

The earth resistance may have a significant influence on the induced voltage amplitude, especially when the lightning (in the case of lightning surge) strike point is in front of a set of arresters. This is due to the fact that, for lower value of earth resistance (Rg), the current that flows to earth (through the surge arresters) increasing the value of the voltage component that, by coupling, reduces the induced voltage in the conductors.

3. Methodology

3.1. Presentation of methods for assessing surge arrester performance

The choice for selecting the arresters is determined by the parameters:

- Surge arrester ratings.
- Impedance division under normal and abnormal conditions to determine whether it will act as low resistive unit under abnormal conditions.
- Selecting station class for its protection capability and distribution class for its restoration capability.
- Also the location of the arrester, determined by the terminal point of the protected equipment.

3.2. Application and Selection of Surge Arrester

The objective of arrester selection is to select the lowest rated surge arrester which will provide adequate overall protection of the equipment insulation and have a satisfactory service when connected to the power system. The arrester with the minimum rating is preferred because it provides the greatest margin of protection for the insulation. A higher rated arrester increases the ability of the arrester to survive on the power system, but reduces the protective margin it provides for a specific insulation level. Both arrester survival and equipment protection must be considered in arrester selection. Generally, surge arrester works on the basis of impedance division. Under normal condition it acts as a highly resistive unit but under stress (voltage surge) it acts as a low resistive unit. The mode of operation is best explained using Thevenin theorem. Thevenin theorem states that when a switch is closed in a circuit, the current that flows can be determined by dividing the voltage across the switch prior to closing, by the impedance seen looking into the circuit at the open switch contacts. Consider the figure 4:

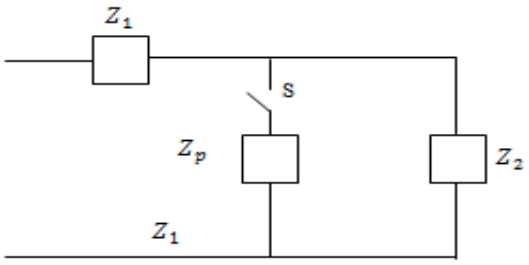


Figure 4. Circuit diagram of surge arrester operation

Suppose Z_1 is the impedance of a system on which a surge is being generated and that Z_2 is a piece of equipment connected to the system then Z_P is the impedance of the arrester. According to Thevenin, when S is closed the current through the arrester is given by:

$$\frac{V}{ZP}+\frac{Z1Z2}{(Z1+Z2)}=\frac{(Z1+Z2)V}{(ZPZ1+ZPZ2+Z1Z2)} \quad (2)$$

And the voltage V_1across Z_P is given as:

$$V1=\frac{ZP(Z1+Z2)V}{(ZPZ1+ZPZ2+Z1Z2)} \quad (3)$$

3.3. Selection of lightning arrester

The lightning arresters are designated by the crest magnitude of the discharge current having $10 \times 20\mu s$ wave shape which the arrester can safely pass without causing damage. The lightning arresters are designated as 8, 10, 20 KA and can safely discharge these current crests. The discharge current from the arrester varies from a few hundred amperes to kilo-amperes Maximum discharge voltage (crest value) and discharge factor for the arrester is defined as the maximum value of voltage which appears across the arrester terminals at the time of discharging if rated current determines its impulse level of protection. The discharge factor is given as in equation 5, to get a result that falls within 2.4 to 3.0:

$$DF=\frac{v_{da}}{v_{ra}} \quad (4)$$

Where, v_{da} = discharge voltage of the arrester

v_{ra} = rated voltage of the arrester

The above ratio for arresters manufactured by different companies varies from 1.6 to 1.8, so, the average value may be taken as 1.7 E rated arrester voltage KV (R.M.S value) and L_S as the minimum impulse insulation level in kV (crest valve) with value after allowing 10% as tolerance factor and 25% as margin factor can be obtained.

$$L_S = \frac{1.25 \times 1.1 \times 1.7 \times \sqrt{2}}{0.8} E = 4.13E \quad (5)$$

In a case of Extra High Voltage (EHV) system, the maximum impulse L_P in kV implies:

L_P = 2.3 × power frequency withstand voltage in kV (RMS)

$= 2.3\ E_L$ (6)

For 75% arrester,

$L_P = 2.37\ E_L$ (7)

For 80% arrester,

$L_P = 2.53 E_L$ (8)

3.4. Surge current ratings

The surge current normally is the largest single surge that a device can withstand without damage, should exceed the largest surge that SPD would experience in service. This work uses the largest surge that can be expected at the service entrance as resistance to multiple 3 kA (8/20μs) surge for panel protectors. At the service entrance SPD rate the surge current per phase which can be calculated by adding the surge current from individual modes of protection L-N and L-G. Thus, surge current ratings for residential should fall within 10 kA to 70 kA per phase with test waveforms other than 8/20 μs also acceptable. This work uses surge current ratings of 20 kA to 70 kA (8/20μs) per phase. For high-lightning area, like Binshua, SPDs with higher surge current ratings in the range of 40 kA to 120 kA is used in order to provide a longer service life and higher reliability.

3.5. Lightning Arrester Design

A station class and intermediate surge arresters are used. The system voltage is 0.415kV and maximum continuous operation voltage is 48 kV rms. The duty cycle rating is 60 kV rms and maximum discharge current is 10 kA.

1. Maximum 0.5μs discharge voltage = 163.5 kV
2. Maximum switching surge protective level=116.4 kV
3. Maximum discharge voltage using an 8/20=148.8 kV Current wave-kV

The maximum discharge voltage for a10kA impulse current wave produces a voltage wave cresting in 0.5μs.

- Specification of Lightning Arrester for Incoming Side

Incoming side of residential, the specifications of lightning arrester are as follow.
System nominal voltage = 0.415kV
Rated normal Voltage = 0.415 × 1.1 = 45.65kV
Continuous Operating Voltage (kV) rms = 48.0kV
Normal Discharge Current (8 /20μs) kA = 10kA
1/50 Impulse Spark over Voltage = 163.5kV
Frequency (Hz) = 50Hz
Type = outdoor

- Specifications of Lightning Arrester for Outgoing Side

The followings are the specifications of lightning arrester for outgoing side of residential.
System nominal voltage = 0.220kV
Rated normal Voltage = 0.22 × 1.1 = 24.20kVkV
Continuous Operating Voltage (kV) rms = 24kV
Normal Discharge Current (8 /20μs) (kA) = 10kA
Frequency (Hz) = 50Hz
Type = outdoor

- Intermediate class arrester

Application is based upon the maximum continuous operation voltage, line to neutral, at the arrester location. For grounded neutral systems (GNS), this is computed as:

$$GNS = \frac{SV_m}{\sqrt{3}} \quad (9)$$

Where, SV_m = maximum system voltage.

For historical comparison, the maximum continuous operating voltage is 81% of the conventional 71% arrester installed on an effectively grounded neutral system.

- Earthing system and Lightning earthing

The frame of every generator, stationary motor, portable motor, and metallic parts of all transformers and the regulating and controlling apparatus connected with supply shall be earthed by separate and distinct connection with earth. Every conductor used on earthing shall be of stranded solid copper or suitable copper alloy, and shall be protected wherever liable to mechanical damage. Also, against corrosion, particular attention being given in these respects to the earthing leads at its point of connection with the earth electrode. The coefficient of earthing is below 80 percent. With earthing coefficient less than 80% in this case, on four wire distribution systems, the transformer neutral is solidly grounded at every voltage level. On high voltage transmission systems the coefficient of earthing does not exceed 75%. In earthing system, lightning arresters with PVC coated wire and cable lug are used.

- 415V Lighting Arrester (70mm^2 PVC Coated Wire) = 50"
- 220V Lighting Arrester (70mm^2 PVC Coated Wire) = 100"
- 70mm^2 Cable Lug = 2

3.5.1. Design Data

For incoming side and outgoing side, the discharge voltage, insulation level, minimum impulse insulation level and power frequency withstand voltage base on 100% arrester are as shown

in Table.1. Ground voltage peak value and switching surge withstand voltage are also shown in Table.1.

Table.1. Design Data Sheet

	System Voltage	
	Incoming Side of 0.415 kV	Outgoing Side of 0.220 kV
Rated Voltage (kV)	45.65	24.20
Discharge Voltage on 100% Arrester (kV)	188.76	93.6
Insulation level for 100% Arrester (kV)	284.78	164.4
Minimum Impulse Insulation level on 100% (kV)	732.389	363.168
Power Frequency withstand Voltage on 100% (kV)	151.8	75.9
Ground Voltage Peak Value (kV)	59.28	29.63
Require Switching Surge withstand Voltage (kV)	215.6	107.8

For the arrester to work properly neither the voltage V across S prior to closing nor the voltage V_1 should exceed the voltage that the equipment is capable of sustaining. The proper selection and application of surge arresters in a system involve impulse test considering the following:

Basic Impulse Insulation Level (BIL), which is the reference insulation level expressed as an impulse crest (or peak) voltage with a standard wave not longer than a 1.2 x 50 microsecond wave, that is, the impulse takes 1.2 microseconds to reach the peak and then decays to 50% of the peak in 50 microseconds. This BIL also have a level that can repeatedly be applied to equipment without flashover, disruptive charge or other electrical failure under test conditions.

Chopped Wave Insulation Level is determined by using impulse waves that are of the same shape as that of the BIL waveform, with the exception that the wave is chopped after 3 microseconds. Generally, it is assumed that the Chopped Wave Level is 1.15 times the BIL level for oil filled equipment such as transformers. However, for dry type equipment, it is assumed that the Chopped Wave Level is equal to the BIL level.

Critical Flashover Voltage, which is the peak voltage for a 50% probability of flashover or disruptive charge.

Impulses Ratio is the ratio of breakdown voltage at surge frequency to breakdown voltage at normal system frequency, (60 Hz).

Coefficient of Earthlings (CE), which is defined as the ratio of highest rms voltage of healthy phase-to-earth to phase-to-phase normal rms voltage and multiplied by 100; that is:

$$CE = \frac{v_{Hrms}}{v_{LLrms}} \times 100 \quad (10)$$

Where, v_{Hrms} = highest rms voltage of healthy line to earth

v_{LLrms} = normal line to line rms voltage

The proper selection also involves decisions in three areas:

- Selecting the arrester voltage rating which is based on whether or not the system is grounded and the method of system grounding.
- Selecting the class of arrester, from the three classes of arresters, in order of protection, capability and cost. The classes are: Station class, Intermediate class and Distribution class. The station class arrester has the best protection capability and is the most expensive.
- Determine where the arrester should be physically located.

The rating of the arrester is defined as the RMS voltage at which the arrester passes the duty-cycle test as defined by the reference standard. Metal oxide arresters are designed and tested in accordance with ANSI/IEEE C62.11. This states that, the lower the arrester voltage rating, the lower the discharge voltage, and the better the protection of the insulation system. The lower rated arresters are also more economical. The challenge of selecting and arrester voltage rating is primarily one of determining the maximum sustained line-to-ground voltage that can occur at a given system location and then choosing the closest rating that is not exceeded by it. This maximum sustained voltage to ground is usually considered to be the maximum voltage on the non-fault phases during a single line-to-ground fault. Hence, the appropriate arrester ratings are dependent upon the manner of system grounding. All of the system parameters need to be considered while choosing an arrester classification. If the actual arrester energy duties are not known and a transient study cannot be performed, then it is suggested that Station class arresters be applied. This is a conservative approach that reduces the chances of misapplication.

The ideal location of arresters from the standpoint of protection is directly at the terminals of the equipment to be protected. At this location, with the arrester grounded directly at the tank, frame or other metallic structure which supports the insulated parts, the surge voltage applied to the insulation will be limited to the discharge voltage of the arrester.

When arresters are connected to the power system they continually monitor the system operating voltage, which is referred to as Maximum Continuous Operating Voltage (MCOV). For each arrester rating, there is a limit to the magnitude of voltage that may be continuously applied to the arrester.

For surge arrester applications the “solidly grounded” classification is usually found in electric utility distribution systems where the system is usually only grounded at the point of supply. These systems can exhibit a wide range of grounding coefficients depending upon the system or location in the system. Accordingly, these systems may require a study to ensure the most economical, secure, arrester rating selection. If this information is not known or available, the ungrounded classification should be used.

The lists arrester ratings, from a manufacturer, that would normally be applied on systems of various line-to-line voltages are as shown in Table 2 [12].

Table 2: Typical Arrester Ratings for System Voltage

Arrester Rating (kV)		
Norminal System L-L Voltage (kV)	Grounded Neutral Circuits	High Impedance Grounded, Ungrounded or Temporarily Ungrounded
2.4	**2.7**	3.0
4.16	3.0	---
	4.5	4.5
		5.1
4.8	4.5	---
	5.1	5.1
	---	6.0
6.9	6.0	---
	---	7.5
	---	8.5
12.47	9.0	---
	10	---
	---	12
	---	15
13.2, 13.8	10	---
	12	---
	---	15
	---	18
23, 24.94	18	---
	21	---
	24	24
	---	27
34.5	27	---
	30	---
	---	36
	---	39
46	39	---
	---	48
69	54	---
	60	---
	---	66
	---	72

3.6. Designing a Home Surge Arrester System

The design of any surge arrester comes under the same principle which is to ground any excess voltage that comes into the electrical or electronic system. This could be achieved by considering the circuit diagram as shown in Figure 5.

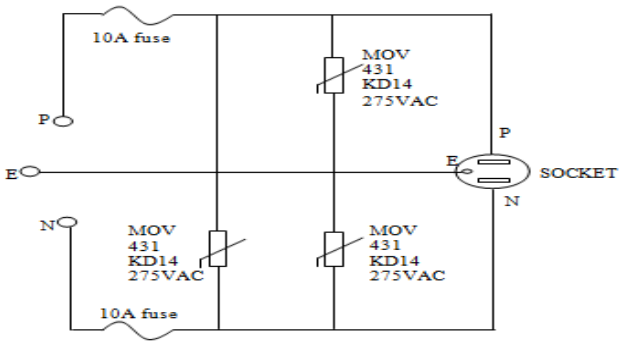


Figure 5. Home Surge Arrester Circuit

From the Figure 4, MOVs are connected in parallel to increase current capability and life expectancy, which, when exposed to a few large transients, or many smaller transients. The MOV degrades, when its triggering voltage falls lower and lower.

When used in power applications, MOVs usually are thermal fused or otherwise protected to avoid persistent short circuits and other fire hazards. In a typical power strip, the visible circuit breaker is distinct from the internal thermal fuse, which is not normally visible to the end user. The circuit breaker has no function related to disconnecting an MOV. A thermal fuse or some equivalent solution protects from MOV generated hazards.

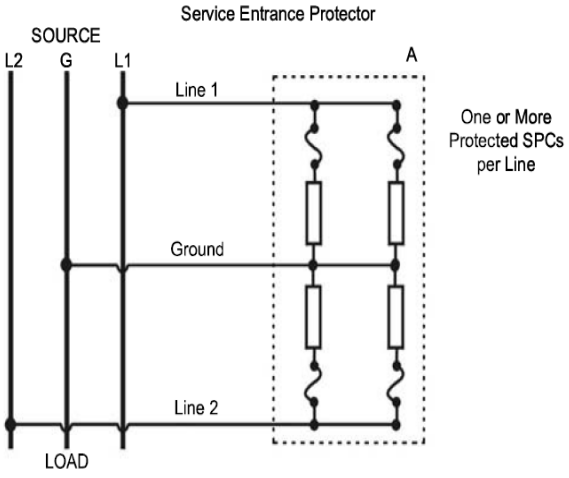
3.7. Modes of Protection

The modes of protection required at the service entrance depend on the configuration of the electric distribution system. Surge is transmitted in the normal mode as line-to-neutral (L-N) or line-to-line (L-L), or in the common mode as line-to-ground (L-G) or neutral-to-ground (N-G). Immediately after a transformer or at the main entrance, L-N or L-G should be the only protection modes that are required, but into the building, L-N, L-G, and N-G should all be the protection modes required. A simplified surge protection device (SPD) circuit showing how the components are connected is shown in Figure 6.

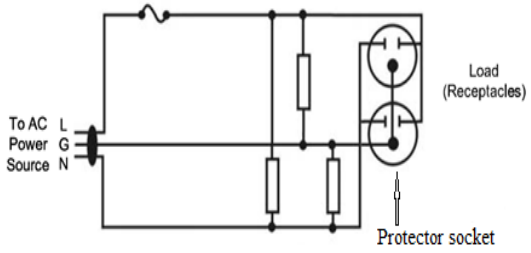
The entrance protector shown in Figure 6a uses two varistor groups to protect the L-G modes. The L-L modes are protected by the two varistor groups in series. This varistor conduct negligible current until a specific limiting voltage across the terminals is reached. Above that voltage, the device starts to conduct, thereby limiting the voltage across the terminals. The point-of-use protectors, as shown in Figure 6b and Figure 6c, uses three varistors to protect all three modes, L-N, L-G, and N-G, as recommended

Since the N and G are directly bonded at the service entrance, SPDs used normally have no need to protect the N-G mode. However, protectors downstream from the service entrance or at the load should protect the N-G mode, since N-G surges might arise from downstream in the building. This work effectively uses two varistors in series, each rated for the L-N supply voltage, to provide the L-L limiting voltage. Although the limiting surge

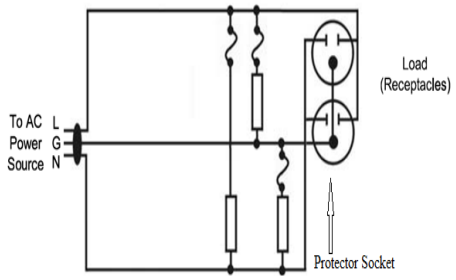
voltage obtained this way may not be as low as that from separate L-L components, the configuration is usually satisfactory.



6(a)



6(b)



6(c)

Figure 6. A simplified surge protection device (SPD) circuit

3.8. Installation of Home Surge Arrester

Today's modern homes are filled to the brim with technology products -- from personal music players, e-readers and laptops to HDTVs to hi-tech refrigerators and washing machines. These appliances and electronics feature delicate circuitry that can easily be destroyed by a simple fluctuation in the home's voltage. This can be caused by lightening or a surge in the power grid.

In order to protect these sensitive items, we usually plug their power cords into surge protectors. But, as more electronics become further integrated into our daily lives, we find ourselves having to protect all of them, which makes having surge protectors for every piece of electronics a problem. One way to avoid surge protector overflow is to install a whole-house surge arrester right in your electrical panel. This device will replace one of the existing double-pole circuit breakers in the electrical panel and is relatively easy to install. Although the surge arrester will be technically wired to one circuit, it will protect all of the circuits in the panel. Once your home is protected by the surge arrester, it is still recommended to use surge protectors with your most sensitive equipment as an extra safety measure.

4. Conclusion

In this paper, the basis theory of lightning, lightning shielding and design of lightning arrester are presented. Station class and intermediate arresters are used. The type of arrester is outdoor type. The rated voltages of arresters are 46kV and 25kV, the maximum discharge current is 10kA and MCOV are 48kVrms and 24kVrms. The lightning arrester in this paper is provided for overvoltage protection in a high-lightning area. This paper will help and give the electrical knowledge of the protection system in high-lightning area which coach to the technicians, the professional engineers, the students who are facing the overvoltage condition and protection coordination in high-lightning area.

So many standards are listed but none of them guarantee that a protector will provide proper protection in a given application. Each standard defines what a protector should do or might accomplish, based on standardized tests that may or may not correlate to conditions present in a particular real-world situation. A specialized engineering analysis may be needed to provide sufficient protection, especially in situations of high lightning risk.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] J. Hernandez, GE Specification Engineer; "*Lightning Arresters: A Guide to Selection and application".*
[2] *Lightning arrester.* From Wikipedia, the free encyclopedia www.wikipedia.com.
[3] J.G. Anderson, "T*ransmission line Reference Book, 345Kv and above", 2nd edition*. EPRI, Palo Alto, CA (1982)
[4] Allan Greenwood, "*Electrical transients in power system",* Wiley, 1971 - Technology & Engineering - 544 pages
[5] *Surge arrestor.* From Wikipedia, the free encyclopedia www.wikipedia.com.
[6] L. Pryor, P.E., GE, Sr. Specification Engineer; "*The Application and Selection of Lightning Arresters".*
[7] Surge arrester, G31-A75X-X809. www.epcos.com
[8] DEHN + SÖHNE*, "Lightning protection guide",* 3rd updated edition as of December 2014 www.dehn-international.com
[9] A. Piantini, "*Lightning Protection of overhead power distribution lines*", 29th international conference on lightning protection June 2008; Uppsala, Sweden.
[10] D. Donovan, "Whole-House Surge Arrester: DIY Installation Guide". http://www.stevejenkins.com/
[11] N. K. Htwe, "Analysis and Design Selection of Lightning Arrester for Distribution Substation", World Academy of Science, Engineering and Technology 48 2008.
[12] ANSI/ IEEE 1989, "IEEE Standard for Gapped Silicon-Carbide Surge Arresters for AC Power Circuit".

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 206-213 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

Detailed Analysis of Torque Ripple in High Frequency Signal Injection based Sensor less PMSM Drives

Ravikumar Setty .A*, Kishore Chatterjee

Electrical Engineering Department, IIT Bombay, Mumbai-400076, India.

ARTICLE INFO

Article history:
Received: 23 December, 2016
Accepted: 20 January, 2017
Online: 28 January, 2017

Keywords:
Sensor less control
High frequency signal injection
Torque Ripple
Audible Noise
PMSM

ABSTRACT

High Frequency Signal Injection based techniques are robust and well proven to estimate the rotor position from stand still to low speed. However, Injected high frequency signal introduces, high frequency harmonics in the motor phase currents and results in significant Output Torque ripple. There is no detailed analysis exist in the literature, to study the effect of injected signal frequency on Torque ripple. Objective of this work is to study the Torque Ripple resulting from High Frequency signal injection in PMSM motor drives. Detailed MATLAB/Simulink simulations are carried to quantify the Torque ripple at different Signal frequencies.

NOMENCLATURE

PMSM	Permanent Magnet Synchronous Motor
Ea,Eb,Ec	Back EMF of phase a,b,c
Va,Vb,Vc	Terminal Voltage of phase a,b,c
Ia, Ib,Ic	Currents through Phase a,b,c
I_{ab1}	$I_{\alpha\beta}$ rotated with injected frequency and filtered.
R	Stator winding resistance / Phase
L	Stator winding inductance / Phase
T_e	Electromagnetic Torque
A	3 phase to 2 phase transformation matrix (Clark Transformation)
B	αβ to dq transformation matrix
T	Transpose
v_c, i_c	High frequency carrier voltage and high frequency carrier current response
L_c	High frequency stator inductance vector
αβ	Stator orthogonal coordinate system
dq	Rotor orthogonal coordinate system
dq'	Magnetic Saliency axis on rotor side.
V_{DC}	DC Supply Voltage to Inverter
θ_r or $\lambda_{dq'}$	Rotor Magnetic saliency Position angle reference to rotor orthogonal system.

1. Introduction

Permanent Magnet Synchronous Motors are driven with Vector control drives to improve the dynamic performance of the drive. Vector control requires rotor position to control the drive. Position sensor like resolver or encoder are used to provide the rotor position information [1], such a position sensor reduces the reliability of the system and increase the cost and weight of the drive. Sensor less control techniques are widely used to address the position sensor issues [2]. Sensor less control techniques uses only terminal variables (voltages and currents) to determine the rotor position. Back EMF (Electro Motive Force) based methods are popular and robust to extract rotor position information from medium to high speeds. These techniques require minimum motor speed for proper drive operation, however all the techniques based on back EMF, have serious problem to extract the rotor position from low speed to standstill [3]-[4]. At standstill, motor back EMF is zero, so all Fundamental methods based on back EMF fail to extract rotor position information.

From low speed to standstill machine geometrical or saturation saliency is used to estimate rotor position information. The idea behind saliency based sensor less estimation scheme is that, as the machine winding inductance is function of the rotor position due to saliency, the rotor position is derived from the profile of inductance variation. Position estimation based on machine magnetic saliency or signal injection techniques are powerful from standstill to minimum speed. In order to reveal the position information from machine saliency property, high frequency

*Ravikumar Setty .A,
Honeywell Technology Solutions ,Bangalore, India.
+91-80-26588360 ravikumar.setty@honeywell.com

voltage signal is injected and resulting response current is demodulated to extract position information [5]-[7].

According to the type of injected signal, continuous signal injection consists of a rotating sinusoidal signal injection [8] -[10], a pulsating sinusoidal [11] -[12], a square wave signal injection [13] , or arbitrary injection [14].

- Continuous signal injection
- Transient signal injection
- PWM excitation without additional injection

The implementation of HF injection schemes is well proven and can be made with standard low cost micro controllers. The response to the injected high frequency signal is used to extract non-zero-sequence current [15], non-zero sequence current derivative [16], which provide the rotor position information. Methods injecting a rotating or pulsating carrier, require filtering in order to distinguish the carrier current response from the fundamental current. This introduces considerable amount of delay, which degrades the dynamics of the sensor less control. This problem is addressed in [17], but their proposals require more complex signal processing and are parameter dependent. Other disadvantage with this method is significant current distortions and torque ripple.

Apart from signal injection techniques open-loop startup methods [18] also used, but these methods are not reliable when enough starting torque motor needs. At start up motor can rotate in reverse direction from the desired direction until the rotor aligned to the known pre-defined position.

In transient injection methods the fundamental current is not affected by high frequency component, but phase currents have significant spikes during the test vector instances. Main concern with the transient signal injection method is the minimum acquisition time for the measurement. The current disturbances also cause additional losses and in some cases undesired torque ripple [19]-[20]. The worst problem is however a significant audible noise, which cannot be neglected for many applications.

To overcome this problem, the inherent excitation of the PWM is used in [21]-[23]. Again, a complex signal processing is required. Implementation of the transient voltage methods requires a more complex and advanced micro controller system. Sensor less control techniques in the low speed to stand still region utilizes the magnetic saliency of the machine to extract rotor position information, however the injected signal could be continuous or transient.

Transient signal injection technique requires current sensing with high bandwidth and which makes current sensing system expensive. This limits the popularity of the transient signal injection techniques. On the other hand continuous signal injection techniques doesn't need expensive current sensing system , but the injected amplitude causes torque ripple, audible noise and HF losses[24]-[25] .

This work focus on detailed analysis of torque ripple resulting from rotating high frequency signal injection based sensor less control.

2. Modelling of PMSM in d-q reference frame for Torque Ripple Analysis

Detailed modeling of PM motor drive system is required for proper simulation of the system. The d-q model has been developed on rotor reference frame. There are many variations of rotor configurations are reported in the literature. Two main configurations of PM synchronous motors are surface magnet type where magnets are mounted on the outer surface of the rotor (SPMSM), and the interior magnet type where the magnets are mounted inside the magnetic structure of the rotor (IPMSM). Surface and interior type rotor configurations are shown in Fig.1.

Mounting the magnets to the surface of the rotor is the simplest and cheapest method for construction of PMBL motor. Another method for mounting the magnet is to imbed them in the interior of the rotor. Interior design provides mechanical robustness and smaller air gap, which results in a component of reluctance torque in addition to the developed torque due to PM excitation. This work analyze the torque ripple on IPM type motor based on high frequency signal injection technique.

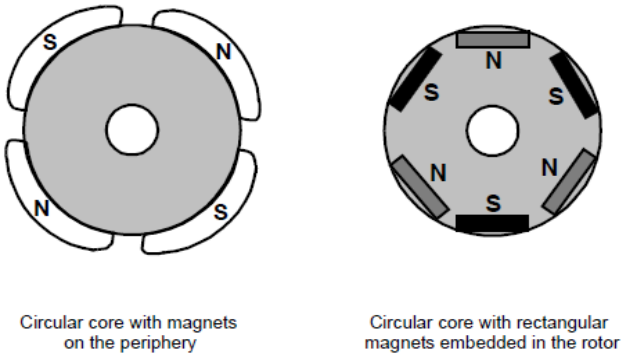


Fig.1. Surface and interior type rotor configurations.

Symmetrical three phase PMSM on stator side is represented with Equations (1) to (3).

$$V_a = R * i_a + L_a \frac{di_a}{dt} - E_a \quad (1)$$

$$V_b = R * i_b + L_b \frac{di_b}{dt} - E_b \quad (2)$$

$$V_c = R * i_c + L_c \frac{di_c}{dt} - E_c \quad (3)$$

Where

$$\begin{bmatrix} E_a \\ E_b \\ E_c \end{bmatrix} = \begin{bmatrix} E \\ E\ (sin\lambda_{dq'} - 2\pi/3) \\ E\ (sin\lambda_{dq'} + 2\pi/3) \end{bmatrix} \quad (4)$$

Using (4) equation (1) to (3) can be re written as (5) to (7).

$$V_a = R * i_a + L_a \frac{di_a}{dt} - E \quad (5)$$

$$V_b = R * i_b + L_b \frac{di_b}{dt} - E\ (sin\lambda_{dq'} - 2\pi/3) \quad (6)$$

$$V_c = R * i_c + L_c \frac{di_c}{dt} - E\ (sin\lambda_{dq'} + 2\pi/3) \quad (7)$$

Fig. 2 (a) shows the cross section of a three phase PMS motor showing three phase stator windings *A, B and C*. The shaded areas of the motor section symbolize areas with a high permeability. It can be seen that the relative air gap width changes with the rotor position *dq'*. The change of the relative air gap changes the values of the resulting three phase equivalent inductances L_a, L_b and L_c as shown in Fig. 2 (b).

Stator winding inductance has two components one is fixed and second one changes with the position of the rotor. First component is the combination of self and mutual inductance, this doesn't include the saliency of the machine and second component depends on the saliency of the machine and rotor position. The three phase equivalent inductances L_a, L_b and L_c are given by (9) to (11), which consist of a constant term L plus a Sinusoidal modulation depending on ΔL and the rotor position $\lambda_{dq'}$. It is assumed that saliency causes a balanced sinusoidal three phase inductance modulation.

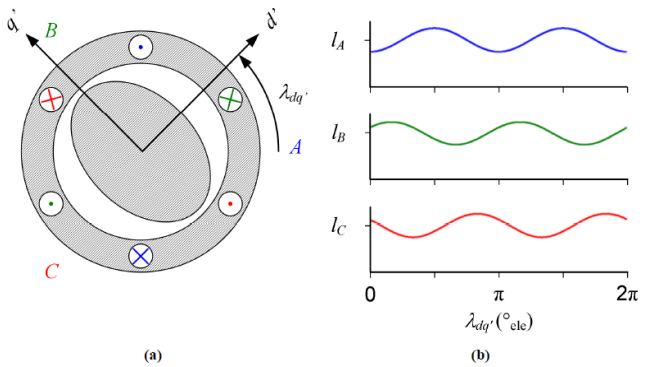


Fig.2. Three phase equivalent inductance modulation because of saliency

$$L_{ABC} = \begin{bmatrix} L_a & 0 & 0 \\ 0 & L_b & 0 \\ 0 & 0 & L_c \end{bmatrix} \quad (8)$$

$$L_a = L - \Delta L * \cos(2\lambda_{dq'}) \quad (9)$$

$$L_b = L - \Delta L * \cos(2\lambda_{dq'} + 2\pi/3) \quad (10)$$

$$L_c = L - \Delta L * \cos(2\lambda_{dq'} - 2\pi/3) \quad (11)$$

Transforming the inductance vector from stationary reference frame to rotating reference frame *dq'* frame gives (13), which shows that the complete saliency of the machine can be described by two variables *Ld'* and *Lq'* , which define the equivalent inductance values into *d'* and *q'* axis.

$$B = \begin{bmatrix} \cos(\lambda_{dq'}) & \sin(\lambda_{dq'}) \\ -\sin(\lambda_{dq'}) & \cos(\lambda_{dq'}) \end{bmatrix} \quad (12)$$

$$L_{dq'} = B * \begin{bmatrix} L - \left[\frac{\Delta L}{2} * \cos(2\lambda_{dq'})\right] & -\left[\frac{\Delta L}{2} * \sin(2\lambda_{dq'})\right] \\ -\left[\frac{\Delta L}{2} * \sin(2\lambda_{dq'})\right] & L + \left[\frac{\Delta L}{2} * \cos(2\lambda_{dq'})\right] \end{bmatrix} * B^{-1}$$

$$= \begin{bmatrix} L - \left[\frac{\Delta L}{2}\right] & 0 \\ 0 & L + \left[\frac{\Delta L}{2}\right] \end{bmatrix} = \begin{bmatrix} L_{d'} & 0 \\ 0 & L_{q'} \end{bmatrix} \quad (13)$$

The stator fixed -reference frame model transformed into a rotor fixed reference frame by using the Park-transformed inductance matrix. PMSM model in rotor reference frame can be expressed as shown in (14) to (15).

$$V_d = R * i_d + L_{d'} \frac{di_d}{dt} - w_r L_{q'} i_q \quad (14)$$

$$V_q = R * i_q + L_{q'} \frac{di_q}{dt} + w_r L_{d'} i_d + w_r \lambda_{pm} \quad (15)$$

Based on (14) and (15) PMSM d-q model can be represented as shown in Fig.3.

Electromagnetic torque (T_e) with stator variables in the rotor reference frame can be expressed as

$$T_e = \left(\frac{3}{2}\right)(p)\left[(\lambda_d * i_q) - (\lambda_q * i_d)\right] \quad (16)$$

Where p is number of pole pairs and λ_d and λ_q can be expressed as below.

$$\lambda_d = \lambda_{pm} + L_{d'} * i_d \quad (17)$$

$$\lambda_q = L_{q'} * i_q \quad (18)$$

Substituting λ_d *and* λ_q terms in to (16) electromagnetic torque (T_e) can be expressed as

$$T_e = \left(\frac{3}{2}\right)(p)\left[(\lambda_{pm} * i_q) - \{(L_{q'} - L_{d'}) * (i_q * i_d)\}\right] = k_T * i_q - \left(\left(\frac{3}{2}\right) * (p) * (\Delta L) * (i_q * i_d)\right) \quad (19)$$

Mechanical model can be expressed as Equation (20), and complete PMSM model is shown in Fig.4

$$T_e - T_l = J * \left(\frac{d\Omega}{dt}\right) + B * \Omega = \frac{J}{p} * \left(\frac{dw_r}{dt}\right) + \frac{B}{p} * w_r \quad (20)$$

3. Position Estimation using High Frequency Signal Injection Technique

This section briefly describe the Vector Controlled Drive operation and position estimation using High Frequency Signal Injection Technique.

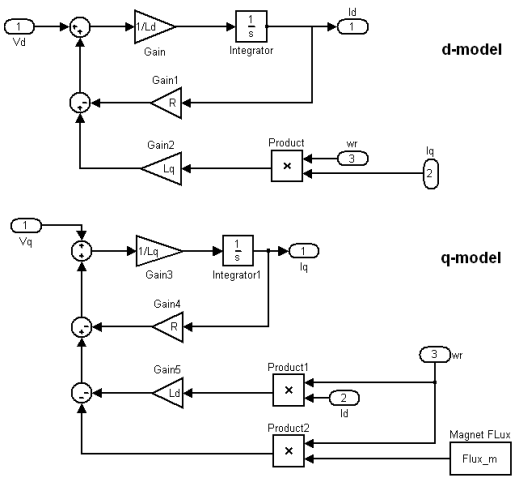


Fig.3. PMSM Electrical model in dq reference frame.

3.1. Vector Controlled PMSM Drive using Position Sensor

Fig.5. depicts the basic block diagram of Vector control of PMS motor drive. Motor phase currents are sensed and then a two phase to three phase transformation is carried out with Clarke transform and Park transform. The motor phase currents i_a and i_b are measured with a current sensor. The Clarke transform is applied to the measured currents to determine the stator current projection in a two coordinate non-rotating frame. The Park co-ordinate transformation is then applied in order to obtain its projection in the (dq) rotating frame.

In order to control the mechanical speed of the motor, an outer speed loop driving the reference current I_{q_ref} is provided. The actual speed is compared with the reference speed and the error is processed with a speed controller. The output of the speed controller provides the reference value of the quadrature axis current for the inner quadrature current loop. The inner quadrature current loop process the quadrature current error through a current controller to provide the reference value of the quadrature axis voltage.

For operation in constant torque region the reference value of i_d is held at zero. The direct axis current is compared with its reference value and the resulting error is processed through the current controller to provide the reference value of the direct axis voltage. The reference direct axis and quadrature axis voltages are fed to the inverse Park transform block to carry out transformation from dq frame to αβ -frame. Thus, the outputs of the current controllers are passed through the inverse Park transform and a new stator voltage vector is impressed to the motor using the Space Vector Modulation technique.

3.2. Sensor less Vector Controlled PMSM Drive using High Frequency Signal Injection Technique

A constant amplitude voltage vector rotating at high frequency (500 to 1500 Hz) is superimposed on the fundamental voltage vector. Therefore a rotating HF current vector arises and which is superimposed on the fundamental current vector. High frequency carrier current response is extracted by a band pass filter from the measured machine currents and filter carrier current is demodulated to reconstruct the rotor position. Negative sequence carrier current contains rotor position information, but all these signals are hidden in the much stronger fundamental stator current and the switching frequency harmonics. Heterodyning demodulation method is popularly used to extract the rotor position signals from the measured stator currents. This work uses heterodyne technique [2] to extract the Rotor Position.

4. Torque Ripple Due to High Frequency Signal Injection

This work deals with the torque ripple analysis of rotating signal injection in αβ -reference frame whose block diagram is shown in Fig.6. A high frequency signal is superimposed to the fundamental voltage vector. The resulting current vector i_c will follow the injected rotating voltage vector with 90 °phase shift. However, the locus of the resulting high frequency carrier current response will not be exactly circular due to the stator inductance difference between d-axis and q-axis. Heterodyne demodulation method shown in Fig.7 is used in this work to derive the rotor position from carrier current response.

High frequency signal is injected to extract the rotor position information; but injected high frequency signal introduces torque ripple. This section derives analytical expressions for the torque ripple due to high frequency signal injection. Phase currents will have two components i.e fundamental component and high frequency component resulted from high frequency voltage injection and these current can be expressed as shown in (21) to (23)

$$i_a = I_f sin(w_r * t) + I_h sin(w_h * t) \quad (21)$$

$$i_b = I_f sin(w_r * t - 2\pi/3) + I_h sin(w_h * t - 2\pi/3) \quad (22)$$

$$i_c = I_f sin(w_r * t - 4\pi/3) + I_h sin(w_h * t - 4\pi/3) \quad (23)$$

Transforming three phase currents in to dq frame rotating at w_r will result in i_q *and* i_d as shown in (24) and (25)

$$i_d = I_h sin((w_h - w_r) * t) \quad (24)$$

$$i_q = I_f + I_h sin((w_h - w_r) * t - \pi/2) \quad (25)$$

Using (24) and (25) in (19) will get electromagnetic torque including high frequency components (26).

$$T_e = (k_T * I_f) + \left(k_T * I_h sin((w_h - w_r) * t - \pi/2)\right) - \left(\frac{3p\Delta L}{2}\left(I_f I_h sin((w_h - w_r) * t) - \frac{I_h^2}{2} sin(2(w_h - w_r) * t)\right)\right) \quad (26)$$

First term in (26) is constant torque component and remaining terms contributes to torque ripple.

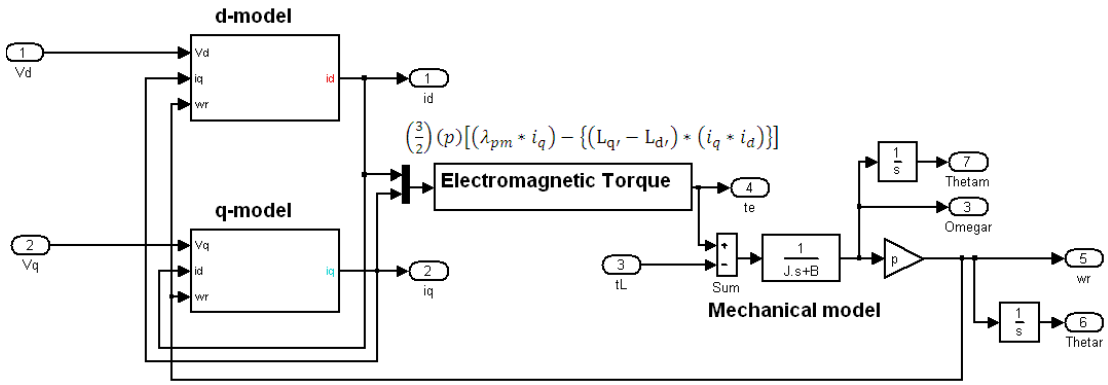
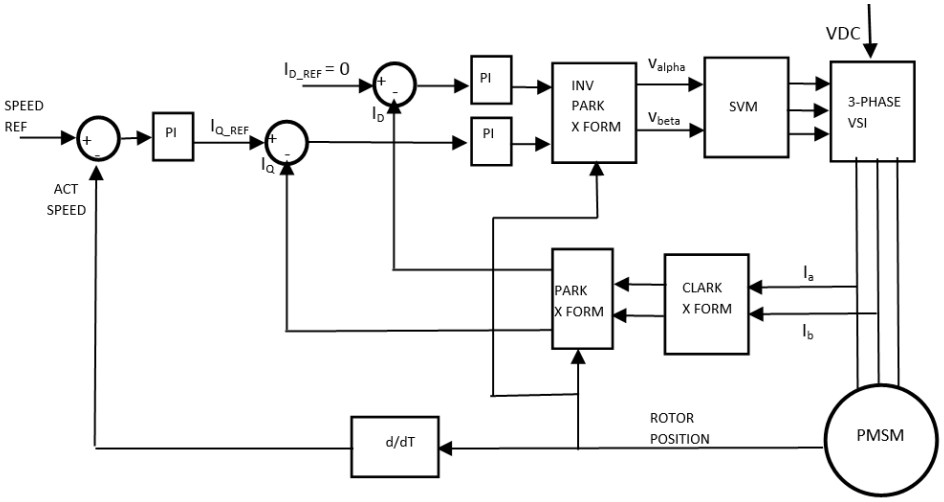
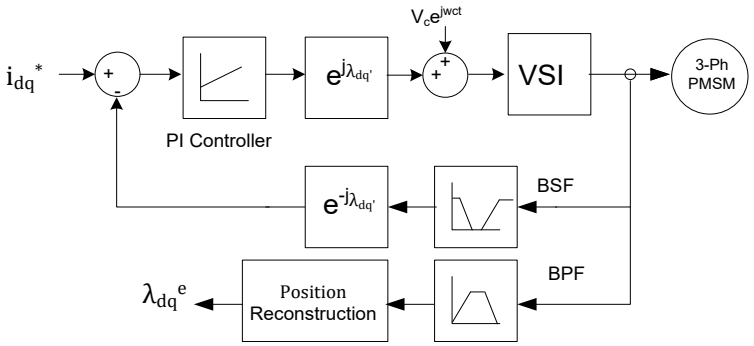


Fig.4. Complete PMSM model in dq reference frame.





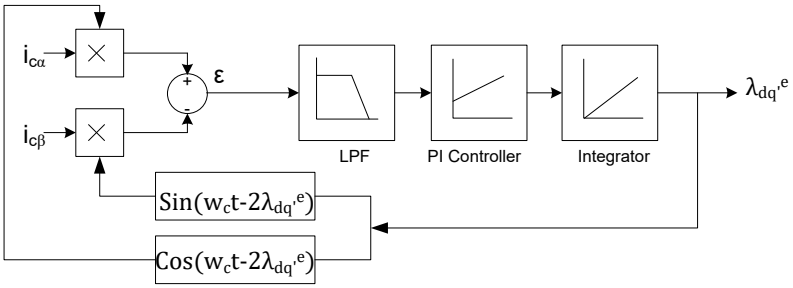


Fig.7.Hetero dyne observer using PI Controller to track rotor position

5. Simulation Results

First sensor less vector control system based on rotating high frequency signal injection is simulated using MATLAB/Simulink TM. , and same platform is used to study the torque ripple at different injection frequencies. Motor parameters used in the simulation are: Ld=16mH, Lq=20mH, 4 poles on rotor and V_{DC} =100V.

Fig.8 shows the phase currents, alpha beta currents, dq currents and output torque at 20Hz speed without signal injection, based on the technique presented in Fig.5. Torque ripple frequency is the difference of injected frequency and fundamental frequency and amplitude fairly direct proportional injection amplitude. Two fundamental currents are considered to compare the, effect of torque ripple at fixed signal injection frequency. Fig.9 and Fig.10 compare the torque ripple frequency at 20Hz and 200Hz fundamental frequency with injection frequency set to 1000Hz. From Fig.9 and Fig.10 it is clear that when the fundamental frequency comes in the order of injection frequency, phase currents distortion is much higher. Higher torque ripple increase the acoustic noise and vibration, so signal injection techniques should be used only for the low speeds.

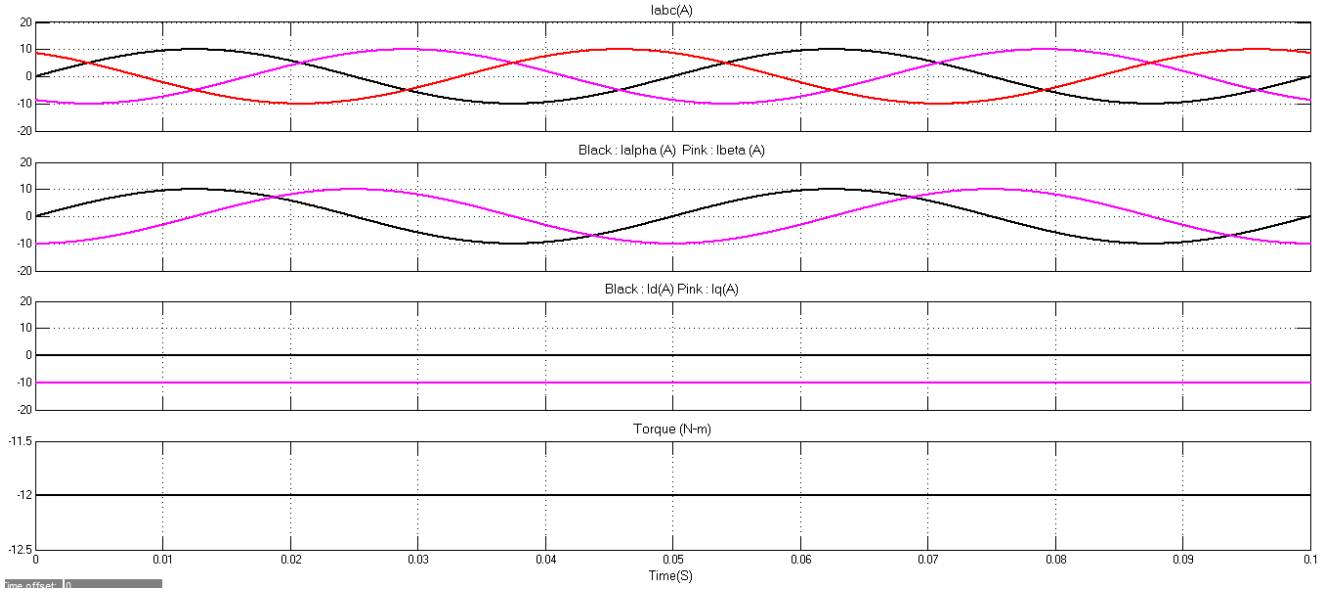


Fig.8. Phase currents, alpha beta currents, dq currents and output torque without any injected signal, stator current frequency = 20Hz

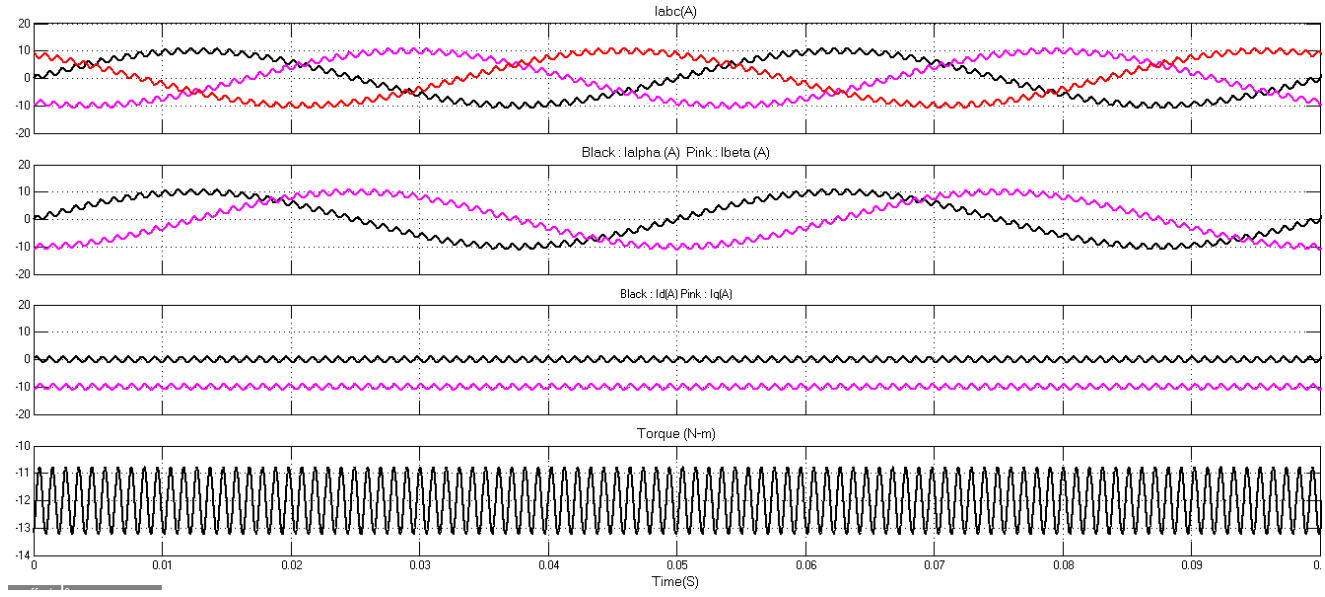


Fig.9. Phase currents, alpha beta currents, dq currents and output torque with injected signal, stator current frequency = 20Hz, Signal Injection frequency = 1000Hz, Signal amplitude = 0.1*Ipeak

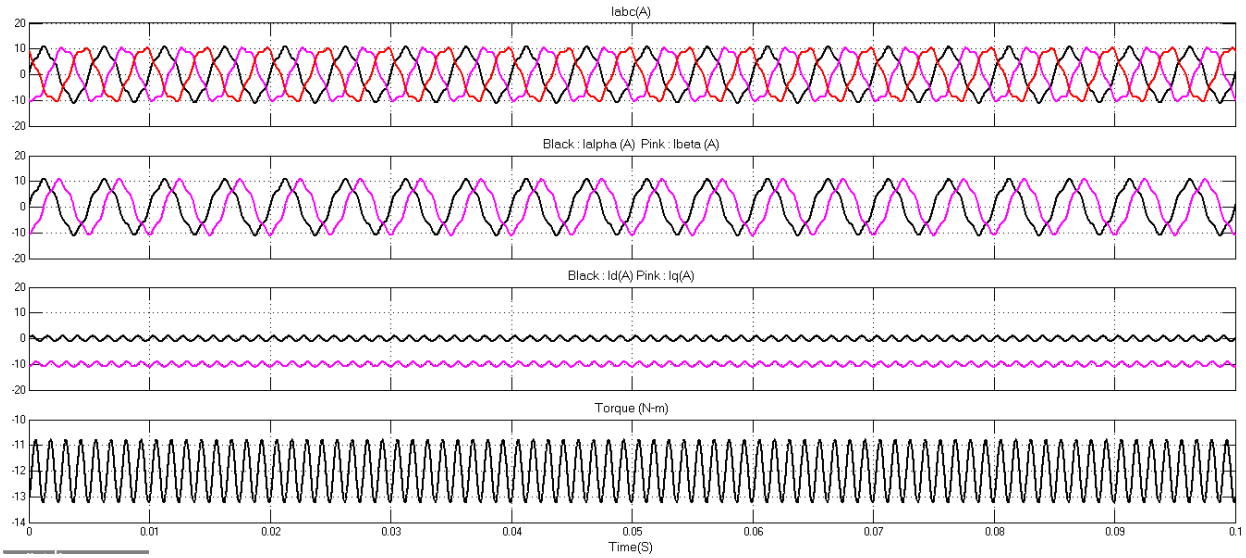


Fig.10. Phase currents, alpha beta currents, dq currents and output torque with injected signal, stator current frequency = 200Hz, Signal Injection frequency = 1000Hz, Signal amplitude = 0.1*Ipeak.

6. Conclusions

This work developed a detailed analysis on Torque Ripple of IPMSM drive based on rotating high frequency signal injection techniques. It is shown that Torque component contains a constant term and high frequency ripple. When the fundamental frequency comes close to the injection frequency, phase currents distortion is much higher and torque ripple also impact the drive operation. This analysis concludes high frequency signal injection method at high speed results in torque ripple and motor vibrations.

Conflict of Interest

The authors declare no conflict of interest.

References

[2] O. Benjak, D. Gerling, "Review of Position Estimation Methods for IPMSM Drives without a Position Sensor Part II: Adaptive Methods", in Conf. Rec., IEEE-ICEM, September 2010, pp. 1-6.

[3] T. Kim, H. Lee, and M. Ehsani, "State of the Art and Future Trends in Position Sensorless Brushless DC Motor/Generator Drives", Proc. of the 31th Annual Conference of the IEEE-IECON, November 2005, pp. 1718-1725.

[4] T. Kim, H. W. Lee, and M. Ehsani, "Position sensorless brushless DC motor/generator drives: Review and future trends," IET Elect. Power Appl., vol. 1, no. 4, pp. 557–564, Jul. 2007.

[5] D. Raca, P. Garcia, D. D. Reigosa, F. Briz, and R. D. Lorenz, "Carriersignal selection for sensorless control of PM synchronous machines at zero and very low speeds," IEEE Trans. Ind. Appl., vol. 46, no. 1, pp. 167–178, Jan./Feb. 2010.

[6] José Carlos Gamazo-Real , Ernesto Vázquez-Sánchez and Jaime Gómez-Gil " Position and Speed Control of Brushless DC Motors Using Sensorless Techniques and Application Trends" Journal of sensors , 2010, 10, 6901-6947.

[7] Roberto Leidhold, "Position Sensorless Control of PM Synchronous Motors Based on Zero-Sequence Carrier Injection" IEEE Transactions on Industrial Electronics, vol. 58, no. 12, December 2011

[8] N. Matsui, "Sensorless PM brushless DC motor drives," IEEE Trans.Industrial Electronics, vol. 43, no. 2, pp. 300-308, 1996.

[9] Hyunbae, K., Kum-Kang, H., Lorenz, R.D., Jahns, T.M.: 'A novel method for initial rotor position estimation for IPM synchronous machine drives', IEEE Trans. Ind. Appl., 2004, 40, (5), pp. 1369–1378

[10] Yu-seok, J., Lorenz, R.D., Jahns, T.M., Seung-Ki, S.: 'Initial rotor position estimation of an interior permanent-magnet synchronous machine using carrier-frequency injection methods', IEEE Trans. Ind. Appl., 2005, 41, (1), pp. 38–45

[11] Yi Li, Z. Q. Zhu, Fellow, IEEE, David Howe, Chris M. Bingham, Member, IEEE, and Dave A. Stone "Improved Rotor-Position Estimation by Signal Injection in Brushless AC Motors, Accounting for Cross-Coupling Magnetic Saturation", IEEE Trans. Ind. Appl, VOL. 45, NO. 5, SEPTEMBER/OCTOBER 2009 1843.

[12] S. Murakami, T. Shiota, M. Ohto, K. Ide, and M. Hisatsune, "Encoderless servo drive with adequately designed IPMSM for pulse-voltageinjection-based position detection," IEEE Trans. Ind. Appl., vol. 48, no. 6, pp. 1922–1930, Nov./Dec. 2012

[13] Sung Park, S. Hun Lee, C. Moon, and Y. Ahn Kwon, "State Observer with Stator Resistance and Back-EMF Constant Estimation for Sensorless PMBL MOTOR", IEEE Region 10 Conference, TENCON 2010, November 2010, pp. 31-36.

[14] F. De Belie, P. Sergeant, and J. A. Melkebeek, "A Sensorless Drive by Applying Test Pulses Without Affecting the Average-Current Samples", IEEE Transactions on Power Electronics, Vol. 25, No. 4, April 2010, pp. 875-888

[15] S. Bolognani, S. Calligaro, R. Petrella, and M. Tursini,"Senseorless Cotnrol of IPM Motors in the Low-Speed Range and at Standstill by HF Injection and DFT Processing", IEEE Trans. Ind. Appl. vol. 47, no. 1, pp. 96-104, Jan./Feb. 2011

[16] E. Robeischl and M. Schroedl, "Optimized INFORM measurement sequence for sensorless PM synchronous motor drives with respect to minimum current distortion," IEEE Trans. Ind. Appl., vol. 40, no. 2, pp. 591–598, 2004.

[17] Toliyat, HA, Hao, L, Shet, DS, and Nondahl, TA "Position-sensorless control of surfacemount permanent-magnet AC (PMAC) motors at low speeds", IEEE Transactions Industrial Electronics,2002, IE-49, No 1, pp 157-164

[18] Ravikumar Setty .A ,Shashank Wekhande and Kishore Chatterjee "Comparison of high frequency signal injection techniques for rotor position estimation at low speed to standstill of PMSM" IICPE 2012.

[19] Y. Hua, "Sensorless control of surface mounted permanent magnet machine using fundamental PWM excitaiton", PhD Thesis, Department of Electrical and Electronic Engineering, University of Nottingham, 2009.

[20] Y.S. Jeong, R.D. Lorenz, T.M. Jahns, and S.K. Sul, "Initial rotor position estimation of an interior permanent-magnet synchronous machine using carrier-frequency injection methods," IEEE Trans. Industry Applications, vol. 41, no. 1, pp. 38–45, 2005.

[21] Ravikumar Setty, A.; Wekhande, Shashank and Chatterjee, Kishore "Adaptive signal amplitude for high frequency signal injection based sensor less PMSM drives" IEEE International Symposium on Sensorless Control for Electrical Drives and Predictive Control of Electrical Drives and Power Electronics (SLED/PRECEDE) Munich, Germany ,17-19 Oct. 2013.

[22] Zhu, Z.Q., Gong, L.M.: 'Investigation of effectiveness of sensorless operation in carrier-signal-injection-based sensorless-control methods', IEEE Trans. Ind. Electron., 2011, 58, (8), pp. 3431–3439

[23] J. Holtz, "Acquisition of position error and magnet polarity for sensorless control of PM synchronous machines," IEEE Trans. Industry Applications,vol. 44, no. 4, pp. 1172–1180, 2008.

[24] Ravikumar Setty, A. ; Wekhande, Shashank and Chatterjee, Kishore "Compensation of rotor position estimation error due to stator winding resistance in signal injection based sensor less PMSM drives" IEEE International Symposium on Sensorless Control for Electrical Drives and Predictive Control of Electrical Drives and Power Electronics (SLED/PRECEDE) Munich, Germany ,17-19 Oct. 2013.

[25] Yan, Y., Zhu, J.G., Guo, Y.G.: 'Initial rotor position estimation and sensorless direct torque control of surface-mounted permanent magnet synchronous motors considering saturation saliency', IET Electr.Power Appl., 2008, 2, (1), pp. 42–48.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 214-226 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

Real Time Implementation of an Improved Hybrid Fuzzy Sliding Mode Observer Estimator

Sorin Mihai Radu[1], Elena-Roxana Tudoroiu[2], Wilhelm Kecs[2], Nicolae Ilias[1], Nicolae Tudoroiu[3*]

[1]*Mechanical and Electrical Engineering Faculty, Mechanical Engineering, University of Petrosani, 332006, Romania*

[2]*Science Faculty, Mathematics and Informatics, University of Petrosani, 332006, Romania*

[3]*John Abbott College, 2127 Lakeshore Road, Sainte-Anne-de-Bellevue, QC, H9X 3L9, Canada*

ARTICLE INFO

Article history:
Received: 18 December, 2016
Accepted: 20 January, 2017
Online: 28 January, 2017

Keywords :
Fuzzy Sliding Mode Observer
Fault Detection and Isolation
Residual Generation
Dc Servomotor Angular Speed

ABSTRACT

This paper extends some of our research results disseminated in the most recent awarded international conference paper concerning the implementation in real time of a sliding mode observer state estimator. For the same case study developed in the conference paper, more precisely a DC servomotor angular speed control system, we extend the proposed concept of sliding mode observer state estimator to a fuzzy sliding mode observer version, more suitable in control applications field such as fault detection of the possible faults that might take place inside the actuators and sensors. The hybrid architecture implemented in a real time MATLAB/SIMULINK simulation environment consists of an integrated control loop structure with a switching bench of two sliding mode observers, one built by using a new approach that improves slightly the proposed sliding mode observer for the conference paper, and second one is an improved intelligent fuzzy version sliding mode observer estimator. The both estimators are implemented in SIMULINK to work independently by using a manual switch. The simulation results for the experimental setup show the effectiveness of the improved fuzzy version of sliding mode observer compared to the standard one, as well as its high accuracy and robustness.

1. Introduction

This paper is an extension of the results disseminated in the most recent international conference paper relating to the real time implementation of a sliding mode observer (SMO) state estimator, integrated in a direct current (DC) servomotor angular speed control system [1]. The main goal of the extended version paper is to investigate new directions for improving the SMO state estimator performance in terms of state estimation accuracy and robustness to the changes in the noise levels, initial conditions values of the estimates, input disturbance and modeling errors uncertainties. The improved version of the SMO estimator is a fuzzy version of SMO (FSMO), a real helpful tool for our future developments in the real time control applications field, namely several real time fault detection and isolation (FDI) control strategies based on various estimation techniques. The purpose of these FDI strategies is to detect and isolate the possible malfunction components inside the actuators and/or faulty sensors as vital control system components frequently prone to errors. The real-time simulations are carried out on the MATLAB/SIMULINK platform, that has special real time implementation features provided by its extensions Real-Time Workshop (RTW) and the Real-Time Windows Target (RTWT). Furthermore, the real-time dc servomotor angular speed control proposed in our case study can be easily interfaced with MATLAB/SIMULINK or AnyLogic multi-paradigms hybrid simulator in the case of Real-Time Unified Modeling Language (UML-RT) implementations with a fast response [2]-[4].

2. Fault Detection and Isolation Schemes Using Real Time State Estimators

The DC servomotor is mostly used as an actuator in feedback closed-loop control systems, but in this research for simulation

*Corresponding Author: John Abbott College, 2127 Lakeshore Road, Sainte-Anne-de-Bellevue, QC, H9X 3L9, Canada, Email: ntudoroiu@gmail.com
www.astesj.com
https://dx.doi.org/10.25046/aj020126

purposes it is considered in the same time as a controlled plant, similar with the one shown in Figure 1 [2]-[4]. One of the main goals of this closed-loop structure might be to control its angular speed or its position either or both. Nowadays, the DC servomotors are extensively used in the common control applications due to their high start torque characteristics, high response performance, and their speed much easier to be controlled by varying the input voltage, compared to those that need expensive frequency drivers [1].

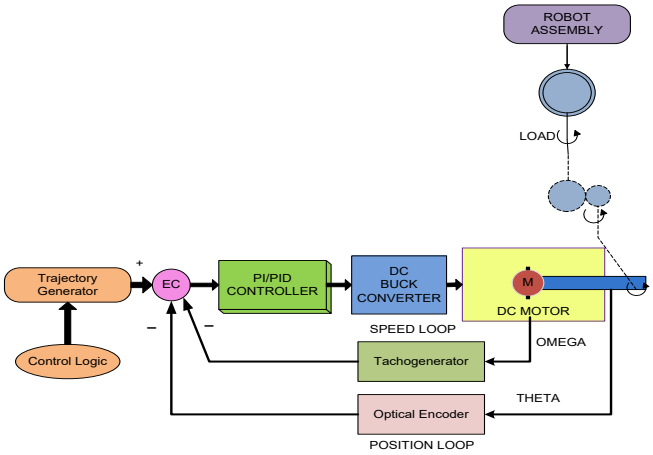


Figure 1: The schematic diagram of the closed-loop control system of the dc servomotor angular speed (Reproduced from [4]).

The components of the DC servomotor actuators during operation experience several possible critical failures that could compromise its performance and cause severe gear damage, such as, armature coil opening, brushes failures, field coil opening, armature static converter short circuit, field static converter short circuit, armature coil short circuit, field coil short circuit, cooling system failure, lack of bearings and bushing lubrication, armature current sensor failure, field current sensor failure, offset on supplied voltage or speed sensor failure [5].

Whenever these critical situations come out the control systems could lose the control, require much more energy, and could operate harmfully. Therefore to operate in real-time at high energy efficiency and to guarantee the equipment safety and reliability it is important to develop suitable FDI strategies capable to detect and diagnose any time every faulty control system components and consequently corrective and reconfiguration actions should be initiated promptly [1]. Fault detection and isolation schemes are implemented as real-time algorithms based on various state estimation techniques that require the input-output measurements data set of the control plant. They are basically used firstly for fault detection to decide whether the plant is in a normal operating condition or in a faulty one, and secondly for fault isolation to point out and identify the kind of the fault, if it is present, among a given faults covering set [1], [5]. Effectively the existing methods to identify and to adjust the equipment failures are mostly labor-intensive task, and consequently sustained, rhythmic and error-prone [1]. In the most of these situations the windings currents are recorded to determine the health of the DC servomotors currents compared to statistical evaluation that necessitates considerable human knowledge, hence error-prone that could generate severely equipment operation [1]-[6]. In these conditions the problem of control systems monitoring and fault diagnosis becomes a critical issue, of high complexity that need to be implemented in real-time environment by using more sophisticated control systems and artificial intelligence estimation strategies [1]. Usually, the control systems design must include FDI issues at their very early design stage with the ultimate goal to reach a fault-tolerant control (FTC) environment [5], [7]. Following the FDI diagnosis, on-line procedures are usually needed for FTC purpose, while off-line procedures could be used for maintenance purpose [5], [7]. To see the strong link between the FDI real time algorithms and the state estimation techniques is worth to know that a FDI strategy is implemented as a two-steps procedure based on the plant data set of input-output measurements. In the first real time procedure step is implemented the fault detection or is generated an alarm to decide whether the system is in a normal operating condition or not using an estimation algorithm, such as Kalman filters estimators [6], [8]-[9] or Luenberger linear and nonlinear sliding mode observers [10]-[16]. In our approach we develop a hybrid structure for state estimation, namely a fuzzy version of sliding mode observer nonlinear estimator, FSMO similar to the approach developed in [17]-[18]. In the second procedure step the isolation tasks or alarms interpretation are required for right decision about which faults chosen from a pre-defined set of faults that cover almost all the possible situations (fault isolation) [5], and therefore to find the faults locations based on an implemented logic, as in our future extension approach (second part of this research paper) that will develop for this purpose an intelligent fuzzy logic classifier [10]. The set of output measurements along with a previously obtained knowledge of the system constitute the algorithm inputs while a set of generated alarms are the algorithm outputs [5]. It is also worth to say that accuracy and the robustness of these estimation techniques are crucial for successful real time implementation of all FDI strategies. Furthermore, these estimation techniques are useful to establish the fault characteristics such as occurrence time, its severity, and the recovery time. The FDI estimation algorithms performance is the ultimate analysis issue that must consider also the accumulate evaluation errors included at every step of the FDI problem solution. This research work is based on our previous experience in control systems implementation, very useful to prove the effectiveness of the improved version of real-time SMO estimator implementation, a simple fuzzy version, FSMO. It will be a new practical approach to build a hybrid control structure that integrates in the same control loop the FDI control strategy together with a FSMO estimator and a fuzzy logic classifier (FLC). Summarizing, the main objective of this research is to improve the performance of the real time SMO estimator developed in the most recent paper conference [1] using a new approach. The new SMO estimator design is an improved fuzzy version of SMO estimator, very simple to be implemented in real-time, more efficient, accurate, robust and consistent. The improved FSMO estimator is a useful tool capable to be applied in our future developments in control applications field, as an extension of these research results to find solutions for different control applications, such as for example the most attractive FDI control strategies.

3. Dynamics Model of Dc Servomotor

For simulations purpose to implement our proposed real time FDI strategy of actuator (input) and sensor (output) faults we consider the same permanently excited DC servomotor with a rated power of $P =$ 550 W at rated speed n = 2500 *rpm*, described in detail in [7]. It consists of two-pair brush commutation, two pole pairs, and an analog tachometer for speed measurement and operates against a hysteresis brake as load, as shown in Figure 2 [7].

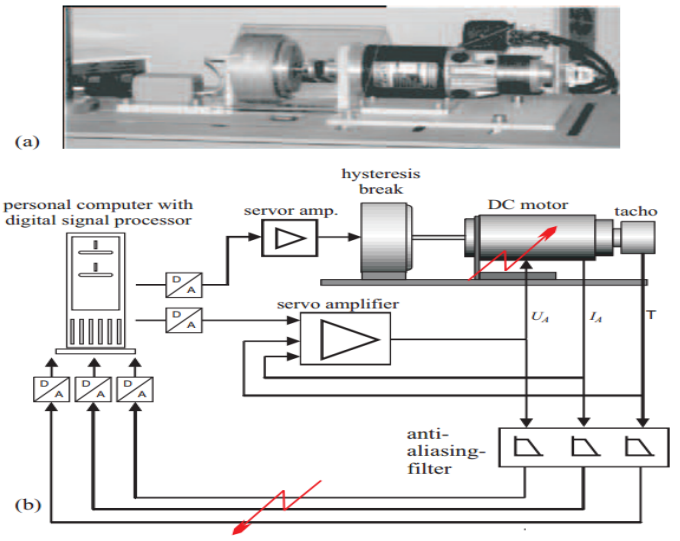


Figure 2: The Dc servomotor test bench with hysteresis brake

a) The experiment set-up b) The equipment block scheme (screenshot from [7])

The measured signals of the angular speed control system are the armature voltage $V[V]$, the armature current $I_a[A]$ and the angular speed $\omega\ [\frac{rad}{s}]$. The closed-loop equipment includes in forward path an analog proportional servo amplifier and the dc servomotor actuator as a controlled plant. In the feedback path is integrated an analog tachometer transducer, three anti-aliasing filters (AFs), three Analog to Digital converters (ADC) as an interface between an input personal computer (PC) and the AF block, a second analog servo amplifier with pulse-width-modulated (PWM) armature voltage as output and speed and armature current as feedback that allows to be built a cascaded speed control system [7]. In Figure 2(b) is shown the flow diagram of the three measured signals that first pass through analog anti-aliasing filters and after they are processed by a digital signal processor (e.g. TXP 32 CP, 32-bits, 50 MHz) and a desktop Intel Pentium host PC [7]. Also the hysteresis brake is controlled by a pulse-width servo amplifier. Typically such dc servomotors can be described by linear dynamic models [1], [7], [20].

However, the simulation results of the experiments carried out have shown that these linear models with constant parameters do not match the process in the entire operational range [7]. For that reason, the block scheme of dc servomotor represented in complex domain ($s \in C$) must include two nonlinearities such that the dynamic model is capable to capture all the process dynamics, as is shown in Figure 3 [7]. The dynamics of the dc servomotor actuator is described by the following input-state-output equations [1], [20]:

$$J\frac{d^2\theta}{dt^2} + b\frac{d\theta}{dt} = k_t \times I_a - T_L \quad (1)$$

$$L_a\frac{dI_a}{dt} + R_a I_a = u_A - e \quad (2)$$

where $T_e = T = k_t I_a$ is the dc servomotor torque developed on the shaft, $T_L = M_L = M_{F_0} sign(\frac{d\theta(t)}{dt})$ is the load torque, Ψ is the magnetic flux, as is shown in Figure 3 and Table 2 [7]. In addition by defining the armature voltage as $u_A = U_A^* - K_B|\frac{d\theta(t)}{dt}|I_a$, as is shown in Figure 3, and the DC servomotor counter electromotive force, e by:

$$e = \psi \times \frac{d\theta}{dt} = k_e \times \omega \quad (3)$$

leads to a linear description of its dynamics given in (1), (2) [7].

Furthermore, in a state-space representation the DC servomotor actuator dynamics is described by following equations:

(i) State Equation [1],[7], [20]:

$$\begin{bmatrix}\frac{dx_1}{dt}\\ \frac{dx_2}{dt}\end{bmatrix} = \begin{bmatrix}-\frac{b}{J} & \frac{k_t}{J}\\ -\frac{k_e}{L_a} & -\frac{R_a}{L_a}\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} + \begin{bmatrix}0\\ \frac{1}{L_a}\end{bmatrix}u + \cdots$$

$$+\begin{bmatrix}-\frac{M_{F_0} sign\left(\frac{d\theta(t)}{dt}\right)}{J} & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}w_1\\ w_2\end{bmatrix} \quad (4)$$

$$x_1 = \omega, x_2 = I_a, u = u_A, x_1(0) = 1\left[\frac{rad}{s}\right], x_2(0) = 0[A] \quad (5)$$

where $w = \begin{bmatrix}w_1\\ w_2\end{bmatrix}$ represents a two-norm bounded noise vector which stands for the uncertainties and the unknown inputs affecting a practical DC servomotor [7].

Output equation:

$$y = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} \quad (6)$$

The measurement vector, $y(t)$ is assumed to be completely known and measurable at each sample time, t. The unknown disturbance, T_L is bounded, i.e. there exists a positive real value, such that the symbols $\|.\|$ and $\overline{T_L}$ stand for the two-norm of noise vector and the upper bound value of T_L, respectively.

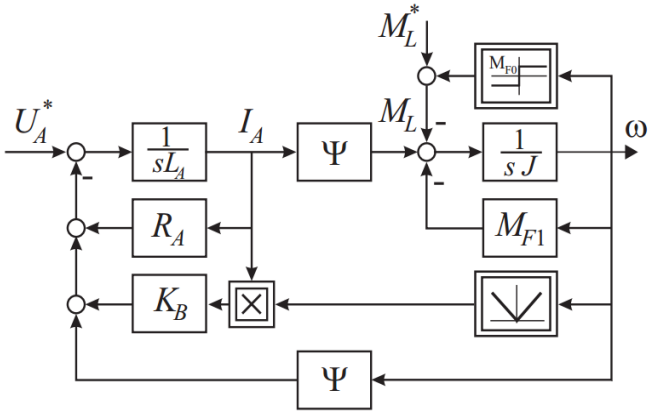


Figure 3: The block scheme of the Dc servomotor in complex domain (screenshot from [7])

3.1. Abbreviations and Acronyms and measure units

To simplify the presentation we include in this subsection the list of all the abbreviations and the acronyms used in the paper text, as is shown in Table 1.

Also in the text we use only the international standard (SI) as primary units, as is shown in Table 2

Table 1: The list with the abbreviations and the acronyms used in text

Item	Acronyms/ Abbreviations	Significance
1	FDI	Fault Detection and Isolation
2	FTC	Fault Tolerant Control
3	PC	Personal Computer
4	PWM	Pulse-Width-Modulation
5	AF	Anti-aliasing Filter
6	ADC	Analog to Digital Converter
7	SMO	Sliding Mode Observer
8	FSMO	Fuzzy Sliding Mode Observer
9	FIS	Fuzzy Inference System
10	FLC	Fuzzy Logic Classifier
11	KF	Kalman Filter
12	RTW	Real Time Workshop
13	RTWT	Real-Time Windows Target
14	UML-RT	Real Time Unified Modeling Language
15	SI	International System of Units
16	DC	Direct Current

Table 2: The list with the SI units abbreviations used in the text

Name	Symbol	SI base Unit	Expressed in other SI units	Name of SI unit
Current	I	A	-	ampere
Voltage	V	V	-	volt
Time	t	s	-	second
Magnetic flux	ψ	W_b	Vs	weber
Viscous friction coefficient (Damping ratio of the mechanical system)	$M_{F_1}(b)$	-	$Nm \times s$	-
Moment of rotor inertia	J	-	kgm^2	-
Dry friction coefficient	M_{F_0}	-	Nm	-
Counter electromotive force coefficient	k_e ψ	-	Nm/A	-
Electromotive force	e	V	-	volt
Proportional coefficient of the motor torque	k_t ψ		Nm/A	-
Electric resistance of the armature	R_a	Ω	V/A	Ohm
Electric inductance of the armature	L_a	H	Ω×s	Henry
Angular speed	ω	-	rad/s	-
Load torque	T_L	-	Nm	-
Motor torque	T_M		Nm	-
Electric Power	P	W	$V \times A$	watt

The nominal values of the DC servomotor model coefficients have the same values as in [7], as are given in Table 3.

Table 3: The nominal values of the dc servomotor dynamic model

Parameter Name	Symbol	Value	Measure Unit
Electric resistance of the armature	R_a	1.52	Ω
Electric inductance of armature	L_a	6.82×10^{-3}	H (Ω×s)
Magnetic flux	ψ	0.33	W_b $(V \times s)$
Voltage drop factor	K_B	2.21×10^{-3}	$V \times s/A$
Moment of rotor inertia	J	1.92×10^{-3}	kgm^2
Viscous friction coefficient	$M_{F_1}(b)$	0.36×10^{-3}	$Nm \times s$
Viscous dry friction	M_{F_0}	0.11	Nm

Replacing the values given in Table 3 in the dc servomotor dynamic model represented by (4)-(6) we get the nominal system equations of the dc servomotor dynamics in a state space representation of the following form:

(i) State Equation [8],[13], [19]:

$$\begin{bmatrix} \frac{dx_1}{dt} \\ \frac{dx_2}{dt} \end{bmatrix} = \begin{bmatrix} -\frac{M_{F_1}}{J} & \frac{\psi}{J} \\ -\frac{\psi}{L_a} & -\frac{R_a}{L_a} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{L_a} \end{bmatrix} u + G(x_1) \begin{bmatrix} W_1 \\ W_2 \end{bmatrix} \quad (7)$$

$$x_1 = \omega, x_2 = I_a, u = u_A, x_1(0) = 1\left[\frac{rad}{s}\right], x_2(0) = 0[A] \quad (8)$$

(ii) Output equation:

$$y = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \omega \\ I_a \end{bmatrix} \quad (9)$$

Or, more compact:

$$\begin{bmatrix} \frac{dx_1}{dt} \\ \frac{dx_2}{dt} \end{bmatrix} = A_{n\times n} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + B_{n\times p} u + G_{n\times n}(x_1) \begin{bmatrix} W_1 \\ W_2 \end{bmatrix} \quad (10)$$

$$y = C_{m\times n} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (11)$$

where n represents the system state dimension, $n = 2$, p is the number of inputs, $p = 1$,and m is the number of outputs, $m = 2$. The load torque T_L is considered as a bounded input uncertainty included in the last term, $G(x_1)\begin{bmatrix} W_1 \\ W_2 \end{bmatrix} = M_{F_0} sign(x_1(t)) \begin{bmatrix} W_1 \\ W_2 \end{bmatrix}$ of (7). The noise vector components $w = \begin{bmatrix} W_1 \\ W_2 \end{bmatrix}$ are generated by a SIMULINK "Band-Limited White Noise" block that generates normally distributed random numbers that are suitable for use in continuous or hybrid systems of different spectral density powers,

e.g., 1 and 0.1 to analyze also the robustness of the hybrid structure SMO - FMSO estimators to the different noise levels.

The matrices triplet $(A_{n\times n}, B_{n\times p}C_{m\times n})$ is defined by:

$$A_{2\times 2} = \begin{bmatrix} -\frac{M_{F_1}}{J} & \frac{\psi}{J} \\ -\frac{\psi}{L_a} & -\frac{R_a}{L_a} \end{bmatrix} = \begin{bmatrix} -0.1875 & 171.8750 \\ -48.3871 & -222.8739 \end{bmatrix},$$

$$B_{2\times 1} = \begin{bmatrix} 0 \\ \frac{1}{L_a} \end{bmatrix} = \begin{bmatrix} 0 \\ 146.6276 \end{bmatrix}, C_{1\times 2} = [1 \quad 0],$$

$$G_{2\times 2}(x_1) = \begin{bmatrix} -57.2917 sign\left(\frac{d\theta(t)}{dt}\right) & 0 \\ 0 & 0 \end{bmatrix}$$

An observability test reveals that both outputs (ω and I_a) can also observe each other, i.e.,

$$\text{rank(Q)} = \text{rank}\left([C_{1\times 2} \;\; C_{1\times 2}A_{2\times 2}]^T\right) = 2 \tag{12}$$

where Q is the observability matrix of full rank. The matrix $A_{2\times 2}$is a stable matrix, i.e. a strictly Hurwitz matrix with its eigenvalues placed to $\chi_1 = -47.6497$ and $\chi_2 = -175.4117$.

4. Sliding Mode Observer for Linear DC Servomotor with Model and Disturbances Uncertainties

Basically the design of any SMO estimator for a nonlinear uncertain system can be simplified by considering its linearized model, similar to those described by (10)-(11), proposed for sate estimation and written in the following most general form [7]:

$$\frac{d\hat{x}}{dt} = A\hat{x} + Bu + L(y - \hat{y}) + Msign(y - \hat{y}) \tag{13}$$

$$\hat{y} = C\hat{x} \tag{14}$$

where, the estimated state vector is denoted by $\hat{x}$, and $(y - \hat{y})$ is the residual of the output signal of the control system. L represents an appropriate designed linear gain matrix, similar to Luenberger Observer gain matrix for linear systems [11], [13] [18] and M is an appropriate designed nonlinear gain matrix that multiplies the common sign function, as in [1] and [16], [18]. In reality they represent two key tuning parameters to control the SMO performance. In Equation (13) its last $Msign(y - \hat{y})$ term is added to compensate the effects of disturbance input (load torque) and modeling uncertainty of the system (10). Therefore, the states of the system can be estimated using data given by the measured input-output data set, $(u(t), y(t))$ of the control system. Several design methods have been used to determine the observer linear gain L, ensuring the stability of the proposed SMO estimator [18]. Among these design methods in [18] is mentioned the most popular Kalman Filter (KF) algorithm that can be applied to find the estimation gain K_L as a good approximation of the gain matrix L for the designed SMO linear estimator (13). Closing, the proposed SMO linear estimator (13) includes two key terms: one of them is the Kalman filter term $K_L(y - \hat{y})$ as a suitable approximation of the Luenberger observer term, $L(y - \hat{y})$, and the other one is the discontinuous *sign* term, $Msign(y - \hat{y})$, where their corresponding gains K_L and M are designed separately [18]. Therefore, the proposed observer design (6) becomes the mixed Kalman Filter - SMO estimator (KF-SMO), whose dynamics is described by the following single input-single output (SISO) output equation:

$$\frac{d\hat{x}}{dt} = A\hat{x} + Bu + K_L(y - \hat{y}) + Msign(y - \hat{y}) \tag{15}$$

$$\hat{y} = C\hat{x} = x_1 = \omega$$

By some manipulations of the matrices A, B, C we can easily find that B, C have a full rank, and also the pair (A, C) we found in section 2 that is observable, as main requirements assumed in [14]-[16], [18].

By replacing (16) in (15), the dynamics of SMO observer is described by the following equation [18]:

$$\frac{d\hat{x}}{dt} = (A - K_L C)\hat{x} + Bu + K_L y + Msign(y - \hat{y}) \tag{16}$$

where the matrix $A_0 = A - K_L C$ concentrates all the dynamics of SMO estimator, so it has to be stable (strictly Hurwitz) such that the state estimate, $\hat{x}$ to converge to a finite value in a finite time [14]-[16], [18]. The matrix A_0 is strictly Hurwitz if all its eigenvalues are situated strictly in the half left complex plane, $\lambda_i \in \mathbb{C}_- | \; \forall i \in Z_+$. The dynamics of the SMO estimator (15) is described by the following first order differential equations:

$$\frac{d\,\hat{x}_1(t)}{dt} = A_{11}\hat{x}_1(t) + A_{12}\hat{x}_2(t) + B_{11}u(t) + K_{L1}e_y(t) + \cdots + M_1\, sign\big(e_y(t)\big) \tag{17}$$

$$\frac{d\hat{x}_2(t)}{dt} = A_{21}\hat{x}_1(t) + A_{22}\hat{x}_2(t) + B_{21}u(t) + K_{L2}e_y(t) + \cdots + M_2 sign\big(e_y(t)\big) \tag{18}$$

where

$$e_y(t) = e_1(t) = x_1(t) - \hat{x}_1(t),\; e_2(t) = x_2(t) - \hat{x}_2(t) \tag{19}$$

represent the states residuals of the DC servomotor. The observer linear gain matrix K_L is chosen in order to make the spectrum of the matrix $(A - K_L C)$ to lie in $\mathbb{C}_-$ [14]-[16], [18]. In simulations we consider two cases, one case for eigenvalues close to origin of the complex plan, for a slow transient, and second case for eigenvalues far from origin of the complex plan, for fast transient, very important in fault detection. Without to lose the generality we can select for the sliding mode observer the following linear gain matrices, setting the matrix components as $K_L(2) = 48.371$, and for $K_L(1)$ any value greater than -0.1875 that guarantees the stability of the matrix $(A - K_L C)$, e.g.,

$$K_L = \begin{bmatrix} 1 \\ 48.371 \end{bmatrix}, \text{ and } K_L = \begin{bmatrix} 100 \\ 48.371 \end{bmatrix}.$$

We analyze in this case the accuracy of the hybrid structure SMO-FSMO estimators corresponding two different poles locations of control system, given as the eigenvalues of the matrix $A_0 = A - K_L C$. The main goal remains now to design the nonlinear gain matrix M such that the discontinuous term, $Msign(y - \hat{y})$ overcomes the parametric uncertainties assuring a stable dynamics of SMO estimator error. The observer nonlinear gain matrix $M = \begin{bmatrix} M_1 \\ M_2 \end{bmatrix}$ has to guarantee a bounded error dynamic for the output residuals (19) satisfying the following gain condition [18]:

$$\bar{M} \geq \bar{G}\,\bar{w} \tag{20}$$

Also the gain condition (20) assures that the SMO residuals e_1, e_2 converge to zero. More precisely, the two-norms of the known term $G(x)$ and the SMO gain matrix M are bounded to corresponding upper bounds, $\bar{G}$ and $\bar{M}$ respectively [18].

According to (20) taking into account the noise levels for gain matrix M we will consider in our simulations the following values $M = \begin{bmatrix}1\\1\end{bmatrix}$, and $M = \begin{bmatrix}10\\10\end{bmatrix}$ respectively.

$$\frac{d\,\hat{x}_1(t)}{dt} = A_{11}\hat{x}_1(t) + A_{12}\hat{x}_2(t) + B_{11}u(t) + K_{L1}e_y(t) + \cdots + \bar{M}sign\big(e_y(t)\big) \quad (21)$$

$$\frac{d\hat{x}_2(t)}{dt} = A_{21}\hat{x}_1(t) + A_{22}\hat{x}_2(t) + B_{21}u(t) + K_{L2}e_y(t) + \cdots + \bar{M}sign\big(e_y(t)\big) \quad (22)$$

4.1. Sliding Mode Observer Simulation Results

For simulation results purpose in the experimental set-up the input voltage profile is set to $u(t) = 84\ [V]$ in the first 0.5 seconds and to $u(t) = 24\ [V]$ on the last 0.4 seconds. Also the spectral density power values of the noise will be set in first case to 1 and for second case will be chosen ten times smaller to 0.1 in order to be able to see the accuracy of the both estimators SMO, and FSMO, The problem design of the sliding mode observer is solved by using one of the most powerful tools, such as MATLAB/SIMULINK software package.

The SIMULINK model of the nominal system is shown in Figure 4, and the evolution of the DC servomotor states, i.e. angular speed (x_1) and armature current (x_2) are shown in Figure 5 and Figure 6. The input profile voltage is shown in Figure 7. In Figure 8 is presented in detail the DC servomotor subsystem nominal model. In Figure 9 is shown the overall view of the SMO SIMULINK diagram of the state estimator. The Figure 10 and Figure 11 show the dynamic evolution of the SMO armature current residual and on the same graph the both model and SMO armature current estimate. Similar for SMO angular speed is shown in Figure 12 and Figure 13. In Figure 14 is shown the switch control function on the sliding surface, to analyze the chattering effects of the sliding mode on the control system efforts to keep the evolution of the states on the trajectory.

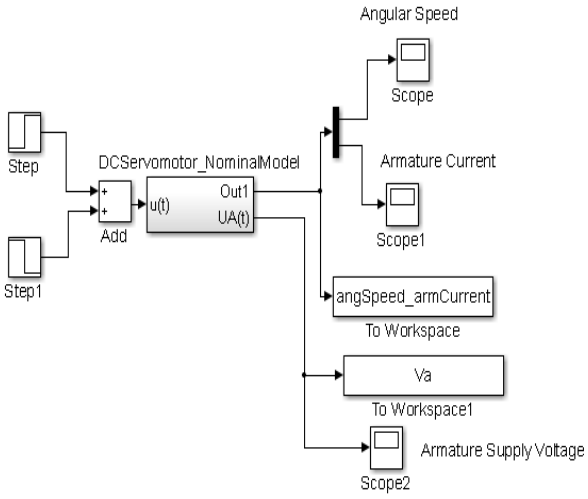


Figure 4: The overall SIMULINK diagram of DC servomotor nominal model

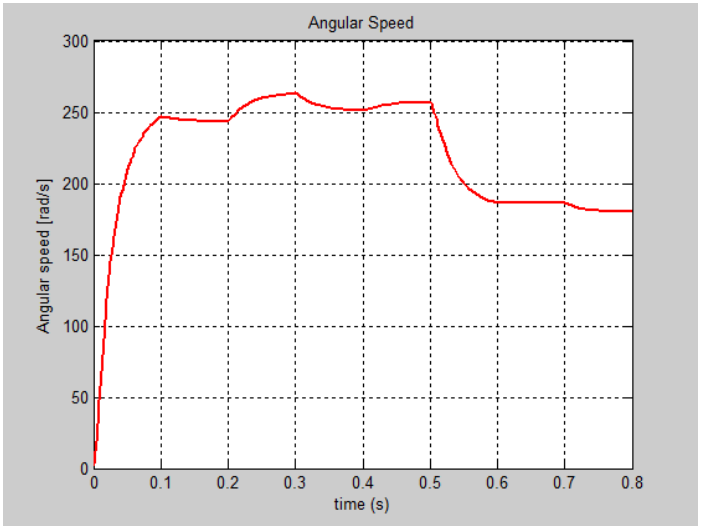


Figure 5: The MATLAB simulations of the evolution of nominal DC Servomotor angular speed

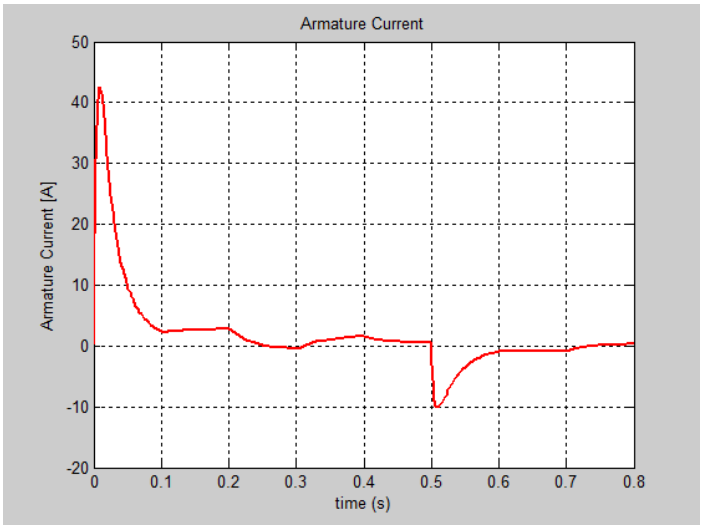


Figure 6: The MATLAB simulations of the evolution of nominal DC Servomotor armature current

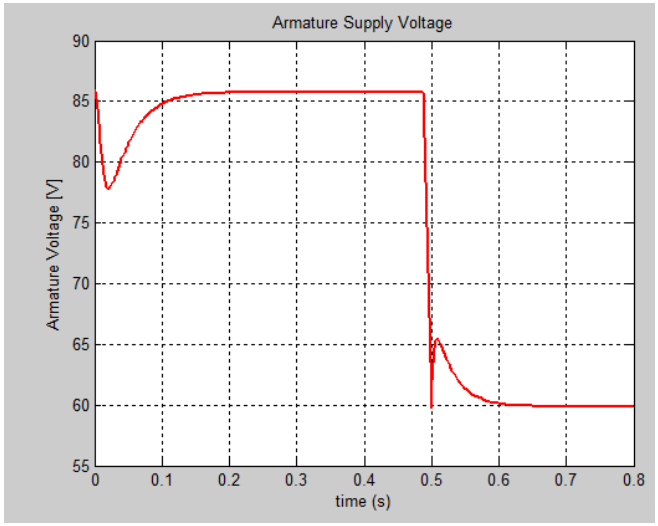


Figure 7: The input Voltage profile in SIMULINK for the DC Servomotor nominal model

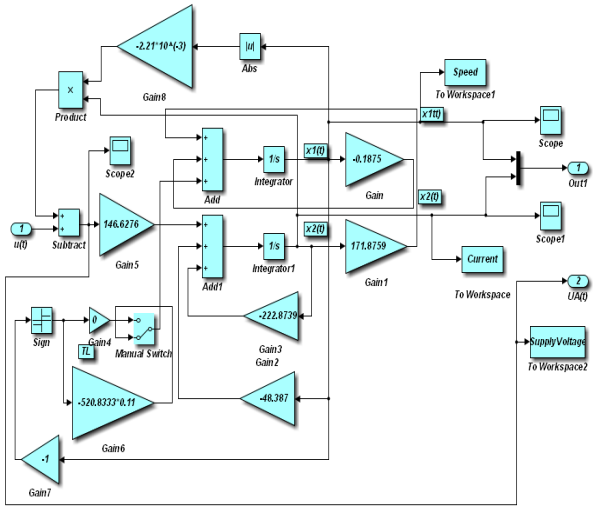


Figure 8: The detailed subsystem SIMULINK diagram of the DC servomotor nominal model

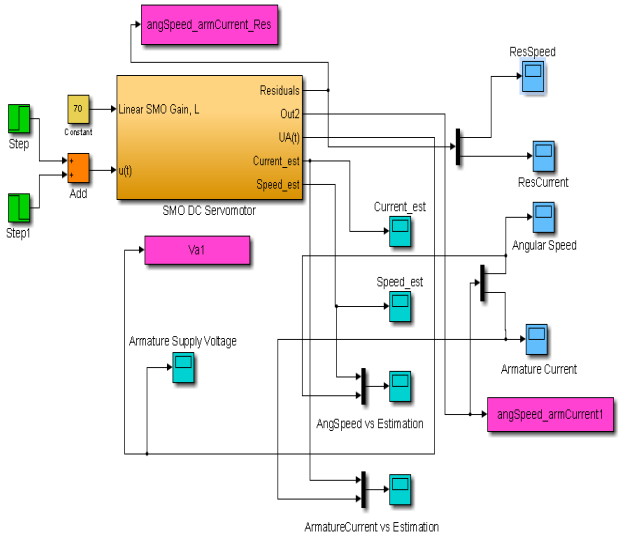


Figure 9: The overall SMO SIMULINK diagram of the DC servomotor model

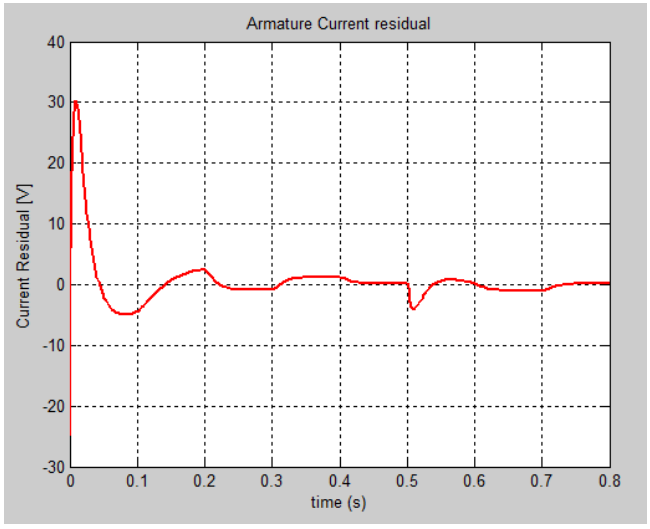


Figure 10: The SMO MATLAB simulations of armature current residual

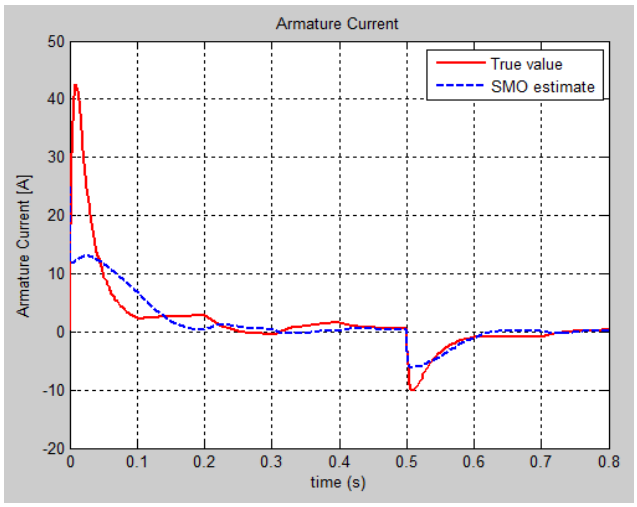


Figure 11: The SMO MATLAB simulations of armature current estimate versus true value

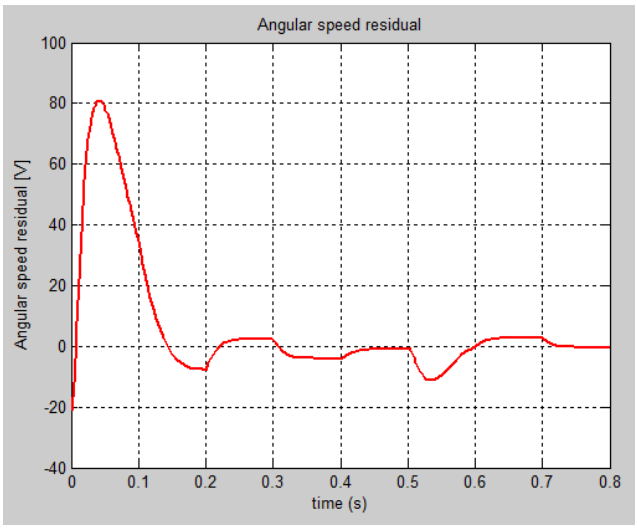


Figure 12: The SMO MATLAB simulations of angular speed residual

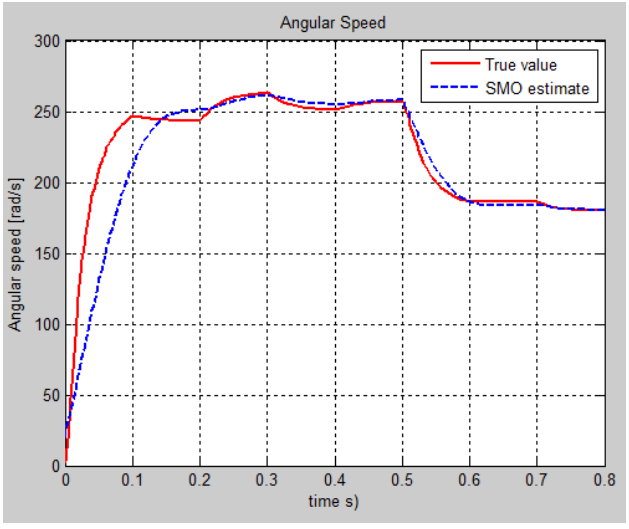


Figure 13: The SMO MATLAB simulations of angular speed estimate versus true value

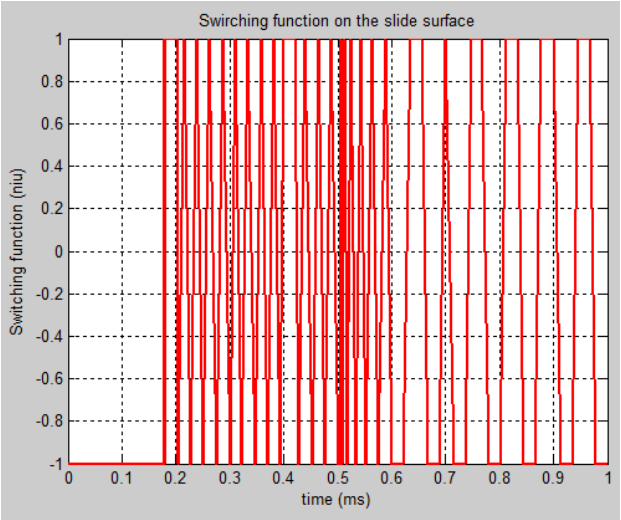


Figure 14: The SMO MATLAB simulations of SMO chattering evolution around the sliding surface

5. Fuzzy Sliding Mode Observer

5.1. The Dynamic Model of Fuzzy Sliding Mode Observer

In the most of practical situations the sliding mode systems experience several difficulties due to the chattering effects; therefore this inconvenient represents one of their main drawbacks [18]. The chattering phenomenon is well visible in Figure 14 that is undesirable because it involves high control activity and furthermore may excite high frequency unmodeled dynamics. The chattering effects induced by SMO estimator design with a high impact on the overall dynamics of the control system may be attenuated using different improvements into the original observer design.

The second method that is a fuzzy SMO approach is mentioned also in [18]. The basic idea is a new interpretation of the chattering effects seen as free estimations that may be achieved using linguistic variables instead of fixed numerical values. Thus, to improve the performance of the standard SMO developed in previous section is required some knowledge provided by an expert, so a new design approach known as Fuzzy logic SMO is developed [18]. The new intelligent FSMO estimation approach has the capability to maintain the robustness property of the clean standard SMO while the chattering phenomenon is significantly decreased. The dynamics of the fuzzy sliding mode observer is described by a similar equation (16) with slight changes by replacing the *sign* term $sign(y-\hat{y})$ by a linguistic variable, so a fuzzy term F_L [18]:

$$\frac{d\hat{x}}{dt} = A\hat{x} + Bu + K_L(y-\hat{y}) + MF_L \tag{23}$$

where the crisp output of the FSMO, F_L is computed through the designed *if - then* rule-base considering the tracking errors:

$$e_y = y - \hat{y} \tag{24}$$

$$\frac{de_y}{dt} = \frac{dy}{dt} - \frac{d\hat{y}}{dt} \tag{25}$$

as input variables for fuzzy inference system (FIS) [18].

Compared to KF-SMO estimator the $sign(y-\hat{y})$ term of the observer (16) is replaced by a fuzzy output variable F_L of the FIS to construct the fuzzy estimator FSMO (23). The hybrid structure of the SMO and FSMO estimators is build in SIMULINK and shown in Figure 15, with the main blocks detailed in the Figure 16 to Figure 19. Since the *if-then* rules of the fuzzy system are generated according to the properties of *sign* term, the FSMO is expected to be a robust observer [18]-[19]. The fuzzy *if-then* rules perform a nonlinear mapping from the input linguistic variables e_y and $\frac{de_y}{dt}$ to the output linguistic variable, F_L as:

$$F_L = FSMO(e_y, \frac{de_y}{dt}) \tag{26}$$

5.2. Fuzzy Logic and Inference System Description Block

Fuzzy Logic (FL) is a computational paradigm that is based on the expert knowledge. Fuzzy Logic looks at the world in vague terms, almost in the same way that our brain takes in information (e.g. pressure is high, speed is slow, concentration is small, temperature is freezing), then responds with precise actions [19]. According to the same reference document… "the human brain can reason with uncertainties, vagueness, and judgments, and the computers can only manipulate precise valuations; therefore, fuzzy logic is an attempt to combine these two techniques". Closing, unlike the false perception that "fuzzy" is a misnomer that has resulted in the mistaken suspicion that fuzzy logic is somehow less exacting than traditional logic however the practice reality has proved that the FL is in fact, a precise problem-solving methodology [19]. It is able also to simultaneously handle numerical data and linguistic knowledge. Fuzzy logic is a technique that facilitates the control of a complicated system without knowledge of its mathematical description, making the development and implementation of the control systems much simpler [19]. It requires no complicated mathematical models, only a practical understanding of the overall system behavior.

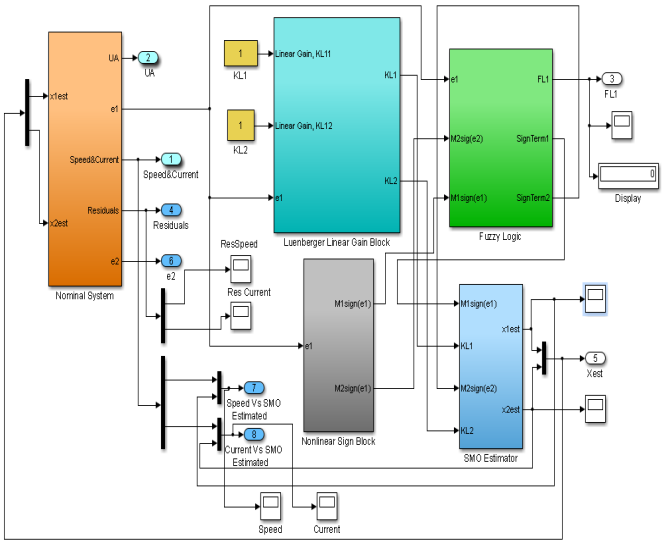


Figure 15: The SIMULINK hybrid structure of SMO and FSMO estimators' models

Fuzzy logic differs from classical logic where an object takes on a value of either zero or one. In fuzzy logic, a statement can assume any real value between 0 and 1, representing the degree to which an element belongs to a given set [19]. Fuzzy Logic mechanisms can also lead to higher accuracy and smoother control, dealing with degrees of truth and degrees of membership. Most things in nature cannot be characterized with simple or convenient shapes or distributions. Membership functions characterize the fuzziness in a fuzzy set, whether the elements in

the set are discrete or continuous, in a graphical form for eventual use in the mathematical formalisms of fuzzy set theory [19]. They can be of different shapes such as triangular, trapezoidal, singleton, sigmoidal function, Gaussian distribution, Wavelet functions, etc. The detailed hybrid structure SMO and FSMO – Fuzzy SIMULINK Block model is shown in Figure 16, including the manual switch from one structure to another one, all other blocks remaining common for the both structures.

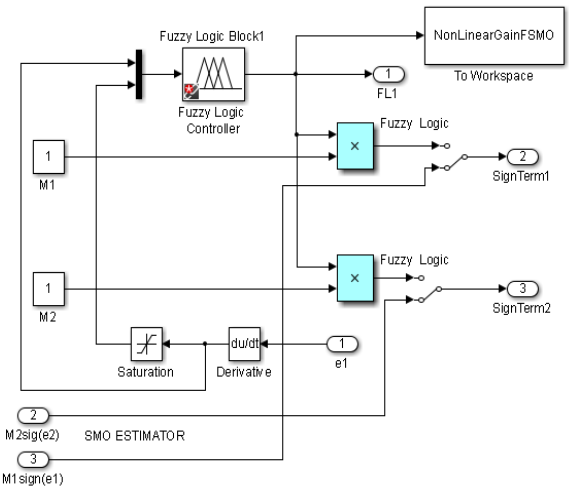


Figure 16: The SIMULINK hybrid structure SMO and FSMO –Fuzzy logic block model

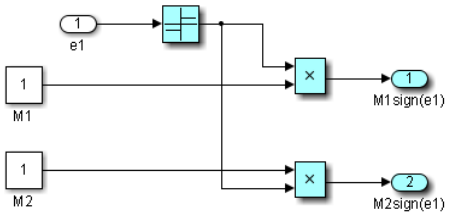


Figure 17: The SIMULINK hybrid structure SMO and FSMO –The gain matrix sign term Block

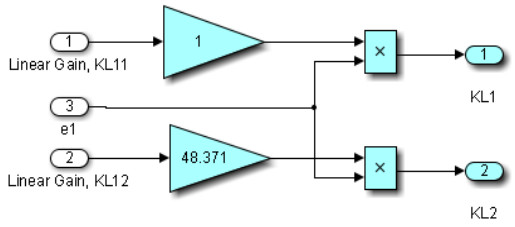


Figure 18: The SIMULINK hybrid structure SMO and FSMO –Linear gain matrix term block

In [18] to design the proposed FSMO estimator is chosen three membership functions of triangular shape corresponding to the input and output fuzzy sets of $e_y, \frac{de_y}{dt}$ and F_L as is shown in Figure 20. The linguistic labels used such as *P*, *N*, *ZE*, *S*, *M* and *L* stand for *positive*, *negative*, *zero*, *small*, *medium* and *large*, respectively. All seven possible combinations of two labels represent the input-output fuzzy set $\Sigma_F = \{NB, NS, NM, ZE, PS, PM, PB\}$ attached to each membership function and denote a *negative-big, negative-small, negative-medium, zero, positive-small, positive-medium, and positive-big*.

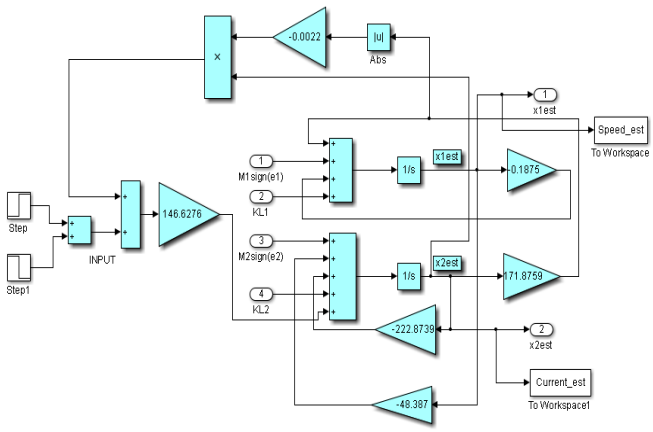


Figure 19: The SIMULINK hybrid structure SMO and FSMO –SMO Estimators Block

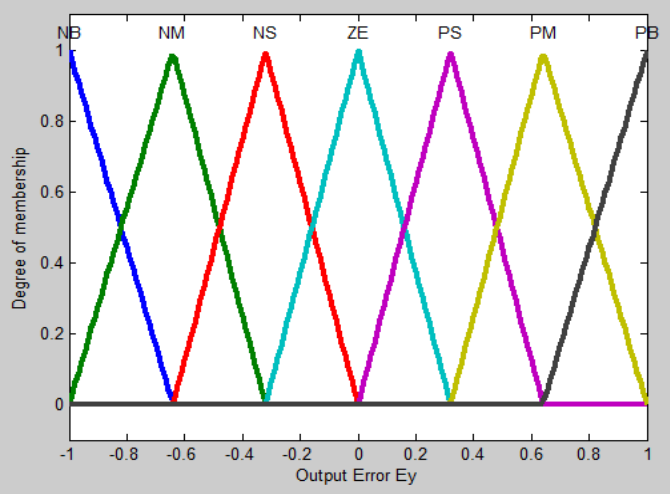


Figure 20: MATLAB Membership functions of input/output fuzzy sets $e_y, \frac{de_y}{dt}$ and F_L [18]

Three main steps are involved to create a Fuzzy Inference System:

- **Fuzzification** that translates inputs into truth values
- **Rule evaluation** to compute output truth values
- **Defuzzyfication** that transfers truth values into output

In fuzzyfication step the inputs from a set of sensors (or features of those sensors) are mapped to values from 0 to 1 using a set of input membership functions, more precisely all the input variables are assigned degrees of membership in various classes [18]-[19]. In the rule evaluation step these inputs are applied to a set of *if-then* control rules. For example, to design the proposed FSMO estimator (23) a simple fuzzy rule Table 4 is constructed considering the following reaching and stability requirements [18]:

1. The output fuzzy set, F_L is normalized in the interval [-1,+1], therefore:

$$|F_L = FSMO(e_y, \frac{de_y}{dt})| \leq 1 \quad (27)$$

2. Anytime when $e_y(\frac{de_y}{dt})$ becomes a positive value, the membership function of F_L is set in such a way that its

sign becomes similar to that of e_y and therefore $e_y \times F_L \geq 0$

3. While $e_y(\frac{de_y}{dt})$ has a negative value, the reaching condition$e_y(\frac{de_y}{dt}) \leq 0$would be satisfied automatically. Therefore, for all the seven membership functions attached to each of two input variables $(e_y, \frac{de_y}{dt})$ of the fuzzy rule base, 49 *if-then* rules in Table 4 are obtained using an expert engineering knowledge in the electric drives machines field and satisfying the reaching conditions 1, 2 and 3. Fuzzy rules are always written in the following form:

If (**input1** is membership function1) *and/or*

(**input2** is membership function2) *and/or* ….

Then (**output** is output membership function).

Table 4: Rule bases of fuzzy logic sliding mode observer [7]

e_y / $\frac{de_y}{dt}$	NB	NM	NS	ZE	PS	PM	PB
NB	NB	NB	NB	NB	NM	NS	ZE
NM	NB	NB	NB	NM	NS	ZE	PS
NS	NB	NB	NM	NS	ZE	PS	PM
ZE	NB	NM	NS	ZE	PS	PM	PB
PS	NM	NS	ZE	PS	PM	PB	PB
PM	NS	ZE	PS	PM	PB	PB	PB
PB	ZE	PS	PM	PB	PB	PB	PB

In many instances, it is desired to come up with a single crisp output from a FIS. This crisp number is obtained in a process known as defuzzification. In defuzzification step the fuzzy outputs are combined into discrete values needed to drive the control mechanism. There are two common techniques for defuzzifying:

- **Center of mass** - This technique takes the output distribution and finds its center of mass to come up with one crisp number, as is shown in [19].
- **Mean of maximum** - This technique takes the output distribution and finds its mean of maxima to come up with one crisp number, as is shown in [19].

More details about Laypunov's stability and how to find an upper bound for the crisp value of output variable F_L for the proposed FSMO (26) are given in [18].

5.3. Fuzzy Logic Sliding Mode Observer Simulation Results

We present in this section only the graphs with the evolution of the true values of the states and their FSMO estimated values, as well as the both residuals needed for developing the future fault detection and isolation strategies. The MATLAB simulations of armature current FSMO estimate versus the true values and its residual are shown in Figure 21 and Figure 24 respectively, and for angular speed in Figure 22 and Figure 23. During the simulations we didn't find significant changes when we switch the gain matrix M values from $M = \begin{bmatrix}1\\1\end{bmatrix}$, and $M = \begin{bmatrix}10\\10\end{bmatrix}$, the simulation results remaining the same. For changes in the noise level from 0.1 spectral density powers to 10, so increased 100 times, we found also that FSMO is very accurate, so very robust to these changes, as we can see from the angular speed armature current residuals, as is shown in Figure 25 and Figure 26 respectively.

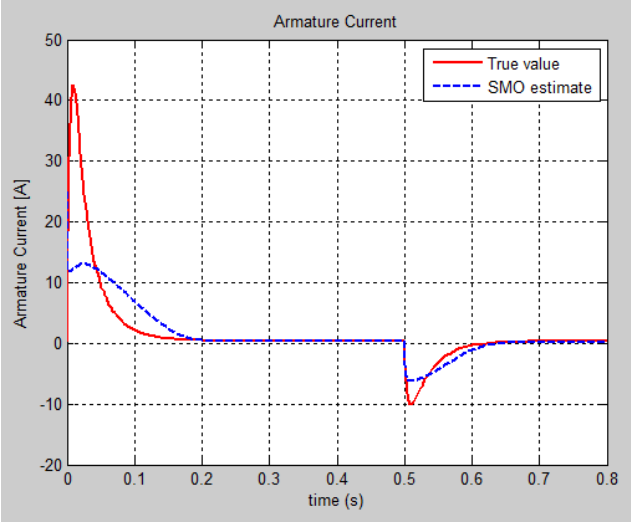


Figure 21: The SMOF simulations of MATLAB armature current estimation

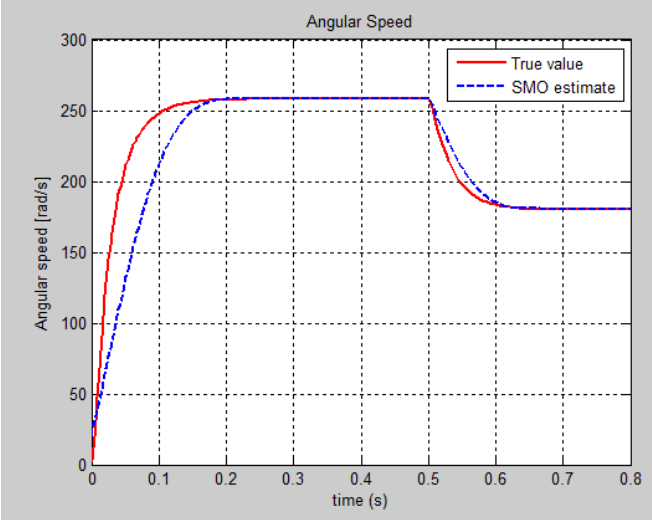


Figure 22: The SMOF simulations of MATLAB angular speed

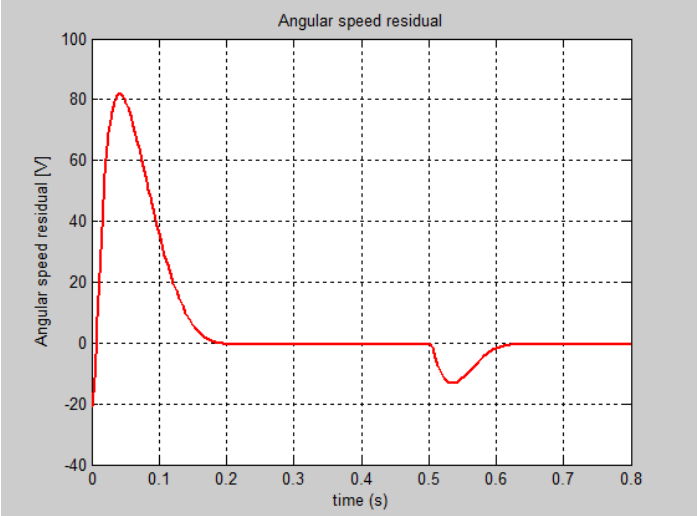


Figure 23: The SMOF MATLAB simulations of angular speed residual

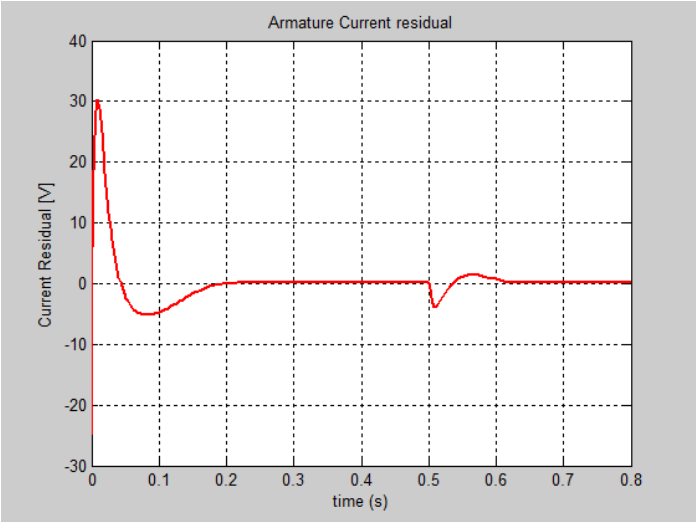


Figure 24: The SMOF MATLAB simulations of armature current residual

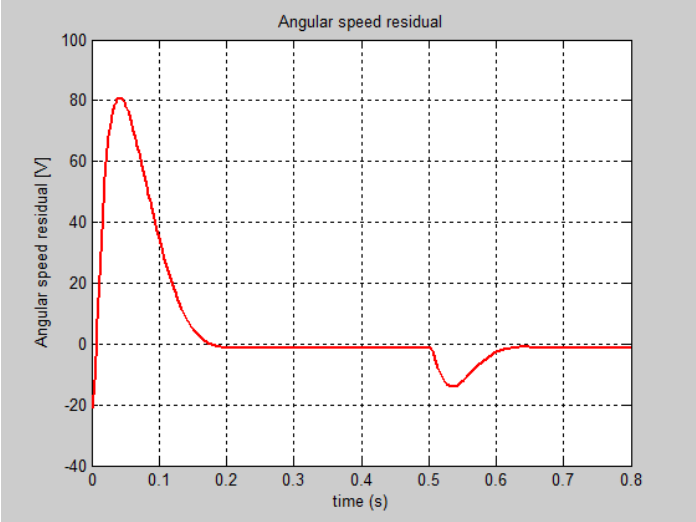


Figure 25: The SMOF MATLAB simulations of angular speed residual

Furthermore we tested the impact of changing the initial condition values of the estimated angular speed from 1$\left[\frac{rad}{s}\right]$ to -20 $\left[\frac{rad}{s}\right]$, and for armature current from 0 $[A]$ to -2.5 $[A]$ and we get the results shown in Figure 26 and Figure 27. We found also that FSOM algorithm is very robust to changes in the initial conditions values of the estimates.

Furthermore we tested also the accuracy of FSMO estimator for changes in the linear gain matrix coefficients from $K_L = \begin{bmatrix} 1 \\ 48.371 \end{bmatrix}$, as in all previous figures from this section, to $K_L = \begin{bmatrix} 100 \\ 48.371 \end{bmatrix}$ that lead to the results shown in Figure 28 and Figure 29. These changes leads also to changes in the estimator dynamics, the poles of the estimators change the location $p_1 = -1.1875$, p_2 = -222.8739 to another location given by $p_1 = -100.1875$, p_2 = -222.8739. In the new location the FSMO estimator becomes more faster.

Closing, a high accuracy for FSMO estimator can be obtained by a suitable tuning of the gain matrices M and K_L.

Compared the simulation results of the both estimators SMO and FSMO it is obviously that FSMO estimator performs better than SMO estimator in terms of accuracy and robustness, so FMSO is an improved version of SMO. This improvement is due to the fact that the chattering effects are reduced considerably, attenuating extremely the effect of the last term of (16), $Msign(y - \hat{y})$.

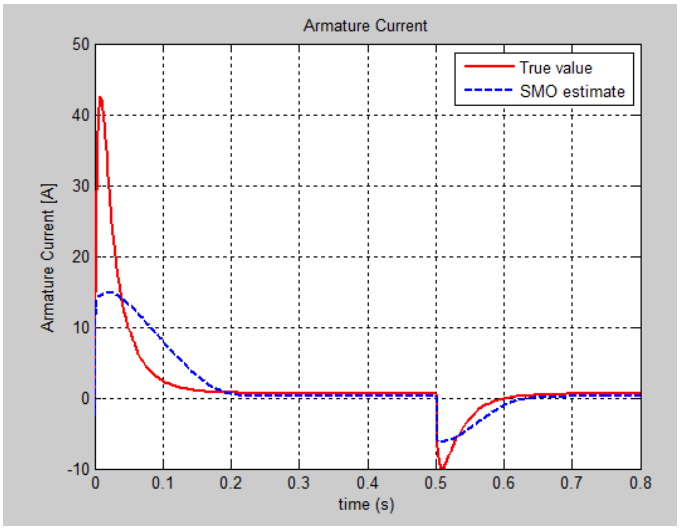


Figure 26: The SMOF simulations of MATLAB armature current estimation with new initial condition for estimated current (Robustness test)

All these results are encouraging for us to investigate a new approach to build more suitable fault detection and isolation control strategies in many control applications, using for estimation the developed hybrid structure with SMO and its improved version, a fuzzy sliding mode observer estimator (FSMO).

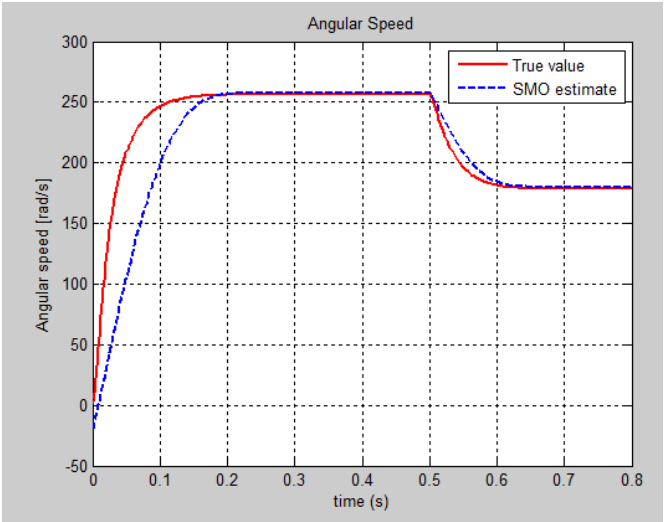


Figure 27: The SMOF simulations of MATLAB angular speed estimation with new initial condition for estimated current (Robustness test)

5.4. Summary of FSO Steps Design

To compute the output F_L (26) of the FIS' FSMO estimator given the inputs, one must go through six steps [19]:

Step1: Determining a set of fuzzy rules

Step2: Fuzzifying the inputs using the input membership functions

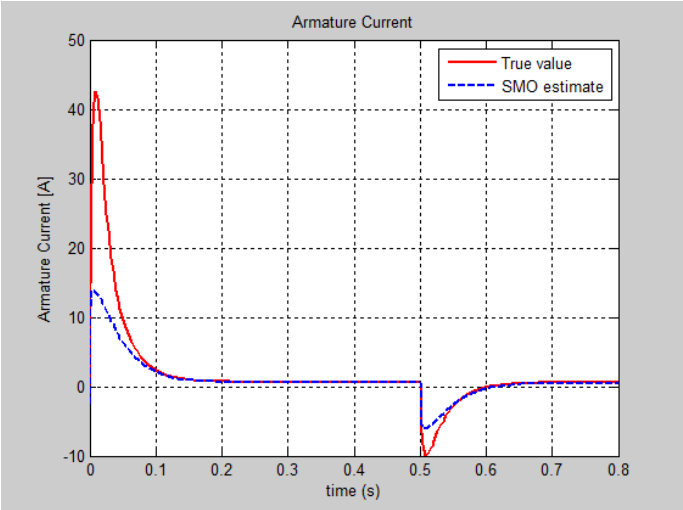


Figure 28: The SMOF simulations of MATLAB armature current estimation by changing the linear gain matrix coefficients (Robustness test)

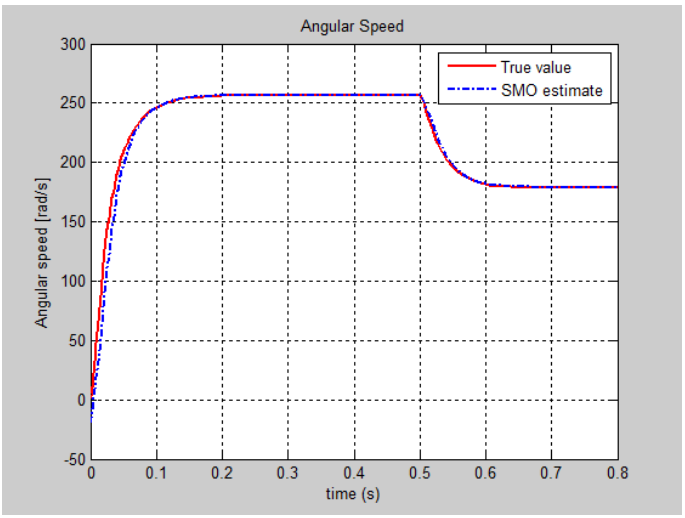


Figure 29: The SMOF simulations of MATLAB angular speed estimation by changing the linear gain matrix coefficients (Robustness test)

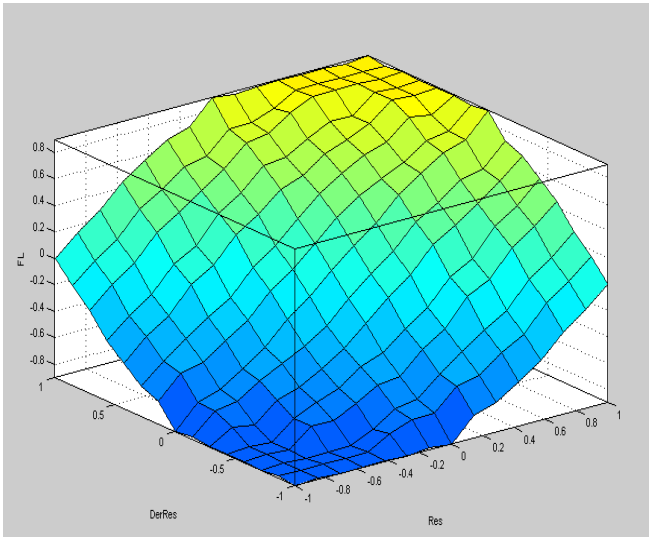


Figure 30: SIMULINK surface view of fuzzy sets $e_y(Res)$, $\frac{de_y}{dt}(DerRes)$ and F_L [18]

Step3: Combining the fuzzified inputs according to the fuzzy rules to establish rule strength

Step4: Finding the consequence of the rule by combining the rule strength and the output membership function (if it's a Mamdani FIS),

Step5: Combining the consequences to get an output distribution, and

Step6: Defuzzifying the output distribution (this step applies only if a crisp output class) is needed), as is shown in Figure 30.

6. Conclusions

In this paper, we developed an improved version of a Sliding Mode Observer strategy design with a considerably changes of the approach followed in the conference paper [1] but using the same study case, namely a DC servomotor actuator with disturbance uncertainty that is integrated in the same control system structure. Encouraged by the results obtained in [1] we have investigated a new approach direction to improve the performance of the proposed SMO in terms of its accuracy, robustness and implementation design. The improved version of the SMO that we developed for the same single-input single-output control system is a fuzzy SMO estimator with a high estimation accuracy of the actuator states, and with a strong capability to attenuate the chattering effects of the sliding mode control design. It is also more robust to the changes in noise levels and in initial conditions values (guess values) of the estimates, and also to the input disturbance and modeling errors. In addition, its implementation design simplicity in real time recommends FSMO estimator without doubt as one of the most useful control strategy tool to be applied for our future developments in fault detection and isolation control applications. The design of new FDI control strategies will be an extension of our research in the future work based on the preliminary results obtained until now in real implementation and control design estimation techniques. Implementation in real time a Sliding Mode Observer for a linear DC Servomotor actuator with bounded disturbance uncertainty.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] R-E. Tudoroiu, W. Kecs, M. Dobritoiu, N. Ilias, S.V. Casavela, N. Tudoroiu, "Real-Time Implementation of DC Servomotor Actuator with Unknown Uncertainty using a Sliding Mode Observer" in Proceedings of the Federated Conference on Computer Science and Information Systems, Gdansk, Poland, DOI: 10.15439/2016F95, ACSIS, ISSN 2300-5963, Vol. 8, pp. 841–848, 2015.

[2] R.E. Tudoroiu, A. Astilean, T. Letia, N. Iliasi, N.Tudoroiu, D. Burdescu, "Wireless UML-RT Simulator for Modeling and Implementing Dynamics Hybrid Structure of a Real Time PI DC Motor Control System", The Journal of Economics and Technologies Knowledge , 1(4), 69-81, 2015.

[3] R.E. Tudoroiu, A. Astilean, T. Letia, V. Cretu, N. Iliasi, N. Tudoroiu, "UML-RT Simulator for Modeling and Implementing Hybrid Structure Dynamics of a Real Time PID DC Servomotor Position Control System Strategy", The Journal of Economics and Technologies Knowledge , 1(5), 68-77, 2015.

[4] R-E Tudoroiu, "Conceiving and Implementing Applications using Real-Time UML", PhD Thesis, Cluj-Napoca Technical University, Romania, 2012.

[5] J. M. Galvez, "A New Fault Detection and Isolation Algorithm Applied to Dc Motor Parameters Supervision", ABCM Symposium Series in Mechatronics , 5, 196-204, 2012, http://www.abcm.org.br/symposium-series/SSM_Vol5/Section_II_Control_Systems/07492.pdf

[6] N. Tudoroiu, K. Khorasani, "Satellite Fault Diagnosis using a Bank of Interacting Kalman Filters", Journal of IEEE Transactions on Aerospace and Electronic Systems, DOI: 10.1109/TAES.2007.4441743, 43(4), 1334-1350, 2007.

[7] R. Isermann, Fault-Diagnosis Applications, Model-Based Condition Monitoring: Actuators, Drives, Machinery, Plants, Sensors, and Fault Tolerant Systems, DOI 10.1007/978-3-642-12767-0_3, Springer-Verlag Berlin Heidelberg, 2011.

[8] S. Haykin, Adaptive Filter Theory, 3rd ed., Prentice-Hall, Upper Saddle River, NJ., 1996

[9] G.L. Plett, G.L., "Extended Kalman filtering for battery management systems of LiPB-based HEV battery packs Part Background, Journal of Power Sources, ELSEVIER, 134, 252–261, 2004.

[10] H. I. Abbas, F.B. Hawraa, "Fault detection and isolation based on hybrid sliding mode observer and fuzzy logic", Kufa Journal of Engineering, Iraq, ISSN 2207-5528, 6(1), 93-102, 2014.

[11] A. Alkaya, I. Eker, "Luenberger observer-based sensor fault detection: online application to DC motor", Turkish Journal of Electrical Engineering and Computer Sciences, DOI: 10.3906/elk-1203-84, 22, 363-370, 2014.

[12] W. Chen, "Robust Fault Diagnosis and Compensation in Nonlinear Systems via Sliding Mode and Iterative Learning Observers", PhD Thesis, Simon Frasier University, Canada, 2003.

[13] R. Loukil, M. Chtourou and T. Damak, "Synthesis of Nonlinear Observers for Actuator Fault Detection and Isolation", International Journal of Computer Applications, 43(4), 27-32, 2012.

[14] S.K. Spurgeon, "Sliding Mode Observers - historical background and basic introduction", Spring School, Aussois, 2015.

[15] S.K. Spurgeon, "Sliding Mode Observers – toward a constructive design framework", Spring School, Aussois, 2015.

[16] X.G. Yan, C. Eduards, "Nonlinear robust fault reconstruction and estimation using a sliding mode observer", Elsevier, ScienceDirect, Automatica, DOI:10.1016/j.automatica.2007.02.008, 43, 1605 – 1614, 2007.

[17] B. Castillo-Toledo, S. Di Gennaro, J. Anzurez-Marin, "On the fault diagnosis problem for non–linear systems: A fuzzy sliding–mode observer approach", Journal of Intelligent and Fuzzy Systems, 20, DOI: 10.3233/IFS-2009-0427, IOS Press, 187–199, 2009.

[18] J. Keighobadi, P. Doostdar "Fuzzy Sliding Mode Observer for Vehicular Attitude Heading Reference System", Journal of Scientific Research, Iran, DOI: 10.4236/pos.2013.43022, 4, 215-226, 2013, http://www.scirp.org/journal/pos

[19] Fuzzy Logic Tutorial, available online on the web site: http://www.massey.ac.nz/~nhreyes/MASSEY/159741/Lectures/Lec2012-3-159741-FuzzyLogic-v.2.pdf

[20] Control Tutorials for MATLAB, Carnegie Mellon Lab, University of Michigan, http://ctms.engin.umich.edu.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 227-233 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

How Effective is Using Lip Movement for Japanese Utterance Training

Miyuki Suganuma*[1], Tomoki Yamamura[2], Yuko Hoshino[2], Mitsuho Yamada[1]

[1]*Graduate school of Information and Telecommunication Engineering, Tokai University, 108-8619, Japan*

[2]*School of Information and Telecommunication Engineering, Tokai University, 108-8619, Japan*

ARTICLE INFO

Article history:
Received: 20 December, 2016
Accepted: 20 January, 2017
Online: 28 January, 2017

Keywords :
Lip movements
utterance
Japanese

ABSTRACT

Lip movements have long been the subject of research. There are many methods of lip movement recognition, such as the calculation of the amount of movement compared to a matching face pattern. In a previous study, we investigated utterance recognition based on the power spectrum by focusing on lip movements, which is one aspect of multimodal voice recognition systems. However, we found that the utterance recognition rate varied depending on the participants throughout our research. For this reason, we propose a training method for the purpose of utterance improvement in Japanese.

1. Introduction

In recent years, the rapid advancement of computer technologies has led to the widespread use of personal information which, when converted into digital data, can be read and changed without place and time restrictions. Therefore, large-scale processing of data is optimal. However, to use paper-medium information as digital information, it must be input by hand using a keyboard. As a way to simplify this work and make it more intuitive, voice recognition is an attractive technology. Data entry using voice recognition is easier to use than keyboard entry because it can be used hands-free. For this reason, many researchers have been researched voice recognition, for example, the non-contact type interface of command input and the individual identification [1,2]. On the other hand, voice recognition technology is not without problems, and the recognition rate can decline when it is used in noisy locations. Therefore, multimodal recognition has been investigated, because it can make use of both voice recognition and face recognition [3,4]. In our research, we have been studied focusing on lip movement recognition [5]. However, we found that the recognition rate of lip movement varies depending on the person who is speaking. In response to this, we are developing ways to improve participants' utterances so that they may be more easily recognized [6]. The purpose of our research is to make participants' utterances more fluent. For the utterance learner, improving their utterance is a great incentive to study Japanese utterance. With regard to studies to improve utterances, the researches that is focused on the tongue movement in vowel utterances not only Japanese also in other language such as English have been conducted [7]. Wilson measured the tongue movement by filling the lower jaw with an ultrasound sensor [7]. Electromyogram is also possible to measure the tongue movement [8], however, it is difficult to specify myoelectric signal along mixing other myoelectric signals with lip movements. Therefore, we propose the utterance training method only using acquired lip image while uttering, expanding the research of our lip feature point collecting application by lip movement because we considered the noninvasive utterance training can be conducted. Using simple training application that we developed in our laboratory enable to do self-utterance learning. Utterance learners can study by themselves until to be satisfied with the result of utterance training. Moreover, we also aim at make it possible for speech- and hearing-impaired individuals to use our lip feature point collection application by themselves.

This paper is an extension of work originally presented in ICCSE 2016 (The 11th international conference on computer science and education) [9]. We explain about speech training system which is developed in our laboratory in second section.

*Corresponding Author Name: Miyuki Suganuma, Graduate school of Information and Telecommunication Engineering, Tokai University, +810334411171, m.suganuma@hope.tokai-u.jp

Third section is method. Fourth section is results and discussions. Finally, we describe conclusion in fifth section.

2. Speech training system

In this research, we used a lip feature point collecting application that was developed in our laboratory [10]. This application was developed with Visual Studio 2010 as a Windows form application in the C++ language. FaceAPI, which is facial recognition software made by SeeingMachines in Australia, was used for the face recognition, as it is able to perform at high speeds and with high precision. The processing speed is estimated to be 0.3 seconds when the head does not move and a Core-2 Duo 2.4 GHz processor is used. Our equipment uses an i7 1.9 GHz processor and has sufficient processing speed for lip movement detection.

Figure 1 is an example of the display used for the acquisition of lip feature points. Figure 2 shows the display for lip movement training. When the subject is recognized by the lip feature point collecting application, the acquired face feature points can then be shown for training. To operate this application, a file for training must first be selected from the “Set up” menu and the file name typed to save data. Next, training is initiated by clicking on the “Start” button. When the subject closes his or her mouth, the application recognizes that person's mouth and the subject can start pronouncing words. If the subject is not recognized by the application, the “Re-recognition” button can be used. The “Stop” button stops recording. The lip movement training display is shown after recording is stopped. The red line on the lip movement training display (Figure. 2) shows the teacher's lip movements, which are used as model data; the black line shows the participant's lip movements. The two sets of lip movements can be compared by clicking on the “Start” button (Figure. 2).

3. Method

The present study was carried out to improve the learner’s Japanese utterance like veteran announcer's utterance by using a lip feature point collecting application that was developed in our laboratory.

The flow of lip movement training is shown in Figure 3. First, a file for training selects from the “Set up” menu. Second, the participant pronounces each sentence with the application by clicking on the “Start” button. Third, the participant compares the announcers’ lip movement data with their own as shown in Figure 2 as the black line. The participant can watch his or her lip movement and the teacher's lip movement any number of times by clicking “Start” in Figure 2. Nevertheless, the participant cannot see a line graph of lip movement. This flow is a set of the training. The training consisted of a total of 10 repetitions. Maximum two repetitions of the training were carried out per day in order to the participants not grow tired.

A model database for Japanese utterances was created in our previous research[7]. The model data were acquired from a television announcer whose lip motion was considered to be correct. The announcers were two veteran announcers, a man and a woman, from the NHK (Japan Broadcasting Corporation) Communications Training Institute. Based on the model data that we created, we trained 4 students from our university (three males and one female: aged 21 to 22). To improve their utterances, we used the lip feature point collecting application that is explained in the preceding section. The experiment was conducted in a studio at our university, and we used a Windows note PC (SONY VAIO PCG-81314N). The web camera screen resolution was 640x480. The set-up for the acquisition of utterance data is shown in Table 1. Japanese data were extracted from the book "Easy training for a good voice" [11], which includes 69 sentences from *"a"* to *"wa"* In order to acquire both long and short sentences, we used the "dandelion” text from a third-grade textbook of the Japanese language that is currently used for announcers’ utterance practice [12].

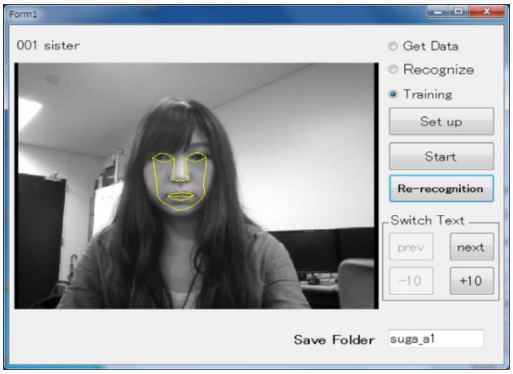


Figure 1. Display for the acquisition of lip future points

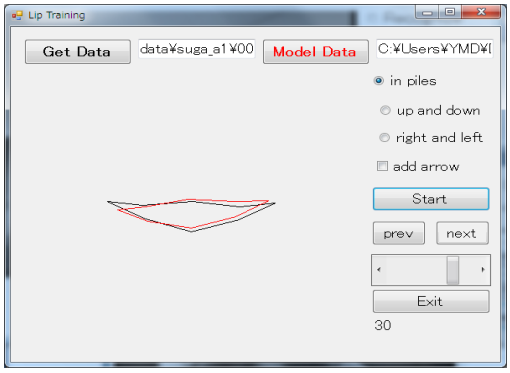


Figure 2. Display for lip movement training

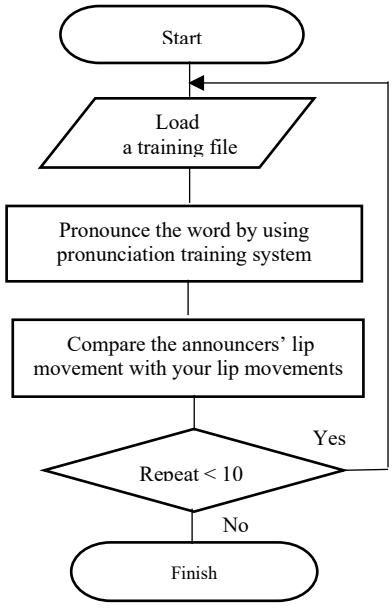


Figure 3. Flow chart for lip movement training

Table 1. Japanese data used

Attara aisouyoku aisatsu shinasai
Akogare-no aite-ni au
Ikigai-wo motomete ikou
Ima ijyou-no omoi-wo ireru
Uta-wo utatte usabarashi
Ukatsu-ni umai uso
Eiyo-yo eikou-yo eien-nare
Erai ekakisan-ga eranda-e
Oishii okashi-wo osusowake
Ookami-no ookina tooboe

4. Results and Discussion

4.1. Lip movement data

We carried out Japanese utterance training using the veteran announcers' model data. The participants were students from our university. We compared the model data and the participants' data to determine their differences. In this section, we show only the results for *"Ookami-no ookina tooboe"* as an example. Figure 4 shows the 5 lip feature points: upper lip, lower lip, left lip, right lip and chin that we collected for the utterance training. Generally speaking, human beings' mouth moves vertically symmetry and lower lip opens more widely than upper lip, so that we used lower lip movement and left lip movement for discussing lip movement.

Figure 5 shows the veteran announcer's lip movements. Figures 6 and 7 show the participants' lip movements in several repetitions of the training (1 repetition, 5 repetitions, 10 repetitions). The vertical axis indicates the amount of displacement (pixel). The screen resolution is 640x480, therefore actual lip opening displacement is about 0.5 cm if an amount of displacement of mouth is 10 pixel. The horizontal axis represents time. The blue line shows the left lip feature points. The red line shows the lower lip feature points.

In the 1st repetition of the training, none of the participants opened their mouths as widely as the veteran announcer. However, in the 5th and 10th repetitions of the training, their utterances became clearer little by little, because the amplitudes of their lip movements became large and they pronounced the sentences precisely by every chunk, such as *"Ookami-no" "ookina" "tooboe"*.

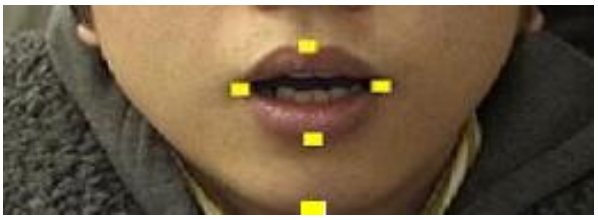


Figure 4. 5 lip feature points for the utterance training
(Upper lip, lower lip, left lip, right lip and chin)

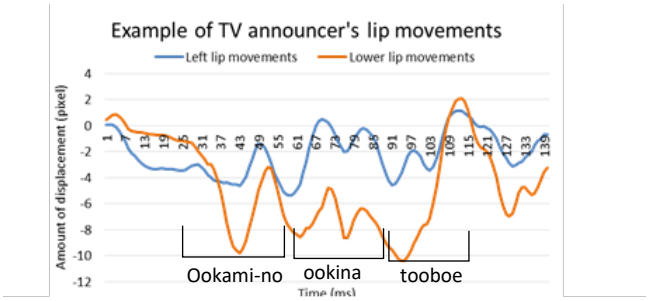


Figure 5. Veteran announcer's lip movement
"Ookami-no ookina toboe"

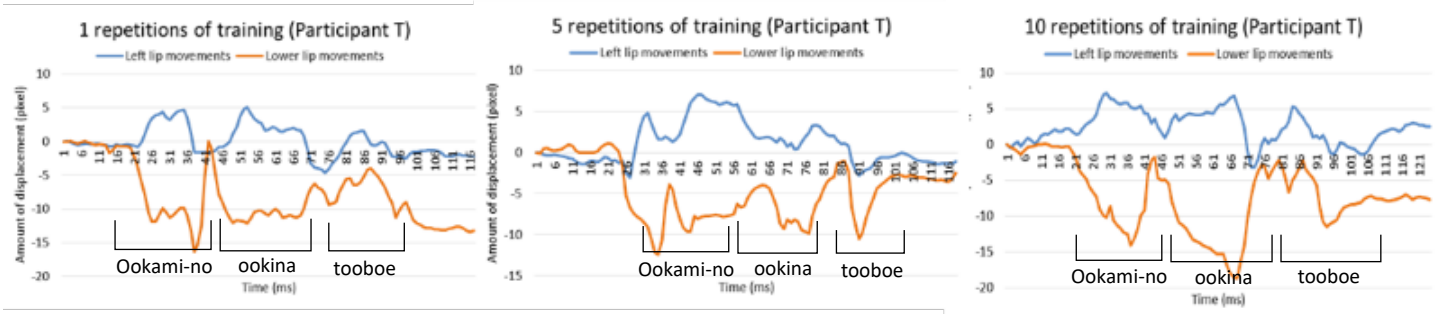


Figure 6. Participant T's lip movement *"Ookami-no ookina toboe"*

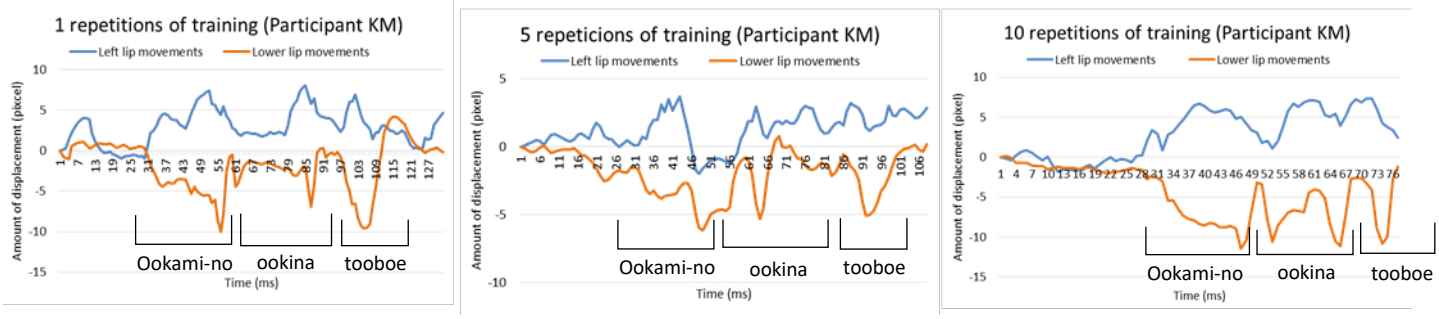


Figure 7. Participant KM's lip movement *"Ookami-no ookina toboe"*

4.2. Voice data

During the training, voice data were also recorded at the same time. Here, we show the voice data results. We used the paired comparison method to evaluate the voice data objectively.

The evaluators were students from our university including the participants who participated in the Japanese utterance training. There were 10 evaluators aged 20 to 22 years old. Participants did not evaluate their own voice data.

4.2.1. Pairwise comparison

Figure 8 shows an example of paired comparisons. We randomly chose the voice data from the 1st, 3rd, 5th, 7th and 10th repetitions of the training and evaluated these data by the paired comparison method. In this method, evaluators alternatively chose which voice is good by focusing on following items: articulation, speed and volume. These three items were uniquely chosen based on the veteran announcers' characteristics which can be able to improve their utterance [13]. In the articulation, evaluators compared sentences whether each sentences uttered clearly by every chunk such as, *"Attara" "aisouyoku" "aisatsu" "shinasai"*. In the voice speed, evaluators chose which sentence is easy to hear, also pronounced appropriate speed. In the voice volume, evaluators genuinely chose the sentence which is louder. At this time, voice volume was set by the evaluators before the training, and we instructed them not to change the volume during the training. These comparison items were explained to evaluators before training. We explained to the evaluators what items they will evaluate before the voice comparison was conducted. Each evaluator evaluated 4 repetitions of voice comparisons per participant. The total number of comparisons was 16.

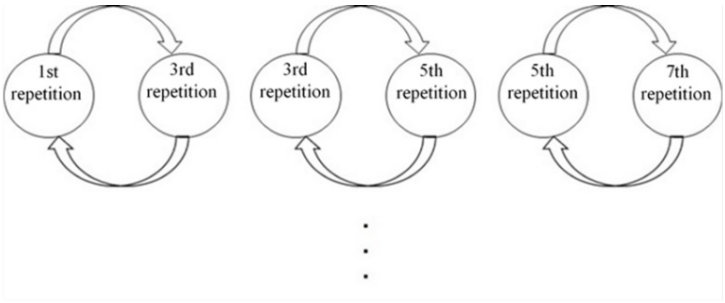


Figure 8. Example of paired comparison

We show the two example sentences. Figure 9 shows the results for *"Attara aisouyoku aisatsu shinasai"*. It shows the average of all comparison results and the standard deviation. First line indicates the articulation. Second line is the voice speed. Third line is the voice volume. Likewise, Figure 10 is the results for *"Ikigai-wo motomete ikou"*. Table 2 shows the summarized results of each evaluation items improvement and the result of t-test in the sentence of *"Attara aisouyoku aisatsu shinasai"*. Table 3 shows the summarized results of each evaluation items improvement and the result of t-test in the sentence of *"Ikigai-wo motomete ikou"* same as Table 2.

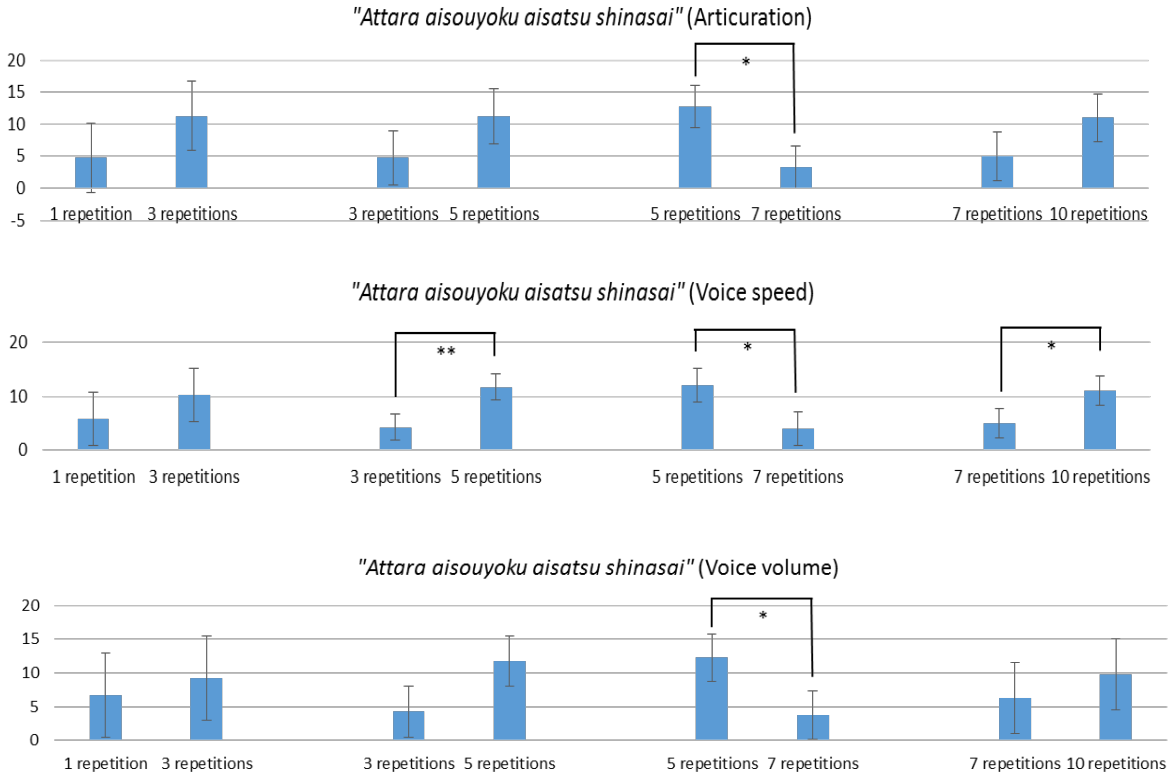


Figure 9. Results of voice comparison *"Attara aisouyoku aisatsu shinasai"*

Table 2. Summarized results of each evaluation items improvement and the result of t-test
"Attara aisouyoku aisatsu shinasai".

Attra aisouyoku aisatsu shinasai						
Set of training	Articulation		Voice speed		Voice volume	
	Improvement (up or down)	Significant differences	Improvement (up or down)	Significant differences	Improvement (up or down)	Significant differences
1-3 repetitions	up	n.s	up	n.s	up	n.s
3-5 repetitions	up	n.s	up	**	up	n.s
5-7 repetitions	down	*	down	*	down	*
7-10 repetitions	up	n.s	up	*	up	n.s

n.s.: non-significant, *: p<0.05, ** p<0.01

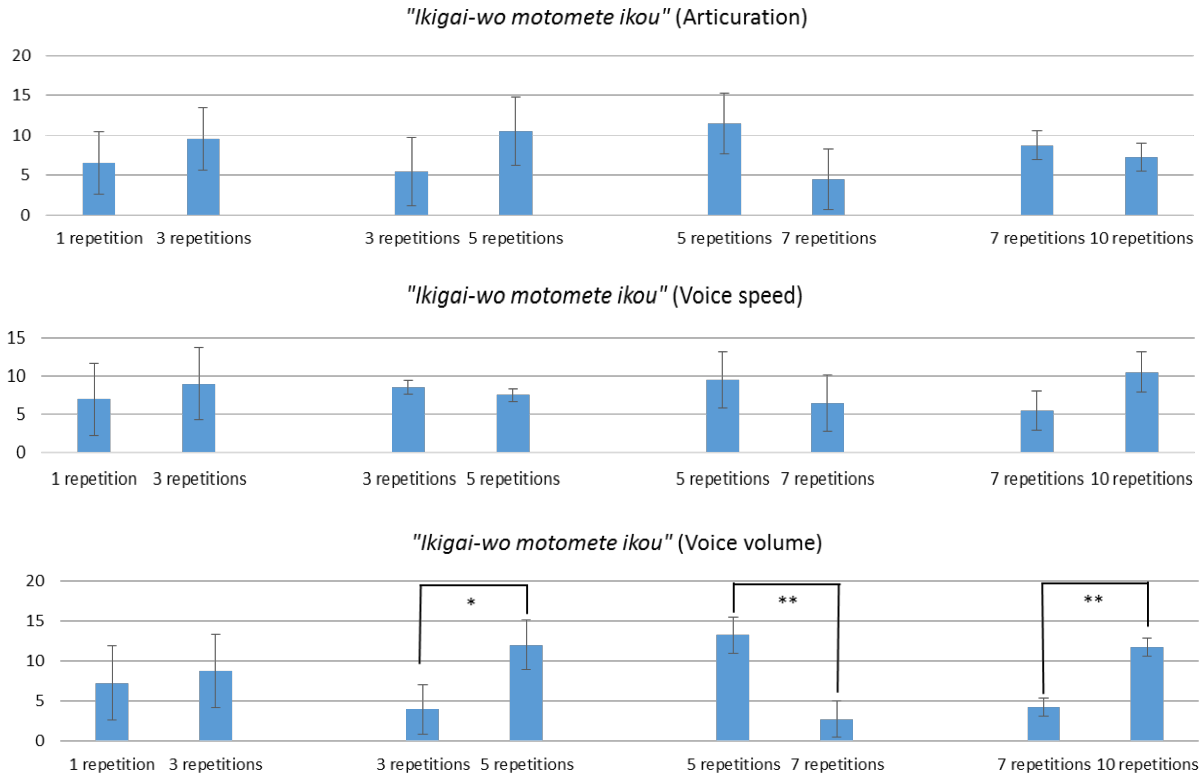


Figure 10. Results of voice comparison *"Ikigai-wo motomete ikou"*

Table 3. Summarized results of each evaluation items improvement and the result of t-test
"Ikigai-wo motomete ikou".

Ikigai-wo motomete ikou						
Set of training	Articulation		Voice speed		Voice volume	
	Improvement (up or down)	Significant differences	Improvement (up or down)	Significant differences	Improvement (up or down)	Significant differences
1-3 repetitions	up	n.s	up	n.s	up	n.s
3-5 repetitions	up	n.s	down	n.s	up	*
5-7 repetitions	down	n.s	down	n.s	down	**
7-10 repetitions	down	n.s	up	n.s	up	**

n.s.: non-significant, *: p<0.05, ** p<0.01

Throughout 10 repetitions of training, transitions of improvement were small from 1st repetition to 3rd repetitions, then the transitions became large from 3rd repetitions to 5th repetitions. However, sometimes improvements decreased from 5 repetitions of training to 7 repetitions of training. After that, improvement tendency became large from 7th repetitions to 10th repetitions. We could see that utterance improvement is not seen in each repetitions of training. Nevertheless, improvement tendency is seen after 5 to 7 repetitions of training because participants came out of a slump. This tendency is in common between all vowels and it is considered to be one of the learning characteristics.

In addition, we carried out an unpaired two-tailed t-test regarding the average of the marking results. The t-test result of paired comparison shows in Figures 9 and 10 (also Tables 2 and 3) in "*" or "**" if there are significant differences. There was significant difference between 3 repetitions of the training and 5 repetitions of the training in voice speed item sentence *"Attara aisouyoku aisatsu shinasai"* ($p=0.008<0.05$) in Figure 9. There were significant difference between 5 repetitions of the training and 7 repetitions of the training, and between 7 repetitions of the training and 10 repetitions of the training in voice volume item sentence *"Ikigai-wo motomete ikou"* (both $p=0.008<0.01$) in Figure 10. All vowel's utterance are improved in all 3 evaluate items from 3rd repetitions to 5th repetitions and from 7th repetitions to 10th repetitions, even though some exception is seen. We also carried out an unpaired two-tailed t-test in other sentences. There were significant differences in /a/, /i/ and /o/ sounds. Whereas, there were no significant differences in /u/ and /e/ sound. When the native Japanese speakers utter /a/, /i/ and /o/ sounds, they usually open their mouth widely so that we consider significant differences could see. From these results, we confirmed that the more training the participants had, the better the evaluations they received.

4.2.2. Formant analysis

We used formant analysis in addition to pairwise comparison method because vowels are commonly expressed in the International Phonetic Association (IPA) vowel chart in English-speaking country. The IPA vowel chart shows the position of the tongue and how the mouth is opened for each sound. If the frequency of the 1st formant is high, the tongue is positioned in the front of the mouth. If the frequency of the 2nd formant is high, the position of the tongue in the mouth is low. [14]. Figure 11 shows average Japanese men's vowel formant. Mori states that the 1st and 2nd formants can classify vowels regardless of race and gender [15]. Otsuka points out that the sound spectrograms of the 1st and 2nd formants are helpful in teaching the pronunciation of English vowels [16]. We have been also researched English pronunciation training using application which was used in this research. We used this evaluation method and we could made close to learners' pronunciation to native English speakers' pronunciation [17].

In this paper, we show the scatter diagram of the formants of the participants about *"Eiyo-yo eikou-yo eien-nare"* from 1 repetition to 10 repetitions of the training as an example in Figure 12 because this sentence includes many /e/ sound. The circle marker indicates 1 repetition of the training. The square marker is 5 repetitions of the training. The diamond marker is 10 repetitions of the training. We can see that each markers gathered throughout the training. Moreover, in compared to the average vowel formant diagram in Figure 12 and the result of the sentence *"Eiyo-yo eikou-yo eien-nare"*, we can see that their formants are significantly close. When the native Japanese speaker move our mouth from *"e"* to *"i"*, they usually move their mouth from lengthwise direction to sideways direction. Therefore, we consider that participants could pay attention to utter the sentence clearly.

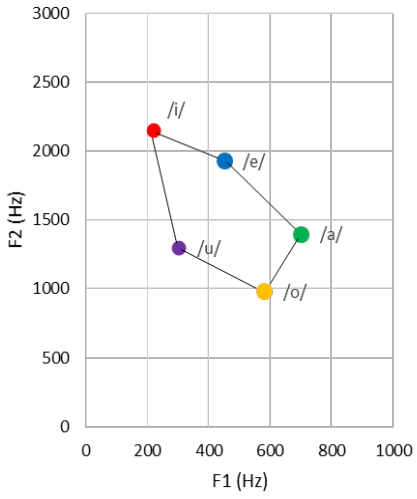


Figure 11. Average Japanese men's vowel formant

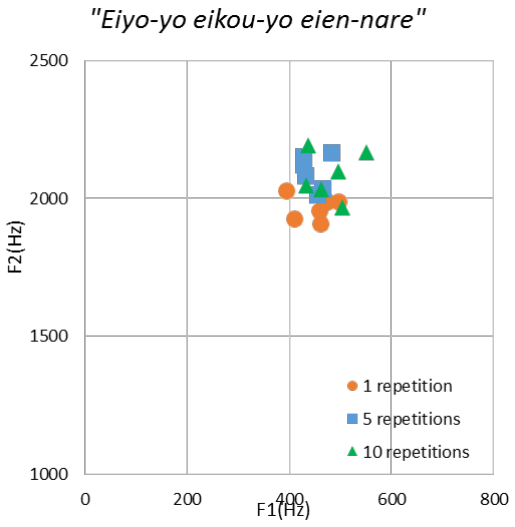


Figure 12. Scatter diagram of the participants about each repetitions of training *"Eiyo-yo eikou-yo eien-nare"*

5. Conclusion

In this research, we evaluated the usefulness of Japanese utterance training using the lip feature point collecting application in order to improve people's utterance ability and prevent a decline in the voice recognition rate. To train participants in making utterances, we used veteran announcers' data as the model data.

We evaluated not only the lip movement results but also the voice data results using the pairwise comparison method and formant analysis. As for the lip movement results, we confirmed that the participants' utterances improved in comparison to the first repetition of training, because the amplitudes of their lip movement's became larger. Also participants' could be able to pronounce clearly by each chunk such as *"Ookami-no" "ookina" "tooboe"*. With regard to the voice comparisons, the evaluation showed that the training seemed have a good effect up through ten repetitions of the training, especially /a/, /i/ and /o/ sounds. Whereas, we could not get a training effect by each repetitions of training because there are training slump or fatigue from 5 repetitions of training to 7 repetitions of training. However, this result indicated that this is one of the learning characteristics. We need to find how to make participants not bored from 5 repetitions of training to 7 repetitions of training to solve this problem. We considered that the more training the participants had the better the evaluations they received, if we can find the solution. In addition, some remarkable improvement is seen in the result of formant analysis, for example, *"Eiyo-yo eikou-yo eien-nare"* because Japanese vowel of /e/ sound in the sentence grow to the Japanese average /e/ sound. These results suggest that Japanese utterance training using the lip feature point collecting application is useful.

In the future, we need let to use the lip feature point collecting application to more participants' and wide range of age groups. We plan to analyze the lip movement data more in detail in order to enhance the reliability of outcomes, for example, an open area of mouth, a count of amplitudes and a calculation of utterance time. Also, we would like to try voice comparison by expanding the sample differences as a future task because our sample differences were small such as 1 repetitions and 3 repetitions of the training. Moreover, we think the relationships between smooth graph and voice quality is also necessary to consider.

Acknowledgment

We are thankful to the television announcers with NHK communications training institute who cooperated with the creation of an utterance database. This work was supported by JSPS KAKENHI Grant Number (25330418) and (16K01566).

References

[1] Y. Sato, Y. Kageyama and M. Nishida, "Proposal of Non-Contact Type Interface of Command Input Using Lip Motion Features" The transactions of the Institute of Electrical Engineers of Japan. C, A publication of Electronics, Information and System Society 129(10), 1865-1873, 2009.

[2] Y. Shirasawa, S. Miura, M. Nishida, Y. Kageyama and S. Kurisu, "Method for Identifying Individuals Using Lip Motion Features" Journal of Image Information and Television Engineering, 60(12), 1964-1970, 2006.

[3] M. Yoshikawa, T. Shinozaki, K Iwano and S. Furui, "Multimodal Speech Recognition Based on Lightweight Visual Features" IEICE Information and Systems, J95-D (3), 618-627, 2012.

[4] S. Hayamizu and T. Takezawa, "Trends in Research on Multimodal Information Integration System" Jornal of the Japanese Society for Artificial Intelligence, 13(2), 206-211, 1998.

[5] E. Wakamatsu, Y. Hoshino and M. Yamada, "Proposal for an utterance training method based on lip movements" in IMQA 2014 (The Seventh International Workshop on Image Media Quality and its Applications), 2014.

[6] M. Suganuma, T. Yamamura, Y. Hoshino and M. Yamada, "How to Evaluate English Pronunciation Learning by Lip Movements" in IMQA 2016, 2016.

[7] I. Wilson, "Using ultrasound for teaching and researching articulation" Acoust. Sci. & Tech., 35 (6), 285-289, 2014.

[8] C. Huang, C. Chen and H. Chung, "The Review of Applications and Measurements in Facial Electrography" Journal of Medical and Biological Engineering, 25(1), 15-20, 2004.

[9] M. Suganuma, T. Yamamura, Y. Hoshino and M. Yamada, "Effect of Japanese utterance training using lip movement" in ICCSE 2016 (The 11th international conference on computer science and education), 2016.

[10] E. Wakamatsu, K. Kikuchi and M. Yamada, "Development and Measurement of an Automatic Word Recognition by Lip Movements" in IMQA 2013, 2013.

[11] E. Fukushima, Easy training for a good voice, Seibido Printing, 2006, in Japanese.

[12] Dandelion, New version of the Japanese elementary school 2nd year textbook, Tokyo Books, 1985, in Japanese.

[13] S. Suzuike, "The study on the significance of the non-verbal elements of the human voice in the communication," Bachelor thesis , Waseda University, 2010.

[14] S. Imaizumi, Phonetics and Linguistics for speech-language-hearing therapist, Igaku -Shoin Ltd., 2009.

[15] T. Mori, "Multilingual vowel sound comparison analysis by the formant in sound, the articulatory phonetics," Bachelor thesis, Nagoya University, 2015.

[16] S. Otsuka, "A Trial in Utilizing Formant Values for Teaching Vowel Pronunciation," Science reports of Tokyo Woman's Christian University, 2014.

[17] M. Suganuma, T. Yamamura, Y. Hoshino and M. Yamada "Proposal of the Way of English Pronunciation Training Evaluation by Lip Movement" Journal of Japan Personal Computer Application Technology Society, 11(2), 2017. (will be published)

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 234-239 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

Wireless Android Based Home Automation System

Muhammad Tanveer Riaz[*,1], Eman Manzoor Ahmed[2], Fariha Durrani[2], Muhammad Asim Mond[2]

[1]*State Key Laboratory of Power Transmission Equipment and System Security, School of Electrical Engineering, Chongqing University, Chongqing, 400044, China*

[2]*Department of Electrical Engineering, University of Lahore, Lahore, 54000, Pakistan*

ARTICLE INFO

Article history:
Received: 15 December, 2016
Accepted: 20 January, 2017
Online: 28 January, 2017

Keywords:
Home automation
Wireless LAN
Wi-Fi
Microcontrollers

ABSTRACT

This manuscript presents a prototype and design implementation of an advance home automation system that uses Wi-Fi technology as a network infrastructure connecting its parts. The proposed system consists of two main components; the first part is the server, which presents system core that manages and controls user's home. Users and system administrator can locally (Local Area Network) or remotely (internet) manage and control the system. Second part is the hardware interface module, which provides appropriate interface to sensors and actuator of home automation system. Unlike most of the available home automation system in the market, the proposed system is scalable that one server can manage many hardware interface modules as long as it exists within network coverage. System supports a wide range of home automation devices like appliances, power management components, and security components. The proposed system is better in terms of the flexibility and scalability than the commercially available home automation systems.

1. Introduction

Home automation is similar to digital home, smart home, e-home and intelligent household. They all mean a high living condition with several smart devices. It is the residential extension of automation which is using telecommunication technology, computer technology and automation technology to give the user a better living environment, security and comfort. It helps people to reduce domestic working and household management by its automation and monitoring system.

A concept on smart home application and development includes several implementation techniques and is never limited. Smart home systems are created based on analysis on client requirements and financial budget for the requirement of system. With technologies available now, better implementation of this system could be achieved.

Now, advancement in smart phones and wireless technology familiarized new ideas such as Bluetooth, the Wi-Fi and Internet network, which has been slowly replacing the old wired technology that required wire bonded interconnection between electrical devices. The main advantage of wireless interlinking includes diminishing the need of wires for connection [1].

[*]Corresponding Author: Muhammad Tanveer Riaz, School of Electrical Engineering, Chongqing University, Chongqing, China.
Email: uetmtr@yahoo.com

Home automation includes atmosphere controls, entryway and window controls, and in addition the control of home appliances, multimedia home theaters, pet sustaining, plant watering etc. Other examples are Heating, Ventilation, and Air Conditioning (HVAC) systems, automatic garage doors, and intruder detection alarms. In short, home automation emphasizes on comforts through ergonomics and simplicity of operation.

Many countries are slowly embracing smart home security control system [2], [3]. There are many platforms over which a home automation system can be implemented. The currently available platforms are Zigbee, RS232, Ethernet, Bluetooth, Infrared, Global System for Mobiles (GSM) and Microcontroller [4-6].

The smart home systems are introduced utilizing wired or remote systems. Wired system expenses are low however the absence of establishment troubles typically makes it the second choice. The cost, control, power consumption and execution are the fundamental determination criteria for remote systems, which incorporate popular technologies like Zigbee, Bluetooth, Wi-Fi module, GSM network etc. [7-9].

Bluetooth technology has very limited range and the one-to-one connection while it enables high data rate and low power consumption. Especially one-to-one connection feature lack its feasibility to use in a smart home system. Infrared too has range

issues and it is restricted to a automate a device within its specific range only. Although Zigbee has high range and low power consumption, its cost restricts the extensive use, it also requires additional platform for Internet communication. The smart home concept on a large scale consists of lot of physical systems which are to be monitored and controlled using communication networks and communication means [10], [11]. Contrary, Wi-Fi module is less expensive than Zigbee module and provides communication between user and home via Internet following 802.11 b/g/n protocol [12].

The proposed system includes Wi-Fi technology in order to automate the system. Though commercial products are available that are using the technology but they are limited to a single device only. The system under design is opted to develop a single android application to control numerous devices. The system would be globally accessible and it will be requiring an Internet Protocol (IP) address of the server and a port number that would be defined and specified in design. The complete IP and port number is collectively termed as a socket.

2. System Architecture

In planning a home automation framework, at least one reasonable platform is utilized as a part of request to construct a solid and adaptable framework that can be effectively worked and adjusted for another new apparatus. Consequently, with the end goal of this venture some particular decisions can be made on the kind of platform, hardware equipment and method of operation of the home automation system.

The designed home automation system uses AT89C51 microcontroller, an android mobile phone, RS232 standard for communication between the microcontroller and server, MAX232 for interfacing the microcontroller, a relay and a driver for interfacing the relay.

As illustrated in the block diagram shown in **Error! Reference source not found.**, when the server receives the required signal, it communicates via the RS232 and MAX232 to the AT89C51.

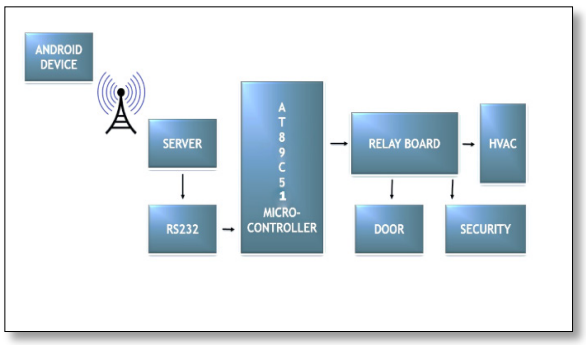


Figure 1:Bock Diagram of System Architecture

The AT89C51 controls the relay state via driver and this in turn determines the state of the connected appliance, whether switched on or off. The connected number of appliances depends upon the number of pins of microcontroller assigned as an output. Depending upon the power source, both Direct Current (DC) and Alternating Current (AC) appliances can be connected. Mostly home automation systems include control of HVAC, security and monitoring systems.

3. Android Application Development

The environment for the Android Application Development is Eclipse. The Eclipse Platform is aimed for building Integrated Development Environments (IDEs), and subjective tools. By Eclipse, it is often meant the Eclipse Software Development Kit (SDK) which is both the leading Java™ IDE and the single best tool available for building products based on the Eclipse Platform. An android application has been developed in order to provide a user friendly interface using Eclipse IDE. Android applications are written in the Java programming language. The Android SDK provides tools for code compilation and packaging data and resource files into an archive file with '.apk' extension called as an Android package. Android devices used the '.apk' file to install the application. Android's application framework allows for the creation of extremely rich feature and novel applications by using a set of reusable components.

The system application is designed to receive an IP address and a port number of the server, to which control circuitry has been connected. An application is capable of communication between the mobile device and the server.

3.1. Manifest file

The Android manifest file is the central configuration file for an Android application. The editor organizes the manifest information into a number of tabs like Manifest, Application, Permissions, Instrumentation and AndroidManifest.xml.

The tabs that are concerned regarding the application are Permissions and AndroidManifest.xml. In Permissions tab, two permissions for the application are used as the system requires internet connection to send instructions to server on the other end. i.e.

- android.permission.INTERNET
- android.permission.NETWORK

AndroidManifest.xml launches different class activities according to their intents. One of these is a launcher activity which in this case is .Pic. When this activity executes, it passes the intent of next activity to manifest file and Android manifest launches the next activity which is .MainActivity.

3.2. Java classes

The application has three Java classes

- Pic.java
- MainActivity.Java
- SocketClient.Java

First class refers to a package followed by imports required for application and the later class defines its functions.

3.3. XML file

There are two xml files in the application of which, one is pic.xml that is called by Pic.Java and other is activity_main.xml called by MainActivity.Java. These xml files give the graphical layout or in other words the interface of activity. They have different attributes in them like buttons, toggle buttons, scrolls, texts, editable texts, layouts, check boxes etc. This has been used in the application as a splash screen that stays for 3 seconds when the application has been opened. The graphics are as given in Figure 2.

Figure 2: Splash Screen

The activity_main file includes editable text, text views and toggle buttons in it. Each button and text has its own id so that it can be accessed in future. In the application, editable text1 is for IP address of the device in which server code is running while editable text3 is for port which is always 4444. Each button is for switching of an appliance like fan, light, door, modem etc.

An on-click listener actually sends a defined data after every button has been toggled. The data that is specified with each button is listed in Table 1. Buttons are ambalished with appropriate pictures which are placed in drawable folder as given in Figure 3.

Toggle button	Port	Data sent to port
Fan on	4444	1
Fan off	4444	2
Light on	4444	3
Light off	4444	4
Door open	4444	5
Door close	4444	6
Alarm on	4444	7
Alarm off	4444	8

Table 1: Data sent by application to the port

Figure 3: Application Layout

Finally when the code has been debugged, eclipse make an apk file of the application which is to be installed on an android device on which it is supposed to run. It can also be licensed by Google and uploaded on Google playstore but this application being very specific in design is to be provided as a complete package along with hardware control chip.

4. Server

The server is a Dynamic Link Library (dll) that is providing an interface between the Wi-Fi and the serial port. The link is developed and hence the messages received from phone are transmitted to serial port where microcontroller has been attached. Its works by running its batch file as shown in Figure 4 .

The COM5 port has been addressed as 4444 that is receiving data and according to the Table, the server has successfully received the sent data. This data is further to be sent to serial port where the microcontroller will be able to get this serially on its receiving port.

```
ib\RXTXcomm.jar serialport.SocketServer
Baud rate :9600
Data bits :8
Stop bit:1
Parity:0
comm port:COM5
Socket Server started. Listening to the port 4444
Stable Library
=========================================
Native lib Version = RXTX-2.1-7
Java lib Version   = RXTX-2.1-7
1
SENDING: 1

2
SENDING: 2

3
SENDING: 3

4
SENDING: 4

5
SENDING: 5

6
SENDING: 6

7
SENDING: 7

8
SENDING: 8
```

Figure 4: Server Operation

5. Hardware Implementation

5.1. Keil µvision IDE

The program code has been written in Keil and debugged here for any errors. The code is built into a target and the hex file is saved that can be imported to Proteus for simulating the circuit. The algorithm of the program is given by the flowchart of the program as given in Figure 5.

The code includes all <reg51.h> library that contains all the necessary information about 8051 microcontroller family.

5.2. Proteus

Proteus has been used for testing the system in a virtual environment. The software includes all the components in its built in library that are being used. The connections are established and hex file from Keil is loaded in the microcontroller. A seven segment display and LEDs are used to test the output. Hence, when any button from the application is pressed, the server receives the data and transmits to the virtual serial port that is further connected to microcontroller. The LEDs operate in accordance to the program that has been load. The simulation results can be shown as in Figure 6. The virtual serial port is

sending commands that are further read by programmed microcontroller. And hence the LEDs acting as loads are switched ON or OFF. The relay is providing 220V supply for the operation.

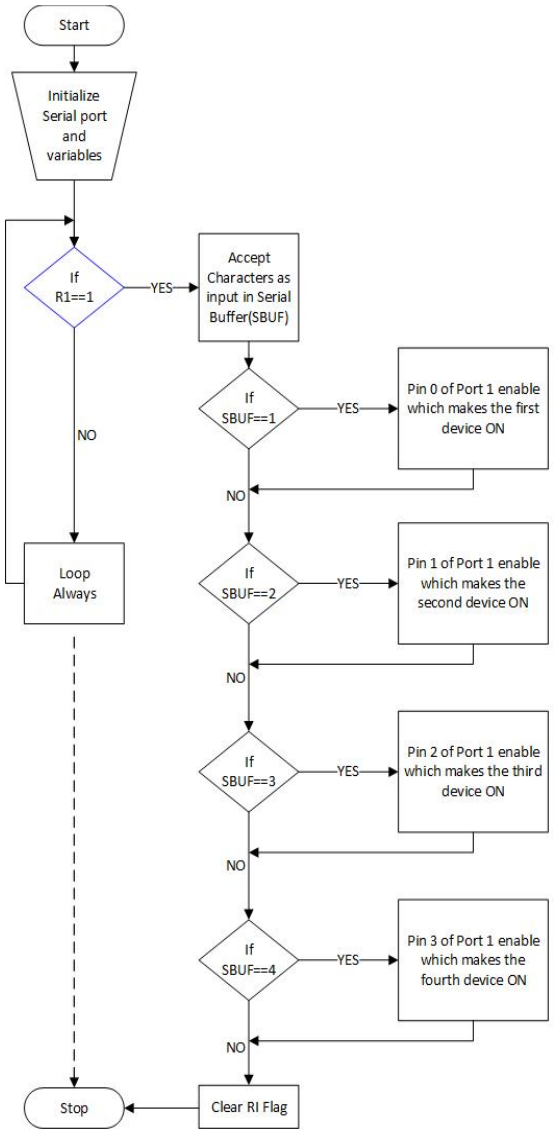


Figure 5: Flow Chart of Microcontroller Program

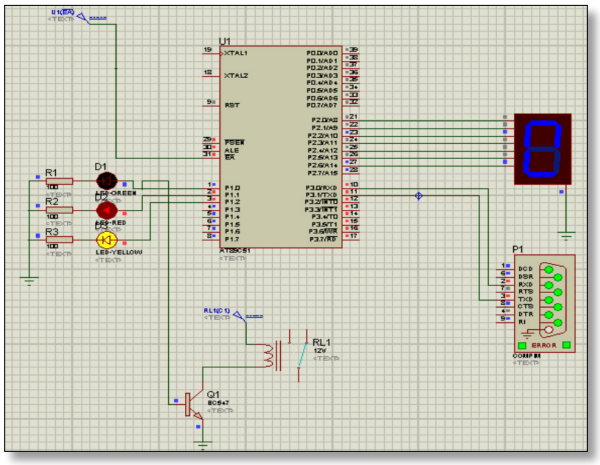


Figure 6: Simulation in Proteus

Diptrace has been used for the Printed Circuit Board (PCB) designing of the circuit needed to do the job. The layout is designed here and printed on the butter paper. It is then imprinted over PCB sheet and afterwards etching is done. The PCB layout at DipTrace is shown below in Figure 7. The complete circuit built on a PCB sheet has been mentioned in the following Figure 8.

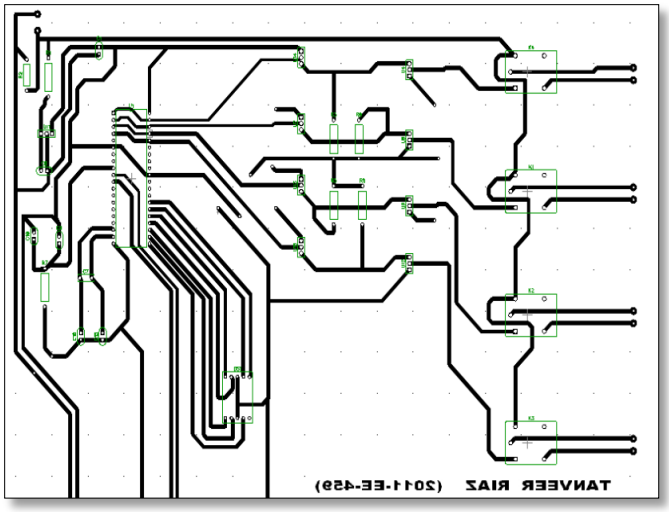


Figure 7: PCB Layout in DipTrace

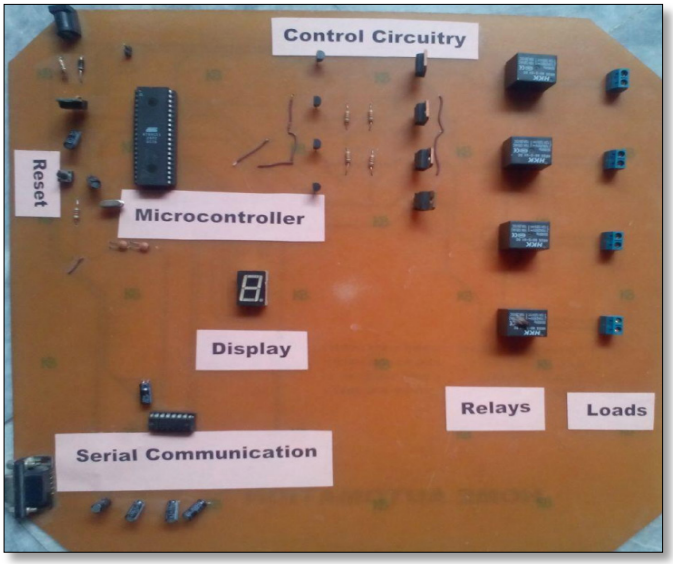


Figure 8: Hardware Circuit

5.3. *mikroC PRO for 8051*

mikroC PRO for 8051 is a full-featured compiler for 8051 devices. It is the best solution for developing code for 8051 devices. It features intuitive IDE, powerful compiler with advanced optimizations, lots of hardware and software libraries, and additional tools that facilitate the work. Libraries are fully-documented and allow a quick start in programming microcontrollers.

mikroC PRO facilitates to communicate with the COMM ports. Data through serial port is sent to a specific COMM port that has been mentioned below in Figure 9.

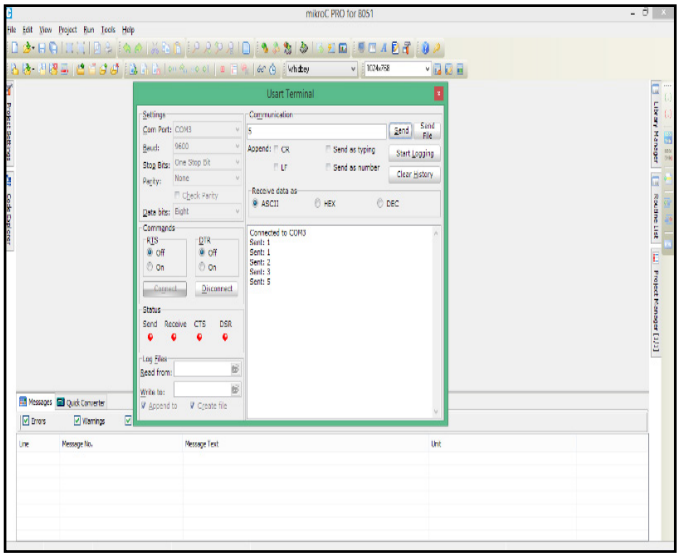


Figure 9: mikroC PRO Programming

5.4. *Genius Programmer 540*

The GENIUS programmer has its own driver that has to be installed. Afterwards the programmer is connected to computer and a window appears as shown in Figure 10 and the specific microcontroller is selected and the hex file of the code from KEIL is loaded into the controller and hence the burning is done.

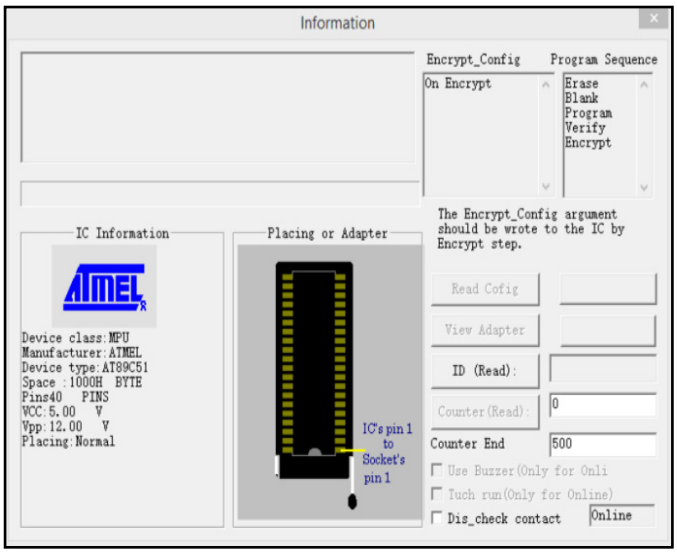


Figure 10: Genius Programmer Operation

6. Results

The designed home automation system was tried and tested several times and verified to control different home appliances used in the lighting system, air conditioning system, heating system, home entertainment system and many more. This follows the condition that the maximum power and current rating of the appliance should not exceed that of the used relay.

Figure 11 shows the installed system in ready state. An android device is ready to send the commands. The server has been installed in a PC having a Wi-Fi connection and the hardware circuit has been attached through a serial port.

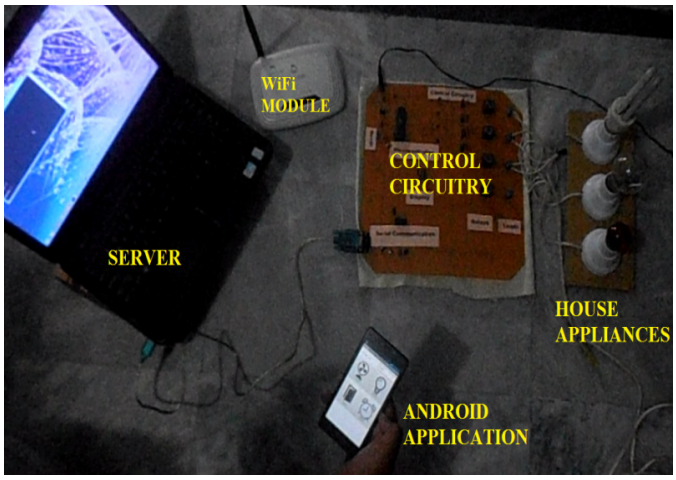


Figure 11: Complete System in Operation

For testing, three bulbs are being used as home appliances. First of all an IP address of the server is to be configured. The IP address is entered into an application. The socket number is fixed that is 4444. As soon as the button has been pressed from the device, the data has been received by PC server and sent to circuit where microcontroller is governing which bulb to be turned on. The turning on of a bulb is shown in Figure 12. For implementation, the bulbs can be replaced by any of the appliance working on 220V. The system was also successfully tested in a room for its fans, lights and air conditioner.

Figure 12: Complete System in Working

7. Conclusions and Recommendation

It is evident from this work that an individual control home automation system can be cheaply made from low-cost easily available components and can be used to control various home appliances ranging from the security lamps, the television to the air conditioning system and even the entire house lighting system. Even better, the parts required are so little and few that they can be bundled into a little slight container.

Finally, this home automation system can be also implemented over Bluetooth, Infrared and WAP connectivity

without much change to the design and yet still be able to control a variety of home appliances. Hence, this system is scalable and flexible.

It is recommended to use a serial Wi-Fi module as a server for ease and where cost is not an issue. Further it would be recommended for the followers to achieve such a system that would also be able to monitor the devices in addition to their control. This can easy be implemented by the usage of TX pin of the microcontroller.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors wish to thank University of Engineering and Technology, Lahore, Pakistan for providing an excellent platform for the work.

References

[1] Reinisch, C. Kastner, W. Neugschwandtner, G. Granzer, "Wireless Technologies in Home and Building Automation," Industrial Informatics, 2007 5th IEEE International Conference on , vol. no 1., pp.93-98, 23-27 June 2007

[2] Jinsoo Han; Chang-Sic Choi; Ilwoo Lee, "More efficient home energy management system based on ZigBee communication and infrared remote controls," Consumer Electronics, IEEE Transactions on , vol.57, no.1, pp.85,89, February 2011.

[3] Erdem, H.; Uner, A., "A multi-channel remote controller for home and office appliances," Consumer Electronics, IEEE Transactions on , vol.55, no.4, pp.2184,2189, November 2009.

[4] R. P. Chaudhari and M. K. Chopade, "Ethernet based field control module for industrial process monitor and control using ATmega328," 2016 International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur, 2016, pp. 2098-2101.

[5] V. Moravcevic, M. Tucic, R. Pavlovic and A. Majdak, "An approach for uniform representation and control of ZigBee devices in home automation software," 2015 IEEE 5th International Conference on Consumer Electronics - Berlin (ICCE-Berlin), Berlin, 2015, pp. 237-239.

[6] V. Puri and A. Nayyar, "Real time smart home automation based on PIC microcontroller, Bluetooth and Android technology," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, pp. 1478-1484.

[7] Lee, Minseok, Younggi Kim, and Younghee Lee. "A home cloud-based home network auto-configuration using SDN." Networking, Sensing and Control (ICNSC), 2015 IEEE 12th International Conference on. IEEE, 2015.

[8] Bian, Jiali, Dengke Fan, and Junming Zhang. "The new intelligent home control system based on the dynamic and intelligent gateway." Broadband Network and Multimedia Technology (IC-BNMT), 2011 4th IEEE International Conference on. IEEE, 2011.

[9] Hua, Liu Yan, and Zhang Ji Xiang. "Smart home based on the Zigbee wireless." Intelligent Networks and Intelligent Systems (ICINIS), 2012 Fifth International Conference on. IEEE, 2012.

[10] L. Mingfu and L. H. Ju , " Design and Implementation of Smart Home Control System Based on Wireless Sensor Networks and Power Line Communications ", IEEE Trans. Ind. Electron. , vol. 62 , no. 7 , pp. 4430 – 4442 , July 2015 .

[11] A. Zanella , N. Bui , A. Castellani , L. Vangelista , and M. Zorzi , " Internet of Things for Smart Cities ", IEEE Internet of Things Jou. , vol. 1 , no. 1 , pp. 22 – 32 , Feb. 2014 .

[12] F. Domínguez, A. Touhafi, J. Tiete and K. Steenhaut, "Coexistence with WiFi for a Home Automation ZigBee product," 2012 19th IEEE Symposium on Communications and Vehicular Technology in the Benelux (SCVT), Eindhoven, 2012, pp. 1-6.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 240-247 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

Internet of Things: An Evolution of Development and Research area topics

Jorge Oliveira e Sá*, João Cacho Sá[1], José Luís Pereira[2], Francisco Pimenta[3], Manuel Monteiro[4]

[1]*joao.cacho@fe.up.pt, MiEEC, Faculty of Engineering of the University of Porto, 4200-465 Porto, Portugal.*

[2]*jlmp@dsi.uminho.pt, Information Systems Department, University of Minho / ALGORITMI Center, 4800-058 Guimarães, Portugal.*

[3]*francisco.pimenta@fe.up.pt, MiEEC, Faculty of Engineering of the University of Porto, 4200-465 Porto, Portugal.*

[4]*manuel.monteiro@hpsl.pt, Hospital Particular São Lucas, Lagoa, Açores, Portugal.*

ARTICLE INFO

Article history:
Received: 13 December, 2016
Accepted: 18 January, 2017
Online: 28 January, 2017

Keywords :
Alliance for the Internet of Things Innovation
European Research Cluster on the Internet of Things
IoT Research Roadmap
Strategic Research Agenda
European Union
Innovation

ABSTRACT

Internet of Things (IoT) is a hot topic in the Europe Union (EU). In the year of 2015 the EU established an Alliance of Internet of Things Innovation (AIOTI) and this alliance included the IoT European Research Cluster (IERC) on the Internet of Things in the work group 01. IERC was established in 2007 and addresses the large potential for Internet of Things-based capabilities in Europe and to coordinate the convergence of ongoing activities. This organization regularly publishes Strategic Research Agendas with short, medium and long term forecasts of development and research topics. The aim of this paper is to study these Strategic Research Agendas published in the years 2010, 2011, 2012, 2014 and 2016 and make an identification of topics, in order to understand its evolution over time in terms of areas that have already been researched (or the research has already started) as well as those not yet explored. This work was performed based on a literature review methodology to identify the evolution in all the Strategic Research Agendas published in the years of 2010, 2011, 2012, 2014 and 2016.

1. Introduction

The creation by the European Union (EU) of the European Research Cluster on the Internet of Things (IERC) in the year of 2007, was provided due to the need to bring together EU-funded projects in this area, thereby trying to define a common vision as well as define Internet of Things (IoT) [1]. However, the objectives of IERC are broader, as it aims to create and develop insight into IoT research activities in Europe, as well as to define a strategy for cooperation with non-European entities in the IoT area.

More recently EU established an Alliance for Internet of Things Innovation (AIOTI) in March 2015 with the purpose of supporting the development of the most dynamic and agile IoT ecosystem and industry in the world. This ecosystem, built on the work of the IoT European Research Cluster (IERC), seeks to spill over innovation across industries and business sectors of IoT transforming ideas into solutions and business models [2].

*Corresponding Author: Jorge Oliveira e Sá, Information Systems Department, University of Minho /ALGORITMI Center, 4800-058 Guimarães, Portugal, jos@dsi.uminho.pt

Other objectives of the AIOTI include: fostering experimentation, replication, and deployment of IoT and supporting convergence and interoperability of IoT standards; gathering evidence on market obstacles for IoT deployment; and mapping and bridging global, EU, and member states' IoT innovation activities.

In this way, IERC has published Strategic Research Agendas (SRAs) since 2010, which intend to give a list of fields to be developed and researched as well as the paths for the future. These agendas have been updated according to the problems found in the investigations carried out. This work intends to make a comparison between the first SRA published in 2010, with the SRAs of 2011, 2012, 2014 and particularly with the SRA published in 2016. This comparative analysis intends to show the evolution of the topics proposed during this time period.

This work has a methodology based on literature review in the several SRAs proposed by IERC, intends to answer the question "What is relevant to research in IoT?"

Thus, in section 2 of this paper the definition of IoT is presented, in section 3 a brief explanation of AIOTI is presented and in section 4 the various SRAs published by IERC will be presented. SRAs were published in the years 2010 SRA [3], 2011 SRA [4], 2012 SRA [5], 2014 SRA [6] and 2016 SRA [2]. In this section the topics related to the area of technological development as well as the topics related to the research area will be mentioned, due to the extension of the goals of each topic it is not possible to detail them, however, the number of changes / evolutions that have suffered is presented and discussed. In section 5 it will be carried out the comparison between the different SRAs, it is intended to identify topics that are interesting to both academic and research communities. Section 6 will contain the conclusions.

2. Internet of Things definition

Although IoT is a recent buzzword, it was nicknamed this way for the first time in 1999 by [7]. However, the concept was developed many years earlier, in the early 1980s, by a group of programmers at Carnegie Melon University who modified a Coca-Cola machine by linking it to the Internet to obtain information on the availability of the drink and the time that the bottle spent inside the machine to ensure that the beverage was chilled and thus prevent them from making the trip in vain [8]. It is a simple example, but it explains the concept.

[7] refers that "today computers – and, therefore, the internet – are almost wholly dependent on human beings for information. Nearly all of the roughly 50 petabytes (a petabyte is 1,024 terabytes) of data available on the Internet were first captured and created by human beings – by typing pressing a record button, taking a digital picture or scanning a bar code. Conventional diagrams of the Internet include servers and routers and so on, but they leave out the most numerous and important routers of all - people. The problem is, people have limited time, attention and accuracy – all of which means they are not very good at capturing data about things in the real world".

The IERC definition states that IoT is "A dynamic global network infrastructure with self-configuring capabilities based on standard and interoperable communication protocols where physical and virtual 'things' have identities, physical attributes, and virtual personalities and use intelligent interfaces, and are seamlessly integrated into the information network" [9].

3. Alliance for Internet of Things Innovation

The AIOTI was formally launched on the 26th of March 2015, in the past year the Alliance has successfully contributed to convergence and interoperability of IoT standards, as well as the Digitizing European Industry policy. To achieve the remainder of its mission the Alliance plans to engage with building IoT Innovation Ecosystems from the ground up, and to establish communication channels with startup ecosystems, early adopter communities, European regions, global IoT initiatives and new potential markets [10].

The AIOTI is divided in the following groups: Management Board (MB), General Assembly (GA), Working Groups (WG) and Steering Groups (SG).

There are thirteen WGs, see Figure 1, divided in four horizontal and nine vertical groups, with a Chair and a Co-Chair, nominated and elected by that WG's members. Every WG elects one representative for the SG.

The SG members are the WG Chairs. They are assigned for a period of two years, to a maximum of six (three periods). The EU Commission is a special member.

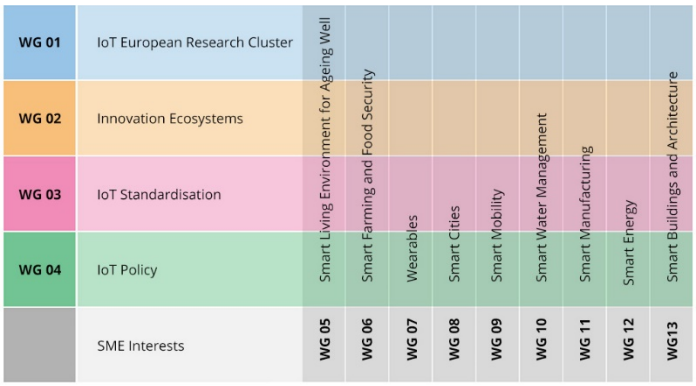


Figure 1 – AIOTI Working Groups [10]

WG 01 – IoT European Research Cluster, the IERC compares EU-funded innovation research and development programs, with the aim of defining a common vision of IoT technology and addressing European research challenges.

WG 02 – Innovation Ecosystems, this WG aims at designing actions to develop innovation ecosystems by stimulating startups, encouraging the use of open IoT platforms, enabling Large Scale Pilots, and linking large and small companies through open innovation.

WG 03 – IoT Standardization, this WG identifies and, where appropriate, makes recommendations to address existing IoT standards, analyses gaps in standardization, and develops strategies and use cases aiming for (1) consolidation of architectural frameworks, reference architectures, and architectural styles in the IoT space, (2) interoperability and (3) personal data & personal data protection to the various categories of stakeholders in the IoT space.

WG 04 – IoT Policy, this WG identifies, and, where appropriate, makes recommendations to address existing and potential barriers that prevent or hamper the take-up of IoT in the context of the Digital Single Market.

WG 05 – Smart living environment for ageing well, the topic for this WG refers to smart homes and smart living environments that can support vulnerable people, such as, but not limited to elderly or disabled people, in staying active, independent and out of institutional care settings, also leading to reduced costs for care systems and better quality of life for vulnerable categories of citizens. The WG deliverables include white papers, recommendation reports and innovative use cases susceptible to improve the quality of life of Elderly people using the latest IoT technologies.

WG 06 – Smart Farming and Food Security, the topic of this WG refers to IoT scenarios/use cases that allow monitoring and control of the plant and animal products life cycle "from farm to fork".

WG 07 – Wearables, the topic for this WG refers to IoT solutions that integrate key technologies (e.g. nano electronics, organic electronics, sensing, actuating, communication, low power computing, visualization and embedded software) into intelligent systems to bring new functionalities into clothes, other fabrics, patches, watches and other body-mounted devices. The WG

focuses its work on healthcare, well-being, safety, security and infotainment applications.

WG 08 – Smart Cities - the topic for this WG refers to IoT solutions used by a city in order to enhance performance, safety and wellbeing, to reduce costs and resource consumption, and to engage more effectively and actively with its citizens. Key 'smart city' sectors may include transport, energy, healthcare, lighting, water, waste and other city related sectors.

WG 09 – Smart Mobility, the topic for this WG refers to IoT solutions that allow for increased multi-modal mobility, more efficient traffic management, a dynamic road infrastructure, automated road tolling, usage based insurance and improved policy making through the analysis of road usage data provided by smart vehicles including autonomous and connected cars.

WG 10 – Smart Water Management, the topic for this WG refers to IoT solutions that improve water management efficiency by monitoring and controlling surface water retention, flooding etc.

WG 11 – Smart Manufacturing, the topic for this WG refers to IoT solutions that bring together information, technology and human ingenuity to achieve a rapid revolution in the development and application of manufacturing intelligence to every aspect of business.

WG 12 – Smart Energy, the topic for this WG refers to IoT solutions deployed by various companies along the value chain (i.e. IoT technology providers, energy companies (in generation, supply, grid and market participants, traders, aggregators, etc.) to allow the performance optimization of their energy asset portfolios (Renewables plants, Grid Substations, Control Rooms, Prosumer Demand Responsive Loads and EV Charging infrastructures).

WG 13 – Smart Building and Architecture, the topic of this WG is the IoT technologies and solutions deployed in buildings and districts of buildings to improve life of the occupant by addressing and optimizing elements such as comfort, light, temperature, air quality, water, nourishment, fitness, and energy usage.

4. European Research Cluster on the Internet of Things

The European Union (EU) started the IERC due to the need to unite projects funded by the EU in the IoT area. The IERCs main purpose is to define a common vision as well as the technologies for IoT [1]. But IERC goals are more embracing because it aims to create and develop a unified vision of all the research on IoT in Europe as well as a cooperation strategy with non-European entities in the IoT field.

The IERC has made available Strategic Research Agendas (SRAs) since 2010 to achieve its goal. These SRAs aim to select research fields as paths for the future.

The SRAs were published in the last years and they have suffered constant updates. The first SRA was released in 2010 and more have been issued in 2011, 2012, 2014 and 2016. Since each SRA has many topics and objectives, it is not possible to describe them in detail but it is possible to define the changes/updates that they suffered in terms of dimension.

The creation of the SRAs by the IERC through the years allowed the identification of a group of areas that are needed to implement successfully IoT in Europe in both short term and long term. Since some of these areas weren't successfully implemented in the SRA they were planned initially, they have been postponed into latter SRAs. The SRAs divide two areas among what needs "Research" and what needs "Development". In this section it will only be given relevance to new topics or new objectives for topics that have been issued in the different published SRAs.

4.1. *2010 SRA*

This was the first SRA to be issued by the IERC two years after the creation of the work team that starter working in the initial concept of IoT. In this first SRA it is done a survey of the areas in which the IoT is or may be present and also identify the areas that deserve greater attention by the practitioners and research community.

The 2010 SRA became the foundation for the remaining SRAs that were published in the following years. The SRA identifies, separately, what needs to be "Development" at the technological level and what needs to be "Research" in order to address existing needs.

In the "Development" area, more precisely in the technological level, there were identified twelve topics, namely: (1) *Identification Technology*, (2) *IoT Architecture Technology*, (3) *Communication Technology*, (4) *Network Technology*, (5) *Software and Algorithms*, (6) *Hardware Devices*, (7) *Data and Signal Processing Technology*, (8) *Discovery and Search Engine Technologies*, (9) *Power and Energy Technologies*, (10) *Security, Privacy & Trust Technologies*, (11) *Material Technology* and (12) *Standardization*.

In the "Research" area there were identified fifteen topics, namely: (1) *Identification Technology*, (2) *IoT Architecture*, (3) *SOA Software Services for IoT*, (4) *IoT Architecture Technology*, (5) *Communication Technology;* (6) *Network Technology;* (7) *Software and Algorithms,* (8) *Hardware Devices,* (9) *Hardware Systems, Circuits and Architectures*, (10) *Data and Signal Processing Technology,* (11) *Discovery and Search Engine Technologies,* (12) *Power and Energy Technologies*, (13) *Security, Privacy & Trust Technologies,* (14) *Material Technology,* (15) *Standardization*.

These topics are divided chronologically by four temporal periods: the first is the time period before 2010, the second is in an interval of years between 2010 and 2015, the third period is in another interval of years between 2015 and 2020 and finally the last period corresponds the years beyond 2020. This SRA is the only one that incorporated the time period before 2010, so this time period is not considered in this study.

4.2. *2011 SRA*

This SRA continues the work carried out by the previous SRA (2010), since only one year passed from its publication. However, the time periods applied in this SRA were adjusted to the date when it was published (2011), thus the first period corresponds to the years from 2011 to 2015, the second from 2015 to 2020 and finally the third time period is beyond 2020.

In this SRA the most of the topics have remained unchanged. However, three topics have undergone changes, one in "Development" at the technological level, the topic (6) *Hardware Devices* has incorporated the addition of another goal for the interval of the years 2011 to 2015, this goal consists of the integration of NFC in mobile phones and Sensors. The other two in "Research", to the time period 2011-2015, the topic (1) *Identification Technology* has incorporated a new goal (Convergence of IP and IDs and addressing scheme) and the topic

(6) *Network Technology* has incorporated another goal, this goal consists of incorporating systems based on RFID sensors.

The three new goals were the only ones to be added in all the time periods present in the 2011 SRA [3].

4.3. *2012 SRA*

In this SRA it is possible to identify some differences in relation to the previous SRAs, mainly due to the development and research in IoT being more mature, since two years have passed since the launching of the 2010 SRA. Once again, the temporal periods used in this SRA have been adjusted to the date it was published (2012). The time periods are: years between 2012 and 2015, years between 2016 and 2020; and maintaining the same period of previous SRAs, i.e., beyond the year 2020.

Since the knowledge about the IoT topic increases, new topics are likely to emerge. Thus, in this SRA, there are seven new topics that are divided as follows: two new topics for "Development" and five new topics for "Research".

However, the 2012 SRA is not only marked by the introduction of new topics, but also by giving great relevance to security technologies for IoT.

In addition to the new topics, in this SRA are added twenty-eight new goals distributed across different areas and topics, in addition to being distributed over different time periods.

In the "Development" area are twenty-four new goals. Of these, twelve arise for the time period from 2012 to 2015 and are distributed by the topics as follows: one in (5) *Software and Algorithms*; two in (6) *Hardware*; one in (12) *Standardization*; and three in (10) *Security, Privacy & Trust Technologies*, the others five goals are in the new topics (13) *IoT Infrastructure* and (14) *IoT Applications*. Eight goals goes to the time period of 2016-2020 and are three to the topic (10), one for the topic (12) and four to the new topics (13) and (14). The remaining four goals goes to the time period 2015-2020, one goes to the topic (10), one goes to the topic (11) *Material Technology*, and two goes to the new topics (13) and (14).

In the "Research" area arise thirty-eight new goals of which twenty of them are present in the time period between 2012 and 2015, namely: one in (9) *Hardware Systems, Circuits and Architectures*; two in (3) *SOA Software Services for IoT*; and seven are present in (13) *Security, Privacy & Trust Technologies*, the remaining ten are in the new topics (16) *IoT Infrastructure*, (17) *IoT Applications*, (18) *Societal Responsibility*, (19) *Governance (Legal Aspects)*, and (20) *Economic*. Twelve goals are present in time period 2016-2020 and one is in the topic (3), another one in topic (9), four in topic (13) and six in the new topics (16), (17), (18) and (19). In the time period beyond 2020, are the remaining six goals, one in topic (3), two in topic (13), one in topic (14) *Material Technology*, and the last two in the new topics (16) and (17) [4].

4.4. *2014 SRA*

In this SRA arouse a new topic named *Interoperability.* It appeared in both the "Development" area in which it is numbered as the topic (15) and in the "Research" area, in which it becomes topic (21).

This SRA has only two time periods, one is the time period 2015-2020 and the other is beyond 2020. Thus is more difficult to compare this SRA with the previous. However each topic has different number of goals from the 2012 SRA.

In the "Development" area the topic (1) *Identification Technology* had seven goals in 2012 and now has six. Topic (2) *IoT Architecture Technology* changed from eight to six goals. Topic (3) *Communication Technology* augmented from seven to eight goals. Topic (4) *Network Technology* diminished from eight to seven goals. Topic (5) *Software and Algorithms* changed from twelve to thirteen goals. Topic (6) *Hardware* reduced from nine to five goals. Topic (7) *Data and Signal Processing Technology* changed from four to three goals. Topic (8) *Discovery and Search Engine Technology* changed from five to four goals. Topic (9) *Power and Energy Storage Technology* changed from nine to seven goals. Topic (10) *Security, Privacy and Trust Technology* reduced from fourteen to ten goals. Topic (11) *Material Technology* changed from five to four goals. Topic (12) *Standardization* decreased from seven to five goals. Topic (13) *IoT Infrastructure* changed from seven to six goals. Topic (14) *IoT Applications* increased from four to eight goals. The new topic (15) *Interoperability* has five goals.

In the "Research" area the topic (1) *Identification Technology* has seven goals in 2012 and now has six. Topic (2) *IoT Architecture* change from two to one goal. Topic (3) *SOA Software Services for IoT* reduced from six to four goals. Topic (4) *IoT Architecture Technology* change from thirteen to twelve goals. Topic (5) *Communication Technology* augment from eight to eleven goals. Topic (6) *Network Technology* change from fourteen to thirteen goals. Topic (7) *Software and Algorithms* maintain the same fourteen goals. Topic (8) *Hardware Devices* reduce from twenty-three to seventeen goals. Topic (9) *Hardware Systems, Circuits and Architectures* reduced from twenty to fourteen goals. Topic (10) *Data and Signal Processing Technology* maintain the same five goals. Topic (11) *Discovery and Search Engine Technology* maintain the same seven goals. Topic (12) *Power and Energy Storage Technology* reduce from ten to six goals. Topic (13) *Security, Privacy and Trust Technology* reduced from twenty to sixteen goals. Topic (14) *Material Technology* change from seven to four goals. Topic (15) *Standardization* has zero goals, so this topic is no long be considered. Topic (16) *IoT Infrastructure* change from six to five goals. Topic (17) *IoT Applications* increase from five to nine goals. Topic (18) *Societal Responsibility* has zero goals, so this topic is no long be considered. Topic (19) *Governance (Legal Aspects)* maintains the same three goals. Topic (20) *Economic* change from one to two goals. Finally, the new topic (21) *Interoperability* has three goals [5].

4.5. *2016 SRA*

This is the last SRA published. This SRA the topic *Material Technology* in "Development" – topic (11) and "Research" – topic (14) has zero goals.

In "Development" area, to the time period of 2016-2020, four topics increase in seven the number of goals, one in the topic (2) *IoT Architecture Technology,* one in topic (5) *Software and Algorithms*, two in topic (3) *Communication Technology* and three in topic (14) *IoT Applications*. For the time period beyond 2020, five topics increase the number of goals, topic (1) *Identification Technology* has a new goal, topic (3) has one new goal, topic (10) *Security, Privacy and Trust Technology* has three new goals, topic (14) has ten new goals and topic (15) *Interoperability* has one more goals.

In "Research" area, to the time period of 2016-2020, eight topics change their number of goals, five new goals to topic (3) *SOA Software Services for IoT,* reduction from twelve to ten goals in topic (7) *Software and Algorithms*, two more goals in topic (8) *Hardware Devices*, one more goal in topic (9) *Hardware Systems, Circuits and Architectures*, two more goals in topic (10) *Data and Signal Processing Technology*, reduction of one goal in topic (17) *IoT Applications*, increase of two goals in topic (19) *Governance (Legal Aspects)*, and one more goal in topic (20) *Economic* [2].

5. Discussion of SRA evolution

To the "Development" area, all topics and number of goals are presented in Table 1. So, in this table, it is possible to see the evolution of the goals that each topic underwent throughout the various SRAs. The first line represents all the topics published, the columns represent the fifteen topics and the lines represent the five SRAs (2010, 2011, 2012, 2014 and 2016) to each one it is presented the various time periods. Each cell on the table contains a number that represents the number of goals that each topic had in each SRA (time period), and a hyphen (-) means that the topic was not covered in the specific SRA and time period.

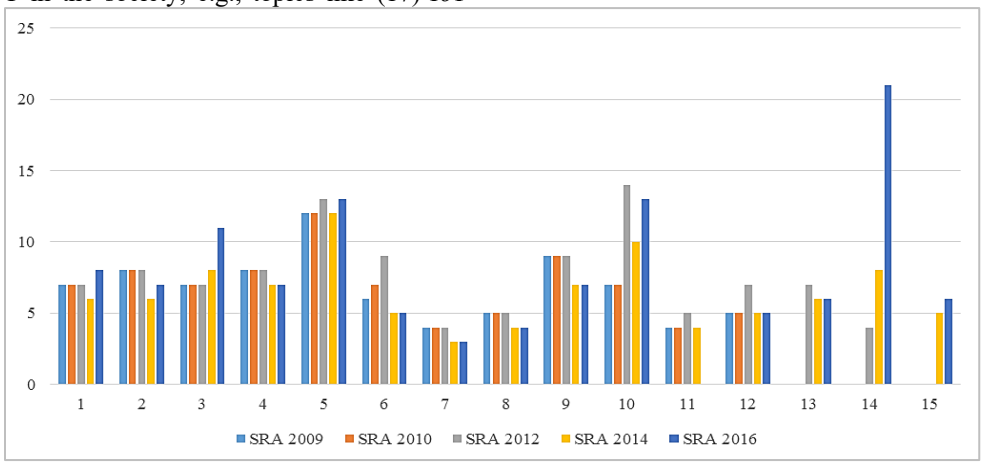
In Table 1 the topic (14) *IoT Applications* is to be highlighted because this topic has only been included in the 2012 SRA with four goals and in the 2016 SRA this topic had twenty-one goals, which is the biggest value in the last SRA. There are other two topics with high importance, topic (3) *Communication Technology* with a growth in 2016 from eight to twelve goals and topic (10) *Security, Privacy and Trust Technology* that presented fourteen goals in 2012, now presents a growth from ten to thirteen goals since the 2014 SRA. Topic (5) *Software and Algorithms* has a high and constant number (12-13 goals) in all SRAs. The other topics have an almost constant number of goals in the five SRAs, see Figure 2.

Analyzing Table 2, that concerns the "Research" area, comparing the 2010 SRA to the 2011 SRA it is verified that the 2011 SRA presents a greater concern in defining some topics that were not defined in the year of 2010, as is the case of topic (6) *Network Technology* that was adapted to the reality of the area of IoT [2,3].

Comparing now the 2011 SRA and 2012 SRA it can be concluded that the 2012 SRA presented five new topics. When analyzing the topics that are being added we can verify the existence of the first concerns regarding the impact of the implementation of IoT in the society, e.g., topics like (17) *IoT Applications*, (18) *Societal Responsibility*, (19) *Governance (Legal Aspects)* and (20) *Economics*. Through the observation of the goals that appear in this SRA, it can be concluded that the goals are related to two subjects, namely: processes, with the creation of infrastructures / standards that allow the integration of IoT; and the use in an industrial environment through the use of sensors. Related to topic (13) *Security, Privacy and Trust Technology*, there are a greater number of goals that focus on privacy techniques to remain anonymous, techniques that ensure users' privacy as well as their data, new methods for assessing trust in devices and data and methods to ensure platform and data security. However, in 2012 SRA there is a situation that needs to be emphasized, which consists in the suppression of the topic (15) *Standardization*, this is justified by the redefinition and new placement of the topic (15) goals to other topics [3,4].

Table 1 – "Development" Area

SRAs	Topic	Identification Technology	IoT Architecture Technology	Communication Technology	Network Technology	Software and Algorithms	Hardware	Data and Signal Processing Technology	Discovery and Search Engine Technology	Power and Energy Storage Technology	Security, Privacy and Trust Technology	Material Technology	Standardization	IoT Infrastructure	IoT Applications	Interoperability
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
SRA 2010	2010-2015	3	4	5	5	4	3	2	2	3	3	3	3	-	-	-
	2015-2020	3	2	1	1	3	2	1	1	4	3	1	1	-	-	-
	Beyond 2020	1	2	1	2	5	1	1	2	2	1	-	1	-	-	-
SRA 2011	2011-2015	3	4	5	5	4	4	2	2	3	3	3	3	-	-	-
	2015-2020	3	2	1	1	3	2	1	1	4	3	1	1	-	-	-
	Beyond 2020	1	2	1	2	5	1	1	2	2	1	-	1	-	-	-
SRA 2012	2012-2015	3	4	5	5	5	6	2	2	3	6	3	4	3	2	-
	2016-2020	3	2	1	1	3	2	1	1	4	6	1	2	3	1	-
	Beyond 2020	1	2	1	2	5	1	1	2	2	2	1	1	1	1	-
SRA 2014	2015-2020	5	4	6	4	6	4	2	2	5	8	2	4	4	6	4
	Beyond 2020	1	2	2	3	6	1	1	2	2	2	2	1	2	2	1
SRA 2016	2016-2020	5	5	8	4	7	4	2	2	5	8	-	4	4	9	4
	Beyond 2020	3	2	3	3	6	1	1	2	2	5	-	1	2	12	2

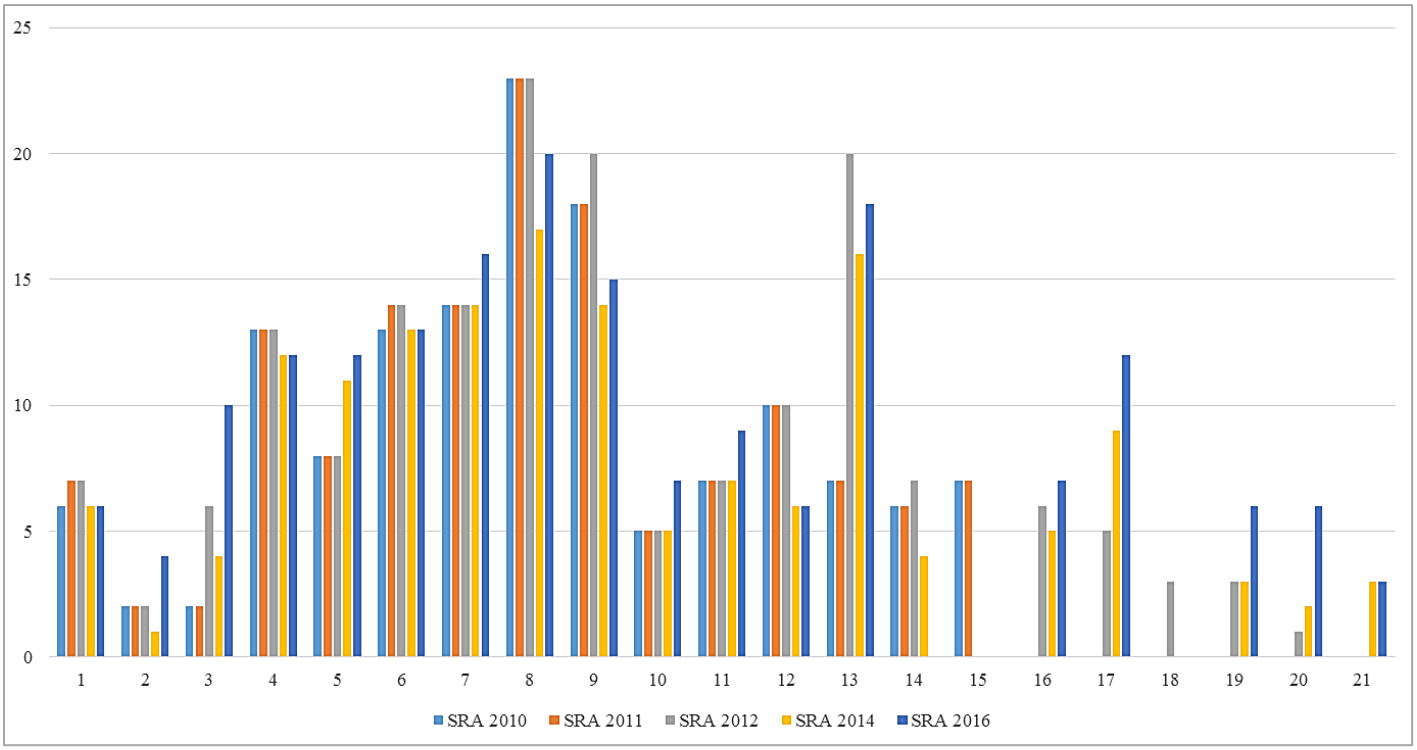


The comparison between years 2012 and 2014 will be carried out, the latter is already focused on the time period from 2015 to 2020. In this comparison the number of goals that are delayed between the 2012 SRA and 2014 SRA will be verified. In the 2014 SRA a total of one hundred and twenty-two goals (122) are identified. These goals are spread over nineteen topics for the time period between 2015 and 2020, with forty-seven of these goals being carried over from the previous time period, because they were not fulfilled in a timely manner, and twenty-eight new goals have emerged. There are two topics that stand out: (17) *IoT Applications*, which includes mobile applications for IoT; and (5) *Communications Technology* which includes the new networks, as is the case of 5G referenced in the goals of the 2014 SRA. In the 2014 SRA the topic (18) *Social Responsibility* disappears, however its goals were not distributed to other topics with the justification that they cannot be applied while IoT does not reach a higher maturity [4,5].

Lastly, it will be carried out the comparison between the years 2014 and 2016. The first highlight is that the number of goals increased from one hundred forty-eight (148) to one hundred eighty-two (182) that represents a growth of thirty-four goals. This happens essentially in the time period of beyond 2020. Generally all topics increased the number of goals for all time periods, but in the time period of beyond 2020 the number of goals increased from thirty in 2014 SRA to fifth-three in 2016 SRA, which represents an increase rate of 77%. However, the topic (14) *Material Technology* was suppressed. It is concluded that the 2016 SRA presents a higher number of goals comparing with the previous SRAs and the justification for this growth is the higher number of goals for almost all topics in the time period of beyond 2020 [5,6]. These numbers may be seen at Figure 3.

Table 2 – "Research" Area

SRAs	Topic	Identification Technology	IoT Architecture	SOA Software Services for IoT	IoT Architecture Technology	Communication Technology	Network Technology	Software and Algorithms	Hardware Devices	Hardware Systems, Circuits and Architectures	Data and Signal Processing Technology	Discovery and Search Engine Technology	Power and Energy Storage Technology	Security, Privacy and Trust Technology	Material Technology	Standardization	IoT Infrastructure	IoT Applications	Societal Responsibility	Governance (Legal Aspects)	Economic	Interoperability
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
SRA 2010	2010-2015	4	1	1	7	4	6	8	12	10	2	4	6	2	5	3	-	-	-	-	-	-
	2015-2020	1	1	1	5	3	3	4	9	6	2	2	3	4	1	2	-	-	-	-	-	-
	Beyond 2020	1	-	-	1	1	4	2	2	2	1	1	1	1	-	2	-	-	-	-	-	-
SRA 2011	2011-2015	5	1	1	7	4	7	8	12	10	2	4	6	2	5	3	-	-	-	-	-	-
	2015-2020	1	1	1	5	3	3	4	9	6	2	2	3	4	1	2	-	-	-	-	-	-
	Beyond 2020	1	-	-	1	1	4	2	2	2	1	1	1	1	-	2	-	-	-	-	-	-
SRA 2012	2012-2016	5	1	3	7	4	7	8	12	11	2	4	6	9	5	-	2	3	2	2	1	-
	2016-2020	1	1	2	5	3	3	4	9	7	2	2	3	8	1	-	3	1	1	1	-	-
	Beyond 2020	1	-	1	1	1	4	2	2	2	1	1	1	3	1	-	1	1	-	-	-	-
SRA 2014	2015-2020	5	1	3	8	10	8	12	15	12	4	6	5	13	3	-	4	7	-	2	2	2
	Beyond 2020	1	-	1	4	1	5	2	2	2	1	1	1	3	1	-	1	2	-	1	-	1
SRA 2016	2016-2020	5	1	8	8	10	8	10	17	13	6	6	5	13	-	-	4	6	-	4	3	2
	Beyond 2020	1	3	2	4	2	5	6	3	2	1	3	1	5	-	-	3	6	-	2	3	1



The Table 3 compares the numerical representation of goals in all topics in both areas "Development" and "Research". The interpretation of this table is similar to Table 1 and 2. When a topic does not exist in one area (i.e., not has any goal) it is represented by a hyphen (-). If the topic has goals, it is represented by two numbers, the first number belongs to the "Development" area and the second number concerns the "Research" area, e.g., (3/1).

From this study it is identified that there are relevant topics for those who want to research in IoT subject. It is not intended to carry out research on "Development" area, since it is considered that the market already responds or will respond to the needs of IoT. In the "Research" area, the topic (21) *Interoperability* is one of the most important in IoT, because "things" need protocols to communicate with other "things", to accomplish this there is a need for an *IoT Architecture* (topic 2). In this IoT ecosystem there is the need for providing services (topic 3 *SOA Software Services for IoT*). These services are delivered by *IoT Applications* (topic 17). These applications will be executed on *Hardware devices* (topic 8) in a secure way, (topic 13 *Security, Privacy & Trust Technologies*) and should return value (topic 20 *Economic*). It is considered that these topics may be relevant to research, especially in the actual state of maturity of IoT. Regarding the topics (18) *Societal Responsibility* and (19) *Governance (legal aspects)* could also be considered, but as mentioned above IoT has not yet reached the state of maturity necessary to consider these topics [5,6].

Table 3 – Development and Research Area

SRAs / Topic		Identification Technology	IoT Architecture	SOA Software Services for IoT	IoT Architecture Technology	Communication Technology	Network Technology	Software and Algorithms	Hardware (Devices)	Hardware Systems, Circuits and Architectures	Data and Signal Processing Technology	Discovery and Search Engine Technology	Power and Energy Storage Technology	Security, Privacy and Trust Technology	Material Technology	Standardization	IoT Infrastructure	IoT Applications	Social Responsability	Governance (Legal Aspects)	Economic	Interoperability
Development/Research		1/1	-/2	-/3	2/4	3/5	4/6	5/7	6/8	-/9	7/10	8/11	9/12	10/13	11/14	12/15	13/16	14/17	-/18	-/19	-/20	15/21
SRA 2010	2010-2015	3/4	-/1	-/1	4/7	5/4	5/6	4/8	3/12	-/10	2/2	2/4	3/6	3/2	3/5	3/3	-/-	-/-	-/-	-/-	-/-	-/-
	2015-2020	3/1	-/1	-/1	2/5	1/3	1/3	3/4	2/9	-/6	1/2	1/2	4/3	3/4	1/1	1/2	-/-	-/-	-/-	-/-	-/-	-/-
	Beyond 2020	1/1	-/-	-/-	2/1	1/1	2/4	5/2	1/2	-/2	1/1	2/1	2/1	1/1	-/-	1/2	-/-	-/-	-/-	-/-	-/-	-/-
SRA 2011	2011-2015	3/5	-/1	-/1	4/7	5/4	5/7	4/8	4/12	-/10	2/2	2/4	3/6	3/2	3/5	3/3	-/-	-/-	-/-	-/-	-/-	-/-
	2015-2020	3/1	-/1	-/1	2/5	1/3	1/3	3/4	2/9	-/6	1/2	1/2	4/3	3/4	1/1	1/2	-/-	-/-	-/-	-/-	-/-	-/-
	Beyond 2020	1/1	-/-	-/-	2/1	1/1	2/4	5/2	1/2	-/2	1/1	2/1	2/1	1/1	-/-	1/2	-/-	-/-	-/-	-/-	-/-	-/-
SRA 2012	2012-2016	3/5	-/1	-/3	4/7	5/4	5/7	5/8	6/12	-/11	2/2	2/4	3/6	6/9	3/5	4/-	3/2	2/3	-/2	-/2	-/1	-/-
	2016-2020	3/1	-/1	-/2	2/5	1/3	1/3	3/4	2/9	-/7	1/2	1/2	4/3	6/8	1/1	2/-	3/3	1/1	-/1	-/1	-/-	-/-
	Beyond 2020	1/1	-/-	-/1	2/1	1/1	2/4	5/2	1/2	-/2	1/1	2/1	2/1	2/3	1/1	1/-	1/1	1/1	-/-	-/-	-/-	-/-
SRA 2014	2015-2020	5/5	-/1	-/3	4/8	6/10	4/8	6/12	4/15	-/12	2/4	2/6	5/5	8/13	2/3	4/-	4/4	6/7	-/-	-/2	-/2	4/2
	Beyond 2020	1/1	-/-	-/1	2/4	2/1	3/5	6/2	1/2	-/2	1/1	2/1	2/1	2/3	2/1	1/-	2/1	2/2	-/-	-/1	-/-	1/1
SRA 2016	2016-2020	5/5	-/1	-/8	5/8	8/10	4/8	7/10	4/17	-/13	2/6	2/6	5/5	8/13	-/-	4/-	4/4	9/6	-/-	-/4	-/3	4/2
	Beyond 2020	3/1	-/3	-/2	2/4	3/2	3/5	6/6	1/3	-/2	1/1	2/3	2/1	5/5	-/-	1/-	2/3	12/6	-/-	-/2	-/3	2/1

6. Conclusions

The years between 2010 and 2016 of the IoT cannot be considered successful years, because it is verified that the goals published in the first SRA in 2010 for the time period of 2010 to 2015, only 49.57% of the goals were fulfilled (fifty-seven of the one hundred and fifteen proposed), and this is equally distributed by both "Development" and "Research" areas.

However, some effort has been made in determining the conditions for the practicality of actual IoT implementations, most of this effort is directed towards the creation of frameworks that allow the implementation of IoT in any type of real environment.

The SRAs evolution is based on the results of development and research projects with cooperation between EU and other countries, while overseen by IERC. There are ten projects from FP7 Call8, e.g., Clout Project, that was funded in 2.3M€ from EU and in 1.6M€ from Japan. There are sixteen projects from FP7 Call5 and a long list of projects (about thirty-five) from FP7, FP6 and FP5. All the projects are finished and this is a signal of success, but there is a big area of topics and goals to be covered.

develop the most dynamic European Internet of Things ecosystem and to become a global influencer on IoT technology. The AIOTI has become a formal organization in the 2016 summer.

IoT is a hot subject all around the globe and the AIOTI's train is leaving the platform and nobody want to miss it.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work has been supported by COMPETE: POCI-01-0145-FEDER-007043 and FCT (Fundação para a Ciência e Tecnologia) within the Project Scope: UID/CEC/00319/2013.

References

[1] "About IERC," 2015, retrieved 2-12-2016: http://www.internet-of-things-research.eu/about_ierc.htm.

[2] O. Vermesan, P. Friess, P. Guillemin, M. Serrano, et al. "IoT Digital Value Chain Connecting Research, Innovation and Deployment", IERC Cluster, SRA, pp. 15-128, 2016.

[3] O. Vermesan, P. Friess, P. Guillemin, S. Gusmeroli, et al. "Internet of Things: Strategic Research Roadmap," Internet Things Strateg. Res. Roadmap, pp. 1–50, 2009.

[4] O. Vermesan, P. Friess, P. Guillemin, S. Gusmeroli, et al. "Internet of Things Strategic Research Roadmap," *Internet Things Strateg. Res. Roadmap*, pp. 9–52, 2010.

[5] M. Lerner, "Internet of Things 2012 New Horizons," p. 360, 2012.

[6] O. Vermesan and P. Friess, Internet of Things Applications - From Research and Innovation to Market Deployment. 2014.

[7] Ashton, K., That 'internet of things' thing, RFiD Journal, vol. 22, nº 7, pp. 97-114, 2009.

[8] Coke Machine, 1998, retrieved 2-12-2016 from http://www.cs.cmu.edu/~coke

[9] O. Vermesan, P. Friess, P. Guillemin, S. Gusmeroli, et al. "Internet of Things Strategic Research Agenda", Chapter 2 in Internet of Things – Global Technological and Societal Trends, River Publishers, 2011.

[10] AIOTI, "The Alliance for Internet of Things Innovation", 2016, retrieved 2-12-2016 from http://www.aioti.org/

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 248-254 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

Using Naming Patterns for Identifying Architectural Technical Debt

Paul Mendoza del Carpio*

Research Professor, Department of Software Engineering, Universidad La Salle, Peru

ARTICLE INFO

Article history:
Received: 10 December, 2016
Accepted: 18 January, 2017
Online: 28 January, 2017

Keywords:
Architectural technical debt
Naming pattern
Code analysis

ABSTRACT

Hasty software development can produce immediate implementations with source code unnecessarily complex and hardly readable. These small kinds of software decay generate a technical debt that could be big enough to seriously affect future maintenance activities. This work presents an analysis technique for identifying architectural technical debt related to non-uniformity of naming patterns; the technique is based on term frequency over package hierarchies. The proposal has been evaluated on projects of two popular organizations, Apache and Eclipse. The results have shown that most of the projects have frequent occurrences of the proposed naming patterns, and using a graph model and aggregated data could enable the elaboration of simple queries for debt identification. The technique has features that favor its applicability on emergent architectures and agile software development.

1. Introduction

This paper is an extension of work originally presented in the 8th Euro American Conference on Telematics and Information Systems (EATIS) [38]. Taking an easy solution on short-term in an activity of any phase of software development (i.e., requirements, design, implementation), can generate an accumulated technical debt, which, in a given period of time, can become big enough to affect future deliveries, making hard getting a successful outcome [6,24,37]. The debt comprises any aspect known as inappropiate which has not been addressed in due time (e.g., complex source code that needs refactoring) [24]. This debt is a topic whose interest has been increased over the years [36]. Frequently the technical debt, when is inserted, is less visible for decision makers in the software development [5]. The development of techniques for identifying and monitoring incidences of technical debt, is important for making explicit the debt and it could be resolved in due time [3,11,22,24,35,37].

The technical debt can be inserted by not complying the architectural design, or by not using conventions or standards of programming [35]. Including this as a decision factor inside the software development, requires information about the incidences of technical debt in the software system, where these are located, and their magnitude; such information can be gotten through source code analysis [5].

The objective of this work is to present:

1. An analysis technique for identifying architectural technical debt by non-uniformity of patterns.
2. A set of naming patterns across the package hierarchy of the software system.

2. Architectural Technical Debt (ATD)

ATD is a kind of technical debt which comprises sub-optimal solutions regarding internal or external quality attributes defined in the intended architecture, mainly compromising the attributes of maintainability and evolvability [2,11].

Changes related to design qualities but not related directly to external behavior of the system, are frequently postponed or neglected to reduce delivery time of the software system [3], increasing the incidences of ATD.

ATD is a debt very related to source code [24], however, in practice, is hard to be identified because this does not provide observable behavior to final users [11,36], and can change with time due to information gotten from implementation details [2]. Therefore, the ATD cannot be completely identified at an initial stage [2].

*Corresponding Author: p.mendozadc@ulasalle.edu.pe

In [2], a set of ATD is introduced. Among them, ATD by non-uniformity of patterns is related to name conventions applied in part of the system which are not followed in another parts [2]. This instance of ATD is addressed in this work.

Furthermore, several agile approaches consider the architecture as an emergent feature where there is no early design; but the source code is refactored and the architectural elements are refined [21]. The refactoring is a regular practice used in agile approaches, and is often applied on source code [1]; this contributes to the emergence of a successful architecture, improving the internal structure of the application, making the architectural elements more comprehensible, and avoiding the architecture decay, specially in them defined slightly [15,21]. Performing an incomplete refactoring is a cause of ATD that can insert part of ATD and generates new debt [2]. The refactoring can be performed manually, or semi or fully automatic. The fully automatic approach carry out the identification and transformation of code elements, nevertheless a human commits modifications [1,16]. This work enables a fully automatic refactoring, taking into account the identification by the proposed analysis, and applying a transformation through a renaming of classes. The last is a kind of global refactoring (i.e., affects classes in more than one package) [10] with API level (Application Programming Interface) [30], which is often used automatically in programming environments [8,30] with aims of organization and conceptualization [25], standing out over other refactoring forms by supporting the software traceability [1].

3. Naming Patterns

As a software evolves, its code becomes a source of information that is up to date and contains relevant information about the application domain [14]. Complex code is a major source of technical debt [22]; the correct use of naming conventions defined by the architecture accelerates and makes easy the activities of software comprehension [34]. Nevertheless, these conventions could not be followed throughout the software system. Such phenomenon can be amplified in agile teams [2]; where the teams are empowered in terms of design, different development teams working in parallel accumulates differences in design and architecture, and naming policies are not always defined explicitly and formally, arising divergences and requiring effort [2].

The relevance of class names lies in determining the code legibility, portability, maintainability, and accessibility to new team members, and relating the source code to the problem domain [19]. Also, industry experts highlight the importance of identifier names in software [12,28,31]. Therefore, such importance can reach architectural analysis levels, where identifying component terms is a task less complicated when identifiers are comprised by complete words or meaning acronyms [9,14]. The following subsections present a set of naming patterns inspired on the organization of source code through packages; the patterns are defined taking into account the frequent use of terms in class names inside the subjacent package hierarchy. Examples are taken from several real projects of the organizations Apache and Eclipse.

3.1. Pattern: Package

In this pattern the term is often used by classes included in a same package. As an example, figure 1 shows packages of Apache MyFaces. f is defined as a value of minimal frequency; T is the set of terms used in class names; P is the set of packages; C(p) is the set of classes of p ∈ P; and C(p,t) is the set of classes of p which have names with the term t ∈ T. The terms t of this pattern are such that (| C(p,t) | / | C(p) |) ≥ f, and | C(p) | > 2.

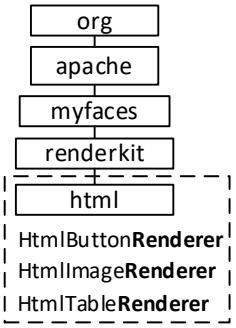


Figure 1: Example of Pattern Package

3.2. Pattern: Package Name

The term is often used by classes included in packages with same name. Figure 3 shows packages of Eclipse EGit. M is defined as the set of names of packages; F(m) is the set of packages with name m ∈ M; and F(m,t) is the set of packages with name m which contain classes having the term t in their names. The terms t of this pattern are such that (| F(m,t) | / | F(m) |) ≥ f, and | F(m) | > 2.

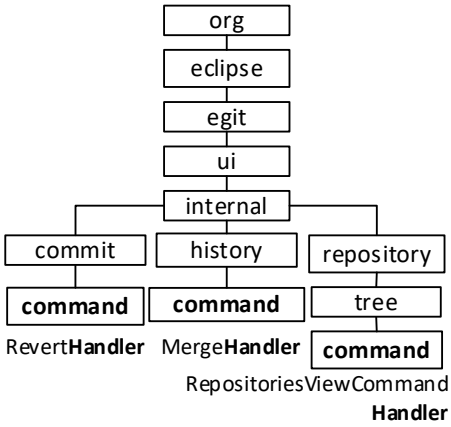


Figure 2: Example of Pattern Package Name

3.3. Pattern: Package Name and Level

The term is often used by classes included in packages with same name at same level of the package hierarchy. As an example, figure 3 shows packages of Apache Hadoop. N is defined as the set of package levels; G(n,m) is the set of packages with name m which are located at level n ∈ N; and G(n,m,t) is the set of packages with name m, at level n, which contain classes having the term t in their names. The terms t of this pattern are such that (| G(n,m,t) | / | G(n,m) |) ≥ f, and | G(n,m) | > 2.

3.4. Pattern: Package immediately superior

The term is often used by classes included in packages that are located in the same superior package. Figure 5 shows packages of Eclipse BPMN2. H(p) is defined as the set of packages located in package p ∈ P; and H(p,t) is the set of packages located in p which contain classes using the term t in their names. The terms t of this pattern are such that (| H(p,t) | / | H(p) |) ≥ f, and | H(p) | > 2.

4. Analysis Procedure

The analysis procedure performs the following steps: reading of packages and classes; creation of a graph of packages and classes; creation of a graph of terms with aggregated nodes; and querying of frequent terms and their frequency in the graph. The following subsections provide major detail about the relevant features.

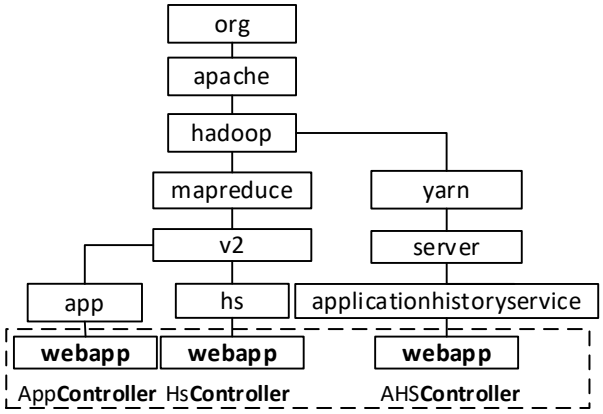


Figure 3: Example of Pattern Package Name and Level

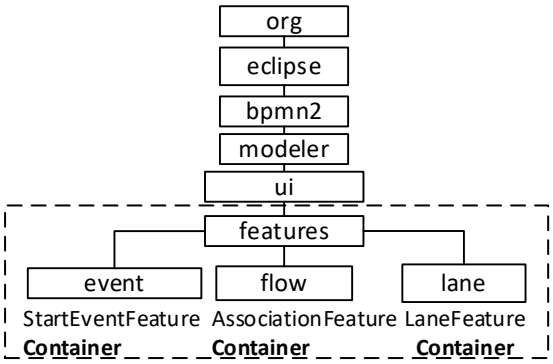


Figure 4: Pattern Package Immediately Superior

4.1. Graph based storage

The gotten terms are stored in a graph based database, such model was chosen due to its visualization capabilities, its ease of adding labels to nodes and creating nodes with aggregated data.

CQL (i.e.; Code Query Language) has been developed to perform exhaustive analysis on source code [27]. However, querying the source code directly without aggregating data, could affects response time. In this work, the graph query language is used as CQL with aims of taking the most of the database query mechanisms, which are developed to manage considerable amounts of data, and visualizing the results graphically. Moreover, having a graph enables software architects to query and visualize the data for purposes beyond this work.

4.2. Analysis of Term Frequency

The procedure of identifying frequent terms uses an analysis based on term frequency with collection range [32], frequency related to the number of times that a term occurs in a collection (e.g.; names of classes organized in packages). The frequency is computed by taking the percentage of term occurrences in same package (pattern Package) or in several packages. For each occurrence, the term position inside the name is considered (e.g., for ClientProtocol, the term Protocol is located in the second position).

The creation of the graph of terms is performed querying the names of classes and storing the occurrence of terms. The new nodes are created aggregating the number of occurrences for each naming pattern: by package, by package name, by package name and level, and by package immediately superior (these nodes will be denominated “aggregated nodes”); such data aggregation enables the simplification of graph queries. Then, aggregated nodes are labeled as frequent terms when they reach a minimal frequency.

5. Results

This work can be considered a valid proposal for ATD, because it corresponds to ATD by non-uniformity of patterns [2], and it takes into account a debt that affects maintainability and evolvability of software, without been included in not accepted topics as technical debt [37]. The approach of this proposal gives relevance to class names, and these determines the maintainability and legibility of software, between others [4,7,18,28,29,33,34]. Looking at the standard ISO/IEC FDIS 25010, maintainability includes the following quality attributes: modularity, reusability, analyzability, modifiability and testability. Considering that identifying components is less complicated task when the identifiers are comprised by significant terms [9,14], the presented analysis can support the analyzability and modifiability, getting significant terms by their frequent use (been representatives). Furthermore, if the naming patterns are not found in a software implementation, it could evidence poor choices of design and implementation with regard to used terms, affecting the test case artifacts [17]; in this sense, the analysis can also support the testability.

Table 1 shows some data about the projects considered henceforth: LOC (lines of code), QF (quantity of files), QP (quantity of packages), and QT (quantity of terms).

For evaluating the proposed analysis technique, an application was implemented to getting the terms used in class name following the CamelCase coding style (predominant style due to its ease of writing and adoption [7,13]), storing terms in a Neo4j database (standard graph database in the industry [26]). The application was executed on twenty projects of the organizations Apache and Eclipse (see table 1). All the source code was gotten from the repositories of Apache and Eclipse in GitHub (https://github.com/). Some project names were simplified to be shown; their names in GitHub are: eclipselink.runtime, hudson.core, scout.rt, servicemix-components.

This evaluation employs a minimal frequency of 0.8 to find frequent terms. Tables 2, 3, 4, and 5 show the following data for patterns 1, 2, 3, 4 (i.e., their order in section Naming Patterns) respectively: N (quantity of frequent terms), Min (minimal frequency found in frequent terms), Max (maximal frequency found), Avg (average frecuency), Stdv (standard deviation of frequency), TN (quantity of terms with a frequency lesser than 1).

Pattern Package is the most used pattern in the set; and Pattern Package Name and Level is the most restrictive and less used. The quantity of projects which does not have occurrences for any pattern is very low. In general, the frequent terms complies some pattern in more than ninety percent of their occurrences (i.e.; average value of 0.9), having cases with one hundred percent.

Table 1: Evaluated projects

	Proyecto	LOC	CA	CP	CT
Apache	JMeter	156900	842	102	526
	Hadoop	860382	4246	404	1403
	MyFaces	161973	816	76	419
	Camel	543222	4070	518	1172
	OpenJPA	324139	1385	54	619
	Wicket	288225	1814	277	804
	ActiveMQ	315613	2122	151	656
	OpenEJB	432266	2758	204	1050
	Geronimo	133804	1087	120	621
	ServiceMix	91837	621	132	314
Eclipse	Birt	1883941	7743	746	1384
	Egit	137718	775	78	402
	BPMN2	190873	1109	96	408
	Scout	415805	3021	691	812
	Xtext	396344	2699	360	1011
	OSEE	593489	6141	815	1496
	EclipseLink	890456	3643	324	994
	Hudson	146540	904	83	687
	EMF	478261	1228	179	475
	Jetty	259521	1315	151	639

Table 2: Frequency of terms for Pattern Package

Proyecto	N	Min	Max	Avg	Stdv	TN
JMeter	15	0.8	1	0.930	0.086	9
Hadoop	72	0.8	1	0.972	0.057	25
MyFaces	17	0.8	1	0.956	0.068	8
Camel	188	0.8	1	0.970	0.062	51
OpenJPA	9	0.8	1	0.914	0.096	5
Wicket	54	0.8	1	0.941	0.082	25
ActiveMQ	34	0.8	1	0.959	0.065	13
OpenEJB	28	0.8	1	0.942	0.077	12
Geronimo	16	0.8	1	0.921	0.078	9
ServiceMix	31	0.8	1	0.960	0.077	10
Birt	90	0.8	1	0.946	0.074	51
EGit	11	0.8	1	0.933	0.080	7
BPMN2	21	0.8	1	0.909	0.079	18
Scout	71	0.8	1	0.945	0.083	26
Xtext	36	0.8	1	0.951	0.074	16
OSEE	125	0.8	1	0.937	0.079	70
EclipseLink	71	0.8	1	0.958	0.066	32
Hudson	18	0.8	1	0.981	0.056	2
EMF	35	0.8	1	0.946	0.075	23
Jetty	34	0.8	1	0.954	0.071	13

TN values show the quantity of ATD incidences by non-uniformity of patterns. The percentage of TN in N shows the percentage of frequent terms, which were not applied uniformly. The maximal accepted value for this percentage can be defined by the development team, in accordance with the degree of use of naming conventions and how well defined is the architecture.

With aims to show the simplicity of queries, figure 6 shows the following query in Cypher language, which gets frequent terms with their respective packages for the pattern Package.

```
MATCH(t:FrequentTerm:PPackage),
(p:Package {fullName:t.packageFullName}) RETURN t,p
```

Code conventions can often be expressed as common practices which follows certain consensus before than as imposed rules [19]. The proposed analysis enables identifying a consensus of terms in following the naming patterns. Taking into account that refactoring can insert poor choices of design and implementation, evidencing such emergent consensus in the source code is useful before performing refactoring [2,19].

Table 6 shows some frequent terms which can be highlighted by their matching with concepts used in popular designs and architectures; showing that is possible to getting emergent and significant concepts from names of source code artifacts. The following query gets the TN terms for all naming patterns.

```
MATCH (t:FrequentTerm) WHERE t.percentage < 1
RETURN DISTINCT t.term
```

Similar works to this proposal were searched in the following digital libraries: ACM, IEEE Xplore and ScienceDirect; the search queries are shown.

For ACM:

```
recordAbstract:(+"technical debt" name names naming identifier identifiers)
```

For IEEE Xplore:

```
("Abstract":technical debt) AND ("Abstract":name OR "Abstract":names OR "Abstract":naming OR "Abstract":identifier OR "Abstract":identifiers)
```

For ScienceDirect:

```
ABS("technical debt") AND (ABS(name) OR ABS(names) OR ABS(naming) OR ABS(identifier) OR ABS(identifiers))
```

The quantities of gotten results for ACM, IEEE Xplore and ScienceDirect are 64, 0, and 1, respectively. The result gotten in ScienceDirect is a book chapter about refactoring advices. Many of the results from ACM are studies about the scope, causes, impact, and features of the technical debt; a few results are slightly related to this work, they address static analysis of source code at a low level, inspecting the source code content (i.e., operations and code sentences). Consequently, it can be affirmed that there is not similar proposals to this work, which is focused in naming of source code artifacts.

Table 3: Frequency of terms for Pattern Package Name

Proyecto	N	Min	Max	Avg	Stdv	TN
JMeter	2	0.938	1	0.969	0.044	1
Hadoop	34	0.800	1	0.965	0.073	7
MyFaces	2	1.000	1	1.000	0.000	0
Camel	21	0.800	1	0.949	0.085	8
OpenJPA	6	1.000	1	1.000	0.000	0
Wicket	9	1.000	1	1.000	0.000	0
ActiveMQ	1	1.000	1	1.000	0.000	0
OpenEJB	1	0.800	1	0.900	0.141	1
Geronimo	0	0.000	0	0.000	0.000	0
ServiceMix	10	0.800	1	0.911	0.090	5
Birt	34	0.800	1	0.951	0.081	11
EGit	5	0.800	1	0.960	0.089	1
BPMN2	3	1.000	1	1.000	0.000	0
Scout	59	0.800	1	0.935	0.089	26
Xtext	10	0.833	1	0.933	0.086	4
OSEE	41	0.800	1	0.963	0.072	10
EclipseLink	17	0.800	1	0.980	0.060	2
Hudson	1	1.000	1	1.000	0.000	0
EMF	17	0.857	1	0.960	0.056	7
Jetty	4	0.889	1	0.944	0.064	2

Table 4: Frequency of terms for Pattern Package Name and Level

Proyecto	N	Min	Max	Avg	Stdv	TN
JMeter	2	0.857	1	0.952	0.082	1
Hadoop	17	0.800	1	0.945	0.082	6
MyFaces	0	0.000	0	0.000	0.000	0
Camel	18	0.833	1	0.991	0.038	1
OpenJPA	0	0.000	0	0.000	0.000	0
Wicket	2	1.000	1	1.000	0.000	0
ActiveMQ	0	0.000	0	0.000	0.000	0
OpenEJB	0	0.000	0	0.000	0.000	0
Geronimo	0	0.000	0	0.000	0.000	0
ServiceMix	9	0.875	1	0.963	0.060	2
Birt	18	0.833	1	0.995	0.029	1
EGit	5	1.000	1	1.000	0.000	0
BPMN2	2	1.000	1	1.000	0.000	0
Scout	63	0.800	1	0.977	0.063	8
Xtext	2	0.800	0.8	0.800	0.000	2
OSEE	20	0.800	1	0.974	0.069	3
EclipseLink	1	1.000	1	1.000	0.000	0
Hudson	1	1.000	1	1.000	0.000	0
EMF	17	0.833	1	0.967	0.063	6
Jetty	4	1.000	1	1.000	0.000	0

Table 5: Frequency of terms for Pattern Immediate Superior

Proyecto	N	Min	Max	Avg	Stdv	TN
JMeter	3	1.000	1	1.000	0.000	0
Hadoop	6	0.800	1	0.967	0.082	1
MyFaces	0	0.000	0	0.000	0.000	0
Camel	26	0.800	1	0.954	0.075	10
OpenJPA	0	0.000	0	0.000	0.000	0
Wicket	2	0.896	1	0.965	0.060	1
ActiveMQ	60	0.800	1	0.915	0.037	57
OpenEJB	2	0.875	0.889	0.882	0.008	2
Geronimo	3	0.857	1	0.952	0.082	1
ServiceMix	6	0.906	1	0.974	0.042	2
Birt	41	0.800	1	0.974	0.059	8
EGit	1	1.000	1	1.000	0.000	0
BPMN2	5	0.800	1	0.922	0.078	4
Scout	20	0.800	1	0.950	0.075	9
Xtext	7	0.800	1	0.919	0.089	4
OSEE	29	0.800	1	0.952	0.084	9
EclipseLink	33	0.800	1	0.977	0.062	7
Hudson	0	0.000	0	0.000	0.000	0
EMF	16	0.800	1	0.947	0.085	3
Jetty	5	0.833	1	0.900	0.091	3

Table 6: Frequent terms

Proyecto	Términos
JMeter	Controller, Converter, Editor, Gui, JDBC, Meter.
Hadoop	Chain, Client, Container, Event, Scheduler.
MyFaces	Handler, Html, Impl, Implicit, Renderer, Tag.
Camel	Bean, Cache, Command, Filter, Task, Yammer.
OpenJPA	Concurrent, Distributed, Identifier, Managed.
Wicket	Bean, Checker, Handler, Resolver, Socket.
ActiveMQ	Adapter, Bridge, Broker, Command, Factory.
OpenEJB	Binding, Command, Entity, Factory, Thread.
Geronimo	Command, Deployment, Manager, Validation.
ServiceMix	Component, Factory, Filter, Interceptor, Ws
Birt	Action, Adapter, Filter, Handler, Validator.
EGit	Blame, Command, Git, Handler, Index, Node
BPMN2	Adapter, Editor, Event, Flow, Task, Validator.
Scout	Activity, Browser, Inspector, Job, Page, Service,
Xtext	Facet, Fragment, Module, Page, Resource, Ui
OSEE	Action, Command, Client, Service, Word.
EclipseLink	Accesor, Converter, Query, Resource, Table.
Hudson	Team, X
EMF	Action, Adapter, Command, Factory, Model.
Jetty	Bean, M, Response, Socket, Web

6. Conclusions

The naming patterns presented frequent ccurrences in several projects of the organizations Apache and Eclipse, showing that most of the frequent terms complies each pattern by ninety percent of their occurrences.

The proposed analysis identifies architectural technical debt by non-uniformity of naming patterns; which are applied frequently, but not followed in all the system. The used approach, based on naming patterns of source code artifacts, differs from other approaches which uses the source code content (e.g.; operations, sentences) for identifying technical debt.

The use of a graph based database was relevant, to enable using the database query capabilities as CQL, avoiding the limitations that could present a conventional CQL tool [27]; performing data aggregation in new nodes and making easy the elaboration of queries, which could be more complex or hard to be defined with a conventional CQL.

The proposal is applicable under an agile approach, which promotes focusing on product features and taking care about uncertainty in respect of ATD [2]. The analysis performed on source code does not require an architecture specification as input, and could be automatic through the continuous execution of queries during the software development, enabling the tracking of ATD. Additionally, the frequent terms, which were discovered, can be useful for identifying new emergent concepts in the software architecture.

References

[1] A. Mahmoud, N. Niu, "Supporting requirements to code traceability through refactoring", Requir. Eng., 19(3), 309-329, 2014.

[2] A. Martini, J. Bosch, M. Chaudron. "Investigating Architectural Technical Debt accumulation and refactoring over time", Inf. Softw. Technol. 67, 237-253, 2015.

[3] A. Ampatzoglou, A. Ampatzoglou, A. Chatzigeorgiou, P. Avgeriou, "The financial aspect of managing technical debt", Inf. Softw. Technol. 64, 52-73, 2015.

[4] B. Liblit, A. Begel, E. Sweetser, "Cognitive perspectives on the role of naming in computer programs" in Proceedings of the 18th Annual Psychology of Programming Workshop, 2006.

[5] C. Seaman, Y. Guo, C. Izurieta, Y. Cai, N. Zazworka, F. Shull, A. Vetrò, "Using technical debt data in decision making: potential decision approaches" in Proceedings of the Third International Workshop on Managing Technical Debt (MTD '12), IEEE Press, Piscataway, NJ, USA, 45-48, 2012.

[6] C. Sterling, Managing Software Debt: Building for Inevitable Change (1st ed.), Addison-Wesley Professional, 2010.

[7] D. Binkley, M. Davis, D. Lawrie, J. I. Maletic, C. Morrell, B. Sharif. "The impact of identifier style on effort and comprehension", Empirical Softw. Engg. 18, 2 (April 2013), 219-276, 2013

[8] E. Murphy-Hill, C. Parnin, and A. P. Black, "How We Refactor, and How We Know It", IEEE Trans. Softw. Eng., 38(1), 5-18, 2012.

[9] F. Deissenboeck, M. Pizka, "Concise and Consistent Naming" in Proceedings of the 13th International Workshop on Program Comprehension (IWPC '05), IEEE Computer Society, Washington, DC, USA, 97-106, 2005.

[10] G. Soares, R. Gheyi, E. Murphy-Hill, B. Johnson, "Comparing approaches to analyze refactoring activity on software repositories", J. Syst. Softw. 86(4), 1006-1022, 2013.

[11] I. Mistrik, R. Bahsoon, R. Kazman, Y. Zhang, Economics-Driven Software Architecture (1st ed.), Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2014.

[12] K. Beck, Implementation patterns, Addison Wesley, 2008.

[13] L. Guerrouj, "Normalizing source code vocabulary to support program comprehension and software quality", in Proceedings of the 2013 International Conference on Software Engineering (ICSE '13), IEEE Press, Piscataway, NJ, USA, 1385-1388, 2013.

[14] L. Guerrouj, M. Penta, Y. Guéhéneuc, G. Antoniol, "An experimental investigation on the effects of context on source code identifiers splitting and expansion", Empirical Softw. Engg., 19(6), 1706-1753, 2014.

[15] L. Chen, M. Ali Babar, "Towards an Evidence-Based Understanding of Emergence of Architecture through Continuous Refactoring in Agile Software Development" in Proceedings of the 2014 IEEE/IFIP Conference on Software Architecture (WICSA '14), IEEE Computer Society, Washington, DC, USA, 195-204, 2014

[16] M. Katić, K. Fertalj, "Towards an appropriate software refactoring tool support" in Proceedings of the 9th WSEAS international conference on Applied computer science (ACS'09), 2009.

[17] M. Tufano, F. Palomba, G. Bavota, R. Oliveto, M. Di Penta, A. De Lucia, D. Poshyvanyk, "When and why your code starts to smell bad", in Proceedings of the 37th International Conference on Software Engineering (ICSE '15), Vol. 1, IEEE Press, Piscataway, NJ, USA, 403-414, 2015.

[18] M. Allamanis, E. T. Barr, C. Bird, C. Sutton, "Suggesting accurate method and class names" in Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering (ESEC/FSE 2015), ACM, New York, NY, USA, 38-49, 2015.

[19] M. Allamanis, E. T. Barr, C. Bird, C. Sutton, "Learning natural coding conventions", in Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering (FSE 2014), ACM, New York, NY, USA, 281-293, 2014.

[20] M. Vakilian, N. Chen, S. Negara, B. A. Rajkumar, B. P. Bailey, R. E. Johnson, "Use, disuse, and misuse of automated refactorings" in Proceedings of the 34th International Conference on Software Engineering (ICSE '12), IEEE Press, Piscataway, NJ, USA, 233-243, 2012.

[21] M. A. Babar, A. W. Brown, and I. Mistrik, Agile Software Architecture: Aligning Agile Processes and Software Architectures (1st ed.), Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2013.

[22] N. A. Ernst, S. Bellomo, I. Ozkaya, R. L. Nord, I. Gorton, "Measure it? Manage it? Ignore it? software practitioners and technical debt" in Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering (ESEC/FSE 2015), ACM, New York, NY, USA, 50-60, 2015.

[23] N. Zazworka, M. A. Shaw, F. Shull, C. Seaman, "Investigating the impact of design debt on software quality" in Proceedings of the 2nd Workshop on Managing Technical Debt (MTD '11), ACM, New York, NY, USA, 17-23, 2011.

[24] N. Alves, T. Mendes, M. de Mendonça, R. Spínola, F. Shull, C. Seaman, "Identification and management of technical debt: A systematic mapping study", Information and Software Technology, 70, 100-121, 2016.

[25] N. Tsantalis, V. Guana, E. Stroulia, A. Hindle, "A multidimensional empirical study on refactoring activity" in Proceedings of the 2013 Conference of the Center for Advanced Studies on Collaborative Research (CASCON '13), IBM Corp., Riverton, NJ, USA, 132-146, 2013.

[26] P. Macko, D. Margo, M. Seltzer, "Performance introspection of graph databases" in Proceedings of the 6th International Systems and Storage Conference (SYSTOR '13), ACM, New York, NY, USA, 2013.

[27] R. Urma, A. Mycroft, "Programming language evolution via source code query languages" in Proceedings of the ACM 4th annual workshop on Evaluation and usability of programming languages and tools (PLATEAU '12), ACM, New York, NY, USA, 35-38, 2012.

[28] R. C. Martin, Clean Code: A Handbook of Agile Software Craftsmanship (1 ed.), Prentice Hall PTR, Upper Saddle River, NJ, USA, 2008.

[29] S. Butler, M. Wermelinger, Y. Yu, H. Sharp, "Exploring the Influence of Identifier Names on Code Quality: An Empirical Study, in Proceedings of the 2010 14th European Conference on Software Maintenance and Reengineering (CSMR '10), IEEE Computer Society, Washington, DC, USA, 156-165, 2010.

[30] S. Negara, N. Chen, M. Vakilian, R. E. Johnson, D. Dig, "A comparative study of manual and automated refactorings" in Proceedings of the 27th European conference on Object-Oriented Programming (ECOOP'13), Springer-Verlag, Berlin, Heidelberg, 552-576, 2013.

[31] S. McConnell, Code Complete, Second Edition, Microsoft Press, Redmond, WA, USA, 2004.

[32] T. Roelleke, Information Retrieval Models: Foundations and Relationships (1st ed.), Morgan & Claypool Publishers, 2013.

[33] V. Arnaoudova, M. Di Penta, G. Antoniol, "Linguistic antipatterns: What they are and how developers perceive them", Empirical Software Engineering, 1-55, 2015.

[34] W. Maalej, R. Tiarks, T. Roehm, R. Koschke. 2014. On the Comprehension of Program Comprehension, ACM Trans. Softw. Eng. Methodol., 23(4), 2014.

[35] Z. Codabux, B. J. Williams, N. Niu, in Proceedings of the International Conference on Software Engineering Research and Practice (SERP'14), 2014.

[36] Z. Li, P. Liang, P. Avgeriou, N. Guelfi, A. Ampatzoglou, "An empirical investigation of modularity metrics for indicating architectural technical debt" in Proceedings of the 10th international ACM Sigsoft conference on Quality of software architectures (QoSA '14), ACM, New York, NY, USA, 119-128, 2014.

[37] Z. Li, P. Avgeriou, P. Liang, "A systematic mapping study on technical debt and its management", J. Syst. Softw., 101, 193-220, 2015.

[38] P. Mendoza del Carpio, "Identification of architectural technical debt: An analysis based on naming patterns", in Proceedings of the 2016 8th Euro American Conference on Telematics and Information Systems (EATIS), IEEE Computer Society, Washington, DC, USA, 10, 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 1, 255-262 (2017)
www.astesj.com

ASTES Journal
ISSN: 2415-6698

Power-Energy Simulation for Multi-Core Processors in Benchmarking

Mona A. Abou-Of[*,1], Amr A. Sedky[2], Ahmed H. Taha[2]

[1]*Assistant Professor, Department of Computer Engineering, Pharos University in Alexandria, Alexandria, Egypt*

[2]*Undergraduate Student, Department of Computer Engineering, Pharos University in Alexandria, Alexandria, Egypt*

ARTICLEINFO

Article history:
Received: 21 December, 2016
Accepted: 19 January, 2017
Online: 28 January 2017

Keywords :
Multi-Core Processors
Energy simulator
Power Profiling Systems
Performance Benchmarks
Convexity Rule
System Clock Speed

ABSTRACT

At Microarchitectural level, multi-core processor, as a complex System on Chip, has sophisticated on-chip components including cores, shared caches, interconnects and system controllers such as memory and ethernet controllers. At technological level, architects should consider the device types forecast in the International Technology Roadmap for Semiconductors (ITRS).
Energy simulation enables architects to study two important metrics simultaneously. Timing is a key element of the CPU performance that imposes constraints on the CPU target clock frequency. Power and the resulting heat impose more severe design constraints, such as core clustering, while semiconductor industry is providing more transistors in the die area in pace with Moore's law. Energy simulators provide a solution for such serious challenge.
Energy is modelled either by combining performance benchmarking tool with a power simulator or by an integrated framework of both performance simulator and power profiling system.
This article presents and asses trade-offs between different architectures using four cores battery-powered mobile systems by running a custom-made and a standard benchmark tools. The experimental results assure the Energy/ Frequency convexity rule over a range of frequency settings on different number of enabled cores.
The reported results show that increasing the number of cores has a great effect on increasing the power consumption. However, a minimum energy dissipation will occur at a lower frequency which reduces the power consumption. Despite that, increasing the number of cores will also increase the effective cores value which will reflect a better processor performance.

1 Introduction

Microprocessor performance has helmed its industry for four decades. Reducing power consumption has become a stringent design principle especially for battery-driven devices. Limiting the increase in CPU clock frequency, because of low-power constraints and high energy efficiency, has become a real challenge for improving microprocessor performance over the next generation. So, other aspects in microprocessor architecture (Instruction Set) and compilers optimizations have to be considered in order to optimize the offered workload. In addition, other factors in microprocessor hardware implementation must be taken into account in order to speed up this workload execution time such as using many cores.

In this paper, we make the case for exploring the trade-off between low power and energy efficiency over a wide range of clock frequencies. We do the experiments on different battery-powered Laptops and Smartphones in [1] on a single core. We enface two problems: the choice of power measure-

[*]Mona A. Abou-Of, mona.abouof@pua.edu.eg

https://dx.doi.org/10.25046/aj020131

ment tools and the choice of performance benchmark tools. An accurate reliable power measurement software has to be selected in such a way to be running on Linux platform for Laptop devices like Powerstat (Power consumption calculator for Ubuntu Linux. Available: http://www.hecticgeek.com), or running on Android platform for Smartphones like Powertutor [2]. To capture the transitions between power states, two different finite state machines (FSM) based power modeling scheme [3] are implemented: The standard CoreMark benchmark (Industry-standard benchmarks for embedded systems. Available: http://www.eembc.org/coremark), executed on Linux OS, represents a disk with tail power state model that writes the running power on a disk file and stays at high power state for a period after the active I/O activity. The custom-made Fibonacci benchmark, written with Java on Android, represents a free model that returns to the base state without inactivity period.
In summary, this paper makes the following contributions:

- We make laboratory experiments for exploring the relationship between processor performance, power consumption and energy efficiency over a range of clock frequencies on different number of enabled cores.
- We represent the experimental setup in order to obtain reliable results.
- We represent a detailed implementation on different laptops and Smartphones operating systems.
- The plotted results assure that minimum energy dissipation is always achieved even with different workloads, and at a certain clock frequency but with a limited performance, lower power consumption and without optimization realization.
- We have proved the Energy/ Frequency convexity rule on multi-core (instead of one core) processors [4].
- Such observations can be fed into an intelligent DVFS scheduling, power management module of an operating system, on multi-core processors, which can achieve energy and power savings without impacting the performance.
- We have proved that increasing number of cores has a great effect on increasing the power consumption. However, a minimum energy dissipation will occur at a lower frequency which reduces the power consumption. Despite that, increasing the number of cores will also increase the effective cores value which will reflect a better processor performance.

The rest of this paper is organized as follows: Section 2 presents the existing energy modeling approaches. Section 3 formulates the problem with some equations. In section 4 the experimental results are evaluated and analyzed. Finally, section 5 concludes the paper.

2 Related Work

Most of the existing system energy modeling approaches combine between power profiling systems and performance benchmark tools. SPEC has developed SPECpower_ssj2008 (S.P.E. Corporation. specpower ssj2008 benchmark suite. Available: http://www.spec.org/power ssj2008) focusing on server computer consumption, and EEMBC has introduced EnergyBench establishing a framework for adding energy to the metrics of the EEMBC's performance benchmarks (E.T.E.M.B. Consortium. energybench version 1.0 power/energy benchmarks. Available: http://www.eembc.org/benchmark/power_sl.php).
McPAT [5] is a fully-integrated power, area and timing modeling framework. It models all types of power dissipation and provides an integrated solution for multithreaded and multi-core processors. McPAT power modeling is combined with Sniper performance simulation in [6].

2.1 Power Profiling Systems

Existing power measurement methods are limited in two ways. First, some systems [3,7,8] and Monsoon power monitor (Available: http://www.msoon.com/LabEquipment/PowerMonitor) generate their models by using external hardware lab equipments like sensors, meters, and data acquisition devices. Second, other systems like Powerstat, [2,9,10] are self-modeling. They construct their models without external circuitry. They use built-in battery sensors or the smart battery interface fuel gauge IC; or read system files available on mobile systems. Integrated sensors are provided on CPUs [11] such as Intel processors [12] and AMD processors [13], on GPU cards [14], or on motherboards equipped with a Baseboard Management Controller (BMC) monitoring chip [15].
Some of this systems are Event-based as in [3,5] or per-component power measurements in addition to the total power as in [5,7,9,16,17]. Others modeled power measurements by applications as in [2].
Industry simulators are typically cycle-accurate that run at a speed of 1 to 10 kHz. Academic simulators, such as [18,19] are not truly cycle-accurate compared to real hardware, and therefore they are faster, with simulation speeds in the tens to hundreds of KIPS (kilo simulated instructions per second) range. They do not scale well to large multi-core systems.

2.2 Performance Benchmark Tools

SPECpower_ssj2008 benchmark and the Apache benchmarking tool (ab - apache benchmarking tool.

Available: http://httpd.apache.org/docs/2.2/programs /ab.html) are used for HTTP server traffics. The SPECpower_ssj2008 is the first industry standard SPEC benchmark that evaluates the power and performance characteristics of volume server class and multi-node class computers.The widespread used benchmark in industry and academia is SPEC CPU2006 [20]. EEMBC has benchmarks for general-purpose performance analysis including CoreMark, MultiBench(multicore), and FPMark (floating-point).

3 Problem Formulation

The basic relationships among computer performance, power consumption and energy efficiency are expressed as follows:

$$Processor\ Performance = 1/CPU\ Execution\ time \quad (1)$$

$$Energy\ Efficiency = 1/dissipated\ Energy \quad (2)$$

$$Energy = Power * CPU\ Execution\ time \quad (3)$$

As shown in [21], the power consumed by a processor is directly proportional with the clock frequency (f).
In order to study the impact of clock speed on the processor performance without DVFS scheduling, the CPU Execution time (t_x) is computed as:

$$t_x = Instruction\ Count * CPI * Tcycle \quad (4)$$

where $Tcycle$ equals $1/f$ and CPI is the average number of cycles per instruction.
i.e. t_x is function of $(1/f)$, and improving the performance requires decreasing t_x and speeding up the CPU frequency. Or

$$t_x = Number\ of\ clock\ cycles/f \quad (5)$$

in case of single core. And

$$t_x = Number\ of\ clock\ cycles/(f * c_e) \quad (6)$$

in case of multi-core where c_e is the effective cores parameter which reflects the degree of the execution parallelization achievement.

Equation (3) shows that, in order to minimize the energy, power should be reduced. This can be achieved by using low clock frequency. On the other side, reducing t_x requires high clock frequency. This trade-off between lower power and better performance leads to the existence of an optimum point for minimal energy usage with a tight performance improvement at a certain specific CPU frequency (f_m).
The goal of the presented experiments in this paper is to search for such minimal energy when the CPU frequency is varied and find the optimum frequency f_m for a varied number of cores.

4 Experimental Setup

The presented experiments measure the power and the execution time while running different workloads on specific Dynamic Voltage and Frequency Scaling (DVFS) mobile system settings over a 0.6 GHz to 1.7 GHz range of CPU frequencies.
The variation of CPU frequency settings needs the CPU frequency information of the used mobile device. These settings demand the resetting of the power management policy, the disabling of some cores; and the setting of the only enabled cores with one of its frequency values in parallel with its upper frequency limit.
The experiments are implemented on three different battery-powered mobile systems shown in Table 1: two Intel Laptops (Acer and Dell) and one ARM Smartphone (Samsung A5), on different Operating Systems Ubuntu and Android respectively in [1] and extended to multi-core on the two Laptops only.
The offered workloads were CoreMark, the standard benchmark tool for Laptops and a custom-made Fibonacci benchmark for the Smartphone and also for the Acer and Dell Laptops. The Fibonacci benchmark is implemented, in Java, iteratively for 2E8 iterations. The execution time is measured via those performance benchmark tools.
The power consumed by these performance benchmark tools is measured by different power profiling systems: Powerstat on Linux O.S. and Powertutor [2] on Android. Both systems use the built-in smart battery interface to measure power at rate 1 Hz while the battery is discharging. Powerstat measures the total power while Powertutor measures also an individual power per application. Both power profiling systems have to be running by at least one minute before running the performance benchmark tools giving the chance to the power to be stabilized.

4.1 How to measure power?

For Laptops with Linux platforms, Powerstat is used to measure the power consumed by the running CoreMark. Two factors are considered: Powerstat measures the total power of the Laptops and CoreMark is a disk with tail powerstate model. Steps to measure CoreMark consumed power (P_c):

1. Reset the power management policy.
2. Operate the frequency scaling governor in userspace mode.
3. Enable only i CPU cores and disable the others.
4. Set a certain frequency for all running cores.
5. Run Powerstat.
6. Wait for two minutes until the power is stabilized.
7. Run CoreMark or any performance benchmark and register Start and End of the Execution time.

Table 1: Simulated Mobile Systems Characteristics

Parameter	Acer Aspire 1	Dell Inspiron15	Samsung Galaxy A5
Processor	4xIntel(R) Atom(TM)CPU N2600 @1.60GHz	4x Intel(R) Core(TM) i5-4210U CPU @1.70GHz	Quad-core 1.2GHz Cortex-A53
Memory	2G RAM	4G RAM	2G RAM
Operating System	Ubuntu 14.04.3 LTS	Ubuntu 13.04	Android OS, v4.4.4
Kernel	Linux 3.13.065 generic(i686)	Linux 3.8.0 0-19 generic(i686)	3.10.28-4197997 dpi@SWDD5006-1

8. Examine the Powerstat log file, register the power before and a while after the execution time of the CoreMark until the completion of CoreMark I/O and take the average power without running the CoreMark (P_s). P_s presents the average power of the system and the Powerstat.

9. Compute the average total power between Start and End-time of CoreMark execution by averaging power batches (P_t).

10. Compute $P_c = P_t - P_s$

11. Repeat steps from 5 to 10 in order to get 10 batches and get the average P_c.

12. Repeat steps from 5 to 11 with all available CPU frequencies.

13. Repeat steps from 3 to 12 for i different number of cores (1 to 4)

A sample output of Power measured by Powerstat with DVFS scheduling and another with 1.6 GHz fixed CPU frequency setting are shown by the Instantaneous Power Profiles in Fig. 1.
The resulting power profile shows that the power with DVFS scheduling returns the base state (7.5 watts) 30 seconds earlier than the one with fixed 1.6 GHz CPU frequency setting and also drops about 0.7 watts. This DVFS scheduling saves about 30 sec * 0.7 watts or 21 joules.

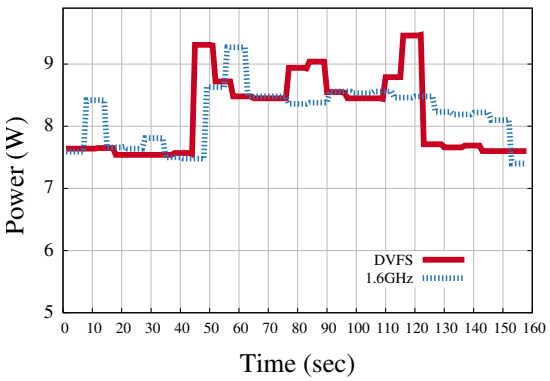


Figure 1: Instantaneous Power Profile when CPU is fixed at 1.6 GHz compared to CPU power with DVFS scheduling.

For Smartphones with Android platforms, Powertutor is used for power management. Referring to the steps described above to measure the CoreMark consumed power, apply the first 7 steps with interchanging PowerStat with Powertutor and CoreMark with Fibonacci Java code. No need to compute the average consumed power P_c for benchmark since Powertutor measures power for each individual application separately and register it in its log file. Then, repeat steps from 5 to 7 with all available frequencies and cores.

5 Experimental Results and Analysis

This section illustrates the relationship between the CPU execution time, the power consumption, and the dissipated energy over a 0.6 GHz to 1.7 GHz range of CPU frequencies. We formulate sixteen experiments on one, two, three and four enabled cores. The half of the experiments run over Linux OS using the CoreMark Benchmark on an Acer and Dell Laptops as shown in Figures 2 and 3. The other half run the custom-made Fibonacci Java code over Linux OS on Acer and Dell Laptops as shown in Figures 4 and 5. The CoreMark offered Workload is set to 200,000 iterations while The Fibonacci offered Workload is set to 2E8 iterations.

Figures 2(a), 3(a), 4(a) and 5(a) plot the time results in seconds of the four different number of enabled cores experiments with the variation of the CPU frequency. The sixteen time curves prove (6). They demonstrate that the CPU execution time t_x decreases at higher clock rates and/or a higher effective cores value (c_e). Although different workloads are offered to the Acer Laptop, they approximately spent the same t_x. This time is much lower when executing the same loads on the higher specifications of Dell Laptop. The Dell Laptop is much faster than the Acer one.

The graphs in Figures 2(b), 3(b), 4(b) and 5(b) plot the power results in Watts for the four varied number of cores experiments with the variation of CPU frequency. The two power graphs of Acer Laptop approximately overlap; while the power consumed by the Dell Laptop is much higher. The power graphs

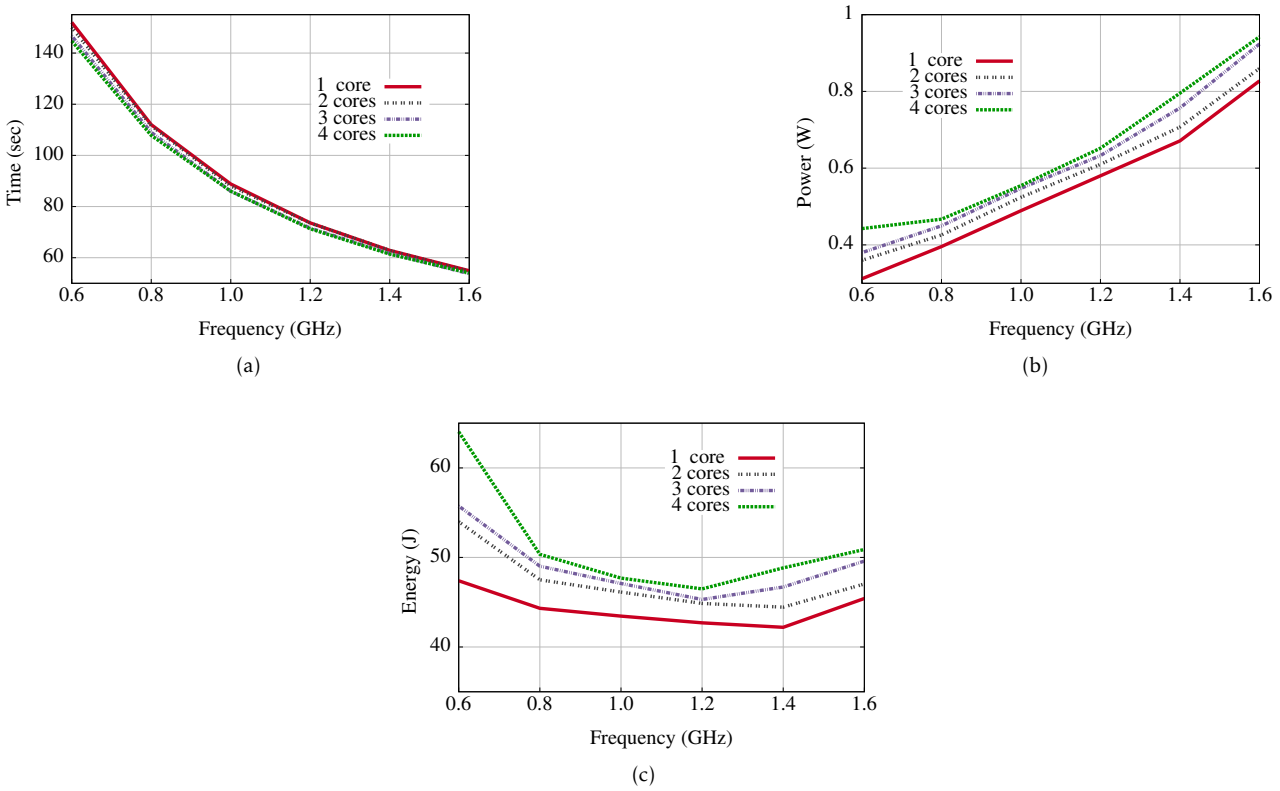


Figure 2: Running CoreMark Benchmark on Acer Laptop. Although increasing number of cores increases the power consumption, there is always an optimal frequency for minimum energy.

of the sixteen experiments ensure that the processor power is proportional to CPU frequency [21]. In addition, incrementing the number of enables cores also increases the power.
All figures (2, 3, 4 and 5) illustrate that increasing frequencies decreases the execution time while increasing the consumed power by the processor. They also show that increasing the number of cores has a great effect on increasing the power consumption.

This trade-off between execution time and power, with the variation of frequencies, leads to the convex energy curves in Figures 2(c), 3(c), 4(c) and 5(c). The energy is computed by (3). The Acer Laptop has minimal energy with the CoreMark benchmark at f_m =1.4 GHz when one or two cores are enabled, and at f_m =1.2 GHz when three or four cores are enabled. While it has a minimal energy at f_m = 1.4 GHz for the Fibonacci benchmark when one, two or three cores are enabled and f_m = 1.2 GHz when four cores are enabled. The advanced Dell Laptop has a minimal energy at f_m = 1.4 GHz with CoreMark benchmark when a single core is enabled, f_m = 1.2 GHz when two or three cores are enabled, and f_m = 1.1 GHz when the four cores are enabled. While it has a minimal energy at f_m = 1.4 GHz with Fibonacci benchmark when one or two cores are enabled and f_m = 1.3 GHz when three or four cores are enabled.

All of the sixteen experiments demonstrate that a minimal energy can be obtained at an optimum frequency f_m. Referring to the execution time in Figures 2(a), 3(a), 4(a) and 5(a) at those f_m frequencies, a tight performance improvement can be achieved: about 75% in Acer and 65% in Dell Laptops from those of the largest frequencies in case of a single core.

Referring to the power consumption in Figures 2(b), 3(b), 4(b) and 5(b) at those f_m frequencies, more power is consumed: about 200% in Acer and 140% in Dell Laptops from those of the smallest frequencies. So other design factors, rather than clock speed, have to be considered for a low-power achievement. In case of multi-core processors, increasing the number of enabled cores shifts f_m to lower frequency and reduces the power but increases the c_e value which reflects a better performance. As an illustrating example, running the CoreMark on Acer Laptop at f_m =1.4 GHz on a single core has the same power as running it at f_m =1.2 GHz on quad-core. From 6, in order to keep the same performance, the c_e value should be at least 1.4/1.2 which is equivalent to 1.16. Therefore, the degree of execution parallelization achievement defined by the effective cores value (c_e) is the dominant factor of the processor performance.

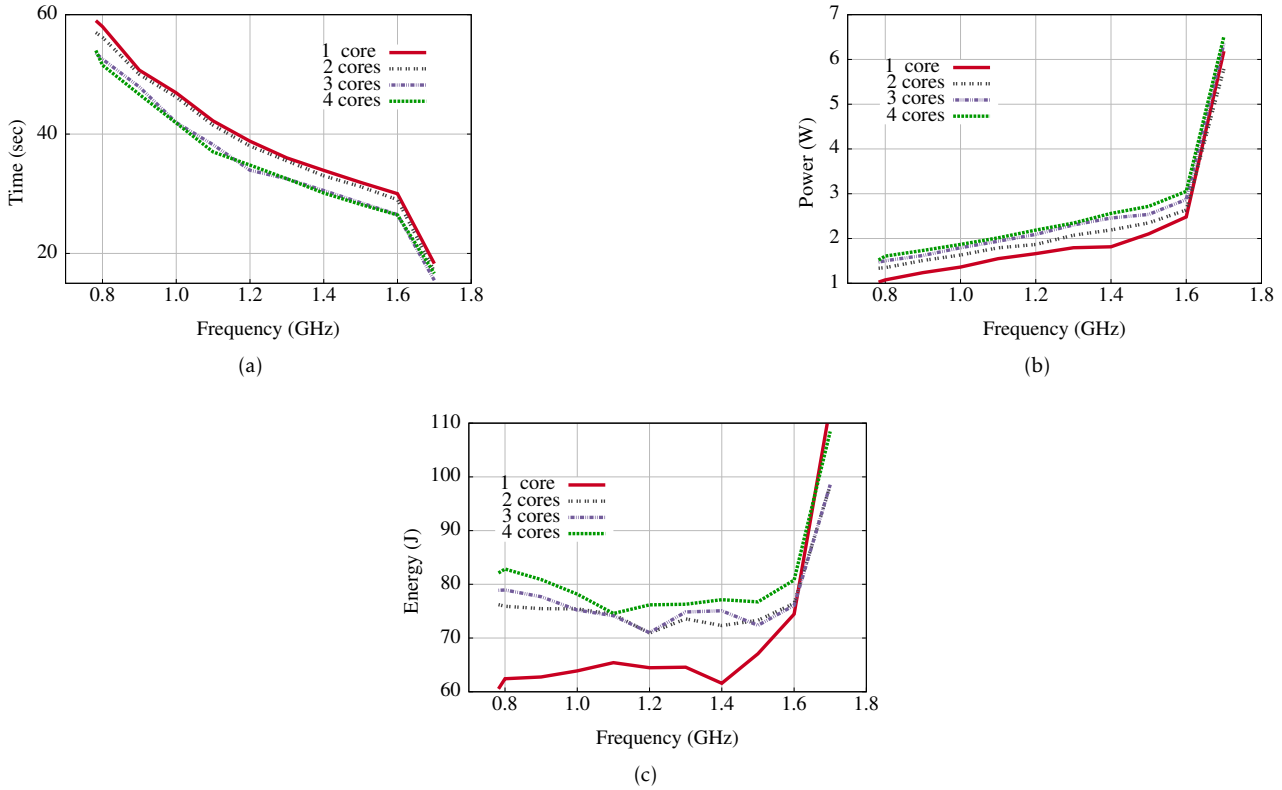


Figure 3: Running CoreMark Benchmark on Dell Laptop. Although increasing number of cores increases the power consumption, there is always an optimal frequency for minimum energy.

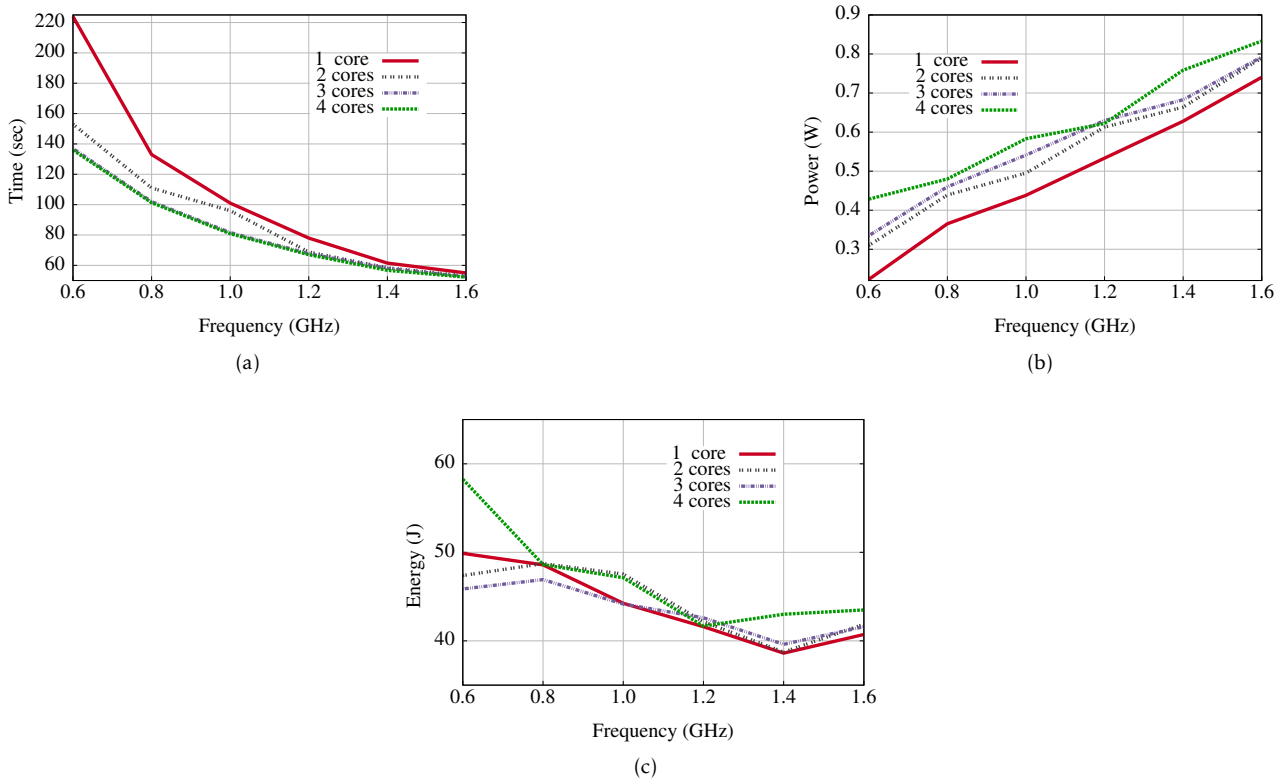


Figure 4: Running Fibonacci Benchmark on Acer Laptop. Although increasing number of cores increases the power consumption, there is always an optimal frequency for minimum energy.

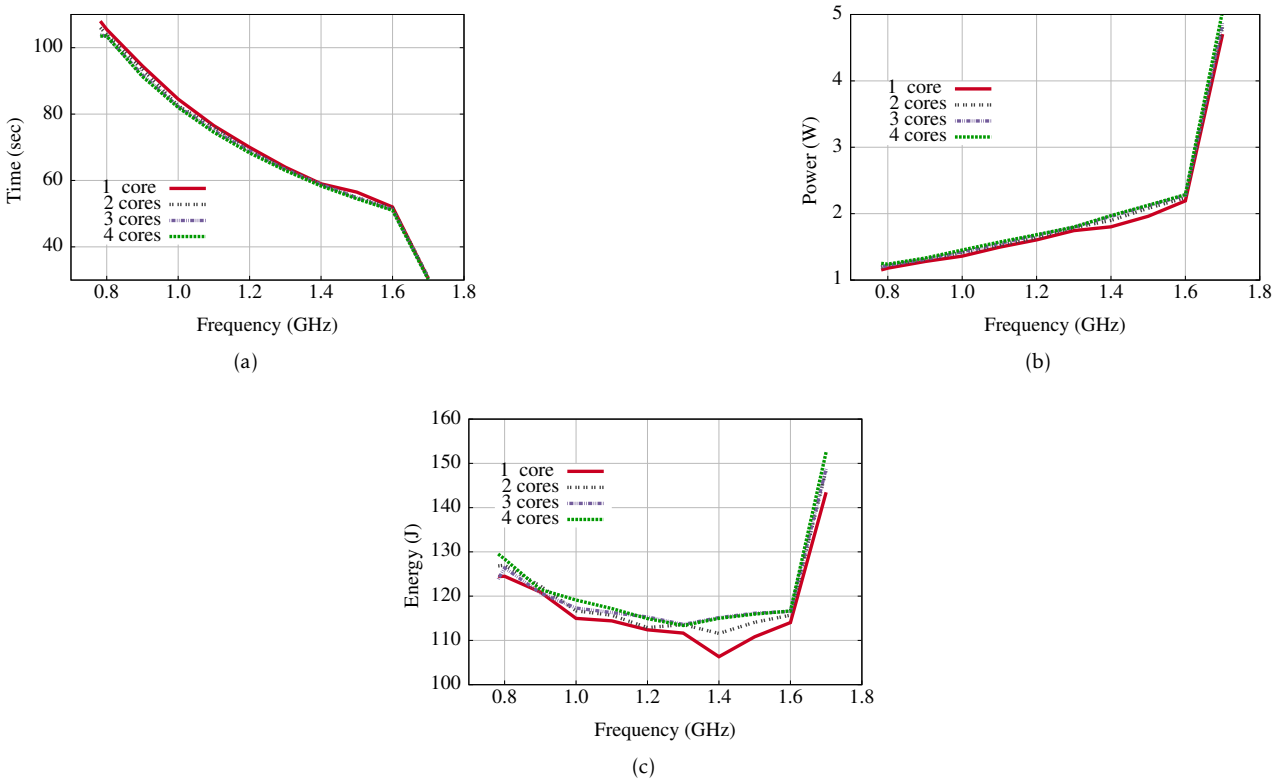


Figure 5: Running Fibonacci Benchmark on Dell Laptop. Although increasing number of cores increases the power consumption, there is always an optimal frequency for minimum energy.

6 Conclusion

Energy efficiency improvement can't be achieved by exploring the hardware implementation of the microprocessor design only. Referring to (4), the CPU performance is also improved by a good design of instruction set architecture (ISA).

ISA optimization decreases the program Instruction Count and the CPI. Such optimization has a direct impact on minimizing the offered workload, consequently it reduces the power by decreasing the CPU utilization.

Improving processor performance by hardware implementation as rising the CPU frequency has a greater side effect on the power. Another factor like CPI has to be considered. High-level of parallelism, including superscaler implementation based on instruction-level parallelism or multi-processing architecture where many core (MTC) are integrated, can achieve a better CPI. Using Multi-core processor, as detected by the experiments, reduces the execution time without extra power while enhancing the energy efficiency.

The demonstrated experiments assure the trade-off between optimizing the energy efficiency and improving the processor performance. Both always affect the power consumption while changing the CPU frequencies. Furthermore, we have proved that increasing number of cores has a great effect on increasing the power consumption. However, a minimum energy dissipation will occur at a lower frequency which reduces the power consumption. Despite that, increasing the number of cores will also increase the effective cores value which will reflect a better processor performance.

Conflict of Interest No conflict of interest.

References

1. MA. Abou-Of, AH. Taha, AA. Sedky, "Trade-off Between Low Power and Energy Efficiency in Benchmarking." in 7th International Conference on Information and Communication Systems. IEEE Conference Publications, 2016.
2. L. Zhang, B. Tiwana, Z. Qian, Z. Wang, RP. Dick, ZM. Morley, L. Yang "Accurate Online Power Estimation and Automatic Battery Behavior Based Power Model Generation for Smartphones." in 8th IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis. ACM, 2010.
3. A. Pathak, YC. Hu, M. Zhang, P. Bahl, YM. Wang "Fine-grained Power Modeling for Smartphones Using System Call Tracing." in 6th Conference on Computer Systems. ACM, 2011.
4. K. De Vogeleer, G. Memmi, P. Jouvelot, F. Coelho "The Energy/Frequency Convexity Rule: Modeling and Experimental Validation on Mobile Devices." in International Conference on Parallel Processing and Applied Mathematics. Springer Berlin Heidelberg, 2013.
5. S. Li, JH. Ahn, RD. Strong, JB. Brockman, DM. Tullsen, NP. Jouppi "McPAT: An Integrated Power, Area, and Timing Modeling Framework for Multicore and Manycore Architectures." in 42nd Annual IEEE/ACM International Symposium on Microarchitecture. ACM, 2009.
6. W. Heirman, S. Sarkar, TE. Carlson, I. Hur, L. Eeckhout "Power-aware Multi-core Simulation for Early Design Stage Hardware/Software Co-optimization." in 21st International

Conference on Parallel Architectures and Compilation Techniques. ACM, 2012.

7. S. Gutiérrez-Verde, O. Benedí-Sánchez, D. Suárez-Gracia, JM. Marín-Herrero, V. Viñals-Yúfera "Processor Energy and Temperature in Computer Architecture Courses: a Hands-on Approach." in Workshop on Computer Architecture Education. 2009.
8. R. Ge, X. Feng, S. Song, HC. Chang, D. Li "Powerpack: Energy profiling and analysis of high-performance systems and applications." IEEE Transactions on Parallel and Distributed Systems, **21**(5), 658-671, 2010.
9. M. Dong, L. Zhong "Self-constructive High-rate System Energy Modeling for Battery-powered Mobile Systems." in 9th International Conference on Mobile systems, Applications, and Services. ACM, 2011.
10. R. Murmuria, J. Medsger, A. Stavrou, JM. Voas "Mobile Application and Device Power Usage Measurements." in 6th International Conference on Software Security and Reliability (SERE). IEEE, 2012.
11. D. Hackenberg, T. Ilsche, J. Schuchart, R. chöne, WE. Nagel, M. Simon, I. Georgiou "Hdeem: High definition energy efficiency monitoring." in Proceedings of the 2Nd International Workshop on Energy Efficient upercomputing, E2SC'14. IEEE Press, 2014.
12. Intel Corporation "Intel® 64 and IA-32 Architectures Software Developer's Manual. Number 253669-053US", 2015.
13. Advanced Micro Devices "AMD BIOS and Kernel Developer's Guide (BKDG) for AMD Family 15h Models 00h-0Fh Processors", 2013.
14. M. Burtscher, I. Zecena, Z. Zong "Measuring GPU power with the K20 builtin sensor." in Proceedings of Workshop on General Purpose Processing Using GPUs, GPGPU-7. ACM, 2014.
15. Intel Corporation "Intelligent Platform Management Interface Spec", 2013.
16. C. Isci, M. Martonosi "Runtime Power Monitoring in High-end Processors: Methodology and Empirical data." in 36th annual IEEE/ACM International Symposium on Microarchitecture. IEEE Computer Society, 2003.
17. A. Shye, B/ Scholbrock, G. Memik "Into the Wild: Studying Real User Activity Patterns to Guide Power Optimizations for Mobile Architectures." in 42nd Annual IEEE/ACM International Symposium on Microarchitecture. ACM, 2009.
18. N. Binkert, B. Beckmann, G. Black, SK. Reinhardt, A.Saidi, A. Basu, J. Hestness, DR. Hower, T. Krishna, S. Sardashti et al. "The gem5 simulator" ACM SIGARCH Computer Architecture News, **39**(2), 17, 2011.
19. MT. Yourst"PTLsim: A Cycle Accurate Full System x86-64 Microarchitectural Simulator." in 2007 IEEE International Symposium on Performance Analysis of Systems & Software. IEEE, 2007.
20. JL. Henning"SPEC CPU2006 benchmark descriptions" ACM SIGARCH Computer Architecture News, **34**(4), 1-17, 2006.
21. JM. Rabaey, AP. Chandrakasan, B. Nikolic "Digital integrated circuits" Englewood Cliffs: Prentice hall, 2002.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 263-268 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

Efficient Resource Management for Uplink Scheduling in IEEE 802.16e Standard

A.R. Rahiman*, Noaman Abduljabbar Ramadhan, Abdullah Muhammed, Zuriati Zulkarnain

Dept. of Communication Technology and Networking, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400, Malaysia

ARTICLE INFO

Article history:
Received: 09 December, 2016
Accepted: 06 January, 2017
Online: 28 January, 2017

Keywords :
Algorithm
Broadband wireless access
Scheduling
QoS
WiMAX

ABSTRACT

The IEEE 802.16e standard, known as mobile Worldwide Interoperability for Microwave Access (WiMAX) becomes the most demanding broadband wireless access (BWA) technology recently. Its main advantage is rapid delivery of services in remote areas due to the cost efficiency factor. The base station (BS) supports data rate up to 70 Mbps, mobile stations with 5–15 km length of coverage, and for the fixed stations the wireless access range up to 50 km. To resolve the bandwidth contention issue and guarantee seamless packet transmission from the subscriber stations (SS) to the BS, the uplink (UL) traffic scheduling must be efficient and reliable. This paper studies the work on the UL scheduling algorithm, namely minimum rest time (MRT). The MRT goal is to strengthen the packet transferring time between the SS and the BS by refining the pre-stipulated expired time and the deadline time of the earliest expiry first (EEF) and earliest deadline first (EDF) hybrid algorithms. These legacy algorithms are inadequate to support the multi-class traffic systems due to the shortage of quality of service (QoS) parameters featuring. Moreover, the algorithms are highly static. Using the Omnet++ with the relevant performance metrics the obtained results confirmed the MRT outperforms effectively from the legacy algorithms.

1. Introduction

The high cost factor in establishing the wired broadband network in rural areas has become the major driving force of the expansion efficient wireless network systems. The present wireless broadband technologies have their own nature in giving the solutions to the challenges and issues imposed by the technology itself. Such challenges include QoS traffic channels, scarce bandwidth resource, interference, error rate, and mobility barrier. Presently, the mobile WiMAX standard becomes the most promising and widely being used for broadband wireless technology [1].

The outstanding features of the mobile WiMAX are wide frequency range coverage, last-mile accessing, and increased QoS traffic supports for various types of applications, especially the multimedia applications [2–3]. It efficiently can serve the metropolitan area (e.g., WMAN) network with high-speed data rates and wide range signal coverage. The two main architecture entities of the standard are BS and SS. The BS provides the air interface for the subscriber (up to 30 miles) while for the mobile stations is between three to ten miles. The SS (referred as customer premises equipment (CPE) can take either indoor or outdoor provides the connectivity among the subscriber's equipment and the BS. Others existing broadband wireless technology solutions are the IEEE 802.11a with data rate 54 Mbps with coverage area up to hundreds of meters, the enhanced data rates for GSM evolution (EDGE) with 384 Kbps and coverage zone up to few kilometers and the code-division multiple access 2000 (CDMA2000), with data rate of two Mbps and coverage zone up to few kilometers.

Figure 1 illustrates the three setup modes for the mobile WiMAX standard, namely i) point-to-point (PTP), ii) point-to-multipoint (PMP), and iii) mesh [4]. The PTP setup mode interconnects two BSs that having different or miscellaneous networks. On the other hand, the PMP mode was designed for the mobility structure where the respective SSs are directly connected to the localized BS. In the mesh mode, the BS and several SSs are connected to each other in an ad hoc manner. Each SS acts as a router that helps and collaborates directly to transmit the data onto the SSs. The transmission schemes for all setup modes take place through two independent channels, either UL channels (from SS to BS) or downlink (DL) channel (from BS to SS) [5-6]. Note that, the UL channel is shared among all SSs while the BS is only using the DL channel scheme.

*Corresponding Author: A.R. Rahiman, Faculty of Computer Science and Information Technology, UPM, Malaysia, Email: amir_r@upm.edu.my

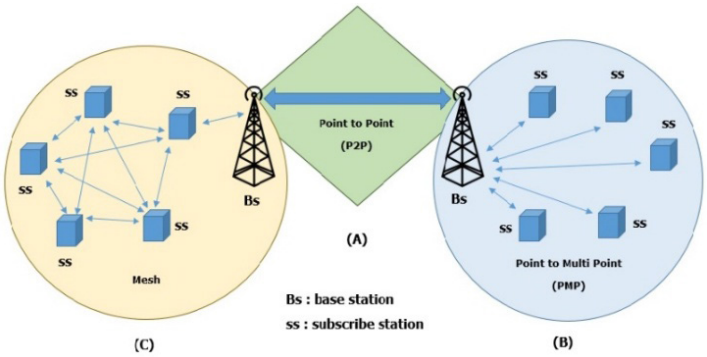


Figure 1: The WiMAX setup modes.

The main attention of this paper is the scheduling algorithms those guarantee the data packet flows for the UL traffic channel. To satisfy user demands and supporting new set of real-time services and applications, a realistic and dynamic resource allocation algorithm is mandatory. The UL scheduling algorithm at the BS should lead its decision with all the SSs whereas the DL channel algorithm is only worried in communicating the decision locally to the BS [7]. But, the UL scheduling algorithm design is a challenging task due to the requirement in different level of QoS classes, fairness and application difficulty [6]. One of the efficient UL hybrid scheduling algorithms is the earliest deadline first (EDF). However, the main problem with the algorithm is when the difference among the deadline is quite large, the lower priority queues have to starve. Moreover, the earliest expiry first (EEF) algorithm has given a solution to solve the packet waiting time avoiding, missing deadline, and delay reducing. However, the EEF algorithm has the limitations when the packet transmission time is not being considered. The packets with high and low priorities are being transmitted together, thus will create a delay in the network. To overcome the limitations of the existing algorithms, this paper studies the optimized UL scheduling, namely minimum rest time (MRT). The algorithm controls and monitors the packets flow using both their rest time and expiry time, respectively. Using comprehensive simulations, the rest and the expiry time parameters have shown a significant impact to QoS of the packets transmission as compared to the legacy UL scheduling algorithms.

The remainder of this paper is organized as follows. Section 2 discusses the background of mobile WiMAX UL scheduling algorithms and their related works. Section 3 provides the detailed explanations about the proposed MRT hybrid-scheduling algorithm. The performance of the proposed algorithm is being evaluated in Section 4. Finally, this paper concludes with Section 5.

2. Background and Related Works

2.1. WiMAX Uplink Scheduling

The main issues raised by many researchers in providing efficient mobile WiMAX services is a packet scheduling. The scheduling issue is constantly emerging the relation to the demands for WiMAX-related services. Thus, the scheduling algorithm should take into account the diverse WiMAX QoS classes and the service requirements. The UL scheduling algorithms are essential in solving the QoS bandwidth allocation contention among the users with diverse service classes of traffic. The examples of the traffic classes are best effort (BE), non-real-time polling service (nrtPS), unsolicited grant service (UGS), and real-time polling service (rtPS). Note that, IEEE 802.16e standard is the expansion from IEEE 802.16-d standard with additional mobility feature. The feature brings a significant effect on the QoS traffic. Figure 2 shows the expansion of both standards that yield the extended real-time polling service (ertPS) traffic class [8]. The ertPS traffic class takes the advantages of both the UGS and the rtPS classes [10]. Unlike the UGS class, the ertPS class is being designed to support the VoIP traffic with silence detention and the traffic flow that generates the variable sized packets. The packets are coupled with QoS guarantee and its space is generated in a periodical manner [2,7, 16–17].

As indicated by [11–12], there are three categories of the UL scheduling algorithms are i) homogenous, ii) hybrid, and iii) opportunistic.

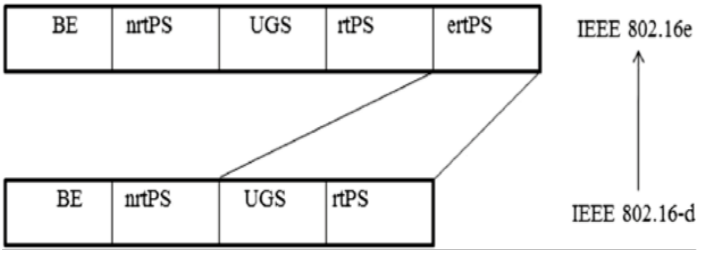


Figure 2: WiMAX QoS traffic classes [15].

The hybrid scheduling is the ideal for the UL since it explicitly caters to all the required QoS parameters associated with the WiMAX traffic classes. This motivates us to mainly focus on optimizing the UL scheduling in guaranteeing efficient mobile WiMAX services.

2.2. WiMAX Uplink Scheduling

Table 1 summarizes the existing hybrid scheduling algorithms with the different QoS traffic classes. In deriving various solutions for optimize the UL scheduling algorithm, most researchers force to create hybrid scheduling [11]. The algorithms combine the legacy algorithms to satisfy the QoS requirements of the multi-class traffic as specified in the IEEE 802.16e standard.

Table 1: Hybrid Scheduling Algorithms

Authors	Algorithms	QoS classes
Oad et. al [15]	EEF + WFQ + FIFO	(EEF → rtPS) (WFQ → nrtPS) (FIFO → BE)
Wongthavarawat and Ganz [19]	EDF + WFQ + FIFO	(EDF → rtPS) (WFQ → nrtPS) (FIFO → BE)
Vinay et. al [14]	EDF + WFQ	(EDF → rtPS) (WFQ → nrtPS, BE)
Settembre et. al [20]	WRR + RR	(WRR → rtPS, nrtPS) (RR → BE)
Gidlund et. al [16]	EDF + WFQ	(EDF → ertPS, rtPS) (WFQ → nrtPS, BE)
Chowdhury et. al [7]	EDF + DFPQ	(EDF → UGS, rtPS) (DFPQ → nrtPS, BE)

The work by [13] has evaluated the various scheduling algorithms used for the UL traffic in WiMAX. Under the hybrid EDF, weighted fair queuing (WFQ) and first in first out (FIFO) schedulers, the UL traffic produces the greatest throughput. In

addition, the hybrid scheduler that appoints multiple legacy algorithms with the established traffic classes have been studied and being evaluated by [14]. For instance, the EDF scheduler for the rtPS, the WFQ scheduler for nrtPS, and the FIFO scheduler for the BE class. The schedulers those serve various traffic classes perform better than the legacy algorithms as they satisfy the multi-class traffic QoS requirements. Further, the hybrid algorithm explicitly caters to all the required QoS parameters associated with the respective traffic classes in the IEEE 802.16e standard.

2.3. EEF+WFQ+FIFO Hybrid Algorithm Scheduling

The algorithm has been designed as an extraction from the EDF hybrid sub-scheduling algorithm [14]. It controls and displays the data packets using the deadline and an individual packet expiry time. Thus, packets are being scheduled based on a mixture of specified deadlines and the on-going knowledgeable delay. Therefore, packets with the smallest time expires will be scheduled first while packets with bigger time expires are always buffered indefinitely. In many instances, those packets will expire and miss their time expire.

Further, the algorithm has catered distinctly each of the diverse requirements of the multiclass traffic supported by the WiMAX. However, in ensuring a preferred service sorted based on traffic constraints, the algorithm is of absolute the chronological preference. For instance, the service is always given to the higher priority traffic class where the lower priority traffic class will be the only service when higher priority has finished. Once the allocation is done, there are no changes until the end of the transmission. This causes rigid and non-flexible resource utilization.

2.4. EDF+WFQ+FIFO Hybrid Algorithm Scheduling

As mentioned by [18], no single inheritance schedulers that can fulfill all the QoS requirements of the WiMAX network applications. The EDF+WFQ+FIFO scheduling algorithm was proposed to provide low-delay and packet loss for real-time applications [19]. In addition, it is designed for handling the diversified QoS traffic classes. The traffic classes create an effective management but they impose higher QoS demands in the networks. The algorithm service is being executed with different QoS schedulers to meet these requirements. For instance, the EDF is being used for the rtPS traffic, the WFQ for the nrtPS traffic, while the BE traffic employs the FIFO scheduler.

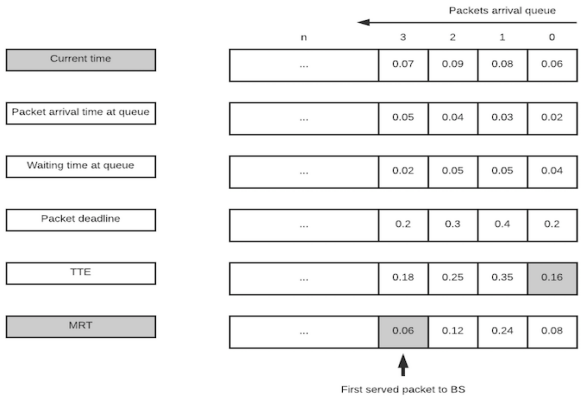
The hybrid EDF+WFQ+FIFO scheduling uses the stern priority service mechanism. That means all the bandwidth for the higher priority SSs are being distributed evenly until they do not have any packets to transmit. The disadvantage is that lower priority SSs will starve in the presence of many higher priority SSs due to the strict priority overall bandwidth allocation. Further, the algorithm does not take into consideration the variable channel conditions of each SS and constraints the flexibility of the WiMAX. In addition, it does not provide different bandwidth grant sizes for different quality of service classes.

3. MRT+WFQ+FIFO Algorithm

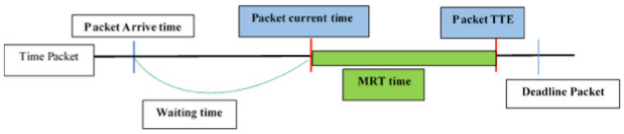
The MRT algorithm has been designed as an abstraction from the EEF hybrid sub-scheduling algorithm. The algorithm has controls and monitors the data packets using their packet rest time and respective expiry time. When compared with the EEF, the packets in the MRT algorithm are being scheduled based on a combination of the specified packet rest time to reduce the happened earlier expiry time. Packets that have shorter MRT are being scheduled first as opposed to those with longer MRT. This means the data packets with a shorter time to expire being served first, but must have a minimum MRT time. On the other hand, packets with maximum MRT has a long expiry time due to all minimum will be served and transferred gradually to serve the longer MRT value packets. To help in this manner, the priority queue is being used to arrange packet arrival time according to the priority. This grants minimum delay time for the earliest packet to be served.

3.1. MRT Algorithm Efficiency

Both diagrams in Figure 3 show how to select the priority packet (e.g., the 4th packet) to be served in the MRT algorithm. The algorithm considers the transmission time of the packet and those packets that are present in the queue are more sensitive to their deadlines. If the expiry time is being crossed, then the packet is useless. To reduce the delay, the MRT algorithm works based on the minimum time for the packet to stay in the queue. Thus, it will lead to the validity of minimum waiting at the packet to be expired.



a) MRT scheduling



b) MRT time calculation

Figure 3: The MRT algorithm

Figure 4 simplifies the MRT scheduling algorithm. To determine the packet to be served in the MRT algorithm, the calculation for the packet as follows: Any packet has the current time (*curr_TMS*) and arrival time (*packet_arr_TMS*), respectively. The waiting time (*wait_TMS*) of the packet in the queue is being calculated as:

$$wait\_TMS = curr\_TMS - packet\_arr\_TMS \quad (1)$$

Each packet has the deadlines which not being crossed each other. The time to expire (*TTE*) is being calculated as:

$$TTE = deadline - wait\_TMS \quad (2)$$

The packet is being transmitted from the source to destination.

Moreover, when it arrives at the queue, it may have minimum time to stay or *TTE*. Thus, the *MRT* time is being calculated as:

$$MRT = (TTE - packet\_arr\_TMS) - curr\_TMS \quad (3)$$

```
1.  Arrival ⇐ arrival( )
2.  if flowType == ertPS then
3.  CheckQueueSize < MaximumQueueSize
4.     ++npa_ertPS
5.     ++Cqsize
6.     PacketArrivalTimeQueue ⇐ simclock
7.     AssignDeadlineArrivalPacket ⇐ deadlineertPS
8.  else if (queue ertPS > 0) || ( queueertPS < threshold1) then
9.              Priority ⇐ deadlineertPS;
10. end else if
11. else DropThePacket then
12.             ++plr
13. end if
14. end arrival
15.
16. Departure ⇐ departure()
17. if CheckQueueSize != 0 then
18.             ++npd_ertPS
19.             Delay ⇐ simclock - PacketArrivalTimeQueue
20.             CurrentDiffernce ⇐ deadlineertPS - Delay
21.             TimeToExpireertPS[currentDiffernce]
22.             RestTime =TimeToExpireertPS –simclock
23.             MiniRestTime ⇐ RestTime
24.             SearchingMinimumValue ⇐ MiniRestTime
25. else CheckQueueSize == 0 then
26.             BSS ⇐ BaseStation_IDLE
27. end if
28. end departure
```

Figure 4: Pseudo code of the MRT Scheduling

4. Experimental Setups and Evaluations

4.1. Simulation Model and Parameter

The traffic rate for each traffic class used in the simulation is tabulated in Table 2 and the proposed MRT algorithm was implemented using the Omnet++ simulation tool.

Table 2: Simulation Parameters

Inputs	Values	Packet Size (bytes)	Rate (Kbps)	Load (λ) (Kbps)
BS	1			
SS	6 to 36			
SS ratio	3:1:1:1			
Simulation time (simClock)	50 secs.			
Service time	20 Mhz			
Voice		23		
Video		150 – 300		
FTP		150		
HTTP		100		
ertPS			64	1000
rtPS			500	3000
nrtPS			500	2500
BE			64	300

The MRT scheduling algorithm performance has been evaluated by conducting extensive simulation experiments. The main components of the developed simulator were oriented on the BS and SSs. The BS representing the scheduling algorithms and each sub-scheduling algorithm within the hybrid QoS classes (i.e., *ertPS*, *rtPS*, *nrtPS* and *BE*) are being executed individually. The simulation run time was set to 50 seconds for ensuring equal comparison algorithm platforms. In this study, six experiments were conducted according to various numbers of SS (e.g., 5, 12, 18, 24, 30, and 36).

4.2. Performance Metrics

The selections of the performance metrics in this study are based on the substantial adoption in UL scheduling for both EEF and EDF algorithms. The performance metrics and their respective derivation are as follows:

$$Avg\_throughput = Packets\ departure\ /\ Simulation\ time \quad (4)$$

$$Avg\_delay = Total\ delays\ /\ Packets\ departure \quad (5)$$

$$Missed\_deadline\ (misdl)\ ratio = tpkt\_misdl\ /\ npa \quad (6)$$

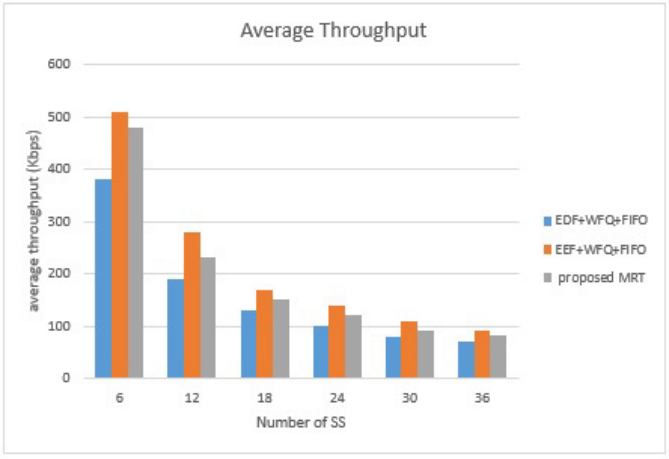
Parameters *tpkt_misdl* refers to the *misdl* overall total while the *npa* is the total number of packets arrived. And

$$Avg\_queue = Total\ queue\ /\ time \quad (7)$$

The parameter time in (7) refers to simulation clock time (e.g., simclock).

4.3. Results and Discussions

The results of the proposed MRT algorithm are presented in comparison with EEF+WFQ+FIFO and EDF+WFQ+FIFO algorithms. The average throughput results are illustrated in Figure 5. As can be seen, the average throughput gradually decreases as the number of SS grew among the algorithms. The proposed MRT algorithm performs better due to the reflection of the dynamic versatile expire time parameter. Compared to the existing algorithms, they depend on the static pre-defined deadline parameter.



Next, since there is many packets waiting in the queue and eventually expired, the MRT algorithm has shown the significant ability in reducing the delays. Figure 6 shows the average delay between the existing and the proposed algorithm. Although the MRT algorithm shows mild increase for the delay, the maximum

value achieved before being constantly maintained (i.e., from the number of SS equivalent to 36) is still lower as compared to existing algorithms.

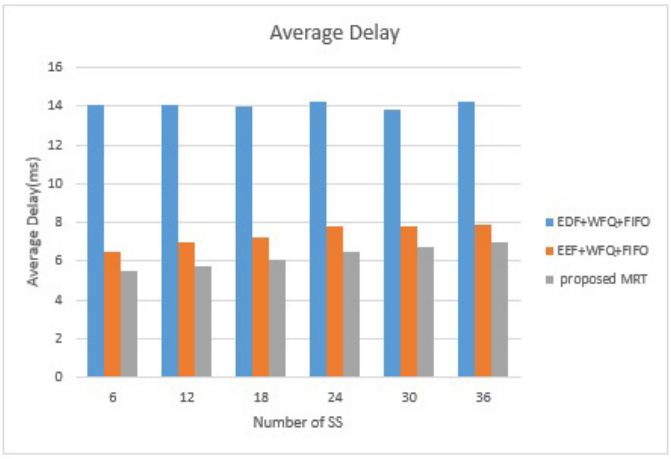


Figure 6: *Avg_delay*

Figure 7 illustrates the packets missed deadline ratio performance among the algorithms where the proposed MRT gives the lowest ratio. By using the MRT time, each of the arrived queued packets which have the earlier waiting will be propagated to have the priority to be served first. Thus, the algorithm performs slightly better the existing UL scheduling algorithms.

Figure 7: *misdl ratio*

In terms of the queue size performance, the percentage of queue size gradually increases when the number of SS increases, see Figure 8. Again, as compared to the MRT algorithm, the algorithm guarantees the lowest average queue size when the packets are being transmitted from the source to destination. Thus, theses simulation results have proven that the proposed MRT algorithm has superior performance.

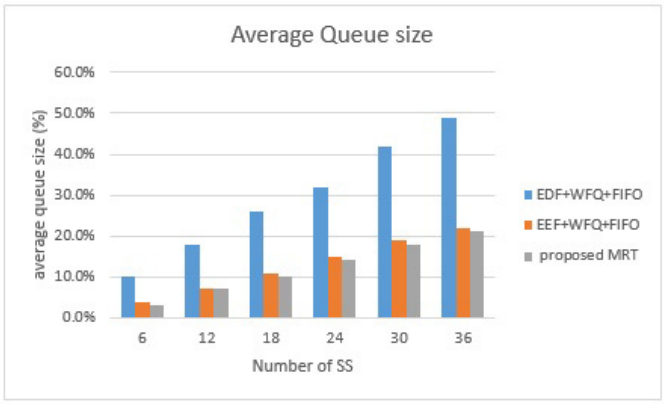


Figure 8: *Avg_queue*

5. Conclusion

The acquired results prove that the proposed MRT algorithm has successfully enhanced by increasing transfer packets arriving at a base station in the UL mode for the mobile WiMAX network system. The results of reviewed scheduling algorithms for the mobile WiMAX have greatly influenced the direction in which the designed and proposed solutions have taken form. In this research, the scheduling algorithm tailored at enhancing the collective performance of hybrid algorithms in the mobile WiMAX domains has been designed and developed. The spectrum of constraints that have been extracted from the hybrid EEF+WFQ+FIFO algorithm includes the static nature by which priorities are assigned and maintained during the entire duration of a transmission time. Reengineering the scheduling mechanics governing the EEF algorithms is the main contribution of this study. The dominance of the pre-stipulated deadline is indeed acknowledged in the proposed and developed enhanced MRT. However, the significance of providing a monitoring mechanism that gauges between the stipulated and the reality of performance is the essential focus of the MRT algorithm. Unlike the legacies algorithms, the MRT algorithm is able to revaluing the packets of interest for the transmission in case when packets decline the service due to a longer deadline. In such cases, the proposed MRT algorithm is being arranged with real-time statistics of the expiry time and will serve as the main denominator of the priority. The results acquired from the conducted performance analysis using the relevant performance metrics have proven the MRT outperforms successfully from the legacy algorithms. The future work that would further provide contributions to the elevation of WiMAX includes the incorporation of mobile SS, correlating the different traffic classes in a holistic admission control and the complementary downlink algorithms analysis.

Acknowledgment

The authors wish to thank the reviewers for their constructive comments and suggestions helped in improving this research paper.

References

[1] A. Singh, A. Vaish, P. K. Keserwani, “Research Issues in Wireless Networks” International Journal of Advanced Research in Computer Science and

Software Engineering (IJARCSSE), 4(2), 572 - 575, 2014.

[2] J. Lakkakorpi, A. Sayenko, J. Moilanen, "Comparison of Different Scheduling Algorithms for WiMAX Base Station" in Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC 2008), Las Vegas Nevada USA, 2008.

[3] N. A. El-fishawy, M. Zahra, M. Ebrahim, M. M El-gamala, "ModifiedCross-layer Scheduling for Mobile WiMAX Networks" in 28th National Radio Science Conference (NRSC), Egypt, 2011.

[4] C. So-In, R. Jain, AK Tamimi, "Scheduling in IEEE 802.16e Mobile WiMAX Networks: Key Issues and a Survey" IEEE Journal on Selected Areas in Communications, 27(2), 156–171, 2009.

[5] K. Komaljot, S. Amitabh. "Bandwidth-Aware Stochastic Uplink Scheduling in WIMAX Networks", International Journal of Advance Research, Ideas and Innovations in Technology (IJARIIT), 2(3), 1-5, 2016.

[6] Ruey-Rong Su, I-Shyan Hwang, Bor-Jiunn Hwang, "A new cross-layer scheme that combines grey relational analysis with multiple attributes and knapsack algorithms for WiMAX uplink bandwidth allocation", EURASIP Journal on Wireless Communications and Networking, 1-11, 2016.

[7] P. Chowdhury, I. S Misra, "A Fair and Efficient Packet Scheduling Scheme for IEEE 802.16 BWA Systems" arXiv preprint arXiv:1009.6091. 2010

[8] M. S. Arhaif, "Comparative Study of Scheduling Algorithms in WiMAX" International Journal of Scientific and Engineering Research, 2(2), 1 – 7, 2011.

[9] L. Nuaymi, WiMAX: Technology for Broadband Wireless Access, John Wiley & Sons, 2007.

[10] M. D. Priya, M. L. Valarmathi, D. Prithviraj. "Performance Analysis of Scheduling Mechanisms in WiMAX Networks", SOP Transaction on Wireless Communications, 1(1), 40 – 49, 2014.

[11] S. R. Puranik, M. Vijayalakshmi, L. Kulkarni, "A Survey and Analysis on Scheduling Algorithms in IEEE 802.16e (WiMAX) Standard", International Journal of Computer Applications, 79(12), 1 – 10, 2013.

[12] A. Kanda, A. K. Gill, A. K. "Comparative Study of Various Scheduling Schemes of Wimax" International Journal of Computer Science and Network Security (IJCSNS), 15(9), 91 – 94, 2015.

[13] P. Dhrona, N. A. Ali, H. Hassanein, "A Performance Study of Scheduling Algorithms in Point-to-Multipoint WiMAX networks" in 33rd IEEE Conference on Local Computer Networks (LCN), Montreal, AB, Canada, 2008.

[14] K. Vinay, N. Sreenivasulu, D. Jayaram, D. Das, "Performance Evaluation of End-to-End Delay by Hybrid Scheduling Algorithm for QoS in IEEE 802.16 Network" in IFIP International Conference on Wireless and Optical Communications Networks, Bangalore, India, 2006.

[15] A. Oad, S. K. Subramaniam, Z. A. Zukarnain, "Enhanced uplink scheduling algorithm for efficient resource management in IEEE 802.16" EURASIP Journal on Wireless Communications and Networking, 2015(1), 1 – 12, 2015.

[16] M. Gidlund, G. Wang, "Uplink Scheduling Algorithms for QoS Support in Broadband Wireless Access Networks" Journal of Communications, 4(2), 133 – 142, 2009.

[17] M. Oktay, H. A. Mantar, "A Real-time Scheduling Architecture for IEEE 802.16 - WiMAX Systems" in IEEE 9th International Symposium on Applied Machine Intelligence and Informatics (SAMI), Smolenice, Slovakia, 2011.

[18] N. A. Ali, P. Dhrona, H. Hassanein, "A Performance Study of Uplink Scheduling Algorithms in Point-to-Multipoint WiMAX Networks" Computer Communications, 32(3), 511 – 521, 2009.

[19] K Wongthavarawat, A Ganz, "Packet Scheduling for QoS Support in IEEE 802.16 Broadband Wireless Access Systems" International Journal of Communication Systems, 16(1), 81 – 96, 2003.

[20] M. Settembre, M. Puleri, S. Garritano, P. Testa, R. Albanese, M. Mancini, V. Lo Curto, "Computer Networks", International Symposium on IEEE, 11 – 16, 2006.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 269-276 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

Proposal of a congestion control technique in LAN networks using an econometric model ARIMA

Joaquín F Sánchez[1], Martha M Cuellar[2]

[1]Docente Ingeniería de Telecomunicaciones, Facultad de Ingeniería, San Mateo Educación Superior, Bogotá Colombia

[2]Docente Negocios Internacionales, Facultad de Negocios Internacionales, San Mateo Educación Superior, Bogotá Colombia

ARTICLE INFO

Article history:
Received: 24 December, 2016
Accepted: 23 January, 2017
Online: 28 January, 2017

Keywords:
LAN Network
Time series
Congestion control
ARIMA model

ABSTRACT

The aim of this paper is to propose a methodology to apply to an econometric model which could be the basis for generating a congestion control strategy in a LAN network. Using time series as analysis tool, ARIMA model (Autoregressive Integrated Moving Average) is used in an econometric model, and this same analysis is transferred to the mechanism of congestion control of the TCP protocol. As a result, a transmission rate reduction parameter is generated in a test environment, where congestion control strategy is applied. It is shown that this procedure works and its application could be further expanded in data networks.

1. Introduction

This paper is an extension of work originally presented in conference 2016 8th Euro American Conference on Telematics and Information Systems (EATIS) "Congestion control LAN networks using an econometric model ARIMA"[1] .In Computer Sciences, especially in Telecommunications, several mathematical tools to make analysis and design of the systems allowing to improve their own performance are used. For instance, time series are used on research and industry fields, and reviewing their applications as in [2]–[5], it was found that stock exchange analysis has some similarities with the telecommunications field. After look over the LAN networks traces as viewed in [5] and [6], network trafficking behavior is as similar as the variation graphs of the shares in the stock exchange. Considering this, the present paper aims to reach commonalities between data network trafficking and stock exchange behavior to 1) design a traffic analysis tool in data networks, and 2) propose a model of congestion control based on an econometric model.

Time series are suitable tools to evaluate a given situation of a performance movement of the analyzed variable, and data congestion in the telecommunications networks in this case. This kind of mathematical resources are appropriate to detect immediate trends or errors related to the volume and movement of analyzed data, providing better correctives and trends which are important for data channels management; therefore, time series afford occurrence predictions for decision-making.

*Corresponding Author: Joaquín F Sánchez, Docente Ingeniería de Telecomunicaciones, Facultad de Ingeniería, San Mateo Educación Superior, Bogotá Colombia Email: js012108@gmail.com

Congestion analysis in data networks may carry out using different parameters and trends, getting estimates for the use, control, and application of statistics measures in decision-making. This could guide predictions of the variables in a given model with their historical behavior, statistics, and features to give the chance to value the future occurrence level [8]. In this way, past data may lead to forecast future values movements if data are established and depicted through graphs and specific indicators given by positive or negative variables trends. It is worth mentioning that delay times and LAN networks bandwidth are the variables considered for this paper.

According to the previous concepts, analysis of one share of a given company will be presented. This shall allow to propose the congestion control model of data obtained from a LAN network implementation. In section 2, a theoretical revision of econometric analysis for one share is made and, in section 3, it is discussed why to design a congestion control strategy is worth, explaining some theoretical elements. Proposal of the model and results are presented in sections 4 and 5, and conclusions are displayed in the last section.

2. ARIMA model (*autoregressive integrated moving average*) to analyze one share

It could be useful to estimate a model capable of making technical analysis related to the behavior of negotiation indicators in the financial market because they might enhance the evaluation of trends, variables, relations of influence, and historical behavior within supply and demand of stock market. In general, the price and the volume of negotiation are the most relevant variables taken

into account mainly because they determine the movement of the variable: the upward or downward trend, or its resistance to change through considered time opening and closing, like relation of values, maximum and minimum, and volatility in periods of time such as days, weeks, months, and years. In this way, information of global market and its necessities is reflected, determining negotiation trends with patterns of behavior, and allowing that information does exhibit a cyclic and repeated activity according to the own market features.

Looking over time series characteristics in stock markets, basic concepts of one analyzed variable could be established by *chartism (analysis of financial graphs)* [9]. In fig. 1, the price of one share is depicted. The downward market is the bold outlined circle and the upward market is the light. The bases (dotted lines) delimit the trend in the bearish movement, giving the chance to rebound towards the bullish, meaning that the price demand stands out compared to the supply. On the other hand, in the upward trend of the market (dashed lines), supply is greater than demand.

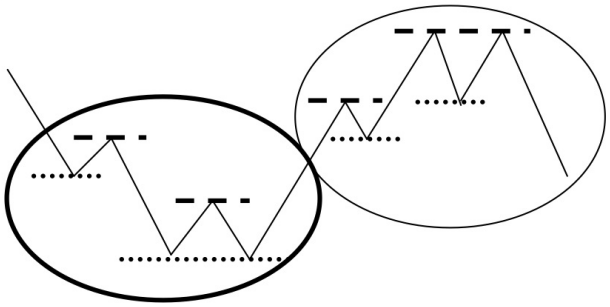


Fig1: Variation concept [2], [8]

The assessments of prices can progress in a specific time $t(n)$ to have a suitable measure in a given period of time. Highlighting that a market is named *bullish* when the price exhibits the trend to be over its own value, the maximum benchmarks come up, (*1, 2, 3, 4* in fig. 2) increasingly higher, and minimum benchmarks (*1., 2., 3., 4.* in fig. 2) are greater than the previous ones. In this way, it is created a trend line which breaks the base of the price trend, giving rise to a new trend, named *pull back* in the stock market environment. On the other hand, the *bearish* market exhibits an opposite behavior compared to the *bullish* market, with the same contextual benchmarks and indicators, associated to the downward trend.

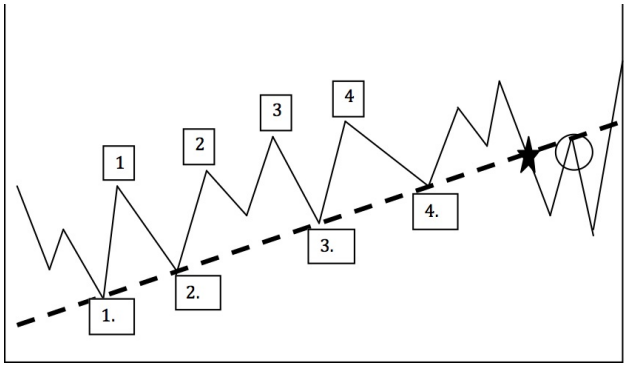


Fig2: Prices trend [2], [8]

ARIMA model can be used to make a *Forecasting* analysis. When the model is proposed, an estimation of values behavior of the shares is obtained, allowing to make a decision to sell or buy. This model comes from a nonstationary homogenous series.

Given the series $(1-B)^d Z_t$ which follows a general stationary process $ARIMA(p,q)$, then:

$$\phi_p(B)(1-B)^d Z_t = \theta_0 + \theta_q(B)a_t \quad (1)$$

Where the stationary operator $AR\ \phi_p(B) = \left(1 - \phi_1 B - \cdots - \phi_p B^p\right)$ and the operator $MA\ \theta_q(B) = \left(1 - \theta_1 - \cdots - \theta_q B^q\right)$ share common factors. The θ_0 parameter is used in different ways when $d = 0$ y $d > 0$. When $d = 0$, the original process is stationary, then θ_0 is related to the mean of the process. When $d \geq 1$, the θ_0 operator is named the term of deterministic trend. The homogenous result of the described model refers to an *autoregressive* model which integrates a moving average unit for an order (p,d,q) and is denoted $ARIMA(p,d,q)$.

Fig. 3 depicts an example of ARIMA model functioning. The time series behavior exhibits the variation of a product with the $ARIMA(2,1,2)$ scheme where:

$$(1 - \phi_1 B - \phi_2 B^2)(1-B)Z_t = (1 + \theta_1 B + \theta_2 B^2)a_t \ (1)$$

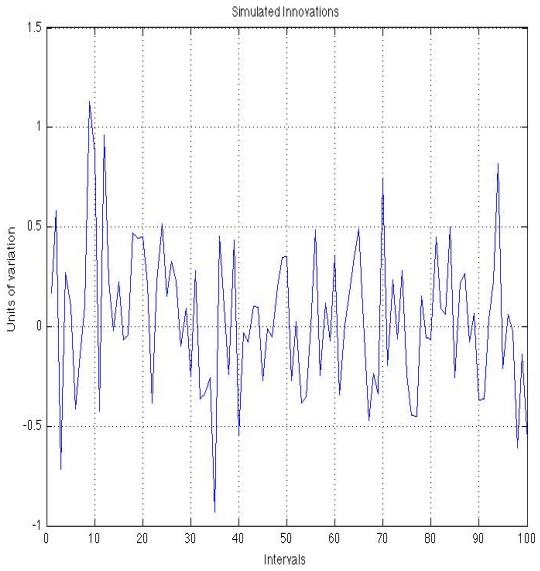


Fig3: ARIMA model (2,1,2)

In other words, $ARIMA(p,d,q)$ model allows to describe one value as a linear function of previous data and errors which come from the random behavior of that data. It works with a stationary series, and the $ARIMA(p,d,q)$ model is used to get an estimation or a price forecast, where the time series is put and mean squared error is calculated. To determine the variance of the series we have:

$$\pi(B)Z_{t+l} = a_{t+b} \quad (2)$$

Where:

$$Z_{t+l} = \sum_{j=1}^{\infty} \pi_j Z_{t+l-j} + a_{t+l} \quad (3)$$

Parameters must be calculated with those equations to get an estimation of the variance trend of the original series; hence, we have the equations as follows:

$$e_n(l) = Z_{n+l} - \hat{Z}_n(l) \quad (4)$$

$$\psi_j = \sum_{i=0}^{j-1} \pi_{j-y}\psi_i \quad (5)$$

Error is calculated with equation 4, and terms of the estimated series are determined with equation 5.

2.1. Graphical analysis and series analysis

Fig. 4 depicts the result of taking stock exchange behavior and time series into account. It is worth thinking in an $ARIMA(p,d,q)$ model which allows to one independent variable (*share #1*) be related to or compared with other independent variable (*share #2*) according to its behavior in the stock exchange.

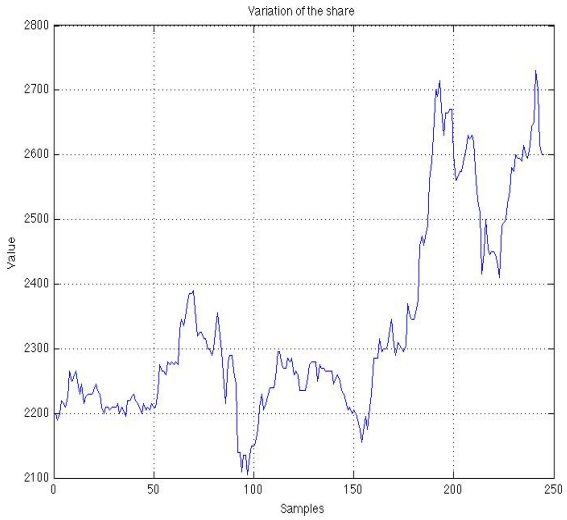


Fig4: Graph of one share behavior

Variable *Share #1* will take error $e_n(l)$ into account within the constant mean, and its ***B* (risk level)** will be reflected, where risk level remains in the trend of its volatility inside of the independent stock market, producing a single one variance. In the same way, behavior of the *share #2* will be the same. Relation between these two variables will produce a covariance which shall describe the behavior inside of the stock market, but behavior itself does not have an impact on the independent variable performance within the market.

Then, we have the mean of the independent variable:

$$x = \frac{\sum_{l-1}^{n} x_i}{N} \quad (6)$$

The variance of the independent variable:

$$\sigma^2 = \frac{(x_1-x)^2+(x_2-x)^2+\cdots+(x_n-x)^2}{N} \quad (7)$$

$$\sigma^2 \frac{\sum_{l-1}^{n}(x_l-x)^2}{N} \quad (8)$$

Covariance of the two independent variables within the stock market:

$$\sigma_{xy} = \frac{\sum f_i x_i y_i}{N} - xy \quad (9)$$

Beta, risk level:

$$\beta = \frac{COV\ AR(Mercado,Accion)}{VAR(Mercado)} \quad (10)$$

Having previous equations, $ARIMA(p,d,q)$ model is used within the stock market [10] and [11] which works to calculate the price of an asset or a liability. This model allows to establish the behavior projection of the share in a specific period of time, and in this paper the congestion volume of a LAN network is the variable; therefore, this kind of analysis let to make an estimation considering historical behavior of data which are being analyzed in periods of measurement.

It is considered that $ARIMA(p,d,q)$ model is defined by:

$$Z_t = Z_{t-1} + \theta_0 + a_t \quad (11)$$

$$Z_t = Z_{t-2} + 2\theta_0 + a_t + a_{t-1} \quad (12)$$

$$Z_t = Z_k + (t-k)\theta_0 + \sum_{j=k+1}^{t} a_j \quad (13)$$

Taking this work as a reference [11]:

- Z_t is the time series of data from expected return of the capital over asset t.
- Z_k is the time series of historical data.
- θ_0 is the parameter which determines the series trend.
- $\sum_{j=k+1}^{t} a_j$ are the rest of the terms of the data series.

3. Congestion control in data networks

Congestion in data networks has become a proper phenomenon of this technology, and it comes up in several parts of the network architecture, beginning in the sizes of the *routers buffers* which are responsible for make the routing of the packets. Those *buffers* are limited, and they can be shaped as a type of queue M/M/1/ N/∞, where the queue size is N which is a limited resource; therefore, when the queue is full, a known consequence starts to come up, denoted packet loss which can be measured as likelihood of blockage.

$$P_B = P_N = \frac{(1-\rho)\rho^N}{1-\rho^{N+1}} \quad (14)$$

In equation 14, P_B represents the likelihood of blockage of requests in a data network, or the likelihood of discarding packets. ρ is the trafficking intensity travelling through connections or flow of packets. N is the number of channels or the size of *buffers* in the switching systems.

Equation 14 develops as follows:

If we have λ = arrival rate of requests y μ = service rate, the next graphic appears:

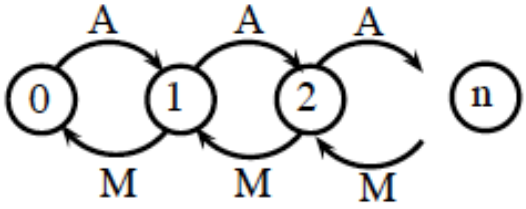


Fig5: Diagram of states for likelihood of blockage

Based on the diagram of state, it is possible to propose equations of balance for different states:

State	Output rate	Input rate
0	$P_0\lambda$	μP_1
1	$P_1\lambda+\mu P1$	$\lambda P_0+\mu P_2$
2	$P_2\lambda+\mu P_2$	$\lambda P_1+\mu P_3$
n	$P_n(\lambda+\mu)$	$P_{n-1}\lambda+P_{n+1}\mu$

Then, we have:

$$P_1 = \frac{P_{0\lambda}}{\mu}$$

$$P_2 = P_0 \left(\frac{\lambda}{\mu}\right)^2$$

$$P_n = P_0 \left(\frac{\lambda}{\mu}\right)^n$$

If it is known that $\sum_{n=o}^{N} P_n = 1$, then:

$$P_n = \sum_{n=0}^{N} P_0 \left(\frac{\lambda}{\mu}\right)^n = 1$$

$$P_0 = 1 \Big/ \sum_{n=0}^{N} \left(\frac{\lambda}{\mu}\right)^n$$

$$P_n = \frac{\left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^N}{1 - \left(\frac{\lambda}{\mu}\right)^{N+1}}$$

Now, it is possible to say that $\frac{\lambda}{\mu} = \rho$ which is the trafficking intensity of requests or packets for a data network.

Equation 14 represents likelihood of blockage of a used model at the beginning of telephone systems. In mathematical terms, it models the value of the service quality that would be given to a network, depending on the size of the *buffer*. In the course of evolution of telecommunications networks, specifically data networks, equation 14 has been adapted to measure resources and fulfill demand needs of the users.

However, the size of the *buffers* is not the only one resource that affects the congestion of data networks, but also the processing level of the network elements, and the capacity of transporting information, such as the connections bandwidth between nodes. Bandwidth is one of the most important resources in data networks that directly affects congestion because it is the main aspect where packet transport becomes tangible and from which is possible to measure service deterioration.

As fig. 6 depicts, congestion can be represented as a downward curve that drops depending on the number of flows that enter the system. If system turns congested, the more flows they enter, the less efficiency of network performance.

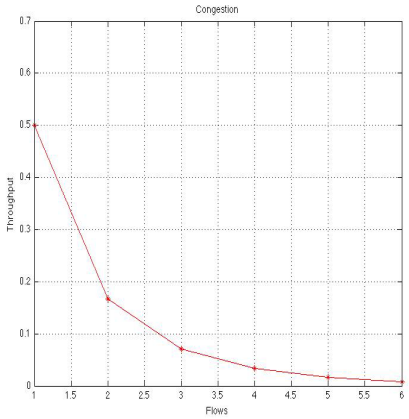


Fig6: Concept of congestion

To demonstrate this behavior, the limit when ρ tends to infinity can be calculated; thus, we have:

$$P_{n=} \lim_{\rho \to 1} \frac{(1-\rho)\rho^N}{1-\rho^{N+1}}$$

$$P_n = \frac{d/_{d\rho}\left((1-\rho)\rho^N\right)}{d/_{d\rho}\left(1-\rho^{N+1}\right)}$$

$$P_n = \frac{1}{N+1}$$

This expression defines the congestion shape in a data channel. This behavior can be replicated not only in connection flows, but also in different junction sections of data connections; for instance, the formed bottlenecks when there is a transfer from distribution network to access network because the mean of transmission has to change due to the architecture, as is seen in HFC networks (*Hibrid Fiber Coaxia*) that mix fiber optic with coaxial cables.

3.1. Congestion control with TCP

Congestion control models have been focused on the transport layer protocol TCP [12] and [13], in such a way that flow control by sliding window protocol, AQM technique (Active Queues Management) [14], admission control, and delay time management are the more used techniques. All of this is made by different changes or versions to the TCP protocol, beginning by ***TCP Tahoe*** and ***TCP Reno***, which modify retransmission of ***ACK*** packets or confirmations.

The main goal of this paper is to propose a model of control without any change to the protocol because it leads to make modifications in the times of transmissions, confirmations, and retransmissions of the network elements, and these changes could worsen the network behavior; therefore, a model without modifications in the protocol structure is proposed to help to treat the behavior of network elements for their adaptation to the flow of produced data from random behavior of users.

3.2. TCP basis

RFC 793[15] describes TCP functioning, where window of congestion control is not dynamic. First variation implemented on TCP protocol is the fitting of transference rate to the level of current congestion presented in data network. The purpose of making this reduction of transference rate is to diminish the number of lost packets in connections. Next steps of change in TCP protocol lead to TCP Tahoe algorithm which modifies *slow start* and *congestion avoidance* algorithms, as well as fast retransmission. In TCP Reno another modifications happens, consisting in TCP Tahoe but with a variation in the fast recovery algorithm [16].

The main functions of slow start and congestion control are the control of transference rates of generated packets and the

determination of available capacity of network connections; however, making a process in congestion window is needed to achieve this. This process is denoted ***cwnd***, and it is defined as the number of transmitted bytes whose confirmation is not required. Another considered element is the recovery window denoted ***rwnd*** which is the number of bytes that receptor is capable of receiving from a given source [17]–[19].

To establish connection between two network elements, activity of information transport can be described as follows:

- Maximum size is assigned to *cwnd* by transmitter.
- The transmitter starts a retransmissions timer.
- Slow start algorithm sends a packet or a section if an ACK is received; then, the *cwnd* value is augmented in one section.
- After that, two sections can be sent and cause two ACK.
- For each ACK, *cwnd* is augmented in one section.
- Growth process of congestion window is exponential and keeps this rhythm as far as the bounded value is reached by *rwnd* and established threshold is attained by ssthresh.
- Congestion avoidance algorithm is executed when ssthresh value is obtained and the *cwnd* value is augmented in one additional section, but this time, by each ACK and RTT.
- When packets loss happens, the process returns to the slow start: *cwnd* is again established to one section and ssthresh to half the window size.

Having knowledge about TCP protocol functioning and how congestion control is made, a congestion control model based on an econometric model denoted ARIMA is pose, which will be fused together with TCP mechanism to measure congestion control of a LAN network.

4. Approach of the model

According to the analysis made in section 2.1 and the congestion control functioning explained in section 3, the relation between ARIMA(p, d, q) model and the congestion control is presented as follows:

- A time series with data of packets circulating in the network must be taken.
- $Z_t$parameters for this time series must be calculated.
- The function of yield rate for Z_t must be estimated.
- Transmission rates must be fitted with obtained results, and the congestion control must be verified.

To attain this, data are taken from a working LAN network through obtained traces from the network interface of a Switch gate. Analogy between compilation of LAN network data and econometric model is the way in which statistics of variables that work in the network are represented. Each feature is taken as one share of a product, leading to analyze the behavior of one or more shares to know how market behavior will be. In this case, market is the whole data trafficking through LAN network. Once the time of network congestion is known, the congestion control operation is made.

Variables taken in network behavior analysis are: **CD** (*amount of connected devices*), **PU** (*percentage of use*), and **RT** (*response time*) which lead to following result:

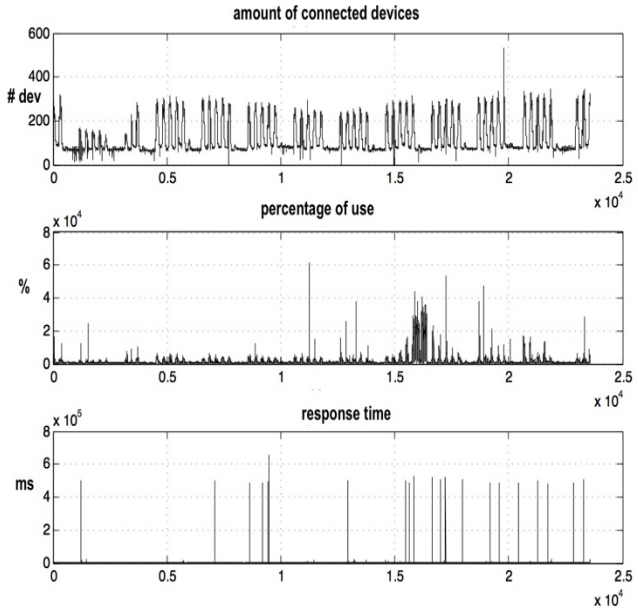


Fig7: Behavior of variables in LAN network

Fig. 7 depicts behavior of variables in LAN network, each one separately. It is worth noting that each variable is affected by the congestion phenomenon; for instance, if congestion comes up, response time begins to increase the same as percentage of use, and the connected devices can undergo complications in different services and applications. Thus, model is applied to detect congestion which is identified by increases in response time or percentage use. When the number is detected, transmission rates will be modified to alleviate the LAN network problems.

Test environment is a LAN network with hierarchical structure whose *core* layer is the most rugged, and devices are *router* of Cisco ® brand. In second layer, i.e. distribution layer, Cisco *switch catalyst* are found, with management capacity. Finally, all devices of the network users are connected to the third layer, i.e. access layer, where congestion control is made through manipulation of the operating system of the *switches* found in this layer. Fig. 8 depicts test environment distribution.

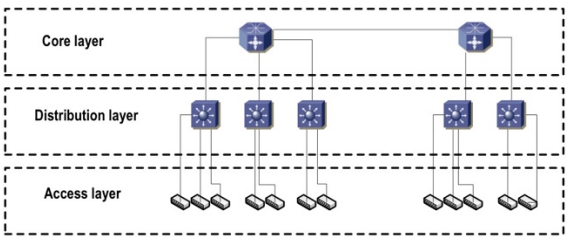


Fig8: LAN network test environment

To apply the model, the following methodology is used: simulation of one transmission control system for TCP is made (fig. 9). This model has a traffic source that feeds the mechanism of TCP; then, RED (*Random Early Detection*) module is found, where transmission rates are modified; after that, feedback module is found which allows to make the modification to the mechanism

of TCP with data being transmitted. This model has been based on previous studies [6], [20], [21]

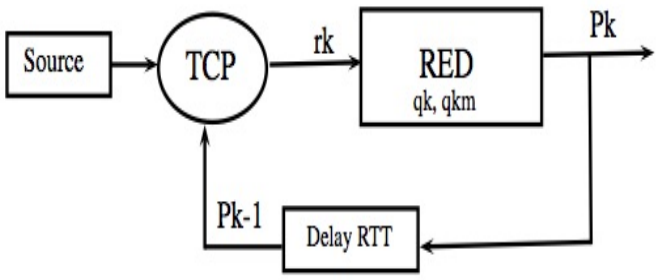


Fig9: TCP control system

TCP control system functioning in fig. 7 is described as follows:

$$r_k = \frac{MK}{RTT\sqrt{P_{k-1}}} \quad (15)$$

$$q_k = \left(\frac{nr_k}{M}RTT - \frac{C}{M}R_0, B, 0\right) \quad (16)$$

$$\bar{q}_k = (1-w)\bar{q}_{k-1} + wq_k \quad (17)$$

For previous equations we have:

- M= packet size
- RTT = *Round Trip Time* value
- p_k = likelihood of dismiss
- B = buffer size
- n = TCP number of flows
- R_0 = RTT minimum value
- C = connection capacity
- K = constant 1.25

During model execution, the modified variable is n. This number is modified by N_t value which is calculated with taken values from the LAN network trace.

Fig. 10 shows N_t series which represents taken response times to analyze one LAN subnetwork.

5. Results of the model

To examine the behavior of the proposed model in the last section, one data series is taken from LAN network. For this data series, Z_t series is created taking as reference the following values:

- θ_0 trend of the series = limit of response times or percentage of use.
- Z_k market yield = LAN network throughput in conditions of no congestion.
- $\sum_{j=k+1}^{t} a_j$ yield of one risk-free asset = throughput of channel capacity.

Taking those variables about TCP control system, we have:

$$q_k = \left(\frac{Z_t r_k}{M}RTT - \frac{C}{M}R_0, B, 0\right) \quad (19)$$

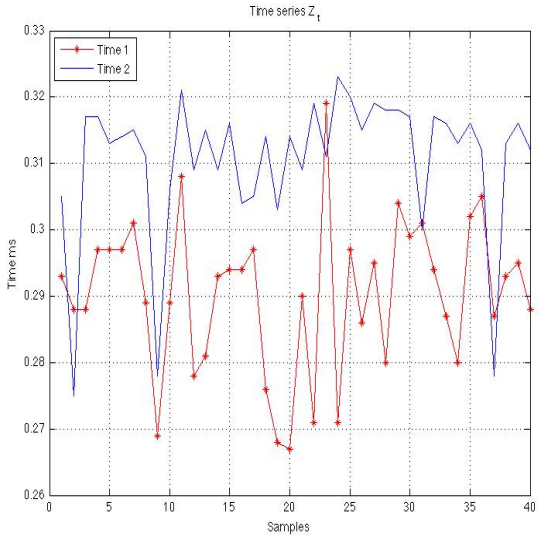


Fig10: Z_t time series

Fig. 11 shows the comparison between behavior of two types of data flows supported by TCP control system. Flow depicted by the continuous line is the normal flow in which congestion control has not been applied, and flow depicted by the dashed line is the flow with q_k modification which is the variable affected by $E(r_i)$.

Fig. 12 shows another way to look at congestion control. The *throughput* measure of network is made, where that network is saturated to observe model's behavior. In this environment, network is saturated by injecting an amount of traffic, and response times are measured. When those are risen, the transmission rate keeps a constant value of 10 Mbps. On the other hand, when network is no longer saturated, transmission rate rises above 10 Mbps because congestion control is no longer applied.

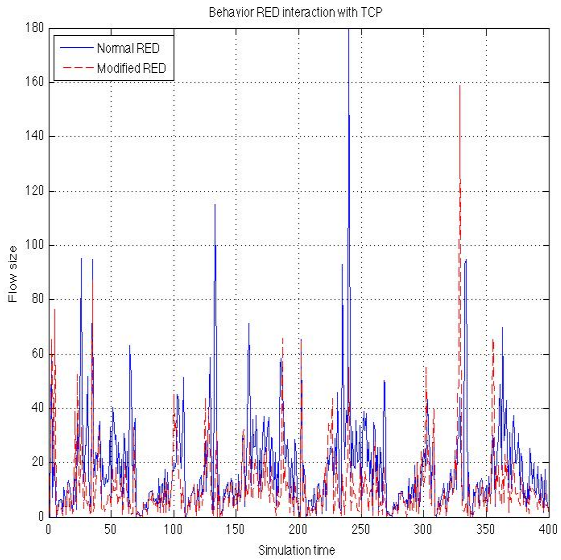


Fig11: Comparison between behavior

Fig. 12 shows a comparison between different subnetworks of the LAN network when congestion experiment is made to saturate subnetwork under observation. An interesting result is that the operation only affects the network where congestion is generated, which when translated into technical work, affects the transmission rates only in the gate where subnetwork is connected and in which the experiment is done. In this comparison, channels independency is shown, and a good functioning of the whole LAN network is guaranteed.

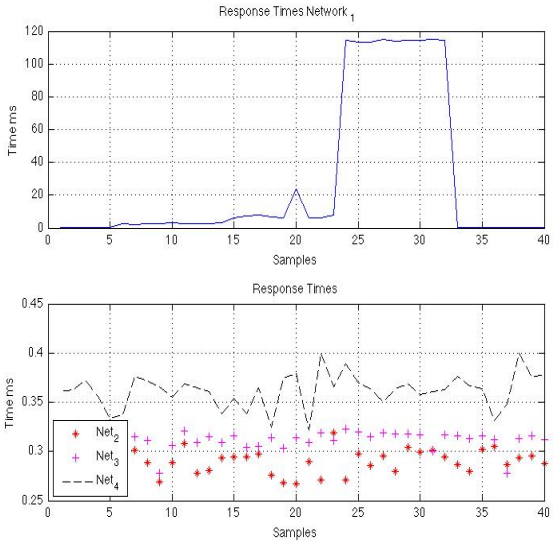


Fig12: Comparison between subnetworks

As a final part of congestion control test on test subnetwork, lost packets are of particular interest because one of the symptoms of congestion is the increase of lost packets in connection. Fig. 13 shows a test subnetwork behavior. At the beginning of the experiment, the number of lost packets rises when the channel is saturated, but then it drops substantially when congestion control is carried out in the switch gate.

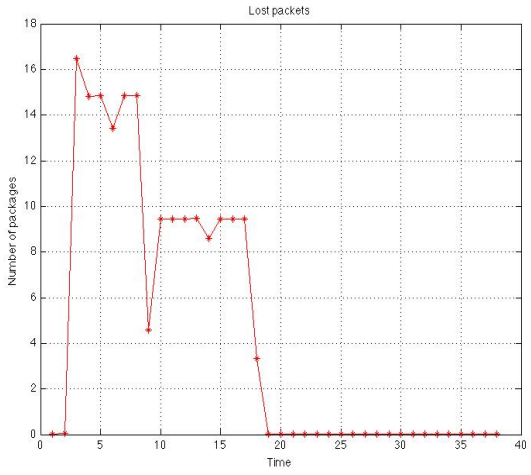


Fig13: Lost packets in a subnetwork

6. Conclusions

This paper has explained a congestion control proposal, where two central concepts are fused together. On the one hand, $ARIMA(p,d,q)$ econometric model works with time series, calculating parameters of a linear function to estimate future values. On the other, TCP conceptual model has been taken to make the congestion control, based on RED (*random early detection*). Following the analogy in which a LAN network is like a stock exchange and the stock market shares are like the network variables that are produced with circulating traffic, a model with a Z_t series is pose, which is the function that represents the estimation for one asset. In this proposal, the modified parameter is n, which is the number of traffic flows in the network. Moreover, three network LAN variables have been taken: RT (*response time*), CD (*connected devices*), and PU (*percentage of use*), and one experiment to saturate LAN network to analyze RT behavior with the congestion control model integrated into the system was carried out.

References

[1] M. M. Cuellar and J. F. Sánchez, "Congestion control LAN networks using an econometric model ARIMA," in *Telematics and Information Systems (EATIS), 2016 8th Euro American Conference on*, 2016, pp. 1–7.

[3] D. Rossi, L. Muscariello, and M. Mellia, "On the properties of TCP flow arrival process," in *Communications, 2004 IEEE International Conference on*, 2004, vol. 4, pp. 2153–2157.

[4] A. Sang and S. Li, "A predictability analysis of network traffic," *Comput. networks*, vol. 39, no. 4, pp. 329–345, 2002.

[5] C. You and K. Chandra, "Time series models for internet data traffic," in *Local Computer Networks, 1999. LCN'99. Conference on*, 1999, pp. 164–171.

[6] J. F. Sánchez and L. A. Cobo, "Theoretical model of congestion control in VANET networks," in *Communications and Computing (COLCOM), 2014 IEEE Colombian Conference on*, 2014, pp. 1–6.

[7] L. F. Niño, E. Ardila, and J. F. Sanchez, "Congestion control model for local ip networks," in *Communications and Computing (COLCOM), 2013 IEEE Colombian Conference on*, 2013, pp. 1–6.

[8] F. X. Diebold, *Elements of forecasting*. Citeseer, 1998.

[9] D. Baur and K. Glover, "A Behavioural Finance Approach with Fundamentalists and Chartists in the Gold Market," *Univ. Technol. Sydney Work. Pap.*, 2011.

[10] I. Dubova, "La validación y aplicabilidad de la teoría de portafolio en el caso colombiano," *Cuad. Adm.*, vol. 18, no. 30, 2005.

[11] J. Durbin and S. J. Koopman, *Time series analysis by state space methods*, no. 38. Oxford University Press, 2012.

[12] B.-J. Chang, S.-Y. Lin, and J.-Y. Jin, "LIAD: Adaptive bandwidth prediction based Logarithmic Increase Adaptive Decrease for TCP congestion control in heterogeneous wireless networks," *Comput. Networks*, vol. 53, no. 14, pp. 2566–2585, 2009.

[13] M. P. Papazoglou, P. Traverso, S. Dustdar, and F. Leymann, "Service-oriented computing: a research roadmap," *Int. J. Coop. Inf. Syst.*, vol. 17, no. 2, pp. 223–255, 2008.

[14] R. Barzamini, M. Shafiee, and A. Dadlani, "Adaptive generalized minimum variance congestion controller for dynamic TCP/AQM

networks," *Comput. Commun.*, vol. 35, no. 2, pp. 170–178, 2012.

[15] F. Gont and A. Yourtchenko, "On the Implementation of the TCP Urgent Mechanism," 2011.

[16] Y. Teng, H. Wang, M. Jing, and Z. Lian, "A study of improved approaches for tcp congestion control in ad hoc networks," *Procedia Eng.*, vol. 29, pp. 1270–1275, 2012.

[17] M. Scharf, "Comparison of end-to-end and network-supported fast startup congestion control schemes," *Comput. Networks*, vol. 55, no. 8, pp. 1921–1940, 2011.

[18] S.-S. Wang and H.-F. Hsiao, "TCP-friendly congestion control for the fair streaming of scalable video," *Comput. Commun.*, vol. 33, no. 14, pp. 1578–1588, 2010.

[19] G. A. Abed, M. Ismail, and K. Jumari, "Exploration and evaluation of traditional TCP congestion control techniques," *J. King Saud Univ. Inf. Sci.*, vol. 24, no. 2, pp. 145–155, 2012.

[20] J. Felix, V. Stavro, M. Aurelio, and A. Monroy, "Determinación de la predecibilidad de trazas de tráfico mediante análisis de recurrencia," 2008.

[21] M. Mejia, N. Peña, J. L. Muñoz, O. Esparza, and M. Alzate, "DECADE: Distributed Emergent Cooperation through ADaptive Evolution in mobile ad hoc networks," *Ad Hoc Networks*, vol. 10, no. 7, pp. 1379–1398, 2012.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 277-287 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

Spatial Sampling Requirements for Received Signal Level Measurements in Cellular Networks of Suburban Area

Abdlmagid Basere*, Ivica Kostanic

Department of Electrical and Computer Engineering, Florida Institute of Technology, Melbourne, FL 32901, USA

ARTICLE INFO

Article history:
Received: 30 December, 2016
Accepted: 23 January, 2017
Online: 28 January, 2017

Kriging
Coverage area estimation
Received signal level
Coverage verification in cellular systems

ABSTRACT

A process for the determination of a required spatial resolution in the collection of the Received Signal Level (RSL) is discussed. This method considers RSL measurements as a three dimensional surface that is sampled through the data collection process. In addition, it is difficult to collect RSL measurements for an entire coverage area because of many obstacles, for example, buildings, lakes, and vegetation. Thus, the estimation of the coverage area is necessary for locations for which it is hard to collect the RSL. Kriging is a well-known technique for estimation of unknown values at specific locations, and it has shown very good results. The distance factor in Kriging has never been studied before for a coverage area. A drive test is used to collect RSL measurements, and they are gathered in different distance sizes, a procedure is known as binning, to form the coverage area. Kriging is used to estimate the entire RSL surface in bin sizes that are 200x200, 100x100, 50x50, and 25x25 meters of resolution. Using a 2-D Fourier analysis, the cutoff frequency for the Fourier transform of the RSL surface is determined. The spatial sampling resolution and the cutoff frequency are linked through a Nyquist sampling criterion. The approach was experimentally verified in a typical suburban environment. The results show that the spatial resolution requirement is between 50 and 100m.

1. Introduction

The cell coverage area is one of the most essential factors in cellular communication system designs [1, 2]. It can be defined as the greatest distance that a cell-phone can be away from the base-station while still receiving a reasonable service [3]. The base station is normally located near the center of the required served area [4].

In cellular networks, the coverage area of a cell is based on the signal strength of its pilot channel. For instance, in the case of a GSM cell, the strength of its Broadcast Control Channel (BCCH) needs to be over the suitable coverage threshold [5]. Also, in W-CDMA, the coverage depends on the signal strength of the Common Pilot Channel (CPICH) [6]. However, in LTE the convenient metric is the Reference Sequence Received Power (RSRP) [7]. While the system is deployed, the coverage area of every cell is tested through a procedure usually referred to as drive testing. This procedure includes measurement of the RSL of the pilot channel within the geographic area that surrounds the cell. Normally, the measurements are conducted using appropriate drive test receivers. These receivers pair measurements of RSL of the pilot channel with the geolocation information. The coverage area of the cell is estimated based on these measurements.

There are two principal practical issues with the RSL measurements. The first issue is associated with signal fading in the wireless environment. It is well understood that as a result of multipath propagation, the measured signal experiences small scale fading. Therefore, estimates of the RSL need to be obtained through some form of statistical averaging of the instantaneous signal readings. The process followed in the averaging of the instantaneous measurements needs to be compliant with a requirement frequently referred to as the Lee criterion. The requirement is described in [8, 9].

The second issue involves coverage area estimation. There are two aspects regarding this issue. First, in every practical scenario, the data may be collected only from a small portion of the area that is accessible to drive testing. Therefore, the coverage estimate for the entire area of interest is obtained through the process of measurement interpolation. An effective way for interpolation that is based on the Kriging method is presented in [1]. However, to make a valid interpolation, the measured data need to be sampled at a sufficient spatial resolution. To the authors' knowledge, the issue of spatial sampling of the RSL has not received sufficient treatment in the available literature, and it is the subject of the work presented in this paper.

*Corresponding Author: Abdlmagid Basere, Department of Electrical and Computer Engineering, Florida Institute of Technology, Melbourne, FL 32901, USA Email: abasere2011@my.fit.edu

1.1. Cell Coverage Area

A cellular system network consists of cells that are linked together to offer radio coverage over a wide geographical area. The cell phone coverage area lets a large number of users communicate with each other, whereas base stations in the cellular network are stationary. These base stations provide a connection to transceivers regardless of their movement, whether they are stationary or moving across cells. The range of cell coverage is based on natural factors, for example, human factors, such as the landscape (urban, suburban, rural), subscriber behavior, and propagation conditions, etc. [10].

1.2. Importance of Understanding Cell Coverage Area

1.2.1 Coverage Planning

The coverage planning process can be divided into three phases. The first phase is named the preplanning phase. The general properties of the future network are examined. In the second phase, a site survey is used to examine the targeted area and an investigation of potential base stations' locations is achieved. Following the initial survey, constant modification is necessary to move the network planning forward. In the last phase, a driving test to collect data is done repeatedly in order to test the coverage area until achieving good coverage. Then, the system is ready to be deployed in the target area to provide service [11].

1.2.2 Concept of Handoff

Handoff is a procedure that allows mobile users to move from one cell to another without interruption of service. The handoff process performance depends on a variety of conditions such as signal strength and interference levels [12]. Having an accurate picture of the coverage in a given region is necessary condition for reliable handoffs.

1.2.3 Interference

Interference results when neighboring cells simultaneously work on the same frequency. One of the main weaknesses that can limit the implementation of a cellular system in a multi-cell network is inter-cell interference. Interference can cause background noise and bad voice quality when a call is in operation. As a result, the determination of coverage area is beneficial in reducing the interference, which leads to higher quality of service [13].

1.2.4 User Location

The mobile telecommunication network consists of several base stations (BSs) that provide mobile telecommunication service to mobile users. Each BS represents a cell that is offering coverage to a specific area, and each cell is divided into sectors. Understanding the coverage of the sectors assists in locating the users that are within the systems of the service area [14].

1.3. Methods for Estimating Cell Coverage

There are many methods of performing the estimation of cellular coverage areas, and each of them has its own advantages and disadvantages, but none of them is 100% accurate [15]. Three implementation estimation methods are explained as follows.

1.3.1 Geometry

The algorithms that use Voronoi diagrams to estimate coverage areas do not take into consideration the features of the propagation environment, and they oversimplify the coverage problem [16]. For this reason, among others, Voronoi diagrams are not being used in this work.

1.3.2 Prediction

Coverage prediction and propagation modeling are among the most fundamental tasks in cellular system design. Over many years a diverse of methods have been used for coverage area prediction using propagation models. These models are useful in predicting path loss or signal attenuation, so they thus enable appropriate reception [17]. Cellular designers generally use advanced planning tools for estimating the network coverage. The estimation of coverage area is dependent on terrain data combined with a propagation model [18].

1.3.3 Measurements

A drive test is usually used to collect the RSL measurements. Performing driving tests is expensive and time consuming, so there have been significant attempts to improve the quality of coverage estimation for locations for which the values are unknown by using values that are obtained from a limited number of driving test measurements. The difficulty with drive tests is that data cannot be collected for the entire region of the network because of many barriers [18] such as lakes, vegetation, and buildings. Therefore, the estimation of the coverage area is significant and must rely on spatial prediction methods for instance, Inverse Artificial Neural Networks, Distance Weighting, and Kriging.

2. Problem Statement

The problem of spatial sampling of the RSL is demonstrated with the assistance of Figure 1. As shown in the figure, the RSL for a specified transmitter site might be seen as a three dimensional surface. During the procedure of data collection, one makes measurements of the "surface height" at different locations in space – i.e. different (x, y) coordinates. The job of interpolation is to reconstruct the surface on the basis of the collected measurements.

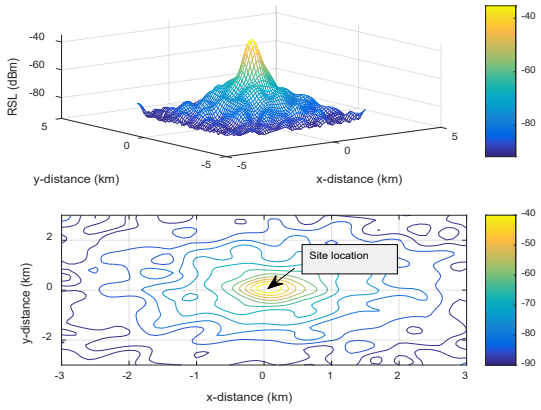


Figure 1: An illustration of the RSL surface

It is clear that the precision of interpolation is basically based on the spatial sampling "frequency" of the measurements. As the number of measurement locations becomes larger, the accuracy of the estimate for the RSL surface becomes higher. Yet, as with any sampling process, the point of diminishing returns is reached.

The objective of the study in this paper is to calculate the required spatial sampling rate for accurate representation of the RSL surface. Shannon's sampling theorem is extended to a two dimensional signal representing the RSL surface in this analysis.

3. Literature Review

3.1. Data Collection

To verify the coverage area of the cellular system it is required that the RSL measurements be conducted by doing spatial averaging of immediate signal power measurements. The data is normally gathered in accordance with Lee sampling criteria. The amplitude of the received signal envelope in the mobile propagation environment is described in equation (1) [9].

$$r(x) = m(x) r_0(x) \tag{1}$$

It is sensible to indicate the received signal by a stationary stochastic process, where $r_0(x)$ denotes the variation of fast fading of a mean signal strength value equal to 1. The mean value of the signal envelope is $m(x)$, and x is the location of the local mean [8].

Getting an estimation that is valid statistically for the local mean, there are two questions that need to be answered. First, how many 'snapshots' of the RSL are required to be conducted over an appropriate distance to make certain the statistical validity of the estimate. Second, what would be a suitable spatial distance to average over. Averaging the space over a distance has to be achieved to guarantee averaging Rayleigh fading effects. Estimation of the local mean can be considered as an average of the power of the signal over a suitable distance (2L) as displayed in equation (2) [19]:

$$p(x) = \frac{1}{2L}\int_{x-L}^{x+L} r(y)^2\, dy = \int_{x-L}^{x+L} m(y)^2 r_0(y)^2\, dy \tag{2}$$

keeping $2L$ small enough so that $m(y) \approx m_x$ over $[x-L, x+L]$

The variance of the power estimate is given as [8].

$$\hat{p}(x) = \frac{m_x^2}{2L}\int_{x-L}^{x+L} r_0^2(y)\, dy \approx m_x^2 \times \langle r_0^2(x) \rangle = \frac{4m_x^2}{\pi} = p_{av} \tag{3}$$

$$\sigma_{\hat{p}}^2 = \frac{1}{2L}\int_0^{2L}\left(1-\frac{x}{2L}\right)\left(m_x^4 R_{r_0^2}(x) - 16\frac{m^4}{\pi^4}\right)dx \tag{4}$$

where $R_{r_0^2}(x)$ is the autocorrelation function of r_0^2 given by

$$R_{r_0^2}(x) = \frac{16}{\pi^2}\left[1 + J_0^2\left(\frac{2\pi x}{\lambda}\right)\right] \tag{5}$$

Equation (4) has been statistically evaluated [2] for different values of L, and the results for 1σ and 2σ spreads are shown in Table 1 [19].

Table 1. Spread of the local mean estimate as a function of the averaging distance

$2L$	$\sigma_{\hat{p}}$	1σ Spread $\log\left(\frac{p_{av}+\sigma_p}{p_{av}-\sigma_p}\right)$	2σ Spread $\log\left(\frac{p_{av}+2\sigma_p}{p_{av}-2\sigma_p}\right)$
5λ	$0.33p_{av}$	2.98 dB	6.5 dB
10λ	$0.24p_{av}$	2.14 dB	3.5 dB
20λ	$0.18p_{av}$	1.55 dB	3.24 dB
40λ	$0.12p_{av}$	1 dB	2.1 dB

Under most practical conditions, terrain variations (shadowing) over a 40λ distance can be presumed to be insignificant and can be considered as negligible. Therefore, the claim that the mean value of the signal is not variable during the integration in equation (3) is valid [19].

"Macroscopic propagation models predict the mean received signal level over a small geographical area called a bin" [19]. The size of the bin is computed as a function of accuracy, computation time, roughness of the terrain, and database terrain resolution, etc. Typically the size ranges from 50m to 500m. The local area mean (LAM) is the mean of the received signal of a bin. Playing averaging of several local means may allow comparison between measurement data with predictions of the propagation model [19].

While driving through a specific bin, the measurement equipment performs the collection of the local means within the bin. The local means tend to comply with a normal distribution in the logarithmic domain with a standard deviation σ and mean value m. The local area mean for the particular bin is calculated as an average value of the local means collected in that specific bin.

$$L_{AM} = \frac{\sum_{k=1}^{N} L_{Mk}}{N} \tag{6}$$

where L_{AM} is the local area mean, L_{Mk} represents the k^{th} local mean and N is the total number of local means collected in the bin [19].

3.2. Spatial Binning

Spatial binning is grouping of measurements that are close to each other. The binning results in a smaller geographically referenced data set, where each point represents an average obtained over a bin area [20].

Incomplete drive test data can cause imprecise estimation of coverage area. Yet, collection of excessive data is time consuming and costly. Grouping measurements using averaging can help reduce the processing load and quantify data to the resolution of a specific terrain which constitutes the bin of the area [19].

3.3. Interpolation Methods

Through interpolation, one-use measurement data samples can be collected and used to estimate measurement data at other points where measurement data are not available. Included in the most common spatial interpolation methods are spline interpolation,

interpolating polynomials, Inverse Distance Weighting (IDW), and Kriging [21]. Amongst those techniques, Ordinary Kriging (OK) and IDW are the most frequently used, and are normally the most commonly suggested interpolation methods [22].

"Kriging is an interpolation technique based on the methods of geostatistics" [23]. Krige in 1951 collected samples over mining fields. The method was used as a optimal interpolation techniques to use in the mining industry. Currently, the Kriging method is being used in groundwater modeling, soil mapping [24, 25], and in several other spatial problems [24].

4. Proposed Application

4.1. Kriging Method

The principle of interpolation equation in the Kriging model is given by [23].

$$\hat{z}(x_0)=\sum_{i=1}^{n}\lambda_i z(x_i) \tag{7}$$

where $z(x_i)$ is the known value at the i^{th} location, λ_i is an unknown weight for the known value at the i^{th} location, and $\hat{z}(x_0)$ is the estimation value.

As can be seen, at the point where the algorithm performs the interpolation, the value is obtained as a weighted average from the points in the immediate neighborhood. The most important mission linked with the implementation of (7) is the calculation of the appropriate weights. The fitted variogram model is used to determine the weights λ_i that are essential for local estimation. The interpolation $\hat{z}(x_0)$ may be unbiased by choosing the weights λ_i. Interpolating many locations may give some values that can be above and below the real values. Therefore, the summation of weights λ_i must be equal to 1 in order to ensure that the estimation of unknown values is unbiased [24].

$$\sum_{i=1}^{n}\lambda_i=1 \tag{8}$$

4.1.1 Semivariogram

"The variogram characterizes the spatial continuity or roughness of a data set. Usually, one dimensional statistics for two data sets might be equal" [23]. Yet, the spatial continuity could be unrelated. The difference between the variogram and the semivariogram is just a factor of 2. The semivariogram equation (9) is described in [24].

$$2\gamma(\mathbf{h})=E\left[Z(\mathbf{x})-Z(\mathbf{x}+\mathbf{h})\right]^2 \tag{9}$$

Equation (9) represents the variogram $2\gamma(\mathbf{h})$ of variations between sites that are functions of distance $\mathbf{h}$ between them, where $Z(\mathbf{x})$ and $Z(\mathbf{x}+\mathbf{h})$ are the values of the random variable Z of interest at locations $(\mathbf{x})$ and $(\mathbf{x}+\mathbf{h})$. The function of interest is $\gamma(\mathbf{h})$ that is called a semi-variogram and is used in Kriging. The statistical expectation operator is $E[\ ]$. Kriging used the empirical variogram $\gamma(\mathbf{h})$ as a theoretic variogram to get the first estimate of the variogram for spatial location [24, 26].

$$\hat{\gamma}(h)=\frac{1}{2n}\sum_{i=1}^{n}\{z(x_i)-z(x_i+h)\}^2 \tag{10}$$

where $Z(x)$ is the value at location x_i, $Z(x_i+h)$ is the value at location x_i+h, and n is the number of pairs of observations of attribute z separated by distance h; the plot of $\hat{\gamma}(h)$ against h is known as the experimental semivariogram. The experimental semivariogram offers beneficial information for optimizing sampling, interpolation, and determining spatial patterns. It is obvious that the number of semi-variogram values is proportional to the pairs of locations. In addition, when datasets are large, the number of pairs of locations will increase quickly. As a result, the binning of the distance is necessary to obtain the variogram parameters, which are the range, the nugget, and the sill [24]. Moreover, in the case of RSL interpolation, the binning of the data is important so that the results of fast fading can be reduced.

4.1.2 Fitting Variogram Models

The fitting model is determined by plotting the empirical semivariance $\gamma(\mathbf{h})$ versus the separation distance of the pairs of locations (lag distance) [27], so the relationship between them provides specific parameters, namely the nugget, the range, and the sill, as shown in Figure 2.

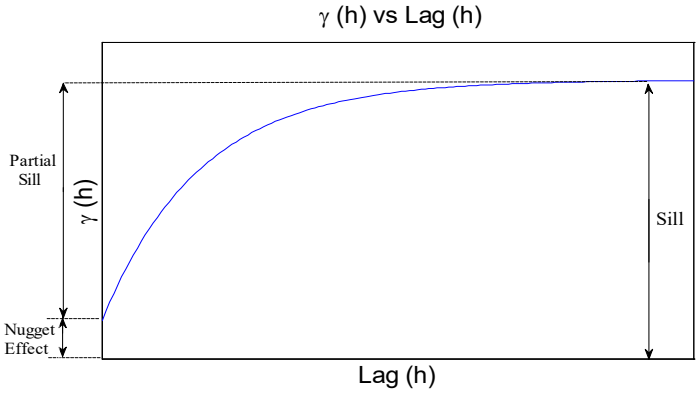
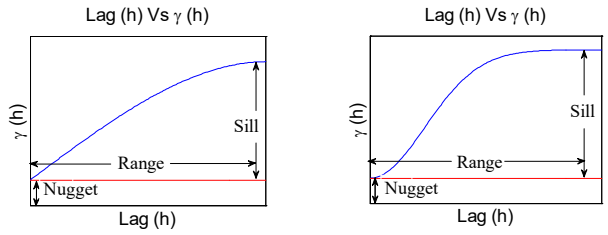
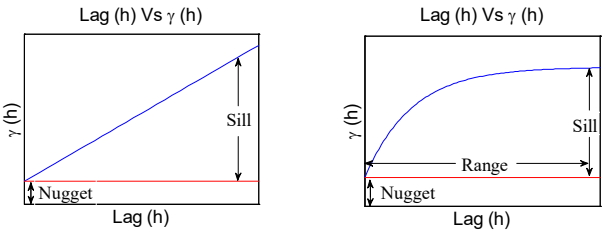


Figure 2: Simple transitional variogram with range, nugget, and sill

A variogram model is a parametric curve fitted to a variogram estimator. There are several semivariogram fitting models that are used in the Kriging method to help estimate data precisely, and the most commonly used semivariogram models are summarized in Figure 3 [24]:



Spherical variogram model Gaussian variogram model



Linear variogram model Exponential variogram model

Figure 3: Examples of most commonly used variogram models

The remainder of this paper is organized as follows: Section V gives an overview of experimental verification, which is divided two parts. In the first part the measurement setup, data collection, and results of interpolation are presented, and the second part has a discussion of the Fourier Spectrum of the RSL surface, and the verification of the approach for determining required sampling distance. In Section VI, a brief summary and conclusions are provided.

5. Experimental Verification

The collection of RSL measurements was performed in an area that is located in Melbourne, FL. The map of the area is presented in Figure 4. Although the study is conducted on a specific RF frequency, the methodology should be widely applicable across all UHF/VHF frequencies used for personal communication systems. The Melbourne area is a suburban area. For the study, the transmitter is located on the rooftop of a multi-story building, making for good wide area coverage. The majority of houses in the targeted area are one to two floors, and their heights range from 4 to 9 meters. Also, there are open areas such as small artificial lakes, parks, and vegetation. The satellite view of the study area is shown in Figure 4.

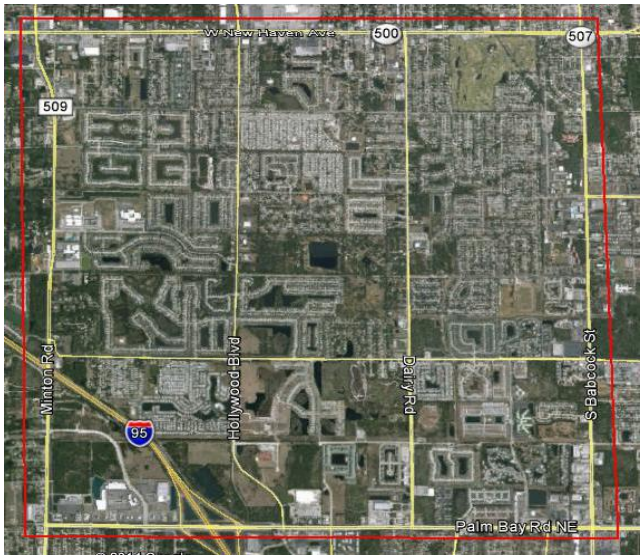


Figure 4: Satellite view of the studied environment

5.1. Experimental Verification For RSL Interpolation

5.1.1 Data Collection

An illustration of the data collection system is presented in Figure 5. The system consists of the following parts:

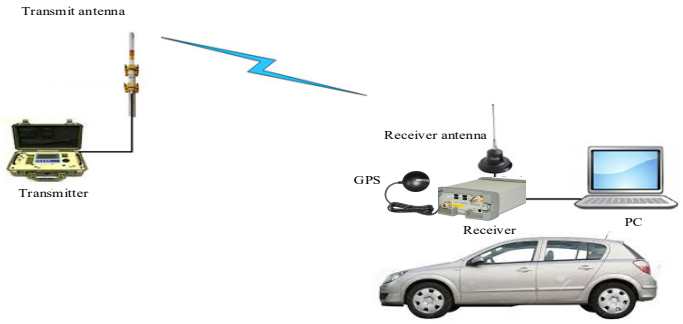


Figure 5: Illustration of the measurement system

- Transmitter: The transmitter is used to evaluate PCS band signal propagation, and the frequency range is 1.85 to 2.1 GHz, where the power is up to 43 dBm. This transmitter allows the transmission of the signal at the radio frequency of 1925MHz.
- Transmit Antenna: The transmitter antenna is used in the transmission of signals to be measured. It is an omnidirectional antenna, and its frequency ranges between 1850 MHz and 1990 MHz. It also radiates in a horizontal plane with 6 dBi gain.
- Receiver: An Agilent Technologies E6474A receiver is used to conduct RSL measurements at different spatial locations.
- Receiver Antenna: The receiver antenna is used to collect RSL measurements. It is an omnidirectional antenna with a frequency range between 1850 and 1990 MHz.
- GPS Antenna: The GlobalSat BU-353-S4 is a USB magnet mount GPS receiver. Its characteristics are low power consumption, an ultra-compact chipset, and extreme sensitivity. It is appropriate for marine navigation, mobile phone navigation, automotive navigation, and personal positioning. It can be connected with a PC in order to allow JDSU Wireless Solutions software to recognize and store the location of measured data in terms of longitude, latitude, and altitude. The recorded position enables the user to get the best interpolation of the coverage area.

5.1.2 Data Collection and Binning

A drive test was undertaken in the selected geographic area, where RSL measurements were collected and stored. The RSL measurements are collected to perform estimation for gaps as shown in Figure 6.

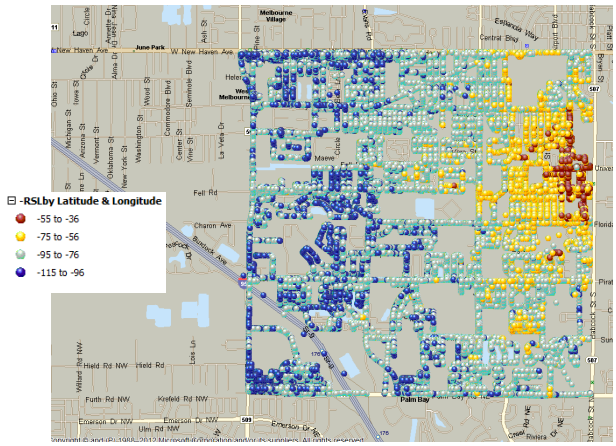


Figure 6: View of the studied environment and measured data

The RSL measurements are used to produce a coverage area map for the targeted area. Every spatial location can be defined as a point of intersection of both the x-axis and the y-axis, where the x-axis represents the longitude and the y-axis represents the latitude. A Matlab code is used to perform the binning. Since the data set is large, the binning is done for the coverage area as shown in Figures 7-10 with different resolutions. The binning is based on a separation of adjacent bins, such as 25x25, 50x50, 100x100, and 200x200 meters.

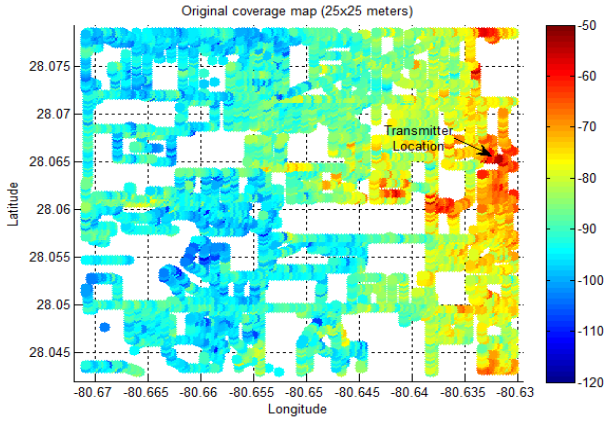


Figure 7: Original coverage area of 25x25 meters resolution

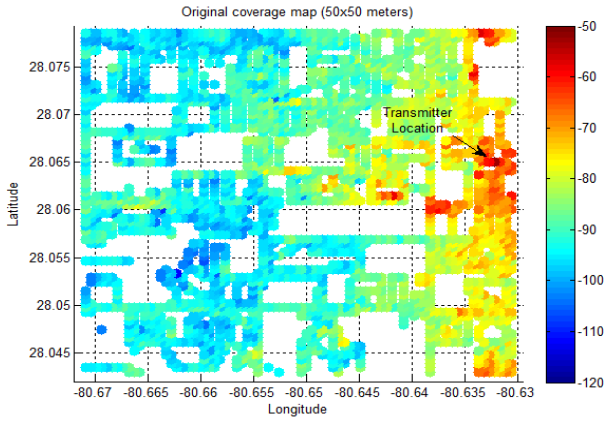


Figure 8: Original coverage area of 50x50 meters resolution

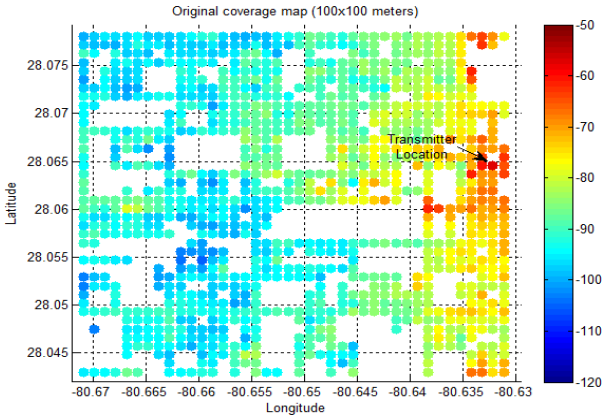


Figure 9: Original coverage area of 100x100 meters resolution

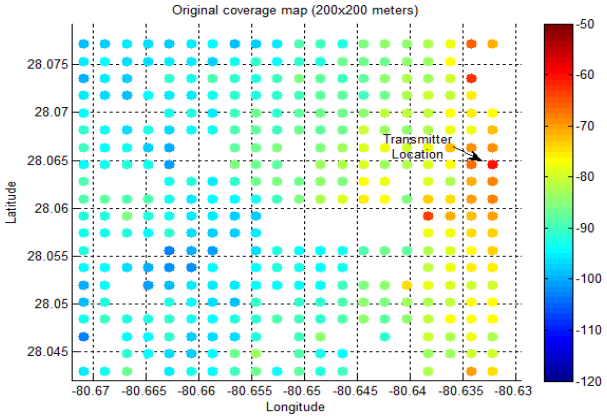


Figure 10: Original coverage area of 200x200 meters resolution

The semivariogram of the fitting model can be determined by plotting the empirical semivariogram versus the distance as shown in Figure 11. It shows that as the dataset becomes large, the number of observations will increase quickly.

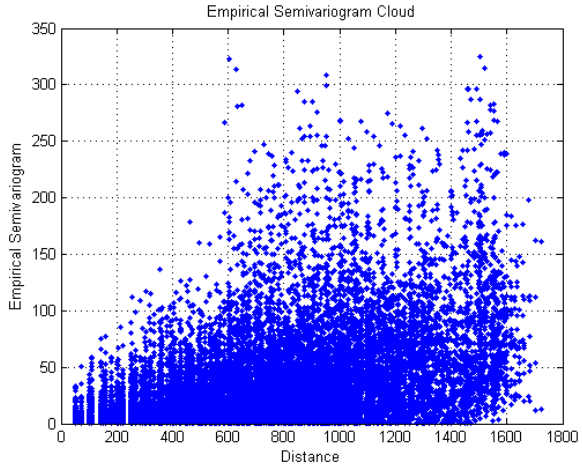


Figure 11: Empirical semivariogram cloud

The difficulty in the computation of the semivariogram fitting model increases as the size of the dataset increases. As a result,

the points of the variogram cloud can be grouped into classes of distances ("bins") to get the variogram fitting model for the area of interest. The RSL measurements demonstrate that the variogram fitting model is linear as shown in Figure 12.

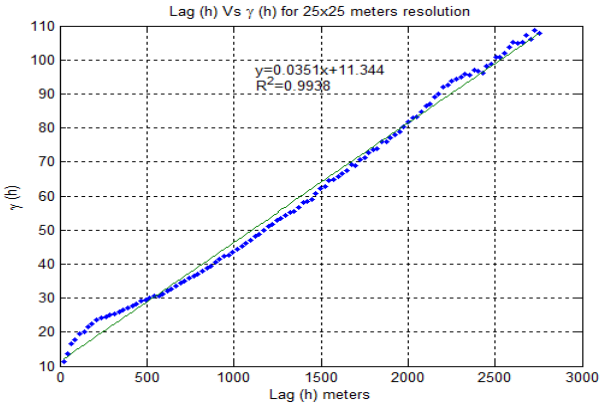


Figure 12: Linear variogram model of 25x25 meters resolution

5.1.3 Data Collection and Binning

The interpolation result of the targeted coverage area for grid sizes of 25x25. 50x50, 100x100, and 200x200 meters resolution are shown in Figures 13-16. The interpolation of the coverage area with 25x25 meters resolution is smooth, and it gives more details about the RSL as shown in Figure 13.

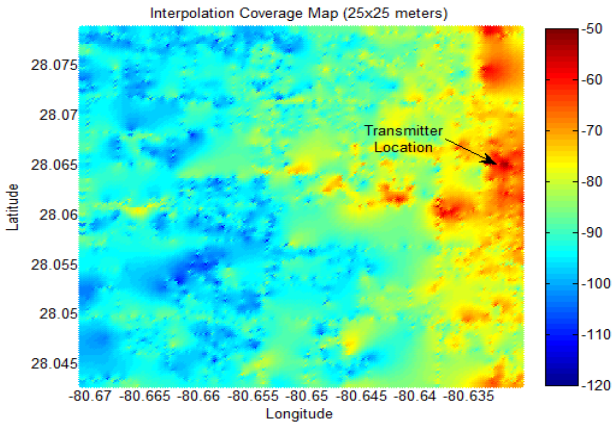


Figure 13: Interpolated coverage area of 25x25 meters resolution

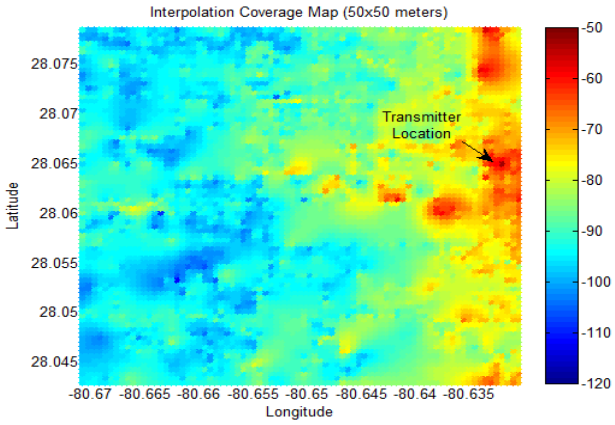


Figure 14: Interpolated coverage area of 50x50 meters resolution

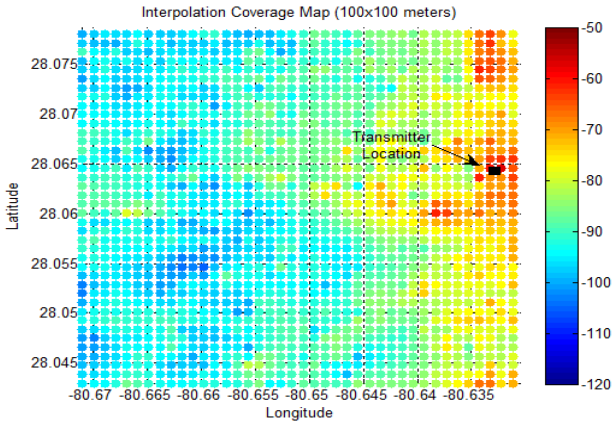


Figure 15: Interpolated coverage area of 100x100 meters resolution

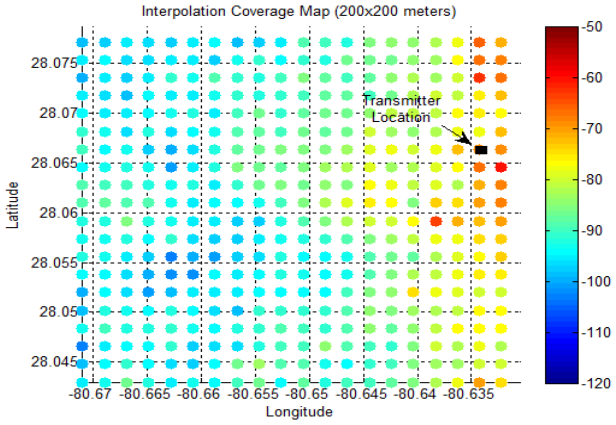


Figure 16: Interpolated coverage area of 200x200 meters resolution

5.2. Experimental Verification For Determining Required Sampling Distance

5.2.1 Fourier Spectrum of The RSL Surface

The two dimensional Fourier analysis is used, with the 2-D spectrum of the RSL surface represented as:

$$F(u,v) = \int_{x=-\infty}^{+\infty} \int_{y=-\infty}^{+\infty} f(x,y) e^{-j2\pi(ux+vy)} dx\,dy \qquad (11)$$

In (11), $f(x,y)$ denotes the RSL at the location (x, y), u and v are spatial domain frequencies, and the transform $F(u,v)$ denotes the 2-D spectrum linked with the "signal" $f(x,y)$.

In the process of the measured data collection, the continuous RSL surface is sampled. Therefore, its 2D spectrum needs to be estimated through its samples using the discrete Fourier transform as:

$$F(u,v) = \sum_{m=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} f[m,n] e^{-j2\pi(um+vn)} \qquad (12)$$

In (12), $f[m,n]$ shows the RSL measurement conducted at the location (mx_0, my_0) and, x_0 and y_0 are sampling distance increments in x and y direction from the transmitting site.

As in any practical application, the RSL surface spans only the finite portion of the x-y plane; and the summation in (12) is performed over a finite amount of samples. This leads to a 2-D spectrum given by:

$$F(u,v) = \sum_{m=0}^{M_0-1} \sum_{n=0}^{N_0-1} f[m,n] e^{-j2\pi(um+vn)} \quad (13)$$

The 2-D spectrum in (13) captures all the frequency components participating in the formulation of build up of the "signal" $f(x,y)$. The spatial sampling results in the periodic extension of the signal's 2-D spectrum. The spatial frequency of the sampling process needs to be sufficiently high so that the aliasing between the spectrum replicas is negligible.

As a further clarification, consider the 2-D magnitude spectrum, i.e. $f(x,y)$, associated with the RSL surface shown in Figure 1, and presented in Figure 17. The RSL is sampled over a 6km by 6km area with the sampling distances $x_0 = 100$ m and $f(x,y)$ m. The spectrum in Figure 17 is generated after the DC component of $f(x,y)$ is removed. The plot in Figure 2 reveals that the RSL surface is a low pass 2-D signal and that one may establish the highest frequency in its 2-D spectrum - $F\max$. The contribution of the spectral terms that are above the highest frequency of the overall energy of $f(x,y)$ becomes negligible. Hence, based on the well-known Nyquist sampling criterion, the sampling frequency needs to be at least two times $F\max$, and the maximum sampling distance in the x-y plane is given as:

$$d_{\max} = \frac{1}{2F_{\max}} \quad (14)$$

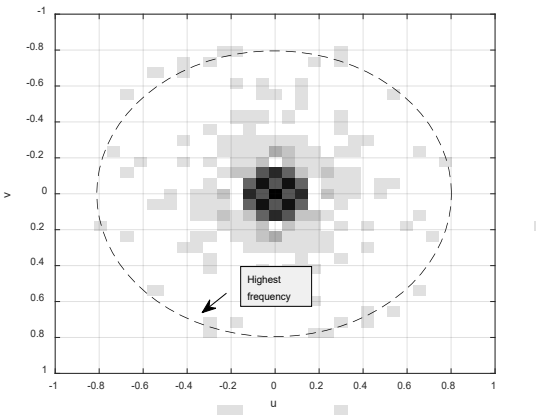


Figure 17: 2D-Spectrum associated with the RSL surface given in Figure 1

Therefore, to determine the furthest sampling distance required for sampling of the RSL surface, one may follow the following procedure:

- Perform sampling of $f(x,y)$ at the smallest possible distance increments x_0 and y_0 that are guaranteed to satisfy the requirement given in (14).
- Use a low pass filter to determine the largest significant frequency in the spectrum of the sampled signal, $F\max$. Once the largest significant frequency is determined, the maximum sampling distance in the spatial domain is given by (14).

Following the above outlined procedure faces several obstacles. Firstly, the value of $F\max$ is directly related to the variability of the RSL caused by the log normal shadowing. The log normal shadowing is environmentally dependent and therefore $F\max$ becomes environmentally dependent as well. Therefore, the initial sampling needs to be done at a fairly fine grid that guarantees that the criterion in (14) is satisfied. Secondly, the measurement can never be collected over the entire area of interest. Therefore one needs to rely on the interpolation to generate a best estimate of the entire RSL surface. This topic is discussed in [1]. Finally, one needs to establish a criterion for determination of $F\max$. By eliminating all the spectrum components outside of the circle in Figure 17, a fraction of the $f(x,y)$ energy is eliminated as well. As a result of this elimination, a filtered version of the 2-D spectrum is obtained as:

$$\hat{F}(u,v) = F(u,v) \cdot \Pi\left(\frac{\sqrt{u^2+v^2}}{F_{\max}}\right) \quad (15)$$

where

$$\Pi(x) = \begin{cases} 1 & , x \le 1 \\ 0 & , x > 1 \end{cases} \quad (16)$$

The RSL surface that corresponds to the filtered version of the 2-D spectrum may be obtained through the inverse Fourier transform of (15). In other words:

$$\hat{f}[m,n] = \hat{f}(mx_0, ny_0) = \frac{1}{N_0 M_0} \sum_{m=0}^{M_0-1} \sum_{n=0}^{N_0-1} \hat{F}(u,v) e^{j2\pi(um+vn)} \quad (17)$$

A natural way of quantifying the difference between $f[m,n]$ and $\hat{f}[m,n]$ are the average Root Mean Square Error (RMSE), and the corresponding Signal to Noise Ratio (SNR) which are defined by:

$$RMSE = \sqrt{\frac{1}{M_0 N_0} \sum_{m=0}^{M_0-1} \sum_{n=0}^{N_0-1} \left[f[m,n] - \hat{f}[m,n] \right]^2} \text{ in dB} \quad (18)$$

$$SNR = 10 \log \left(\frac{\sum_{m=0}^{M_0-1} \sum_{n=0}^{N_0-1} \left[f[m,n] - \bar{f} \right]^2}{\sum_{m=0}^{M_0-1} \sum_{n=0}^{N_0-1} \left[f[m,n] - \hat{f}[m,n] \right]^2} \right) \text{ in dB} \quad (19)$$

where $\bar{f}$ is the average RSL over the complete area of attention. The SNR in (19) based on $F\max$. As $F\max$ decreases, additional high frequency components from the 2-D spectrum are removed. Therefore, the RMSE in (18) rises and thus the SNR in (19) decreases. Based on the study reported in [28], the RSL measurements are repeatable within 3dB. In this study, the difference between $f[m,n]$ and $\hat{f}[m,n]$ that is a result of filtering in (15) is considered as a random variable. The highest significant frequency $F\max$ is computed so that the absolute value of this difference is below 3dB in more than 95% of points.

5.2.2 Verification of The Approach for Determining Required Sampling Distance

The data are collected with a spatial resolution of 25m. In other words, over the drive test route, a measurement is stored each 25m. It is supposed that this resolution is higher than what would be required by the condition in (14). Moreover, as it may be seen, the data are collected over all locations available for drive testing – i.e., every road in the area is tested.

- Baseline Received Signal Level Map: From the measured data shown in Figure 6, a baseline of RSL surface is created. The baseline surface has a bin resolution of 25 by 25 m. The interpolation procedure depends on the Kriging technique explained in [1]. The baseline map generated through the interpolation is shown in Figure 18. It is considered as an accurate representation of the RSL surface.

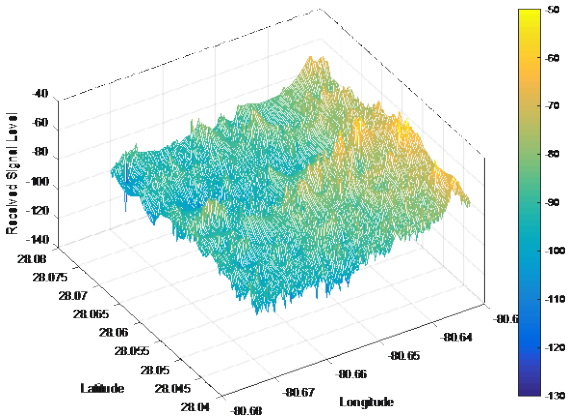


Figure 18: Baseline RSL surface obtained from data in Figure 3 through Kriging interpolation

The algorithm shown in Figure 19 is applied to determine the maximum frequency in the spectrum of the RSL surface. The interpolated RSL surface is transferred into the frequency domain through a 2D Fourier transform. The Fourier spectrum of the surface is low-pass filtered as shown in Figure 17. A filtered version of the spectrum is brought back into the spatial domain through the inverse 2D Fourier transform, and the performance metrics as defined in (18) and (19) are generated. The results of the process are summarized in Table 2. The table shows the following columns:

- $\xi = \frac{F_{\max}}{f_s/2}$ represents the normalized low pass cut off frequency. The initial sampling of the RSL surface through the drive test process is performed at a resolution of 25m. Therefore, the sampling distance $d_s = 25$ m and the sampling frequency is $f_s = 0.04$ m^{-1}.
- The average RMSE is calculated on the basis of (18).
- 95% of RMSE – Difference between filtered RSL surface $f[m,n]$ and the baseline RSL surface $\hat{f}[m,n]$ at a given point (m,n) may be treated as a random variable. The value reported in this column is the 95% percentile of the estimated cumulative distribution. In other words, 95% of the points exhibit absolute difference smaller that the value reported in this column.
- The average SNR is calculated using (19).

d_F sampling distance is calculated as $d_F = d_s / \xi$.

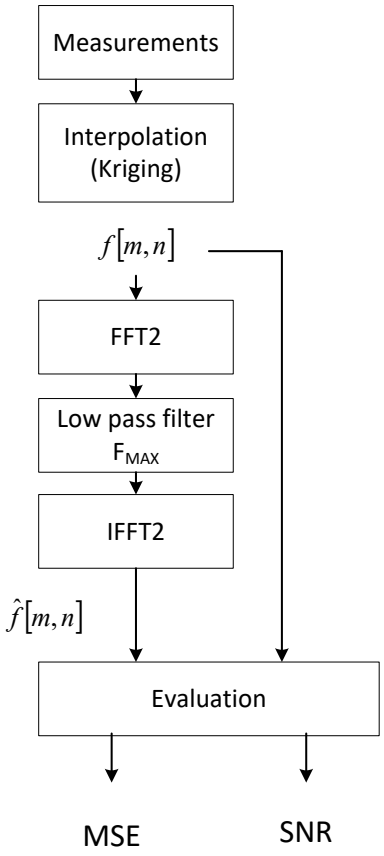


Figure 19: Algorithm for determination of the maximum frequency in the RSL surface spectrum *F max*

As expected, a larger $F\max$ sampling frequency results in a smaller MSE and a larger SNR. At the same time, a larger results in a more frequent sampling of the RSL surface in the spatial domain. For example, the value of $\xi = 0.4$ yields a filtering frequency of $F_{\max} = 0.008$ which corresponds to the sampling distance of 83 meters. This sampling distance introduces an average RMSE of roughly 1.68dB and the 95% value of 2.43dB. The SNR of the approximation is on the order of 13.93dB.

Table 2. Values of MSE and SNR as a function of the filtering frequency

$\xi = \frac{F_{max}}{f_s/2}$	Average RMSE (dB)	95% of RMSE (dB)	SNR (dB)	d_F (m)
0.2	2.42	3.75	10.52	125
0.3	1.97	2.92	12.45	83
0.4	1.68	2.43	13.93	63
0.5	1.47	2.16	15.10	50
0.6	1.30	2.00	16.22	42
0.7	1.14	178	17.34	36
0.8	0.98	1.58	18.65	31
0.9	0.80	1.27	20.43	28

Based on the values in Table 2, and the study in [28], the choice of $\xi \sim 0.3$ seems to be appropriate. This choice corresponds to the sampling distance of 83m and the 95% value of 2.92dB, which is within the bounds of the RSL measurement repeatability.

6. Summary and Conclusion

The RSL measurements were collected in a suburban area. The interpolation of the coverage area for a targeted area was done successfully. Measurements demonstrate that the semivariogram fitting model is linear. The use of Kriging confirms the interpolation to be very efficient. It provides a good result when the coverage area is divided into different bin sizes that are 25x25, 50x50, 100x100, and 200x200 meters of resolution. Furthermore, the importance of exact RSL measurements for the binning process is shown. The approach seems to be promising and is expected to be worth pursuing at a larger scale.

In this study, the Shannon sampling theory was used to determine the appropriate spatial sampling distance in the RSL measurements. It is explained that the RSL surface can be treated as a two-dimensional low pass signal for which a maximum cutoff frequency $F\max$ can be recognized. By means of this experimental work, a cutoff frequency was determined for a typical suburban environment in the US. The value of $F\max$ in the measured environment was 0.006 1/m and the value for the corresponding sampling distance is 83m.

In general, the corresponding sampling distance and the cutoff frequency are functions of the propagation environment. It is expected that environments with higher variability of the path loss, i.e., with more pronounced lognormal shadowing, have a higher value of $F\max$ Thus, the sampling distance in such environments needs to be smaller than the one obtained in this work. Similarly, for environments that are less variable than the suburban environment in this study, the value of $F\max$ should be smaller, and correspondingly, a coarser sampling in the spatial domain may be preferable. For future attempts, a study that establishes an $F\max$ value and corresponding sampling distance across various propagation environments may be undertaken.

Refrences

[1] A. Basere and I. Kostanic, “Cell Coverage Area Estimation From Receive Signal Level (RSL) Measurements,” World Congress on Engineering and Computer Science (WCECS), 2016, in press.

[2] A. Basere and I. Kostanic, “Spatial Sampling Requirements for Received Signal Level Measurements in Cellular Networks,” IEEE Computing and Communication Workshop and Conference (CCWC), 2017, in press.

[3] H. Jiang and C. H. Davis, "Cell-coverage estimation based on duration outage criterion for CDMA cellular systems," IEEE transactions on vehicular technology, vol. 52, pp. 814-822, 2003.

[4] J. D. Parsons and P. J. D. Parsons, "The mobile radio propagation channel," 2000.

[5] S. M. Redl, “An Introduction to GSM,” Atech House Publishers, 1995.

[6] H. Holma and A. Toskala, WCDMA for UMTS, 3rd Ed., Wiley 2004.

[7] E. Dahlman and S. Parkvall, 4G: LTE.LTE-Advanced for Mobile Broadband, Accademic Press, 2011.

[8] W. C. Y. Lee and Y. S. Yeh, Ón the Estimation of the Second-Order Statistics of Log-Normal Fading in Mobile Radio Environment,” IEEE Transaction on Communication, Com-22, June 1974, pp. 869-873.

[9] I. Kostanic, J. Zec and N. Faour, “Sampling Criteria for Wideband Received Signal Measurements,” in proc. of International Conference in Wireless Networks, Las Vegas, NV, June 2003, pp. 24-29.

[10] P. K. Sharma and R. Singh, "Cell coverage area and link budget calculations in GSM system," International Journal of Modern Engineering Research (IJMER) vol, vol. 2, pp. 170-176, 2012.

[11] Y. Sun, "Radio Network Planning for 2G and 3G," Master of Science in Communications Engineering, Munich University of Technology, 2004.

[12] K. R. Manoj, "Coverage estimation for mobile cellular networks from signal strength measurements," Citeseer, 1999.

[13] J.-S. Sheu and W.-H. Sheen, "Characteristics and modelling of inter-cell interference for orthogonal frequency-division multiple access systems in multipath Rayleigh fading channels," IET Communications, vol. 6, pp. 3015-3025, 2012.

[14] L. Arigela, P. Veerendra, S. Anvesh, and K. Satya, "Mobile Phone Tracking & Positioning Techniques," International Journal of Innovative Research in Science, Engineering and Technology, vol. 2, 2013.

[15] I. Akbari, O. Onireti, M. A. Imran, A. Imran, and R. Tafazolli, "Effect of inaccurate position estimation on self-organising coverage estimation in cellular networks," in European Wireless 2014; 20th European Wireless Conference; Proceedings of, 2014, pp. 1-5.

[16] M. Argany, M. A. Mostafavi, F. Karimipour, and C. Gagné, "A GIS based wireless sensor network coverage estimation and optimization: A Voronoi approach," in Transactions on Computational Science XIV, ed: Springer, 2011, pp. 151-172.

[17] S. Sharma and R. Uppal, "RF Coverage Estimation of Cellular Mobile System," Intrnational journal of Engineering and Technology, vol. 3, p. 60, 2011.

[18] B. Sayrac, J. Riihijärvi, P. Mähönen, S. Ben Jemaa, E. Moulines, and S. Grimoud, "Improving coverage estimation for cellular networks with spatial bayesian prediction based on measurements," in Proceedings of the 2012 ACM SIGCOMM workshop on Cellular networks: operations, challenges, and future design, 2012, pp. 43-48.

[19] I. Kostanic, N. Rudic, and M. Austin, "Measurement sampling criteria for optimization of the Lee's macroscopic propagation model," in Vehicular Technology Conference, 1998. VTC 98. 48th IEEE, 1998, pp. 620-624.

[20] O. Hoeber, G. Wilson, S. Harding, R. Enguehard, and R. Devillers, "Visually representing geo-temporal differences," in Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on, 2010, pp. 229-230.

[21] D. Zimmerman, C. Pavlik, A. Ruggles, and M. P. Armstrong, "An experimental comparison of ordinary and universal kriging and inverse distance weighting," Mathematical Geology, vol. 31, pp. 375-390, 1999.

[22] J. Li and A. D. Heap, "A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors," Ecological Informatics, vol. 6, pp. 228-241, 2011.

[23] A. Konak, "Estimating path loss in wireless local area networks using ordinary kriging," in Proceedings of the Winter Simulation Conference, 2010, pp. 2888-2896.

[24] P. A. Burrough, R. A. McDonnell, R. McDonnell, and C. D. Lloyd, Principles of geographical information systems: Oxford University Press, 2015.

[25] D. Kbiob, "A statistical approach to some basic mine valuation problems on the Witwatersrand," Journal of Chemical, Metallurgical, and Mining Society of South Africa, 1951.

[26] H. Wackernagel, Multivariate geostatistics: an introduction with applications: Springer Science & Business Media, 2013.

[27] J. K. Leung and T. Law, "Kriging Analysis on Hong Kong Rainfall Data," HKIE Transactions, vol. 9, pp. 26-31, 2002.

[28] N. Mijatovic and I. Kostanic, "Repeatability of Received Signal Level Measurements in GSM Cellular Networks", in proceedings of 2nd International Symposium on Wireless Pervasive Computing, Feb 2007.