

In the Name of God

Journal of

Information Systems & Telecommunication

Vol. 4, No. 4, October-December 2016, Serial Number 16

Research Institute for Information and Communication Technology
Iranian Association of Information and Communication Technology

Affiliated to: Academic Center for Education, Culture and Research (ACECR)

Manager-in-Charge: Habibollah Asghari, ACECR, Iran

Editor-in-Chief: Masoud Shafiee, Amir Kabir University of Technology, Iran

Editorial Board

Dr. Abdolali Abdipour, Professor, Amirkabir University of Technology, Iran

Dr. Mahmoud Naghibzadeh, Professor, Ferdowsi University, Iran

Dr. Zabih Ghasemlooy, Professor, Northumbria University, UK

Dr. Mahmoud Moghavvemi, Professor, University of Malaya (UM), Malaysia

Dr. Ali Akbar Jalali, Professor, Iran University of Science and Technology, Iran

Dr. Alireza Montazemi, Professor, McMaster University, Canada

Dr. Ramezan Ali Sadeghzadeh, Professor, Khajeh Nasireddin Toosi University of Technology, Iran

Dr. Hamid Reza Sadegh Mohammadi, Associate Professor, ACECR, Iran

Dr. Ahmad Khademzadeh, Associate Professor, CyberSpace Research Institute (CSRI), Iran

Dr. Abbas Ali Lotfi, Associate Professor, ACECR, Iran

Dr. Sha'ban Elahi, Associate Professor, Tarbiat Modares University, Iran

Dr. Ali Mohammad-Djafari, Associate Professor, Le Centre National de la Recherche Scientifique (CNRS), France

Dr. Saeed Ghazi Maghrebi, Assistant Professor, ACECR, Iran

Dr. Rahim Saeidi, Assistant Professor, Aalto University, Finland

Executive Manager: Shirin Gilaki

Executive Assistant: Behnoosh Karimi

Print ISSN: 2322-1437

Online ISSN: 2345-2773

Publication License: 91/13216

Editorial Office Address: No.5, Saeedi Alley, Kalej Intersection., Enghelab Ave., Tehran, Iran,

P.O.Box: 13145-799

Tel: (+9821) 88930150 Fax: (+9821) 88930157

E-mail: info@jst.ir

URL: www.jst.ir

Indexed by:

- | | |
|---|-------------------------|
| - Index Copernicus International | www.indexcopernicus.com |
| - Islamic World Science Citation Center (ISC) | www.isc.gov.ir |
| - Directory of open Access Journals | www.Doaj.org |
| - Scientific Information Database (SID) | www.sid.ir |
| - Regional Information Center for Science and Technology (RICeST) | www.ricest.ac.ir |
| - Iranian Magazines Databases | www.magiran.com |

Publisher:

Regional Information Center for Science and Technology (RICeST)

Islamic World Science Citation Center (ISC)

This Journal is published under scientific support of
Advanced Information Systems (AIS) Research Group and
Digital & Signal Processing Research Group, ICTRC

Acknowledgement

JIST Editorial-Board would like to gratefully appreciate the following distinguished referees for spending their valuable time and expertise in reviewing the manuscripts and their constructive suggestions, which had a great impact on the enhancement of this issue of the JIST Journal.

(A-Z)

- Abdali Mohammadi Fardin, Razi University, Kermanshah, Iran
- Anvaripour, Mohammad, Sahand University, Tabriz, Iran
- Borna, Keyvan, Kharazmi University, Tehran, Iran
- Buyer, Asgar Ali, , Azarbayjan Shahid Madani University, Tabriz, Iran
- Darzi, Mohammad, Academic Center for Education Culture and Research (ACECR), Iran
- Fathi, Abdolhossein, Razi University, Kermanshah, Iran
- Haji Mohammadi, Zeinab, Amir kabir University, Tehran, Iran
- Hamidi, Hojjatollah, khaje Nasir-edin Toosi University, Tehran, Iran
- Hasanzadeh, Reza, Gilan University, Gilan, Iran
- Horri Najafabadi, Abbas, Shahre Kord University, Shahre Kord, Iran
- Izad khah, Habib, Tabriz University, Iran
- Jahd Amal, Nima, jamia Hamdard University, Delhi, India
- Kashef, Seyed Sadra, Tarbiat modares, Tehran
- Keshavarz, Hengameh, University of Sistan & Baluchestan, Zahedan, Iran
- Mansoorizadeh, Muharram, Bu-Ali Sina University, Hamedan, Iran
- Masoudifar, Mina, Ferdowsi University, Mashhad, Iran
- Mosallanejad, Ali, Shahid Beheshti University, Tehran, Iran
- Mohebbi, Keyvan, Islamic Azad University, Mobarakeh Branch, Iran
- Nilforoushan, Zahra, Kharazmi University, Tehran, Iran
- Oveisi, Farid, Academic Center for Education Culture and Research (ACECR), Tehran, Iran
- Sadegh Mohammadi, Hamid Reza, Associate Professor, ACECR, Iran
- Safaei, Ali Asghar, Tarbiat Modares University, Tehran, Iran
- Sajedi, Hedieh, Tehran University, Tehran, Iran
- Sattari Naeini, Vahid, Shahid Bahonar University , Kerman, Iran
- Sedighian Kashi, Saeid, khaje Nasir-edin Toosi University, Tehran, Iran

Table of Contents

• Short Time Price Forecasting for Electricity Market Based on Hybrid Fuzzy Wavelet Transform and Bacteria Foraging Algorithm	210
Keivan Borna and Sepideh Palizdar	
• Identification of a Nonlinear System by Determining of Fuzzy Rules	215
Hodjatollah Hamidi and Atefeh Daraei	
• Automatic Facial Emotion Recognition Method Based on Eye Region Changes	221
Mina Navraan, Nasrollah Moghadam Charkari and Muharram Mansoorizadeh	
• A Semantic Approach to Person Profile Extraction from Farsi Documents	232
Hojjat Emami, Hossein Shirazi and Ahmad Abdollahzadeh Barforoush	
• An Effective Risk Computation Metric for Android Malware Detection	244
Mahmood Deypir and Ehsan Sharifi	
• A New Node Density Based k-edge Connected Topology Control Method: A Desirable QoS Tolerance Approach	255
Mohsen Heydarian	
• High-Resolution Fringe Pattern Phase Extraction, Placing a Focus on Real-Time 3D Imaging ..	265
Amir Hooshang Mazinan and Ali Esmaili	
• Hybrid Task Scheduling Method for Cloud Computing by Genetic and PSO Algorithms	271
Amin Kamalinia and Ali Ghaffari	

Short Time Price Forecasting for Electricity Market Based on Hybrid Fuzzy Wavelet Transform and Bacteria Foraging Algorithm

Keivan Borna*

Department of Computer Science, Faculty of Mathematics and Computer Science, Kharazmi University, Tehran, Iran
borna@khu.ac.ir

Sepideh Palizdar

Department of Computer Engineering, Faculty of Engineering, Kharazmi University, Tehran, Iran
cs.progress@gmail.com

Received: 30/Jan/2016

Revised: 01/Mar/2016

Accepted: 10/Jun/2016

Abstract

Predicting the price of electricity is very important because electricity can not be stored. To this end, parallel methods and adaptive regression have been used in the past. But because dependence on the ambient temperature, there was no good result. In this study, linear prediction methods and neural networks and fuzzy logic have been studied and emulated. An optimized fuzzy-wavelet prediction method is proposed to predict the price of electricity. In this method, in order to have a better prediction, the membership functions of the fuzzy regression along with the type of the wavelet transform filter have been optimized using the E.Coli Bacterial Foraging Optimization Algorithm. Then, to better compare this optimal method with other prediction methods including conventional linear prediction and neural network methods, they were analyzed with the same electricity price data. In fact, our fuzzy-wavelet method has a more desirable solution than previous methods. More precisely by choosing a suitable filter and a multiresolution processing method, the maximum error has improved by 13.6%, and the mean squared error has improved about 17.9%. In comparison with the fuzzy prediction method, our proposed method has a higher computational volume due to the use of wavelet transform as well as double use of fuzzy prediction. Due to the large number of layers and neurons used in it, the neural network method has a much higher computational volume than our fuzzy-wavelet method.

Keywords: Price Prediction; Wavelet Transform; Fuzzy Logic; Bacteria Foraging Algorithm; Electricity Market.

1. Introduction

In the last couple of decades, the electric power industry, in most countries of the world, has undergone radical and fundamental changes which are referred to by different names such as deregulation, restructuring, law revision, etc.. Also in Iran, changes in the electric power industry, began with the launch of Iran's electricity market in November 2003. With Iran Grid Management Company being established in 2004, Iran's electricity market took a more serious shape. Also, considering the approval of the implementing regulations related to paragraph (b) of Article 25 of the Third Development Plan law by the Council of Ministers, the approval of the law related to the independence of distribution companies by the Islamic Consultative Assembly, and the interpretation of Article 44 of the constitution by the supreme leader, Iran's electricity industry will undergo fundamental changes in the coming years [1].

A lot of work has been done with respect to price prediction. In 2002, a paper was presented in which an autoregressive integrated moving average (ARIMA) model has been used to predict the prices in the electricity market. Since the price in the electricity market depends on various factors, by comparing the results of this study with future

studies, it can be seen that this method has not had an appropriate solution. This method does not have a desirable solution especially where the price changes are big.

In 2009, a paper was presented in which a hybrid model, which combined an autoregressive integrated moving average (ARIMA) model and generalized autoregressive conditional heteroskedasticity (GARCH) model, has been used to predict the average daily price in Iran's electricity market. This method had a medium computational volume, and according to the results obtained, it can be seen that this forecasting method has not had a desirable solution where the speed of price changes were high. [1]

Since the time series models did not have a desirable solution in drastic changes in the desired signal, thus a paper was presented in 2008, in which the wavelet transform and a neural network have been used. One of the capabilities of the wavelet transform is that, where drastic changes occur, the length of the window, in which the signal is evaluated, becomes smaller and calculations will be done more carefully, and among the disadvantages of this method an increased computational volume can be mentioned. [5]

Reference [7] has presented a method in which a parallel method has been used to reduce prediction errors,

* Corresponding Author

which has been used to predict the next day price of the electricity market, and in which the data predicted in the previous period are used as data for the next period.

In 2009, a method was presented based on adaptive regression. According to the article, other prediction methods can at most predict up to next 3 or 4 steps, but this method can predict up to next 10 steps. A disadvantage of this method is that it requires the ambient temperature data simultaneously with the electricity price data for prediction. [8].

In the next section we consider linear prediction.

2. Linear Prediction

In this section we will explain linear prediction method. A linear prediction model represents the time series of the signal samples during a given time period. Its usual linear model is as follows [8]:

$$y(t+T) = a_1 \cdot y(t) + a_2 \cdot y(t-T) + \dots + a_m \cdot y(t-(m-1)T) \quad (1)$$

where a_1, a_2, \dots, a_m are linear prediction coefficients, m is the model order, T is the sampling time, $y(t+T)$ is the future sample, and $y(t), y(t-T), \dots, y(t-mT)$ are the present and past observations. In this relation, the function output is a linear combination of the present and past samples, thus this function is called a linear prediction.

Two stages have to be done to achieve a linear prediction of prices in the electricity market using equation (1). According to this equation, first the model order has to be chosen carefully, and then the coefficients a_1, a_2, \dots, a_m have to be calculated from the modeling window. And then the obtained model can be used to predict the price of electricity market in the time steps ahead.

A least-squares error method can be used to estimate the coefficients a_1, a_2, \dots, a_m in equation (1). This error is measured between the estimated value and actual value. In use of least-squares method, the energy in the error signal, falls to its minimum. A solution to this problem is vector B which estimates the unknown vector of parameter β . The least-squares solution is as follows:

$$b = \hat{\beta} = (X^T \cdot X)^{-1} \cdot X^T \cdot y \quad (2)$$

In the next section prediction using the neural network method is studied.

3. Prediction Using the Neural Network Method

In this section we will explain prediction using the neural network method and the type of neural network that we use is considered throughly.

Due to their remarkable abilities to infer results from complex and ambiguous data, neural networks can be used to extract patterns and identify various trends, which are very difficult to identify for humans and computers. In this study, a back-propagation neural network is used, with 5 layers and 5 inputs, all of which are the data of the

price of market electricity from the previous time period. The figure below shows how to choose a neural network.

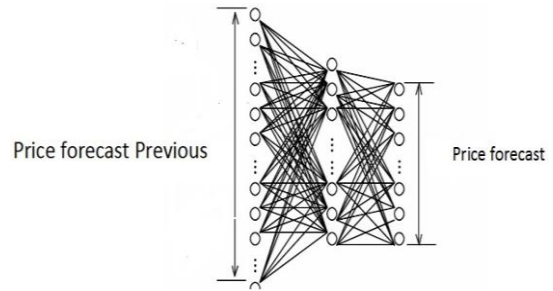


Fig. 1. The structure of a BP neural network for price prediction in the electricity market

4. Prediction Using the Fuzzy Regression Method

In the next section we study the history of fuzzy logic and fuzzy regression.

Dr. Lotfizadeh introduced the theory of fuzzy systems in 1965. In such an atmosphere where the scientists of engineering sciences were seeking for mathematical methods to defeat more difficult problems, the fuzzy theory took steps toward another kind of modeling.

In conventional fuzzy systems, the number and type of membership functions are determined by trial and error. But what is obvious is that, a larger number of membership functions are needed for more complicated systems. On the other hand, as the number of membership functions increases, the number of fuzzy rules usually increases, which ultimately leads to the complexity of implementation. Membership functions can have different shapes, such as triangular, Gaussian, bell, trapezoidal, etc.. In the linear regression, the goal is to find the fuzzy coefficients of the polynomial below. In other words, the goal is to express the linear prediction using the fuzzy coefficients.

Its usual linear model is as follows [8]:

$$y(t+T) = \bar{A}_1 \cdot y(t) + \bar{A}_2 \cdot y(t-T) + \dots + \bar{A}_m \cdot y(t-(m-1)T) \quad (3)$$

Where; (\bar{A}_i) s are fuzzy coefficients which are expressed by fuzzy membership functions. Here the goal is to find (\bar{A}_1) s in a way that the prediction error is minimized.

5. Our Proposed Method: Fuzzy-Wavelet Prediction

In this section the history of wavelet and our proposed method which is combination of fuzzy prediction method and wavelet is presented.

The constant resolution problem in the short-time Fourier transform, has its roots in Heisenberg's uncertainty principle. According to this principle, a time-frequency description of a signal cannot be achieved exactly; that is, it cannot be found out exactly that at a given signal, what frequency components are available at what time intervals, but we can only found out that what

frequency bands are available at what time intervals. This principle directly returns to the concept of resolution. Although the time and frequency resolution problems are results of a physical phenomenon (the Heisenberg uncertainty principle) and exist regardless of the transform used, it is possible to analyze any signal by using an alternative approach called the multiresolution analysis (MRA).

The wavelet transform is a kind of windowing technique with variable-sized windows. The wavelet analysis gives us the possibility to achieve our goal both in a long duration where we require high accuracy at low frequency data, and in shorter durations where we need high-frequency data. The wavelet transform does not convert into a time-frequency region, but rather into a time-scale region.

Using the wavelet transform, the price signal of electricity market can be divided into data with big changes (details) and data with little changes (generalities). Figure 2 shows an overview of this conversion.

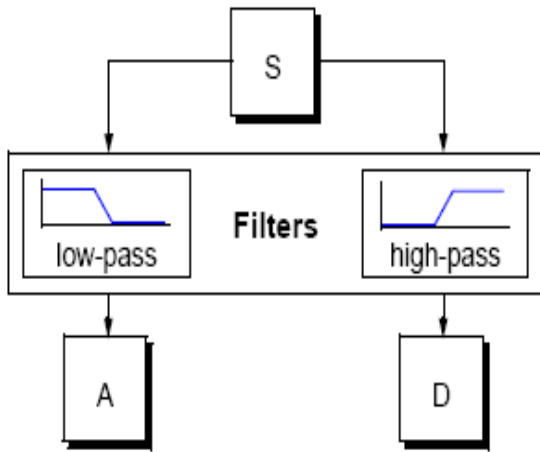


Fig. 2. The structure of the wavelet transform for price prediction in the electricity market.

Since in order to increase the accuracy of predictors when many changes occur, it is necessary to increase the number and/or degrees of membership functions, which increases the volume of calculations, another solution is to use the multiresolution analysis, in this way that; high frequency data (details) are estimated by a predictor and low frequency data (generalities) by another predictor. By doing this, the desired accuracy and less computational volume can be achieved.

This paper uses a combination of wavelet transform and fuzzy regression to predict the price of electricity market.

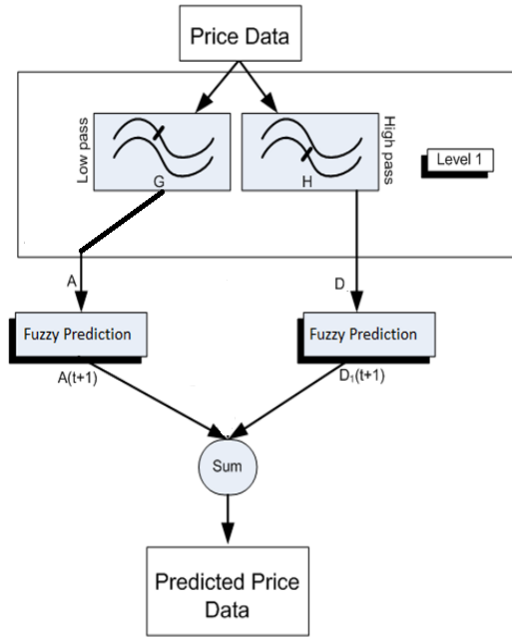


Fig. 3. shows an overview of this system.

Figure 3. The general structure of the proposed fuzzy-wavelet transform, for price prediction in the electricity market.

To compare the prediction methods presented in this paper, some indicators are defined as follows.

The mean absolute percentage error (MAPE): This error is defined as follows:

$$MAPE = \frac{1}{n} \sum_{t=1}^n |PE_t| \tag{4}$$

Where; n is the number of data, and PE represents a relative error and is defined as follows:

$$PE_t = \left(\frac{y_t - F_t}{y_t} \right) \times 100 \tag{5}$$

If y_t is the actual observation for duration t, and F_t a forecast for the same duration, then the error will be defined as in equation (6):

$$e_t = y_t - F_t \tag{6}$$

The maximum forecast error: is the greatest value of error in the prediction for the test data set which is defined as in equation (4-1):

$$ME = \max(\text{Price}_{actual} - \text{Price}_{predicted}) \tag{7}$$

The maximum forecast percentage error is defined as follows:

$$MPE = \max\left(\frac{\text{Price}_{actual} - \text{Price}_{predicted}}{\text{Price}_{actual}}\right) \times 100 \tag{8}$$

Also the mean squared error (MSE) is defined as follows:

$$MSE = \sqrt{\frac{1}{N} \sum_{k=1, \dots, N} |\text{Price}_{actual} - \text{Price}_{predicted}|^2} \tag{8}$$

In the next section the simulation results and comparisons are presented.

6. Simulation and Results

In order to determine the best model in quantitative terms, the three measures of prediction errors: MSE, mean absolute error (MAE), and mean absolute percentage error (MAPE), were used to evaluate and compare the models. For a better evaluation, the results of these forecasting methods as well as their instantaneous errors are presented in Figures 4 to 11.

By taking them into consideration, it can definitely be concluded that the values resulting from the wavelet-fuzzy forecasting method has a better solution than the previous forecasting methods.

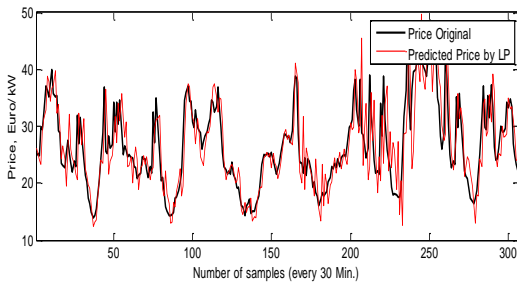


Fig. 4. The results obtained from the linear prediction, for price prediction in the electricity market

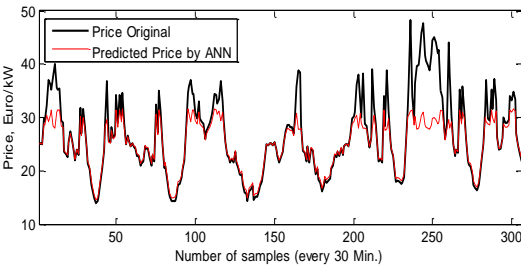


Fig. 5. The results obtained from the neural network prediction, for price prediction in the electricity market

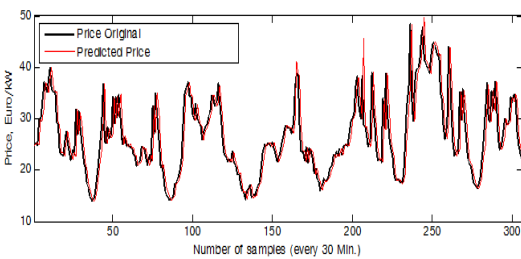


Fig. 6. The results obtained from fuzzy prediction, for price prediction in the electricity market

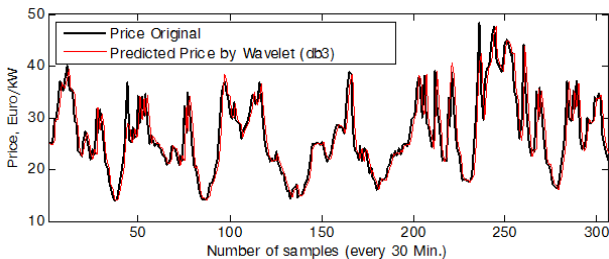


Fig. 7. The results obtained from our fuzzy-wavelet prediction, for price prediction in the electricity market

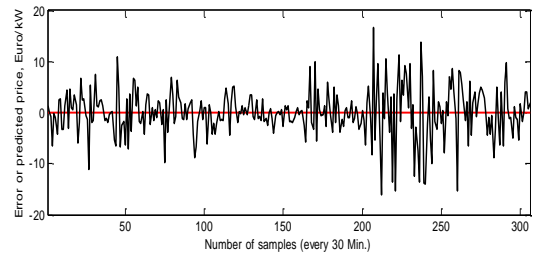


Fig. 8. The instantaneous error for price prediction in the electricity market, using the linear prediction method

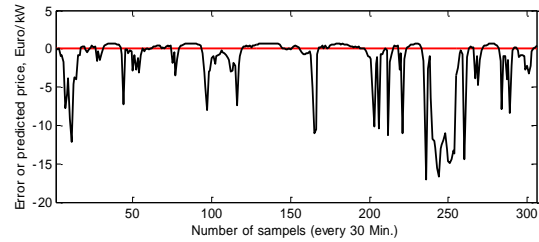


Fig. 9. The instantaneous error for price prediction in the electricity market, using the neural network prediction method

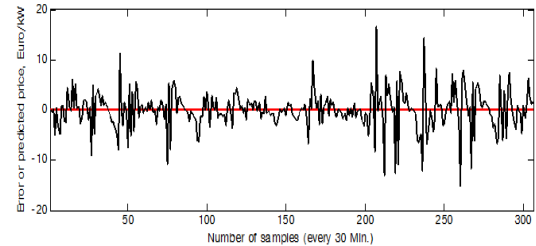


Fig. 10. The instantaneous error for price prediction in the electricity market, using the fuzzy regression prediction method

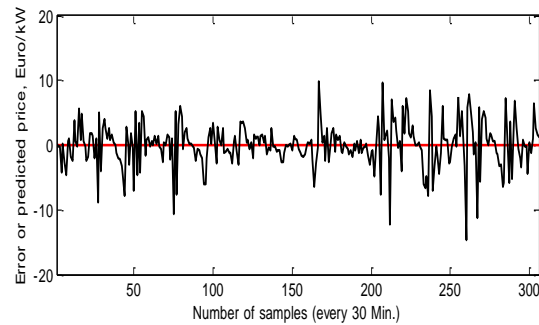


Fig. 11. The instantaneous error for price prediction in the electricity market, using our fuzzy-wavelet prediction method

Table 1. Comparison of errors among the prediction methods presented [13]

Method	MAPE (%)	Maximum (%)	MSE	Volume of calculations
Linear prediction[13]	12.59%	16.7	1.5294	High
Neural network prediction[13]	6.2%	1.346	17.06	Medium
Fuzzy prediction[13]	9.40%	1.2618	16.2	Medium
Fuzzy-wavelet prediction[our method]	8.11%	1.081	12.18	High

7. Conclusion

By evaluating the results obtained for price prediction in Queensland electricity market, it can be seen that use of fuzzy logic-wavelet forecasting method resulted in an improved performance, compared with that of fuzzy logic forecasting method. Also choosing two different types of filters; low-pass and high-pass, in the wavelet transform, increased the efficiency of the predictor in the fuzzy prediction method. To further investigate the presented methods, the results of these methods have been collected in Table 1. As can be seen, the fuzzy-wavelet method has a more desirable solution than the other presented

methods have, also by choosing a suitable filter and a multiresolution processing method, the maximum error has improved by 13.6%, and the mean squared error has improved about 17.9%. But in comparison with the fuzzy prediction method, the proposed method has a higher computational volume due to use of wavelet transform as well as double usage of fuzzy prediction. Due to the large number of layers and neurons used in it, the neural network method has a much higher computational volume than our fuzzy-wavelet method has, but this method, depending on the data used for training, has a greater maximum error than the proposed method has.

References

- [1] International Energy Outlook 2005, Energy Information Administration; <http://www.eia.doe.gov/iea>.
- [2] Patel, M.R., "Electricity price and solar power systems", CRC Press LLC, 1999.
- [3] Global power source, Global electricity price energy council, <http://www.gwec.net/>.
- [4] Cota, A., "A review on the young history of electricity price power short term prediction", Journal on Renewable Energy Review, vol. 12, Issue 6, pp. 1725-1744, 2008.
- [5] Sideratos, G. Hatziaargyriou, N.D. "An Advanced Statistical Method for Electricity price Power Forecasting", IEEE Transaction on power systems, Nat. Tech. Univ. of Athens, vol. 22, Issue 1, pp. 258-265, Feb 2007.
- [6] Hui, L., Hong-Qi, T., Chao, C., Yan-fei, L. "A Hybrid Statistical to Predict Electricity price Speed and Electricity price Power", Renewable energy, Science direct, December 2009.
- [7] Monfared, M., Rastegar, H., Kojabadi, H. M. "On Comparing Three Artificial Neural Networks for Electricity price Speed Forecasting", Applied energy, Science direct, January 2010.
- [8] MATLAB, Mathematical Foundations of Multiple Linear Regressions, R2007a.
- [9] Sahin, A. D., Zekai, S., "First-order Markov chain approach to electricity price speed modeling", Journal of wind Engineering and Industrial Aerodynamics 89 (2001) 263-269.
- [10] Shamshad, A., Bawadi, M.A., Wan Hussin, W.M.A., Majid, T.A., Sanusi, S.A.M., "First and second order Markov chain models for synthetic generation of electricity price speed time series", Energy 30 (2005) 693-708.
- [11] Kennedy, J., and Eberhart, R., "Particle Swarm Optimization", IEEE International Conference on Neural Networks, pp. 1942-1948, 1995.
- [12] M. Monfared, H. Rastegar, H. M. Kojabadi "A New Strategy for Electricity price Speed Forecasting Using Artificial Intelligent Methods", Renewable energy, Science direct, Vol. 34, Issue 3, pp. 845-848, March 2009.
- [13] A. Motamedi, H. Zareipour, W.D. Rosehart, "Electricity Price and Demand Forecasting in Smart Grids", IEEE Transactions on Smart Grid, vol. 3, pp. 664-674, 2012.

Keivan Borna joined the Department of Computer Science at the Faculty of Mathematics and Computer Science of Kharazmi University as an Assistant Professor in 2008. He earned his Ph.D. in Mathematics from the Department of Mathematics, Statistics and Computer Science of the University of Tehran. His research interests include Computer Algebra, Cryptography, Approximation Algorithms, Evolutionary Computations and Computational Geometry. He is the author of the "Advanced Programming in JAVA" (in Persian) and is a life member of "Elite National Foundation of Iran".

Sepideh Palizdar received her M.Sc. degree from Faculty of Engineering at Kharazmi University, Tehran, Iran at 2015. Her research interests include evolutionary computations.

Identification of a Nonlinear System by Determining of Fuzzy Rules

Hodjatollah Hamidi*

Department of Industrial Engineering, K. N. Toosi University of Technology, Tehran, Iran
h_hamidi@kntu.ac.ir

Atefeh Daraei

Department of Industrial Engineering, K. N. Toosi University of Technology, Tehran, Iran
adaraei@mail.kntu.ac.ir

Received: 22/Apr/2016

Revised: 29/Jul/2016

Accepted: 10/Aug/2016

Abstract

In this article the hybrid optimization algorithm of differential evolution and particle swarm is introduced for designing the fuzzy rule base of a fuzzy controller. For a specific number of rules, a hybrid algorithm for optimizing all open parameters was used to reach maximum accuracy in training. The considered hybrid computational approach includes: opposition-based differential evolution algorithm and particle swarm optimization algorithm. To train a fuzzy system which is employed for identification of a nonlinear system, the results show that the proposed hybrid algorithm approach demonstrates a better identification accuracy compared to other educational approaches in identification of the nonlinear system model. The example used in this article is the Mackey-Glass Chaotic System on which the proposed method is finally applied.

Keywords: System Identification; Combined Training; Fuzzy Rules; Database Design.

1. Introduction

System identification is extensively used in a number of programs including control systems [1], communications [2], signal processing [3], controlling chemical processes [4], biological processes [5] and etc. to be exact all problems of the real world are nonlinear per se. Anyhow, we face less computational problems in identifying a linear system and normally we are not faced with simple problems in identifying nonlinear systems. In [6] particle swarm optimization with different particle length is proposed for production of structure and parameters of fuzzy rule database. In [7], the continuous version of ant colony optimization is used for designing of fuzzy rules database. In this work the online method is used for determining the number of fuzzy rules and all open parameters in each fuzzy rule in the continuous space was continually optimized by ant colony optimization algorithm. In [8], system training is presented using two-step swarm intelligence algorithm. This algorithm includes two steps. In the first part structure and initial parameter identification is carried out using online clustering of ant colony optimization algorithm in disrupted space. In the second part particle swarm optimization is used for greater optimization of all open parameters in the continuous space.

Fuzzy systems are suitable for complex systems' modeling, due to the good feature of general approximation especially for systems in which mathematical description is difficult. It is proven that any continuous function can approximate a logical degree of accuracy using a fuzzy system which is trained by meta-

heuristic algorithms. This approximation function can act as a model for a number of functional complex systems. Juang et al, 2014, have shown that fuzzy systems that are trained by algorithm can be effectively used in identification of nonlinear models.

Recently, the differential evolution algorithm (DE) is considered as a modern technique of evolution calculations [9,10] that are used for optimization issues. DE is preferred over other evolution methods like genetic algorithm (GA) [11,12] and particle swarm optimization (PSO) and this is due to its notable characteristics like simple concept, easy execution and rapid convergence [13,14]. Generally all population-based optimization algorithms which also include DE suffer the long calculation period due to their evolutionary-accidental nature.

The concept of opposition-based learning (OBL) is introduced by Tizhoosh [15]. In this article, this concept is used for acceleration of learning in fuzzy systems. The main idea in OBL concept is simultaneous consideration of an estimate and its corresponding opposing estimate. OBL leads to achievement of a better estimation and acceleration the rate of DE convergences. PSO is an algorithm with local search pattern and can be used to fine-tune the present results and faster access to global minimum. Therefore the proposed method in this article is called hybrid opposition-based differential evolution with particle swarm optimization (HODEPSO). ODE utilizes opposing numbers during the start of population and also for production of new population during the evolution process. Here, opposing numbers are used to accelerate the rate of convergences of DE optimizing algorithm. Pure random sampling or selection of solutions from data

* Corresponding Author

population provides for a chance to visit or even inspection of undiscovered regions of the search space. It has been proven that the probability of this incident is less for opposing numbers than purely random numbers. In fact mathematical proof has been used to show that the probability of opposing numbers being closer to desired solutions is higher than completely pure numbers [16-19]. In [17], the benefit of opposing numbers is investigated by replacing them with random numbers and this method has been utilized for initializing the population and generation skipping for different DE versions.

This article presents a new educational sample of fuzzy systems which are combined with meta-heuristic evolutionary algorithm, meaning:

- Use of Opposition-Based Learning concept as simultaneous considering of an estimate and its opposing corresponding estimate which would lead to better estimation and acceleration of the rate of convergence of differential evolution algorithm (DE).
- Combination of ODE and PSO to prevent the probability of getting caught up in local optimum and quicker and more accurate achievement of general optimum.
- Performance of the trained fuzzy system using HODEPSO is shown by comparing the results of some of the present methods in the un-linear system identification. Results of stimulation, shows the suitable performance of the proposed method compared to other methods of identification.

2. Meta-Heuristic Optimization (MHO) Algorithms

Optimization methods are search methods that aim at finding answers to the optimization problem so that the evaluated quantity is optimized. According to evidence and records of results, the best quality and time opposition for fuzzy system optimization is provided using meta-heuristic algorithms.

2.1 Differential Evolution (DE) Algorithm

Differential evolution algorithm (DE) is one of the effective search-based methods [20-33]. Like other evolution algorithms, this one also starts by initializing a population. Then through implementation of agents like combination, mutation and generation convergence, the new-born is formed and in the next step which is called selection, new born generation is compared to parent generation to determine the rate of aptitude which is evaluated by the goal function. Then the best members enter the next round as the next generation. This trend continues until desired results are reached. Different levels of this algorithm are stated here in sequence.

Population Initialization: The number of variables in this algorithm are shown with D. each of these variables hold a high and low limit. Initial population with the size of N_p in D is randomly formed according to equation (1).

$$X_{i_o} = X_i \min + \text{round}(\delta_i \cdot (X_i \max - X_i \min)) \quad (1)$$

$$, i = 1, 2, \dots, N_p$$

Where δ_i is a random number in the (0,1] domain, $X_i \max$ and $X_i \min$ are the high and low limits of the variables and N_p is the number of members.

Mutation and Intersection: in this algorithm five strategies can be utilized for combination and production of new-borns [18]. In this article best person-random person-random person is used for mutation as follows:

$$Z_{i,G} = X_{\text{best},G} + F \cdot (X_{r1,G} - X_{r2,G} + X_{r3,G} - X_{r4,G}) \quad (2)$$

Where F is called standard factor, X_{2S} are randomly selected members and X_{best} is the best member of the present population.

For every variable of each member of the population a random number, K, in the [1, D] domain and a random number, u, in the [0, 1] domain is selected. Intersection is carried out according to the equation below:

$$\text{if } u \leq CR \text{ or } j = k$$

$$\text{then } Z_{i,j} = X_{r1,j} + F(X_{r3,j} - X_{r2,j})$$

$$\text{else } Z_{i,j} = X_{i,j} \quad (3)$$

Where j is the number of any variable from i^{th} member of the population and CR is the intersection constant and is chosen as a number between 0 and 1.

Estimation and selection: at this stage the new-borns and parents are valued according to the goal function and if the newborn has a higher value than the parent, it replaces the parent. Otherwise the parent moves on to the next level with the next generation.

$$Z_{i,g+1} = \arg \max(f(z_i, g), z_i, g + 1) \quad (4)$$

In this equation g stands for generation, $Z_{i,g+1}$ is the new generation population (new-borns) and $Z_{i,g}$ is the previous generation population (parents). F is the goal function of the problem.

Repetition: repeating steps 2 and 3 until maximum repetition or the whole population convergence is reached.

2.2 Opposition-Based Differentiation Evolution (ODE) Algorithm

In optimization approaches of evolution algorithm, a unified random guess for the initial population is considered. In each generation the goal includes movement towards the desired solution and the research trend ends when some of the pre-determined criterion are satisfactory. Calculation time usually depends on initial guess, meaning that the greater the distance between initial guess and desired solution, the more time it takes to reach the end and vice versa. Opposition-based learning increases the chance to start with a better initial population through revision of opposing solutions.

Similar approaches to this can be used not only in initial solutions but also utilized continually in the present population for any solution [19].

2.2.1 Definition of Opposing Number

Suppose $x \in [a, b]$ is a real number. The opposing number is \tilde{x} which is defined by $\tilde{x} = a + b - x$.

Definition of opposing point: suppose $p = (x_1, x_2, \dots, x_d)$ is a point in a D-dimensional space in which $x_1, x_2, \dots, x_d \in R$ and $x_i \in [a_i, b_i]$. Opposing point is:

$$\tilde{p} = (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_d) \text{ where } \tilde{x}_i = a_i + b_i - x_i$$

2.2.2 Opposition-Based Optimization (OBO)

Suppose $p = (x_1, x_2, \dots, x_d)$ is a point in a d-dimensional space, meaning suppose an elective solution. $F(0)$ is a proportion function which is used to measure the proportion of selections. According to definition of opposing points, $\tilde{p} = (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_d)$ opposes $p = (x_1, x_2, \dots, x_d)$. Now, if $(\tilde{p}) \geq f(p)$, then p can be replaced by \tilde{p} otherwise we continue with p . therefore the point is evaluated simultaneously with its opposing point so that we continue the algorithm with the more suitable ones.

2.3 Particle Swarm Optimization (PSO) Algorithm

Particle Swarm Optimization Algorithm (PSO) works according to the social behaviour of birds [20]. For better understanding of this technique, consider the below scenario: "a flock of birds are randomly looking for food in a specific region. There is only one piece of food in this region which the birds are not aware of but are aware of their distance with the food all the time." At this state, a suitable strategy for finding the exact location of food is following the bird that is closer to the food. Actually PSO has been inspired by such a scenario too and presents a solution for optimization problems in PSO each bird is a solution to the problem. All the present responses have a fitness value which is calculated by the defined fitness function for the problem. The aim of this technique is finding a location with the best fitness value in the problem setting. This fitness value has a direct effect on the direction and speed of these birds' movement (solutions to the problem) towards the location of the food (optimal response).

PSO starts with a number of initial response (particles) and looks for optimal response by moving these responses in continuous repetitions. In every repetition two values are determined: P_{Best} and G_{Best} .

P_{Best} : Location of the best P_{Best} fitness value where each [article has reached in its movement,

G_{Best} : Location of the best particle fitness in the present population.

After the above values are calculated, the particles' speed of movement is calculated by equation (4) and each particle's next location is calculated by equation (5).

$$V_{i,t+1} = w.v_{i,t-1} + c_1.r_1.(P_{best_i} - P_{i,t}) + c_2.r_2.(G_{best_i} - P_{i,t}) \quad (5)$$

$$P_{i,t+1} = P_i + v_i \quad (6)$$

In these equations r_1 and r_2 values are random numbers between zero and one and c_1 and c_2 coefficients which are called learning coefficients are usually equalled to two initializations.

In every repetition of algorithm, the speed of particle movement (rate of change for each particle) in every dimension can be limited with a pre-determined V_{max} value. At this state if the speed of each particle in each dimension exceeds this limit, we replace it by V_{max} .

3. Hybrid Opposition-Based Differential Evolution and Particle Swarm Optimization Algorithms

In this article, the hybrid algorithm of opposition-based differential evolution and particle swarm optimization (HODEPSO) is effectively developed. The exact details of steps in hybrid algorithm of opposition-based differential evolution and particle swarm optimization is explained below:

Step 1: Random population initialization and by considering simultaneous Gaussian of opposing initial values.

Step 2: enforcement of opposition-based differential evolution algorithm agents on the initial population.

Step 3: Evaluation of the cost function (which is as RMSE in the solved example in this article) for each particle and updating P_{Best} and G_{Best} .

Step 4: Selection of parents' and Gaussian their opposition and enforcement of opposition-based differential evolution algorithm agents on them.

Step 5: Evaluation of cost function for the new-borns and updating P_{Best} and G_{Best} particle speed.

Step 6: After selection of new-borns from the elected parents, the survivor selection mechanism is performed.

Step 7: updating particle speed and status using equations (5) and (6).

Step 8: Evaluation of cost function for each particle and updating P_{Best} and G_{Best} .

Step 9: If the conditions for ending is established, hybrid algorithm can end otherwise go to step 4.

4. Fuzzy System

In this section, the design of the hybrid algorithm-based fuzzy system is described. The fuzzy system used if of TSK, zero-degree kind in which the i^{th} fuzzy rule is specified with R_i and described as follows:

$$R_i : \text{If } x_1 \text{ is } A_1^i(x_1) \text{ and } x_2 \text{ is } A_2^i(x_2) \text{ then } y \text{ is } B^i \quad (7)$$

Where R_i is the i^{th} fuzzy rule, x_j is the input variable, y is the output variable, $A_{ij}(x_j)$ is the fuzzy set and B is a

certain value. The fuzzy set $A_{ij}(x_j)$ is a Gaussian membership function described by the equation below:

$$A_{ij}(x_j) = \exp \left\{ - \left(\frac{x_j - m_{ij}}{\sigma_{ij}} \right)^2 \right\} \quad (8)$$

Where m is the center and is the width of the Gaussian membership function. For x_1 and x_2 inputs, the meridian or effective weight w_i is calculated as follows:

$$w_i = A_1^i(x_{1k}) + A_2^i(x_{2k}) \quad (9)$$

If the fuzzy system has r rules, the fuzzy system output is calculated with the defuzzication weighted average as follows:

$$y_i = \frac{\sum_{i=1}^r w_i B^i}{\sum_{i=1}^r w_i} \quad (10)$$

The status of each particle is stated with the vector below in the search space:

$$P = [m_{11}, \sigma_{11}, m_{12}, \sigma_{12}, a_1, m_{21}, \sigma_{21}, m_{22}, \sigma_{22}, a_2] \quad (11)$$

Where a is the tally value in each fuzzy law. For example if the fuzzy system has n input variables and r is the rule, the number of vector member, p (number of optimizing variables) would equal $(2n + 1) \times r$.

5. Results of Simulation

In this example the designed fuzzy system is used to predict future values of Mackey-Glass chaotic time series. This time series is produced using the Mackey-Glass delay differential equation as below:

$$\dot{x}(t) = \frac{0.2x(t - \tau)}{1 + x^{10}(t - \tau)} - 0.1x(t) \quad (12)$$

The issue of predicting time series based on Mackey-Glass differential equation is a famous criterion for comparison of capacities of different fuzzy models. 1000 pairs of input-output data were extracted from Mackey-Glass chaotic time series. The first 500 pairs were used for training of the fuzzy system while the remaining 500 were used as test data to evaluate the performance of the fuzzy model in prediction.

$$\begin{aligned} \text{Input 1} &= x(t - 30) \\ \text{Input 2} &= x(t - 18) \\ \text{Input 3} &= x(t - 12) \\ \text{Input 4} &= x(t) \end{aligned} \quad (13)$$

To evaluate the performance of the designed fuzzy model RMSE was used which is defined as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum_{k=1}^n (y_k - \hat{y}_k)^2} \quad (14)$$

Where n is the number of data, y_k is the real output and \hat{y}_k is the fuzzy model output.

To show the efficiency of hybrid algorithm compared to algorithms of opposition-based differential evolution and particle swarm optimization, each algorithm was used individually. The results of the use of these algorithms are shown in table (1) with an average of 50 times use.

According to the results, the designed fuzzy system with the proposed method, in addition to simplicity (reduction in the number of fuzzy rules) shows a better performance compared to fuzzy models presented in [21] and [22] and has achieved a lower RMSE compared to those.

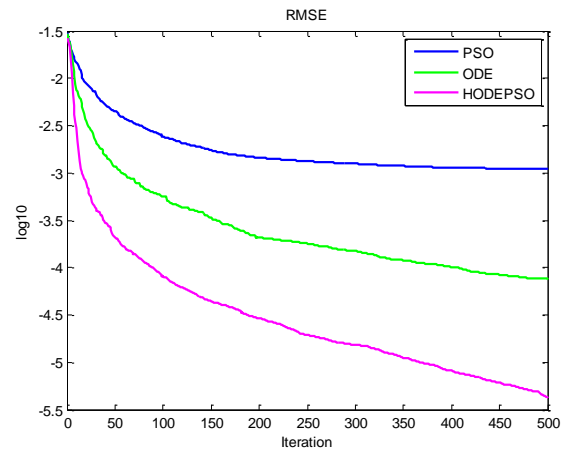


Fig. 1. RMSE values achieved in every repetition by PSO, ODE and HODEPSO.

According to Fig.1 it is observed that the hybrid algorithm converges more quickly to the optimal solution and has better performance compared to individual ODE and PSO algorithms.

The same example is studied in references [21] and [22]. Comparison of the results of these methods and the proposed method are shown in table (1).

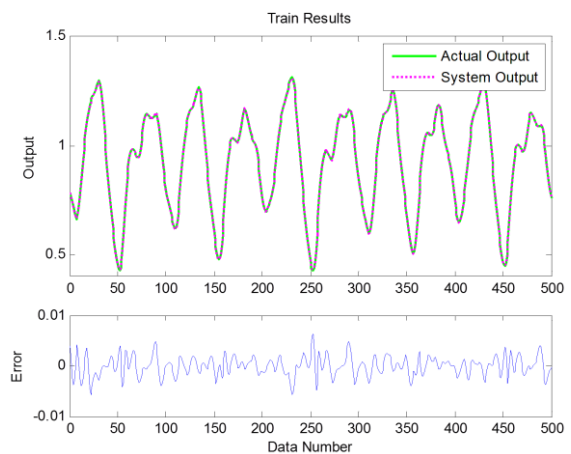


Fig. 2. Comparison of real output and fuzzy model output for train data

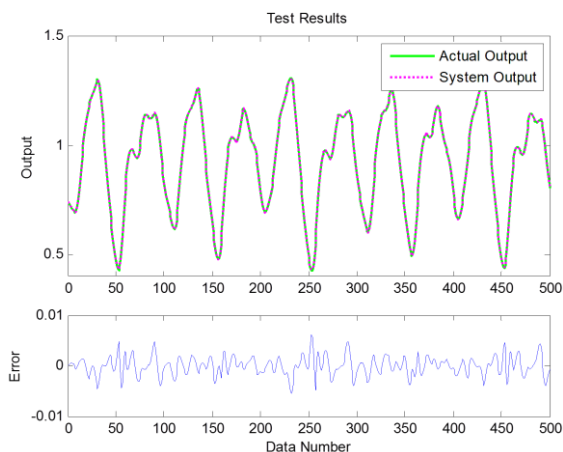


Fig. 3. comparison of real output and fuzzy model output for test data

Table 1: Comparison of results of different methods

Method	No. of rules	RMSE trainig	RMSE Test
PSO	10	1.100E-3	2.300E-3
ODE	10	7.580E-5	9.340E-5
HODEPO	10	4.140E-6	8.360E-6
[21]	4	0.0094	0.0061
[21]	10	-	0.0039
[22]	3	0.00588	0.00587

References

- [1] S. Chen, S.A. Billings, Representation of non-linear systems: the NARMAX model, *Int. J. Control* 49 (1989) 1013–1032.
- [2] H. Hujiberts, H. Nijmeijer, R. Willems, System identification in communication with chaotic systems, *IEEE Trans. Circuits Syst. I* 47 (6), 2000, pp. 800–808.
- [3] M. Adjrad, A. Belouchrani, Estimation of multi component polynomial phase signals impinging on a multisensor array using state-space modeling, *IEEE Trans. Signal Process.* 55 (1), 2007, pp. 32–45.
- [4] K. Watanabe, I. Matsuura, M. Abe, M. Kebota, D.M. Himmelblau, Incipient fault diagnosis of chemical processes via artificial neural networks, *AICHE J.* 35 (11) (1989) 1803–1812.
- [5] Y. Xie, B. Guo, L. Xu, J. Li, P. Stoica, Multistatic adaptive microwave imaging for early breast cancer detection, *IEEE Trans. Biomed. Eng.* 53 (8), 2006, pp. 1647–1657.
- [6] D. Chen, J. Wang, F. Zou, H. Zhang, W. Hou, “Linguistic fuzzy model identification based on PSO with different length of particles”, *Applied Soft Computing* 12, pp. 3390–3400, 2012.
- [7] C. F. Juang, and P. H. Chang, “Designing Fuzzy-Rule-Based Systems Using Continuous Ant-Colony Optimization”, *IEEE Trans on Fuzzy Syst.*, vol. 18, no. 1, pp. 138-149, Feb 2010.
- [8] C. F. Juang, C. Lo, “Zero-order TSK-type fuzzy system learning using a two-phase swarm intelligence algorithm”, *Fuzzy Sets and Systems* 159, 2910 – 2926, 2008.
- [9] R. Storn, System design by constraint adaptation and differential evolution, *IEEE Trans. Evol. Comput.* 3 (1999) 22–34.
- [10] J. Ilonen, J.K. Kamarainen, J. Lampinen, Differential evolution training algorithm for feed forward neural networks, *Neural Proc. Lett.* 17 (2003) 93–105.
- [11] E. Goldberg, J. Richardson, Genetic algorithms with sharing for multimodal function optimization, in: J. Richardson (Ed.), *Genetic Algorithms and their Applications (ICGA’87)*, 1987, pp. 41–49.
- [12] K. Kristinsson, G.A. Dumont, System identification and control using genetic algorithms, *IEEE Trans. Syst. Man Cybernet.* 22 (1992) 1033–1046.
- [13] J. Ilonen, J.K. Kamarainen, J. Lampinen, Differential evolution training algorithm for feed forward neural networks, *Neural Proc. Lett.* 17, 2003, pp.93–105.
- [14] R. Storn, System design by constraint adaptation and differential evolution, *IEEE Trans. Evol. Comput.* 3, 1999, pp. 22–34.
- [15] H.R. Tizhoosh, Opposition-based learning: a new scheme for machine intelligence, in: *Proc. Int. Conf. Comput. Intell. Modeling Control and Autom.*, vol. I, Vienna, Austria, 2005, pp. 695–701.
- [16] S. Rahnamayan, H.R. Tizhoosh, M.M.A. Salama, Opposition Versus Randomness in Soft Computing Techniques, *Elsevier J. Appl. Soft Comput.* 8 (March (2)), 2008, pp. 906–918.
- [17] S. Rahnamayan, H.R. Tizhoosh, M.M.A. Salama, Opposition-based differential evolution, *IEEE Trans. Evol. Comput.* 12 (1), 2008.
- [18] K. V. Price, R. M. Storn, and J. A. Lampinen, “Differential Evolution: A Practical Approach to Global Optimization”, (Kindle Edition). Springer, 2005.
- [19] H.R. Tizhoosh, Opposition-based learning: a new scheme for machine intelligence, in: *Proc. Int. Conf. Comput. Intell. Modeling Control and Autom.*, vol. I, Vienna, Austria, 2005, pp. 695–701.

6. Conclusion

In this article the hybrid optimization algorithm of differential evolution and particle swarm is introduced for designing the fuzzy rule base of a fuzzy controller. For a specific number of rules, a hybrid algorithm for optimizing all open parameters was used to reach maximum accuracy in training. The aim of using this algorithm was to set the parameters of the rule base in the zero-degree fuzzy system (TSK) in order to minimize the performance index (Root Mean Square Error (RMSE)).using the mentioned algorithm, the time-consuming process of parameter adjustment became a simple and quick task. The results show the suitable performance of the proposed model compared to other methods.

- [20] J. Kennedy, R. Eberhart, "Particle swarm optimization", in: Proc. IEEE Internat. Conf. Neural Networks, Perth, Australia, pp. 1942–1948, 1995.
- [21] C. F. Juang, C. W. Hung and C. H. Hsu, "Rule-Based Cooperative Continuous Ant Colony Optimization to Improve the Accuracy of Fuzzy System Design", IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 22, NO. 4, AUGUST 2014
- [22] W. Zhao, Q. Niu, K. Li and G. W. Irwin, "A Hybrid Learning Method for Constructing Compact Rule-Based Fuzzy Models", IEEE TRANSACTIONS ON CYBERNETICS, VOL. 43, NO. 6, DECEMBER 2013.
- [23] H.Hamidi, "A Model for Impact of Organizational Project Benefits Management and its Impact on End User", JOEUC, Volume 29, Issue 1, 2017, pp.50-64.
- [24] K.Mohammadi, H.Hamidi., Modeling and Evolution of Fault-Tolerant Mobile Agents in Distributed System.The Second IEEE and IFIP International Conference on wireless and Optical Communications Networks (WOCN 2005), March 6 –8, 2005.
- [25] S. A.Monadjemi, H.Hamidi, A.Vafaei. Analysis and Evaluation of a New Algorithm Based Fault Tolerance for Computing Systems. International Journal of Grid and High Performance Computing (IJGHPC), 4(1), 2012, pp. 37-51.
- [26] S. A.Monadjemi, H.Hamidi., A.Vafaei. "ANALYSIS AND DESIGN OF AN ABFT AND PARITY-CHECKING TECHNIQUE IN HIGH PERFORMANCE COMPUTING SYSTEMS" Journal of Circuits, Systems, and Computers (JCSC), JCSC Volume 21 Number 3, 2012.
- [27] A.Vafaei., S. A.Monadjemi, H.Hamidi., Evaluation of Fault Tolerant Mobile Agents in Distributed Systems. International Journal of Intelligent Information Technologies (IJIT), 5(1), 2009, pp.43-60.
- [28] A.Vafaei, S. A.Monadjemi, H.Hamidi "Evaluation and Check pointing of Fault Tolerant Mobile Agents Execution in Distributed Systems," Journal of Networks, VOL. 5, NO. 7. 2009.
- [29] H.Hamidi, A New Method for Transformation Techniques in Secure Information Systems Journal of Information Systems and Telecommunication, Vol. 4, No. 1, January-March 2016, pp. 19-26.
- [30] X. Ye, T.Sakurai, "Robust Similarity Measure for Spectral Clustering Based on Shared Neighbors," ETRI Journal, vol. 38, no. 3, June. 2016, pp. 540-550.
- [31] J.Wu, F. Ding, M. Xu, Z. Mo, A..Jin. Investigating the Determinants of Decision-Making on Adoption of Public Cloud Computing in E-government. JGIM, 24(3), 2016, pp. 71-89.
- [32] S. KUMAR, "Performance Evaluation of Novel AMDF-Based Pitch Detection Scheme," ETRI Journal, vol. 38, no. 3, June. 2016, pp. 425-434.
- [33] B. Shadloo, A. Motevalian, V. Rahimi-movaghar, M.A. Esmaeili, V. Sharifi, A. Hajebi, R. Radgoodarzi, M. Hefazi, A. Rahimi- Movaghar, Psychiatric Disorders Are Associated with an Increased Risk of Injuries: Data from the Iranian Mental Health Survey, Iranian Journal of Public Health 45(5), 2016, pp. 623-635.

Hodjatollah Hamidi born 1978, in shazand Arak, Iran, He got his Ph.D in Computer Engineering. His main research interest areas are Information Technology, Fault-Tolerant systems (fault-tolerant computing, error control in digital designs) and applications and reliable and secure distributed systems, Machine learning, Knowledge Discovery and Data Mining. Since 2013 he has been a faculty member at the IT group of K. N. Toosi University of Technology, Tehran Iran. Information Technology Engineering Group, Department of Industrial Engineering, K. N.Toosi University of Technology.

Atefeh Daraei born in Khorram Abad, Lorestan Iran. She received her B.Sc in Information Technology in 2011 from University College of Nabi Akram, Tabriz, Iran. She is currently an M.Sc student in Information Technology (E-Commerce), at K. N. Toosi University of Technology, Tehran Iran. Her research interests include Machine learning, Knowledge Discovery and Data Mining and Customer Relationship Management.

Automatic Facial Emotion Recognition Method Based on Eye Region Changes

Mina Navraan

Faculty of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran
m.navran@modares.ac.ir

Nasrollah Moghadam Charkari*

Faculty of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran
charkari@modares.ac.ir

Muharram Mansoorizadeh

Faculty of Electrical and Computer Engineering, Bu-Ali Sina University, Hamadan, Iran
mansoorm@basu.ac.ir

Received: 19/Apr/2015

Revised: 19/Mar/2016

Accepted: 19/Apr/2016

Abstract

Emotion is expressed via facial muscle movements, speech, body and hand gestures, and various biological signals like heart beating. However, the most natural way that humans display emotion is facial expression. Facial expression recognition is a great challenge in the area of computer vision for the last two decades. This paper focuses on facial expression to identify seven universal human emotions i.e. anger, disgust, fear, happiness, sadness, surprise, and neutral. Unlike the majority of other approaches which use the whole face or interested regions of face, we restrict our facial emotion recognition (FER) method to analyze human emotional states based on eye region changes. The reason of using this region is that eye region is one of the most informative regions to represent facial expression. Furthermore, it leads to lower feature dimension as well as lower computational complexity. The facial expressions are described by appearance features obtained from texture encoded with Gabor filter and geometric features. The Support Vector Machine with RBF and poly-kernel functions is used for proper classification of different types of emotions. The Facial Expressions and Emotion Database (FG-Net), which contains spontaneous emotions and Cohn-Kanade(CK) Database with posed emotions have been used in experiments. The proposed method was trained on two databases separately and achieved the accuracy rate of 96.63% for spontaneous emotions recognition and 96.6% for posed expression recognition, respectively.

Keywords: Facial Emotion Recognition; Gabor Filter; Support Vector Machine (SVM); Eye Region.

1. Introduction

Emotion recognition methods can be classified into different categories along a number of dimensions: speech emotion recognition vs. facial emotion recognition; machine learning methods vs. statistic methods. Furthermore, facial expression method can also be classified based on input data to a dynamic image sequences or static image.

Scientists and psychologists classify the emotions of people in seven different expressions: Anger, Disgust, Fear, Happiness, Neutral, Sadness, and Surprise. According to their studies, human emotions could be recognized via different ways like human expression, appearance, biological signals etc. Among them, analyzing speech and facial expressions to recognize human emotion are most popular approaches. Speech analysis is based on the vocal information whereas facial expression analysis deals with the changes and movement of the facial muscles.

According to [4], speech contributes to the emotions of the speaker by 50% for the spoken word and by 38% for the voice; whereas, the facial expression is affected by 55%. There are some interesting algorithms which have

been introduced to analyze speech or facial expressions in order to recognize human emotion. According to the previous studies, both the aforementioned approaches succeed in classifying the emotions. However, the facial expression approaches have revealed more precise results than the speech ones. Furthermore, some studies used hybrid methods (i.e. speech and facial expressions) which yield more accurate results [5]. Other type of hybrid methods for emotion recognition used two different models for facial and hand gesture recognition by separate classifier in advance. Then, the results of both classifiers are combined using a third classifier to recognize the emotion [51].

In general, emotion recognition is not a difficult task for the majority of human beings. However, it is a very challenging issue for computer based methods. The method that is designed for automatic analysis of facial expression is usually called Facial Expression Recognition method (FER). Some studies made their effort to detect facial expression based on action units (AUs) activation according to Facial Action Coding System (FACS) [24, 42, and 44].

Building a standardized database for FER method is very crucial since expressions can be posed (on purpose

* Corresponding Author

as a voluntary action) or spontaneous (unconsciously). Experimental results show that posed and spontaneous expressions vary widely in terms of configuration, their characteristics, temporal dynamics and timings [26].

Volitional facial movements originate in the motor cortex, whereas the involuntary facial actions, originate in the sub-cortical areas of the brain [27]. Therefore, while some facial movements might be easier to make voluntarily, many actions are done spontaneously. This may make it difficult to collect posed expression data of all possible facial movement. In this regard, many of the posed expressions that researchers have used in evaluation of their FER methods are highly overstated in compare with spontaneous expressions datasets.

Fig. 1. Show exaggeration in expressing fear expression in posed Cohn-Kanade dataset and its comparison with fear expression in spontaneous FG-Net dataset. The research community shifts their focus from posed to spontaneous expression recognition.

The main characteristic of FER method is that it should be effortless and efficient. It is also preferred to FER methods to be real-time which is especially important in both: human-computer interaction (HCI) and human-robot interaction applications. Other characteristics of an efficient FER methods are, the capability to work with video as well as images, the ability to simultaneously recognize spontaneous expressions and posed expression, robustness against the changes of lighting conditions and view angles, properly working in the presence of occlusions, invariant to facial hair, glasses, makeup etc.

More importantly, good FER methods are person independent and work on people from different cultures, different skin colors, and different ages (in particular, recognize expressions of both infants, adults and the elderly). Finally, they must be able to work with videos and images of different resolution.

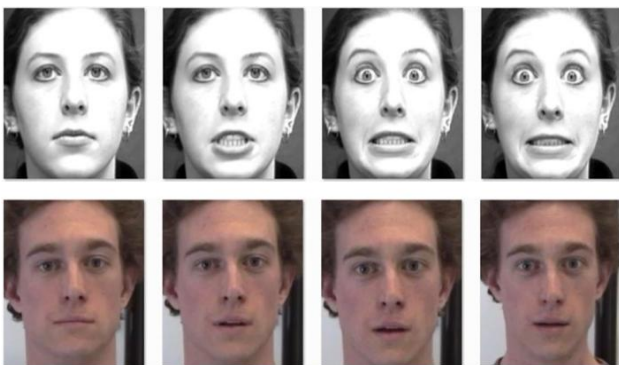


Fig. 1. Differences between the posed and spontaneous video frames in facial emotion recognition. First row: fear exaggeration in posed Cohn-Kanade dataset [55], Second row: fear expression in spontaneous FG-Net dataset[56].

Feature extraction is a crucial step in facial expression recognition and largely defines the effectiveness of the performance. Therefore, selecting a suitable type of features for facial expression representation plays an essential role in modeling FER methods. Features can be

classified into geometric and appearance (or texture) where each category has its own strength and weakness.

The final stage of any facial expression recognition method is the classification module based on the extracted features. Some works have been focused their works on proper classification by different classifiers. [15,16] studied static classifiers like the Naïve Bayes (NB), Tree Augmented Naïve Bayes (TAN), Stochastic Structure Search (SSS), and dynamic classifiers like Single Hidden Markov Models (HMM) and Multi-Level Hidden Markov Models (ML-HMM).

1.1 Motivation

According to the literature studies, the majority of facial expression recognition methods use Facial Action Coding System and action units (AUs) detection [52]. However, these methods generally suffer from time-intensive processes which result in high time complexity on video data. On the other hand, the importance of developing an approach to automatically differentiate between posed and spontaneous facial expression has been considered as one of the interesting issues within the last decade. In this regard, the main contributions of this work are summarized as follow;

It has been found that both eyes and mouth regions are parts of the face with the maximum amount of information in every facial expression [49]. So, we restrict the proposed method on extracting features from eye regions in facial expression recognition. This leads to a lower time complexity of each video frame with high accuracy rate. These regions are of great importance in differentiating between different facial expressions since the majority of facial activities particularly occur in upper face region.

The method is able to efficiently work on video and has the ability to recognize spontaneous and posed expression in low time complexity. To evaluate the method two different datasets (posed and spontaneous) have been employed in this paper. The Proposed FER method has been developed to process the video of facial emotion as well as to recognize the displayed actions in terms of seven basic emotions. In this regard, we have used both type of features (i.e. geometric and texture). We have employed spatial information of eyes for extracting geometric features whereas textural features are found by applying Gabor filter to eyes regions. Since features of eye region are extracted, feature space dimension will reduce dramatically. Furthermore, expression classifiers can be trained in much less time. Consequently, this leads to lower computational complexity where the emotional recognition rate is kept high in value. Finally, among different classification methods, we chose SVM because of its simplicity, flexibility and resistance to noise and outliers [42].

The rest of the paper is organized as follows: Section II reviews some related work. Our approach for FER method is presented in Section III and experimental results are reported in Section IV. Finally, Section V draws some conclusions.

2. Related Work

Numerous approaches have been presented to automatically analysis facial expressions from static image to dynamic image sequences. The early works have been summarized by [7, 8, 9, and 10]. Recent advances on automatic facial emotion recognition have been studied in [6]. Most of the existing FER methods employ various pattern recognition methods and emotions classification are based on 2D geometric and appearance facial features.

The typical procedure for almost all the emotion recognition method is depicted in Fig. 2. The first stage is face detection and tracking. It involves the process of locating face regions from the input data; align the face to a common coordinate method, and tracking the face region in every frames of video. The second stage is facial feature extraction and representation which is responsible for extracting and representing the facial variations caused by facial expressions. Finally, the last component is facial expression classification.

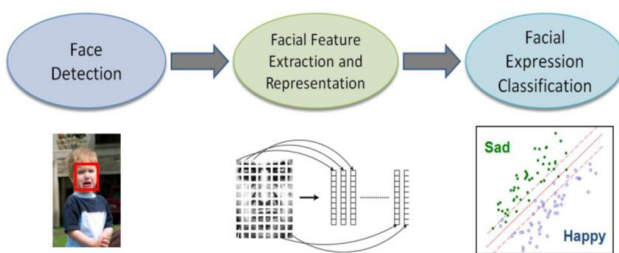


Fig. 2. Three main components of a typical facial expression recognition method [1].

2.1 Face Detection

Identifying the presence of faces on an image or frames of video and determining the locations and scales of them are the first step in facial emotion recognition. The accuracy of this stage is particularly significant in realistic condition.

A typical face detection algorithm performed the detection processes in the following steps. Given a set of training images acquired in a fixed pose (e.g. frontal, near-frontal, or profile view), histogram equalization or standardization is performed to dominate the effects of illumination. After this step, some face samples are extracted with knowledge based [28] or learning based methods [29, 30, and 31]. Here, the knowledge based methods model the face patterns by some definitive rules, such as facial components, face textures or skin color; the learning based methods model the face patterns by learning from a set of features with some discriminate functions [1]. With the extracted face patterns, the face detector method scans through the entire image to locate and detect the faces.

We have concentrated on the cascade based face detectors for its good performance. The AdaBoost based face detector by Viola and Jones is the most commonly used face detector in automatic face recognition and expression analysis [1].

The main distinctive idea of the Viola and Jones detector is to train a cascade classifier for haar-like rectangular features. The haar-like rectangular features can be efficiently computed with integral images [2], which facilitate the approach to gain real time detection approach. To further increase the detection speed while retaining the accuracy, AdaBoost was used to select the representative haar-like features [32]. In cascade based face detector, instead of training a single strong classifier, a number of weak classifiers are constructed and then combined into a cascade. The classifiers at the beginning of the cascade can efficiently reject the non-face regions, while the stronger classifiers later in the cascade simply need to classify the more face-like regions [1].

Several extensions have been made to Viola and Jones detector for detecting faces from different views [17, 18, and 32].

There are a few other face detection approaches in the recent literature, including the energy-based method that detects faces region and estimates the poses simultaneously [45], the face detectors using support vector machines (SVM) [46], the face detectors trained with only positive images [47], and the component-based face detector using Naive Bayes classifiers [48].

2.2 Feature Extraction

After face detection and tracking, the next step is to extract the representative and distinctive features of facial expression. Extracting effective features from the detected face image is crucial for successful facial expression recognition. Optimal feature extraction should be done in a way that minimizes within-class variations of expression while maximizing across-class variations. Features are generally divided into four groups:

Geometric features: Represent the shape and location of facial components or predefined facial feature points, which are extracted to form a feature vector to represent the face geometry [19, 20, 33, 34, 35]. Deformations between neutral state and current frame are parameters of facial expression [25].

The automatic and efficient detection and tracking of facial features point is still an open problem in many computer vision applications. This motivates the use of appearance (texture) based features for facial expression analysis.

Appearance features: Appearance based features measure the appearance changes which are mainly based on texture analysis. Typical appearance-based approaches use different image filters such as Gabor or linear filters, to extract feature vectors such as texture, correlation and gradient from whole face, specific regions, or regions of interest (ROI).

Gabor filters [50] have been found to be powerful in face expression analysis and widely used to extract the facial appearance changes as a set of multi scales and multi orientation coefficients for analyzing the texture. In problems with time and memory limitation another method called Local Binary Patterns (LBP) gains more popularity in facial texture analysis [38, 39, 40, and 41].

Fusion the features: Geometric feature capture macro-variation of facial structure while appearance-based approaches are capable of identifying local subtle changes [3]. By combining these two kinds of features the performance of facial emotion recognition method would be improved. Active Appearance Model (AAM) is a combination of appearance and texture information to construct a parameterized model of facial features. Relationships between AAM displacement and the image difference would be analyzed for facial expression detection. AAM has been widely used for facial feature extraction.

Temporal features: The main idea of extracting this kind of features is achieving the temporal information about the movements of facial components in video frames for facial expression [22].

2.3 Classification

The final stage of the FER method is based on machine learning theory; precisely it is the classification module. The input to the classifier is a set of features which has been retrieved from face region in feature extraction stage. The feature vector is formed to describe the facial expression. The first part of Classification module is training. The training set of classifier consists of labeled data. Once the classifier is trained, it can recognize input images by assigning them a particular expression class label.

All approaches for classification of facial expression can be divided into two groups: frame based recognition which only relies on a single frame; image sequence based approaches exploited the temporal behaviors of facial expressions.

In different researches, various classifiers have been applied such as , Neural Network, Bayesian Network (BN), Support Vector Machine (SVM), rule-based classifiers, and Hidden Markov Models (HMM) [21]. Some studies like [24, 42, and 44] in facial expression analysis, made their effort to classify action units (AU). Bartlett. et al. [42, 44] studied AU activations based on Gabor filter responses on whole face region. Other studies like [14, 23, 26, and 43] classified each emotional state based on the extracted features.

In [43] Gabor filters with different frequencies and orientations were applied on face region, and SVM classification method used for recognizing four basic emotion. Zhan. et al. [26] proposed a method to recognize seven basic spontaneous expressions using FG-Net dataset. For feature selection, Gabor filters with different frequencies and orientations were applied only to a set of facial landmark positions. [23] Used Adaboost to choose a subset of feature extracted from Gabor filters. The SVMs are then used for classification. Developed from statistical learning theory, is a widely used classifier for facial expression classification.

3. Proposed Method

In this section, we describe the details of our facial expression recognition method. The section is divided into three subsections. In Section A, we describe how face image are automatically processed for feature extraction. In section B, those features which have efficiently been employed to extract human emotions will be discussed. Finally, in section C, we adopt SVM for classifying anger, disgust, fear, happy sad, and surprise expressions. The main components of the proposed FER method are shown in Fig. 3.

3.1 Face Detection

In the first stage of FER method, after converting video to image sequences, facial regions are detected and tracked. For each frame of the video, an approximate region of a face is found by the Viola–Jones face detector. In order to perform facial emotion recognition, some processes are done in preprocessing step. Firstly, face region on each frame is extracted by means of Viola–Jones algorithm from the background and resized to 860*860 pixel. Second, key frame detector is performed to extract some video frames. Third, tracking process is conducted. Forth, Viola–Jones algorithm is employed for detecting the eyes region on key frames. Fig. 4. Show the extracted face and the detected eye regions, respectively.

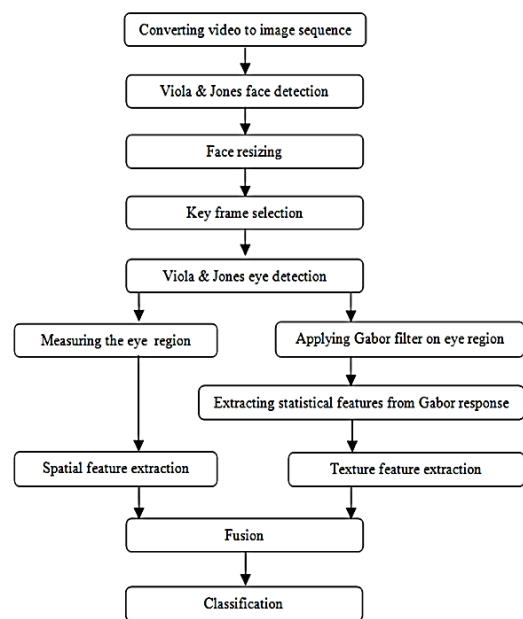


Fig. 3. overview of the proposed FER method

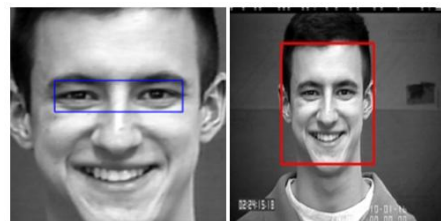











Fig. 4. output of Viola–Jones detector for detecting face and eyes regions in our work.

3.2 Feature Vector

The detected regions are used to extract two types of features: geometric and texture. Previous facial expression recognition methods typically used either geometric or appearance (texture) features [10]. We have employed both features since they provide complementary information. Two features are extracted separately and combined in the classification stage.

To define geometric features, after detecting eyes regions, their sizes are found. The changes in size of eyes regions form geometric feature and can be coped well with the variations in skin patterns or dermatoglyphics. One example of size changes in eye region during disgust expression in our work is shown in TABLE I.

Table 1. one example of size changes in eye region during disgust expression from Cohn dataset in our work.

#	Eye region	height	width
1		137	560
2		139	568
3		141	575
4		140	570
5		138	560
6		144	585
7		144	587
8		143	582
9		143	583

To extract texture features, Gabor filter has been applied. We convolve the extracted eye regions with Gabor filter banks with 1 spatial frequency that properly provides edges, Wrinkles and furrows, and 9 different orientations from 0 to 180 degrees with a 20degree steps. An example of Gabor response in one of the experiments is shown in Fig. 5.



Fig. 5. gabor response on eye region in our work

The Gabor responses from the whole image as features lead to huge dimensionality problem. To cope with these problem, different approaches like feature selection and employing Gabor filter on Regions-of-Interest (ROIs) have been suggested [11, 12, 36 and 37]. In this work, we have found the response of Gabor filter for every detected eye region.

To decrease dimensionality of feature space we use two strategies in Gabor feature extraction:

1) We restrict Gabor filter bank to one spatial frequency which properly provides edges, Wrinkles and furrows instead of using different spatial frequencies.

The reason is that, all the spatial frequencies don't provide useful information on image. The low spatial frequencies result in blurred edges in Gabor response where the high spatial frequencies lead to some broken and not continuous edges of Gabor response. Examples of Gabor filter responses with low and high spatial frequencies are shown in Fig. 6.

By applying Gabor filter banks with 9 different spatial frequencies from $1/2$ to $1/32$ in units, and 9 different orientations from 0 to 180 degrees with a 20 degree steps from the eye region with the size of $144 * 585$ pixels, dimensionality becomes huge ($9 * 9 * 144 * 585 = 6,823,440$). By restricting the filter bank to 1 spatial frequency and 9 orientations, the dimensionality of the features reduce to $1 * 9 * 144 * 585 = 758,160$ whilst better response with the most useful information would be obtained.

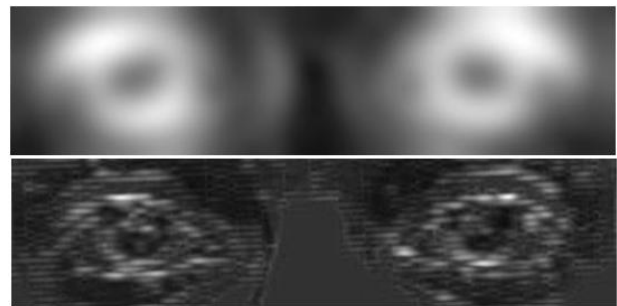


Fig. 6. gabor response with low spatial ferequcy in above picture and gabor response with high spatial ferequcy in below picture.

2) After getting the response of Gabor filter, we extract some statistical information such as histogram, mean, and standard deviation from every response of Gabor filter. This process reduces the dimensionality of the features dramatically.

Accordingly, histogram is defined as the number of pixels of each brightness level in the image. Mean is defined

as the sum of the pixel values in Gabor response divided by the number of values. Finally, standard deviation is used to measure the amount of variation from the average value.

Horizontal & vertical profile: Data in a Gabor response matrix can be profiled. We get the vertical, horizontal or arbitrary profiles of the matrix data. The columns summation is defined as vertical profile. Similarly, horizontal profile is the rows summation.

3.3 Classification

In the final step, classification is employed. Facial emotion recognition requires classifier training with a set of images with particular emotions displayed. In the proposed FER method, the Support Vector Machine (SVM) is used as a classifier. We choose SVM because of its simplicity, flexibility and resistance to noise and outliers and having the advantage of solving non-linear, or high dimensional classification problem. The SVM is a classifier which receives labeled training data and transforms it into higher dimensional feature space. The SVM classification method computes separating hyper plane between classes. SVM determines the best separation between classes with respect to margin maximization. Regarding the kernels tested, we compare the most commonly used ones in the literature i.e. polynomial and rbf.

4. Experimental Results

There is still no standard method for evaluation of automatic FER methods. However, the majority of studies have reported the performance of their method on one or more available emotional face datasets. Many works have reported the results of their method on Cohn-kanade, FG-Net datasets, or both.

To investigate the efficiency of the proposed FER method, two facial expression datasets are used in this paper. The first dataset is the FG-Net Facial Expression which consists of MPEG video files with spontaneous emotions recorded. It contains some expression examples gathered from 18 different subjects (9 female and 9 male).

Facial expressions of subjects are captured during watching videos. The dataset formed from FG-Net dataset consists of 1542 images of seven states, neutral and emotional (anger, disgust, fear, happy, sadness, and surprise). Accordingly, 222 frames for neutral, 221 frames for anger expression, 222 frames for disgust expression, 220 frames for fear expression, 218 frames for happiness, 220 frames for sadness, and 219 frames for surprise expression are used.

Second dataset is the Cohn-Kanade Facial Expression Dataset which contains image sequences displayed by 97 posers. The sequence displays the emotion from the start (neutral state) to the peak. The dataset contains 1532 images divided into seven classes: neutral, anger, disgust, fear, happiness, sadness, and surprise. Accordingly, 220 frames for neutral, 218 frames for anger, 218 frames for disgust, 218 frames for fear, 220 frames of happiness, 220 frames of sadness, and 218 frames for surprise expressions are used.

TABLE II. gives an overview of the existing methods where Gabor filter response is part of their extracted features, for expression analyzes.

Our work consists of two set of experiments: binary classification and multiclass classification. Evaluation criteria for classification problem are:

F-Measure: Combination of precision and recall by their harmonic mean. It varies between zero and one; when the value is closed to one it indicates a better performance of classification (1).

$$F - \text{Measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (1)$$

False positive rate: It is also known as the false alarm ratio refers to the rate of occurrence of positive test results in samples known to be negative for which an individual is being tested (2). TN is the number of negative results which have detected correctly as negative results.

$$\text{false positive rate} = \frac{FP}{FP + TN} \quad (2)$$

Table 2. Comparison of some facial expression analyses

Reference	Database	Feature type	Classification	Performance
2001 Tian.[13]	Cohn-Kanade	Permanent features: Gabor response on nasolabial region, nasal root, and eye corner transient features: canny edge detection	ANN	Recognition of upper face action units: 96.4% Recognition of lower face action units: 96.7% Average to independent databases:93.3%
2003 Bartlett.[23]	Cohn-Kanade	Gabor filter responses selected by Adaboost for each expression	SVM	Across 7 prototypic expression: AdaSVM+RBF: 90.7%
2008 Kotsia.[14]	Cohn-Kanade	Gabor filter output vectors on whole face have been concatenated to form a new long feature vector	SVM&MLP	Across 6 prototypic expression: Cohn : 91.6%,
2008 Zhan.[26]	FG-Net	Gabor filters with different frequencies and orientations are applied only to a set of facial landmark positions	SVM	Accuracy rate for 7 expressions : 82%
2010 Wu.[53]	Cohn-Kanade	Gabor motion energy filters (GME) output vectors on whole face	SVM	Average ROC over 6 expressions: on onset sequences classification: 78.56% on apex sequences classification: 97.81%
2014 L.Zhang.[54]	Cohn-Kanade	Gabor filters with different frequencies and orientations are applied on face region	SVM	Across 7 prototypic expression Cohn : 95.3%,
This work	Cohn-Kanade FG-Net	Gabor: statistical information from every response of Gabor filter Geometry: size of eye region	SVM	Average Measure for classifying 7 prototypic expression: Cohn: 96.6% FG-Net : 96.63%

4.1 Binary Classification

In the first experiment, we have conducted our method to detect emotional state from neutral face. So, we have used binary classification to determine the efficiency of the method.

In each video, the initial frames start from neutral state and the remaining frames rise to a complete emotional states. The results of the experiment are shown in TABLE III, IV, V, VI, VII, and VIII. In binary classification, the best result on posed Cohn-Kanade and spontaneous FG-Net datasets found for disgust expression with 98.64% and 99.22% correctly classified instances, respectively. The worst result on posed Cohn-Kanade dataset is for anger expression with 97.09% correctly classified instances. Similarly, on spontaneous FG-Net, the worst result is for surprise expression with 97.94% correctly classified instances.

Table 3. SVM classification of neutral & anger expression with poly & rbf kernels

	Kernel	FP Rate	%F-Measure	% average of Correctly Classified	dataset
neutral	Poly	0.05	96.67	96.68	Cohn-Kanade
anger		0.017	96.66		
average		0.033	96.67		
neutral	RBF	0.058	97.2	97.09	
anger		0	97		
average		0.029	97.1		
neutral	Poly	0.031	98.5	98.44	FG-Net
anger		0	98.4		
average		0.016	98.4		
neutral	RBF	0.023	98.9	98.83	
anger		0	98.8		
average		0.012	98.8		

Table 4. SVM classification of neutral & disgust expression with poly & rbf kernel

	Kernel	FP Rate	%F-Measure	% average of Correctly Classified	dataset
neutral	Poly	0.037	98.2	98.18	Cohn-Kanade
disgust		0	98.1		
average		0.019	98.2		
neutral	RBF	0.028	98.7	98.64	
disgust		0	98.6		
average		0.014	98.6		
neutral	Poly	0.031	98.5	98.45	FG-Net
disgust		0	98.4		
average		0.016	98.4		
neutral	RBF	0.016	99.2	99.22	
disgust		0	99.2		
average		0.008	99.2		

Table 5. SVM classification of neutral & fear expression with poly & rbf kernel

	Kernel	FP Rate	%F-Measure	% average of Correctly Classified	dataset
neutral	Poly	0.035	98.2	98.23	Cohn-Kanade
fear		0	98.2		
average		0.017	98.2		
neutral	RBF	0.044	97.4	97.34	
fear		0.009	97.3		
average		0.026	97.3		
neutral	Poly	0.036	97.9	97.84	FG-Net
fear		0.007	97.8		
average		0.021	97.8		
neutral	RBF	0.029	98.2	98.2	
fear		0.007	98.2		
average		0.018	98.2		

Table 6. SVM classification of neutral & happiness expression with poly & rbf kernel

	Kernel	FP Rate	%F-Measure	% average of Correctly Classified	dataset
neutral	Poly	0.035	98.2	98.23	Cohn-Kanade
happy		0	98.2		
average		0.017	98.2		
neutral	RBF	0.043	97.8	97.79	
happy		0	97.8		
average		0.021	97.8		
neutral	Poly	0.026	98.7	98.68	FG-Net
happy		0	98.7		
average		0.013	98.7		
neutral	RBF	0.033	98.4	98.35	
happy		0	98.3		
average		0.016	98.3		

Table 7. SVM classification of neutral & sadness expression with poly & rbf kernel

	Kernel	FP Rate	%F-Measure	% average of Correctly Classified	dataset
neutral	Poly	0.046	97.3	97.25	Cohn-K
sad		0.009	97.2		
average		0.028	97.2		
neutral	RBF	0.056	96.4	96.33	
sad		0.018	96.2		
average		0.037	96.3		
neutral	Poly	0.014	99	99.06	FG-Net
sad		0.005	99.1		
average		0.009	99.1		
neutral	RBF	0.009	98.8	98.83	
sad		0.015	98.9		
average		0.012	98.8		

Table 8. SVM classification of neutral & surprise expression with poly & rbf kernel

	Kernel	FP Rate	%F-Measure	% average of Correctly Classified	dataset
neutral	Poly	0.046	97.8	97.73	Cohn-Kanade
surprise		0	97.7		
average		0.023	97.7		
neutral	RBF	0.046	97.8	97.73	
surprise		0	97.7		
average		0.023	97.7		
neutral	Poly	0.049	97.2	97.12	FG-Net
surprise		0.008	97.1		
average		0.029	97.1		
neutral	RBF	0.033	98	97.94	
surprise		0.008	97.9		
average		0.02	97.9		

4.2 Multiclass Classification

Some studies like [24, 42, and 44] in facial expression analysis, made their effort to classify action units (AU). Other studies like [14, 23, 26, and 43] classified each emotional state based on the extracted features.

In this experiment, our objective is to detect and to distinguish anger, disgust, fear, happiness, sadness, surprise and neutral frames from each other. So, we face with 7- label classification problem. We have studied two popular strategies to solve such classifications problems; One-vs.-the-Rest and One-vs.-One.

One-Vs.-The-Rest: The strategy consists in fitting one classifier per class. For each classifier, the class is fitted against all the other classes. In addition to its computational efficiency, another advantage of this

approach is its interpretability. In this strategy, for each class, we have one and only one classifier. So this feature makes it possible to gain the knowledge about the class by inspecting the corresponding classifier.

One-Vs-One: It constructs one classifier per pair of classes. The class which receives the most votes at prediction step would be selected. This method is beneficial for algorithms such as kernel ones which don't scale well with $n_samples$. This is because each individual learning problem only involves a small subset of data whereas, with one-vs.-the-rest, the complete dataset is used $n_classes$ times.

We have employed One-Vs-One scheme for solving our multi-classification problem. The results of the experiments are shown in TABLE IX, X.

Table 9. SVM classification of neutral, anger, disgust, fear, happiness, sadness & surprise expression with poly & rbf kernel in cohn dataset.

	Kernel	FP Rate	%F-Measure	% average of Correctly Classified	dataset
neutral	Poly	0.026	86.9	95.76	Cohn-Kanade
anger		0.002	98.9		
disgust		0.002	98.4		
fear		0.006	96.3		
happy		0.004	96.6		
sad		0.006	96.8		
surprise		0.004	96.8		
average		0.007	95.8		
neutral		RBF	0.02		
anger	0.001		99.5		
disgust	0.002		98.4		
fear	0.004		97.2		
happy	0.004		97.5		
sad	0.005		97.5		
surprise	0.005		97.2		
average	0.006		96.6		

Table 10. SVM classification of neutral, anger, disgust, fear, happiness, sadness & surprise expression with poly & rbf kernel in fg-net dataset.

	Kernel	FP Rate	%F-Measure	% average of Correctly Classified	dataset
neutral	Poly	0.015	92	95.91	FG-Net
anger		0.005	97.8		
disgust		0.002	97.7		
fear		0.011	94.1		
happy		0.002	98.4		
sad		0.004	97.5		
surprise		0.008	94		
average		0.007	95.9		
neutral		RBF	0.012		
anger	0.005		98		
disgust	0.002		98.2		
fear	0.008		95.7		
happy	0.004		98.2		
sad	0.002		97.9		
surprise	0.007		95.7		
average	0.006		96.6		

To validate the proposed FER method we use 10-fold cross validation method. In 10-fold cross validation, the original sample is randomly partitioned into 10 equal size subsamples. Of the 10 subsamples, a single subsample is retained as the validation data for testing the model, and the remaining 9 subsamples are used as training data. The cross-validation process is then repeated 10 times, with each of the 10 subsamples used exactly once as the

validation data. The 10 results from the folds would then be averaged to produce a single estimation.

The confusion matrices in TABLE XI, XII indicates the performance of 10-fold cross validation on the posed and spontaneous datasets, respectively. The rows represent the emotion that was intended by the subject and the columns represent the emotion determined by the classifier. The total number of images that have been classified for each intended emotion is shown on the last column of the table.

For Cohn-kanade dataset, the overall percentages of correctly classified emotions were: 89.54% for intended neutral, 99.54% for anger, 97.71% for disgust, 96.79% for fear, 97.27% for happiness, 98.18% for sadness, and 97.25% for surprise.

Similarly, for FG-Net dataset, the overall percentages of correctly classified emotions were: 92.79% for intended neutral, 98.64% for anger, 97.75% for disgust, 96.36% for fear, 98.62% for happiness, 96.82% for sadness, and 95.43% for surprise.

Our proposed FER method has some drawbacks in spatial feature extraction. Despite the good performance of Viola Jones eye detection algorithm, this method doesn't provide the changes in the size of eye regions as expected. For example, during disgust expression the height of eye region decreased. However, as we have shown in TABLE I. viola Jones doesn't show height changes as expected.

5. Conclusion

The research community shifts their focus from posed to spontaneous expression recognition. Since spontaneous expression are more natural in compare with posed expression. Recognizing expressions which contain exaggeration in expressing facial emotion aren't useful in the real world applications. In future robust spontaneous expression recognizers will be developed and deployed in real-time methods and used in building many real-world applications especially in HCI applications.

The fully automated nature of the FER method allows us to perform facial expression analysis in many applications of HCI. By creating computer methods that can understand emotion, we enhance the communications that exists between humans and computers.

In this paper we introduced a fully automatic facial expression recognition method. Contrary to other approaches in emotional recognition which employ the whole face region or some interesting regions of faces, we restrict our FER method to analyze human emotional states based on eye region changes. We showed that how this region plays significant role for further analysis of facial expression process. Since the features of eye region are extracted, the size of feature dimension reduces dramatically. Therefore, expression classifiers can be trained in much less time. Consequently, this leads to lower computational complexity where the emotional

recognition rate is kept high in value. We evaluate the proposed FER method on posed (Cohn-Kanade) and spontaneous (FG-Net) datasets. For feature extraction we use a fusion based approach. The classification rate on spontaneous dataset increased to 96.63%. On posed dataset the accuracy of FER method is 96.6%. The results are comparable to the existing facial expression recognition methods.

Proper recognition rate of the proposed FER method is because of extracting useful informative features during expressing emotions. We achieve this goal by restricting our method to analyze eye region changes during expressing facial emotion.

Table 11. The confusion matrices of 10-fold cross validation for multi-class expression classifier when tested on cohn-k dataset.

CLASSIFIED AS →	Neutral	Anger	disgust	Fear	Happy	Sad	Surprise	total
Intended Neutral	197	1	2	4	5	7	4	220
Intended Anger	1	217	0	0	0	0	0	218
Intended Disgust	5	0	213	0	0	0	0	218
Intended Fear	5	0	0	211	0	0	2	218
Intended Happy	6	0	0	0	214	0	0	220
Intended Sad	4	0	0	0	0	216	0	220
Intended Surprise	5	0	0	1	0	0	212	218

Table 12. The confusion matrices of 10-fold cross validation for multi-class expression classifier when tested on fg_net dataset.

CLASSIFIED AS →	Neutral	Anger	disgust	Fear	Happy	Sad	Surprise	total
Intended Neutral	206	4	1	1	5	1	4	222
Intended Anger	3	218	0	0	0	0	0	221
Intended Disgust	3	0	217	1	0	0	1	222
Intended Fear	3	0	1	212	0	0	4	220
Intended Happy	2	0	0	1	215	0	0	218
Intended Sad	3	1	1	2	0	213	0	220
Intended Surprise	2	1	0	6	0	1	209	219

References

- [1] Kaimin Yu, "Toward Realistic Facial Expression Recognition", Doctoral dissertation, School of Information Technologies at The University of Sydney, 2013.
- [2] P. Viola, M.J. Jones, "Robust Real-Time Face Detection", International Journal of Computer Vision, 57(2):137{154, May 2004.
- [3] Li, Peiyao, "Adaptive feature extraction and selection for robust facial expression", Master of Engineering by Research thesis, University of Wollongong, School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, 2010. <http://ro.uow.edu.au/theses/3268>
- [4] A. Mehrabian, "Communication without words". Psychology today, vol.2, no.4, pp.53-56, 1968.
- [5] M. Mansoorizadeh, N. Moghaddam Charkari, "Multimodal information fusion application to human emotion recognition from face and speech" Springer Science, Multimedia Tools Appl 2010.
- [6] Z. Zeng, M. Pantic, G.I. Rosman, and T.S. Huang. "A Survey of Affection Recognition Methods: Audio, Visual, and Spontaneous Expressions". IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(1):39{58, January 2009.
- [7] A. Samal, A. Lyengar. "Automatic recognition and analysis of human faces and facial expressions: a survey". Pattern Recognition, 25(1):65{77, January 1992.
- [8] M. Pantic, J.M. Rothkrantz. "Automatic Analysis of Facial Expressions: The State of the Art". IEEE Transaction on Pattern Analysis and Machine Intelligence, 22(12):1424{1445, December 2000.
- [9] B. Fasel, J. Luettin. "Automatic facial expression analysis: a Survey". Pattern Recognition, 36(1):259{275, January 2003.
- [10] Y. Tian, Takeo Kanade, F. Cohn, "Facial Expression Analysis. In Handbook of Face Recognition", Springer New York, 2005.
- [11] Y. Tian, T. Kanade, and J. Cohn. "Eye-state action unit detection by Gabor wavelets". In Proceedings of International Conference on Multi-modal Interfaces, pages 143{150, 2000.
- [12] Y. Tian, T. Kanade, and J. Cohn. "Evaluation of gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity". In IEEE International Conference on Automatic Face and Gesture Recognition, page 229, 2002.
- [13] Y. Tian, T. Kanade and J. Cohn, "Recognizing Action Units for Facial Expression Analysis," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 2, pp. 97-115, 2001.
- [14] I. Kotsia, I. Buciu and I. Pitas, "An Analysis of Facial Expression Recognition under Partial Facial Image Occlusion," Image and Vision Computing, vol. 26, no. 7, pp. 1052-1067, July 2008.
- [15] I. Cohn, N. Sebe, F. Cozman, M. Cirelo, and T. Huang, "Learning Bayesian Network Classifiers for Facial Expression Recognition Using Both Labeled and Unlabeled Data," Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 1-595-I-604, 2003.
- [16] I. Cohn, N. Sebe, A. Garg, L.S. Chen, and T.S. Huang, "Facial Expression Recognition From Video Sequences: Temporal and Static Modeling", Computer Vision and Image Understanding, vol. 91, pp. 160-187, 2003.

- [17] M. Jones, P. Viola. "Fast Multi-view Face Detection", Technical report, Mitsubishi Electric Research Laboratories, 2003.
- [18] J. Wu, S.C. Brubaker, M. Mullin, and J. Rehg. "Fast asymmetric learning for cascade face detection", IEEE Transaction on Pattern Analysis and Machine Intelligence, 30(3):369{382, 2008.
- [19] M. Pantic, M. Rothkrantz. "Facial action recognition for facial expression analysis from static face images", IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 34(3):1449{1461, June 2004.
- [20] M. Pantic, I. Patras, "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences", IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 36(2):433{449, April 2006.
- [21] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski. "Classifying facial actions", IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(10):974{989, 1999.
- [22] S. Koelstra, M. Pantic, "A Dynamic Texture-Based Approach to Recognition of Facial Actions and Their Temporal Models", IEEE TRANSACTIONS on Pattern Analysis and Machine Intelligence, VOL. 32, NO. 11, NOVEMBER 2010.
- [23] M.S. Bartlett, G. Littlewort, I. Fasel, and R. Movellan, "Real Time Face Detection and Facial Expression Recognition: Development and Application to Human Computer Interaction," Proc. CVPR Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction, vol. 5, 2003.
- [24] A. Savran, B. Sankur, M. Taha Bilge, "Regression-based intensity estimation of facial action units", Image and Vision Computing 2012.
- [25] M. Pantic and L. Rothkrantz, "Expert System for Automatic Analysis of Facial Expression", Image and Vision Computing J., Vol. 18, No. 11, p. 881-905, 2000.
- [26] C. Zhan, W. Li, P. Ogunbona, F. Safaei, A Real-Time Facial Expression Recognition System for Online Games, International Journal of Computer Games Technology, Volume 2008.
- [27] Ekman, P. and Friesen, W. V, 'Detecting deception from body or face.' Journal of Nonverb Behav., (1974).
- [28] D. Reisfeld and Y. Yeshurun, "Robust detection of facial features by generalized symmetry", in Proceeding of 11th IAPR International Conference on Pattern Recognition, PP. 117-120, 1992.
- [29] G.A. Ramirez and O. Fuentes, "Face detection using combination of classifiers", in Proceeding of The Canadian Conference on Computer and Robot Vision, PP. 610-615, 2005.
- [30] P. Zhang, "A video-based face detection and recognition system using cascade face verification modules", in Proceeding of 37th IEEE Applied Imagery Pattern Recognition Workshop, PP. 1-8, 2008.
- [31] B. Yipp, "Face and eye rectification in video conference using artificial neural network", in IEEE International Conference on Multimedia and Expo, PP. 690-693, 2005.
- [32] P. Viola, M.J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance", in Proceeding of Ninth IEEE International Conference on Computer Vision, Vol. 2, PP. 734-741, 2002.
- [33] Z. Zhang, M.J. Lyons, M. Schuster, and S. Akamatsu. "Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron". In IEEE International Conference on Automatic Face & Gesture Recognition, 1998.
- [34] R.E. Kaliouby and P. Robinson. "Real-time inference of complex mental states from facial expressions and head gestures". In IEEE CVPR Workshop on Real-time Vision for Human-Computer Interaction, 2004.
- [35] Michel Valstar, I. Patras, and M. Pantic. "Facial Action Unit Detection using Probabilistic Actively Learned Support Vector Machines on Tracked Facial Point". In IEEE Conference on Computer Vision and Pattern Recognition, page 76, 2005.
- [36] G. Ford. "Tutorial on gabor filters". Technical report, Machine Perception Lab, Institute of Neural Computation, 2002.
- [37] Ying-Li Tian, T. Kanade, and J. Cohn. "Eye-state action unit detection by Gabor wavelets". In Proceedings of International Conference on Multi-modal Interfaces, pages 143{150, 2000.
- [38] T. Ojala, M. Pietikainen, and D. Harwood. "A comparative study of texture measures with classification based on featured distribution". Pattern Recognition, 29(1):51{59, 1996.
- [39] T. Ahonen, A. Hadid, and M. Pietikainen. "Face recognition with local binary Patterns". In European Conference on Computer Vision, 2004.
- [40] X. Feng, A. Hadid, and M. Pietikainen. "A coarse-to-fine classification scheme for facial expression recognition". In International Conference on Image Analysis and Recognition, 2004.
- [41] Caifeng Shan, Shaogang Gong, and Peter W. McOwan. "Facial expression recognition based on Local Binary Patterns: A comprehensive study". Image and Vision Computing, 27(6):803{816, May 2009.
- [42] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. "Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior". In Computer Vision and Pattern Recognition, 2005.
- [43] P. Lekshmi. V. I., Dr. M. Sasikumar, "Analysis of Facial Expression using Gabor and SVM", International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009.
- [44] M.S. Bartlett, G. Littlewort, M.G. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Fully Automatic Facial Action Recognition in Spontaneous Behavior," Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition (AFGR '06), pp. 223-230, 2006.
- [45] M. Osadchy, Y. LeCun, and M. Miller. "Synergistic face detection and pose estimation with energy-based models". Journal of Machine Learning Research, 8:1197{1215, May 2007.
- [46] B. Heisele, T. Serre, and T. Poggio. "A component-based framework for face detection and identification." International Journal of Computer Vision, 74(2):167{181, 2007.
- [47] D. Keren, M. Osadchy, and C. Gotsman. "Antifaces: A novel fast method for image detection". IEEE Transaction on Pattern Analysis and Machine Intelligence, 23(7):747{761, 2001.
- [48] H. Schneiderman and T. Kanade. "Object detection using the statistics of parts". International Journal of Computer Vision, 56(3):151{177, 2004.
- [49] M. Pardo and A. Bonafonte, "Facial animation parameters extraction and expression recognition using Hidden Markov Models," Signal Processing: Image Communication, vol. 17, pp. 675-688, 2002.
- [50] M.J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. "Coding facial expression with Gabor wavelets". In IEEE International Conference on Automatic Face & Gesture Recognition, pages 200{205, 1998.

- [51] Priya Metri, Jayshree Ghorpade, Ayesha Butalia, " Facial Emotion Recognition Using Context Based Multimodal Approach", International Journal of Emerging Sciences, 2(1), 171-182, March 2012.
- [52] Ekman, Paul, and Wallace V. Friesen. "Manual for the facial action coding system". Consulting Psychologists Press, 1978.
- [53] Wu, Tingfan, Marian S. Bartlett, and Javier R. Movellan. "Facial expression recognition using gabor motion energy filters." In Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, pp. 42-47. IEEE, 2010.
- [54] Zhang, Ligang, Dian Tjondronegoro, and Vinod Chandran. "Random Gabor based templates for facial expression recognition in images with facial occlusion." Neurocomputing 145 (2014): 451-464.
- [55] Kanade, Takeo, Jeffrey F. Cohn, and Yingli Tian. "Comprehensive database for facial expression analysis." In Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on, pp. 46-53. IEEE, 2000.
- [56] Wallhoff, Frank. "Facial expressions and emotion database." Technische Universität München (2006).

Mina Navran received her B.Sc. from Islamic Azad University of Qazvin in 2008 and M.Sc. in Computer Engineering from Tarbiat Modares University of Tehran in 2014, respectively. Her main interests are Affect Computing, image analysis, and data mining.

Nasrollah Moghaddam Charkari received his B.Sc. in Computer Engineering from Shaheed Beheshti University of Tehran in 1987, and his M.Sc. and Ph.D. in Computer Science and Information System Engineering from Yamanashi University of Japan in 1992 and 1995, respectively. He is currently an Associate Professor at School of Electrical and Computer Engineering at Tarbiat Modares University of Tehran, Iran. His main research interests are image analysis and understanding, complex networks, Parallel Algorithms and Processing, and bioinformatics.

Muharram Mansoorizadeh received his B.Sc. from University of Isfahan in 2000, and his M.Sc. and Ph.D. from Tarbiat Modares University of Tehran in 2002, and 2008 in Computer Science, respectively. He is currently an Assistant Professor at Bu-Ali Sina university of Hamadan, Iran. His main research interests are Affective Computing, Machine Learning, and Machine Vision.

A Semantic Approach to Person Profile Extraction from Farsi Documents

Hojjat Emami*

Department of Information and Communication Technology (ICT), Malek-Ashtar University of Technology, Tehran, Iran
h_emami@mut.ac.ir

Hossein Shirazi

Department of Information and Communication Technology (ICT), Malek-Ashtar University of Technology, Tehran, Iran
shirazi@mut.ac.ir

Ahmad Abdollahzadeh Barforoush

Computer Engineering Department, Amir Kabir University of Technology, Tehran, Iran,
ahmadaku@aut.ac.ir

Received: 14/Jun/2016

Revised: 29/Aug/2016

Accepted: 29/Aug/2016

Abstract

Entity profiling (EP) as an important task of Web mining and information extraction (IE) is the process of extracting entities in question and their related information from given text resources. From computational viewpoint, the Farsi language is one of the less-studied and less-resourced languages, and suffers from the lack of high quality language processing tools. This problem emphasizes the necessity of developing Farsi text processing systems. As an element of EP research, we present a semantic approach to extract profile of person entities from Farsi Web documents. Our approach includes three major components: (i) pre-processing, (ii) semantic analysis and (iii) attribute extraction. First, our system takes as input the raw text, and annotates the text using existing pre-processing tools. In semantic analysis stage, we analyze the pre-processed text syntactically and semantically and enrich the local processed information with semantic information obtained from a distant knowledge base. We then use a semantic rule-based approach to extract the related information of the persons in question. We show the effectiveness of our approach by testing it on a small Farsi corpus. The experimental results are encouraging and show that the proposed method outperforms baseline methods.

Keywords: Web Mining; Information Extraction; Person Profiling; Farsi Language.

1. Introduction

Entity profiling (EP) is an active research topic in Web data mining and information extraction (IE). EP aims to gather, infer, refine and group unobservable information of a given entity from observable data about it. There are many ongoing researches on the problem of EP in different languages. However, one of the less studied languages in EP is Farsi. Farsi speaking people constitute 1.5% of the world's population. They spend hours daily in the Internet and easily publish data on their homepages, news articles, blog entries, item reviews, comments, micro-posts, and social networks. This results a huge volume of valuable Farsi contents on the Internet, which a significant part of them are unstructured, free-text documents. Farsi contents constituent 1% of all the digital contents on the Internet. The volume of Farsi content has increased at a steady rate over the past years (blue line in Figure 1), and this growth is expected to be continuing for the future (orange line in Figure 1). Due to the special and different nature of Farsi such as linguistic phenomena, lack of appropriate natural language processing (NLP) tools and underlying linguistic resources, processing Farsi content is more serious [1],[2]. These challenges highlight the necessity of developing high quality IE approaches in Farsi. In this article, we present an approach to extract

profile of persons in Farsi, and report an evaluation of that. Person profiling is a specific variant of the general EP problem. In person profiling, we are given a collection of Web pages about different persons. The process is to extract profile of each given person from his/her relevant Web pages. A central task in person profiling is *attribute extraction*. In recent years, a few efforts have been made to automatically process the Farsi text and to extract attributes of entities. However, these approaches suffer from several fundamental issues.

The first is that many existing work (e.g., [3]-[5]) relying on syntactic information and used pre-specified lexico-syntactic rules or specific machine learning approaches. These methods cannot entirely solve the problems of *synonymy* and *polysemy* that need deep understanding of text. The second problem is that the resulting attributes are surface textual facts and are not linked to an ontology.

* Corresponding Author

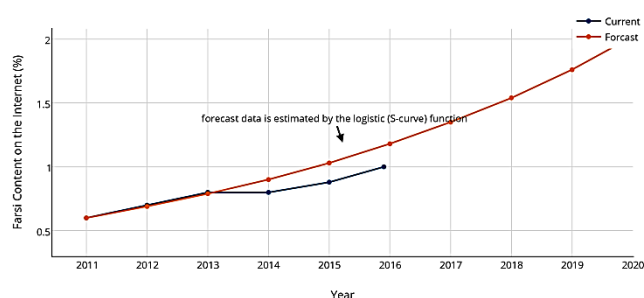


Fig. 1. Farsi content's growth on the Internet; the blue curve shows the growth rate of Farsi content from 2011 to 2016¹, and orange line shows the forecast growth rate up to 2020.

The third is the problem of syntactic variation, presenting the same meaning with different surface linguistic expression forms. This makes it hard for any Web AE to cover all variations of writing patterns. These observations promoted us to developing a new semantic AE approach to overcome the challenges of syntactic variation, and extracting semantic, meaningful attributes.

We use two types of AE methods to extracting profile information of an entity: *verb-based AE* and *noun-based AE*. We use verb-based AE to extract attribute values from semantic role (SRL) frames which their verb predicates serves as an indicator of a given attribute class. We show how verb-based AE can improve the quality of IE, partially solve the syntactic variation. To extract attribute values from noun-based constructions, we use noun-based AE. For noun-based attributes, we map each sentence into a semantic boosted dependency graph, and then use dependency-based patterns to extract the target attributes. The observation underlying our approach is that understanding the text semantically can improve the results. Our approach addresses the problems of *synonymy* and *polysemy*, and makes full use of the merits of both syntactic and semantic analysis of the text. Our approach links the resulting textual surface facts contained in profiles to their possible meaning in a distant ontology. This is very helpful for multi-lingual text processing. The resulting profiles are structured and machine-readable and can simply translate to other more-studied and more-resourced languages, and facilitate understanding and processing of Farsi text documents.

To summarize, our contributions in this article are as follows:

- For verb-based AE, we present a semantic approach, which is effective for rich profile IE from SRL frames. This is helpful to solve the problem of syntactic variation in expressing the same meaning.
- For noun-based AE, we improve the robustness of IMPLIE system [6] by incorporating co-reference information. We apply our extended algorithm on semantic enriched dependency graph to extract both single-value and multiple-value attributes.
- To evaluate the performance of our method, we create a small Farsi dataset drawn from Wikipedia articles. We compare our approach with baseline

methods. Our experiments demonstrate that our method is an encouraging approach, and it can extract high quality information about entities.

The remainder of this paper is organized as follows. After a brief survey of related work in Section 2, we describe our personal EP approach in Section 3. Section 4 describes the experiments we performed to evaluate our approach. Finally, we draw some conclusions, and identify future work in Section 5.

2. Related Work

The task of extracting structured and relevant information about entities from text documents is a long-standing task in the IE and NLP community. Early IE systems were based on lexico-syntactic rule-based methods and are domain dependent. For example, Li et al. [7] used a multi-level rule-based approach, which relies on various linguistic IE patterns to extract entity-centric relations from English corpora. However, such approaches to IE are limited by the availability of domain knowledge, the difficulty in designing rules for all types of text, and less accurate results under noisy setting. Later systems to achieve robustness under noisy setting and to extract arbitrary entity-centric relations use probabilistic [8],[9] and statistical methods [10],[11]. However, these approaches do not discover the semantic information contained in text entirely. Recently, machine learning methods are widely used for entity-centric relation extraction. Supervised learning methods achieved high performance in relation extraction, but they need more hand labelled training data in order to be effective [12]. Due to the lacking of high quality of labelled training data, and the low performance of supervised methods for extracting arbitrary relations from large-scale corpora such as Web, semi-supervised learning methods [12],[13] bootstrapping methods [14], self-supervision approaches [15], distant supervision methods [16], and unsupervised clustering methods [17] are developed. However, each of these methods suffers from several challenges. For example, bootstrapping methods suffers from semantic drift problem, and distant supervision suffers from noisy training data. The output of unsupervised learning methods often does not resemble ontological relations and the resulting relations are hard to map to a domain ontology.

Open IE [18] as a new emerged IE methodology aims to extract arbitrary domain-independent relations in the text without a pre-determined set of relations and with no domain-specific knowledge engineering effort. Open IE extractions are surface text and do not resemble domain-specific ontological relations [19]. Recently, a few approaches focused on adapting Open IE extractions to domain-specific ontology [6],[19]-[21]. Soderland et al. [19] propose a two-step approach for adapting open domain relational tuples to domain-specific ontological relations. In the first stage, the tuples are annotated by a domain concept recognizer, and then a number of relation-mapping rules are learned by using a cover

¹ The data for draw the current status of Farsi content on the Internet are driven from "Usage of content languages for Websites", [www.W3Techs.com], Retrieved 10 March 2016

learning algorithm to map the tuples to domain relations. Since machine learning approach to learning high precision mapping rules need more training data, and such a high volume data are not available, the authors in [20] chose to create mapping rules manually rather than adopting machine learning approaches. In IMPLIE [6] syntactic dependency rules are used to find relations that are beyond the scope of Open IE extractions. IMPLIE begins with user-collected semantic taggers for a set of target attribute classes, and then uses dependency parse rules to find noun phrase that are modified by terms of a target class. We borrow the idea from their work for extracting noun-based attributes, but we enrich the syntactic information with semantic information obtained from a distant ontology to alleviate the errors produced by syntactic parsing. Exploiting syntactic dependencies for relation extraction is not a new idea and studied in early work. For example, [22] formulates the entity-centric relation extraction as the problem of finding the shortest paths between entities on dependency graph. Some other information extraction works rely on shallow semantic analysis of text [23]-[25]. For example, Surdeanu et al. [23] proposed a rule-based approach, which contains a number of mapping rules to map SRL frames to relations in question. However, these approaches have not been addressed entirely some challenging linguistic phenomena such as *synonymy* and *polysemy*.

Some other works integrate syntactic dependencies and semantic information derived from distant knowledge bases to address the challenges like *synonymy* and *polysemy*. For example, Moro and Navigli [26] combined syntactic dependencies and distributional semantic information to extract ontologized relations. However, the resulting relations are still bound to surface text, lacking actual semantic content. Bovi et al. [27] developed DEFIE system, which extracts semantic relations from Web text through deep syntactic and semantic analysis of the text. They obtained syntactic information from dependency parser, and semantic information from Babelfy [28]. They mainly focused on verb-based relations. We borrow the idea of enriching local document-level information with semantic information derived from distant knowledge base Babelfy from the work of Bovi et al. [27]. We (a) focus on Farsi language; and (b) integrate SRL frames with semantic information obtained from Babelfy to extract verb-based relations, and (c) in addition to verb-based relations, we focused on extracting noun-based relations by adopting and extending the rule-based approach presented in [20].

There are also some relation extraction works in Farsi. Relation extraction is a central task in entity profiling and focuses on learning atomic facts about entities. We notice that in contrast to work in relation extraction, our work addresses the problem of entity profiling, which extracts a richer information structure about a given entity. One of the first approaches to semantic relation extraction in Farsi is based on hand-crafted rules, which uses the syntactic and lexical information [3]. A similar approach

is done by Moradi et al. [4]. They adopt the Hearst's approach [29] to relation extraction in Farsi. In their approach, relations are extracted by matching some pre-defined patterns over the text. Their approach has syntactic nature and does not analyze the text semantically. In other work, [5] uses a set of semantic and lexico-syntactic patterns and templates for extracting taxonomic and non-taxonomic relations and axioms from Farsi text. Our own earlier work on personal information extraction in Farsi includes [30]. In the paper [30], we used syntactic-based patterns and attribute-specific gazetteers to extract personal attributes. However, the limitations to our previous work [30] and some existing work in Farsi are that (i) the resulting attributes are surface textual facts and are not linked to an ontology, and (ii) they did not address the language phenomena of *synonymy* and *polysemy* that need deep understanding of the text, and (iii) they suffer from the problem of syntactic variation. In this paper, we use different semantic-based approaches to improve the quality of attribute extraction, alleviate the problem of syntactic variation and the challenges of *synonymy* and *polysemy*, which was not studied before in previous work. Our approach relies on deep semantic analysis of the text, and enriches local entity-centric information with semantic information obtained from a distant knowledge base. The resulting profile attributes are meaningful and are linked to their possible meaning in a distant ontology. This is greatly helpful for the multi-lingual text processing such that the resulting profiles can simply translate to other language.

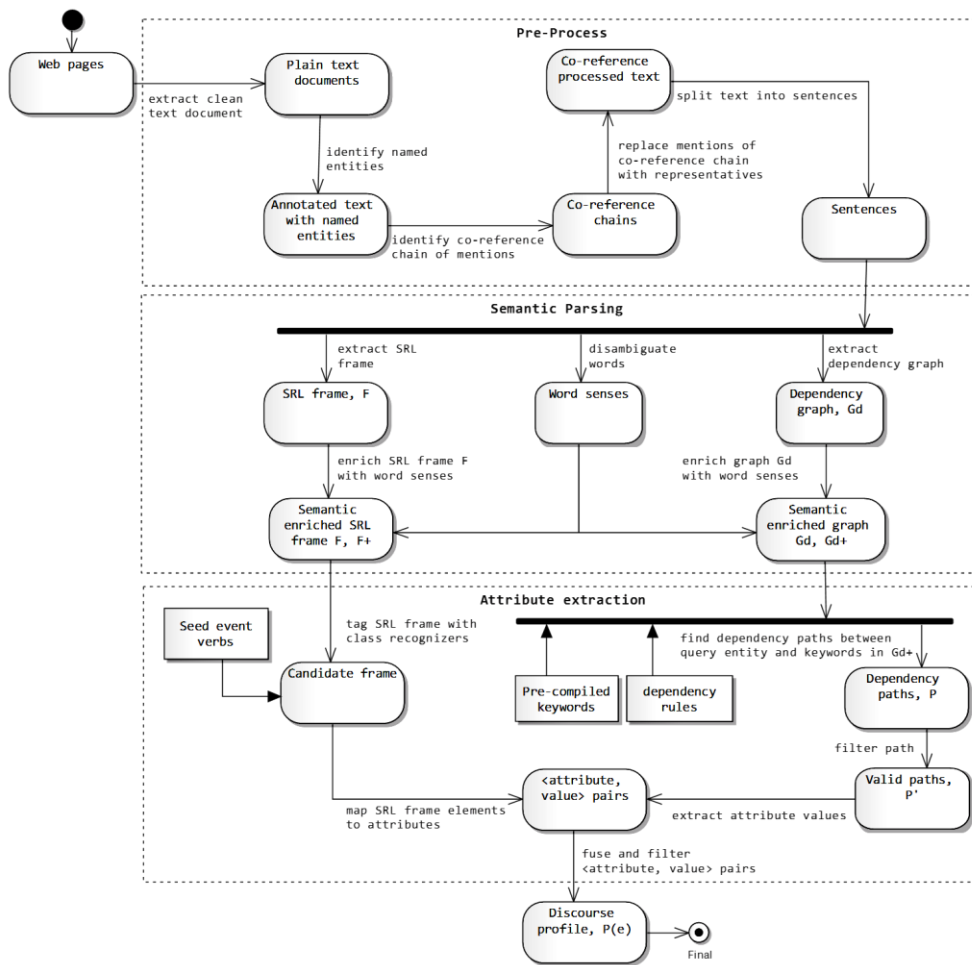


Fig. 2. The state diagram of our EP approach

3. Our Proposed Method

Our profile extraction system takes as input a query entity e , and extracts its discourse profile, which contains a number of <attribute, value> pairs, each of which represents a certain characteristic of the entity e . Formally, we define the discourse profile of the entity e , $P(e)$ as follows:

$$P(e) = \{(a, v) | a \in A, v \in V(a)\} \quad (1)$$

where a is a given attribute, v is a value for attribute a , A is the vocabulary of attributes that can be used to describe characteristics of the entity e ; and $V(a)$ represent a set of filler values for attribute $a \in A$. Similarly, we define the entity profiling problem as follows:

$$\text{Given } \{D, A, e\}, \text{ design a system } \varphi, P(e) \leftarrow \varphi(D, e) \quad (2)$$

This formulation says that given a text document D , a query entity e , and a vocabulary of attributes A , our goal is to design a profiling system φ to extract structured information $P(e)$ related to entity e from document D . Figure 2 shows the state diagram of our proposed method. We decompose person profile extraction problem as three major subtasks: (i)

pre-processing, (ii) semantic analysis, and (iii) attribute extraction. Pre-processing provides the input text as system's desired format using existing pre-processing tools. Semantic analysis takes a pre-processed sentence as input to produce its semantic representation. This component extracts the syntactic information (dependency graph) and semantic information (SRL frame) from pre-processed text, and enriches syntactic dependency graph and SRL frames with word senses and disambiguated entity mentions. The attribute extraction component, is given the query entity e , and a vocabulary of attributes A , and must find a set of filler values V for each attribute $a \in A$ from annotated text generated by semantic analysis component. The extracted <attribute, value> pairs are validated and integrated to form discourse profile of entities in question. In the following, we describe these tasks in more detail.

3.1 Pre-Processing

In this article, we focus on the textual part of the Web pages, because the majority of the information about entities on the Web is often expressed by natural language text. Web pages need to be pre-processed and prepared according to system's desired format. Pre-processing includes four main stages: (i) html tag removal, (ii)

named entity recognition, (iii) co-reference resolution, and (iv) sentence splitting. First, for each Web page, Jsoup (Java HTML Parser, [https://jsoup.org]) is run to cast it into plain text document. We then use a multi-lingual named entity tagger [31] to annotate the text for coarse-grained lexical entity types including person, location, and organization. The annotated text documents are passed to a rule-based co-reference resolution module to identify co-reference chains for all the entities mentioned in each document. The mentions in every co-reference chain of interest are then replaced with their corresponding representative mentions. Next, for the co-reference chain of interest within each document, we split the document to sentences. We note that, in our implementations, we focused on formal-style sentences. A formal-style sentence follows prescribed writing standards, and prepared for a fairly broad audience [32], [33] A formal-style sentence is often complete and contains a subject, verb and an object. The pre-processing may produce errors, which propagate to the latter stages. However, improving the pre-processing tools is beyond the scope of this paper. The remainder of the processing described in the following use this pre-processed text.

3.2 Semantic Analysis

The semantic analysis component takes as input the pre-processed formal-style text, extracts SRL frames and dependency graphs from the sentences of pre-processed text, and augments them with word senses derived from a distant knowledge base. Semantic parsing consists of four subtasks: (i) word sense disambiguation, (ii) SRL frame extraction, (iii) dependency parsing, and (iv) semantic enrichment. In the following we describe these subtasks in more detail.

3.2.1 Word Sense Disambiguation

Word sense disambiguation (WSD) provides a sense mapping from surface words and entity mentions in a

given text to concepts and named entities in an ontology. In WSD stage, we first disambiguate the word senses using Babelify [28], a state-of-the-art entity linking and word sense disambiguation system. We filter the resulting senses by pruning the senses corresponding to short-tail mentions that covered by other long-tail mentions. We then map surface textual words and entity mentions to word senses and named entities in BabelNet ontology [28]. Figure 3(a) shows the WSD result for a sample sentence. In Figure 3(a), notation bn:i refers to the i-th BabelNet sense for the given word.

3.2.2 SRL Frame Extraction

In the stage of SRL frame extraction, we assign a “who did what to whom, when, where, why, and how” structure to the sentences of text. We use a rule-based semantic role labeller system [34] to annotate constituents of the sentences with semantic roles. We extract SRL frames from the output of semantic role labeller system. An SRL frame consists of a verb predicate and a number of semantic role elements. Let $F = \{p, E_1, \dots, E_m\}$ be an SRL frame in which p denotes the verb predicate, and E_i is the i th SRF element in the frame. Each element E_i is a (s, g) pair, where s indicates the type of semantic role, and g denotes the value for the underlying argument. We note that there may be multiple SRL frame in a sentence depending on the number of verbs in the sentence. In our implementations, types of semantic roles in the output of semantic role labelling system follow the annotation guideline in VerbNet [35]. Figure 3(b) shows a sample sentence along with SRL frames extracted by semantic role labelling system. Semantic role labelling system produces errors (e.g., incorrect argument boundary, or incorrect associated semantic role labels to words), which propagate to the later stages. However, improving the semantic role labelling system is orthogonal to our problem and out of the scope of this paper.

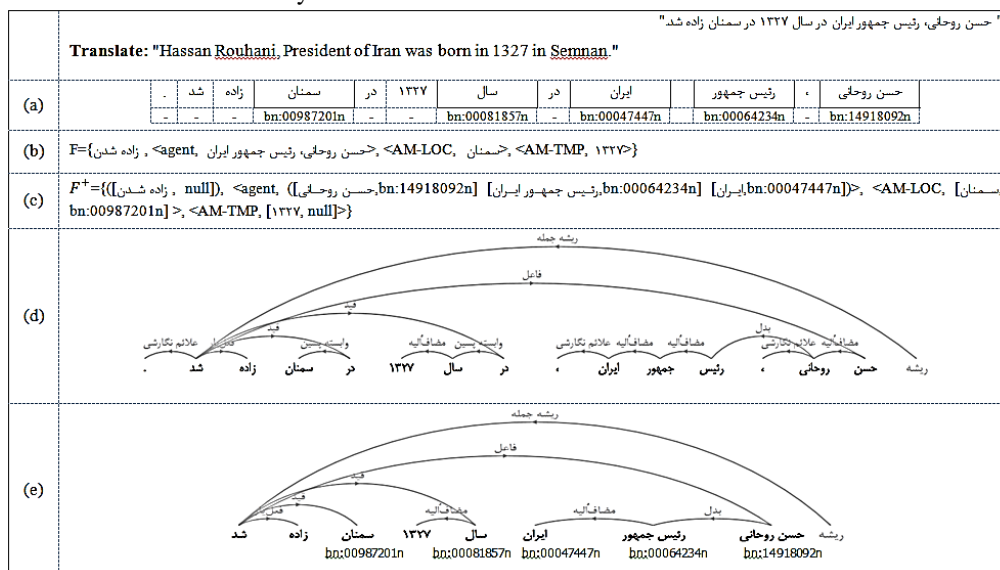


Fig. 3. semantic analysis for a sample sentence; (a) disambiguated word senses; (b) SRL frame F; (c) semantic enriched SRL frame; (d) dependency graph G_d ; (e) syntactic-semantic graph G_{sd}

3.2.3 Dependency Parsing

In dependency parsing stage, we first parse each sentence of the text to obtain corresponding dependency graph G_d . In G_d , each single word figured as a node and word-word dependencies are represented as directed edges between nodes. In other words, G_d represents binary relationships between words of a sentence, in which words are connected with their parent words with a unique edge labelled with a syntactic function. The definition of syntactic functions is given in [36]. Figure 3(d) shows the dependency graph G_d for a sample sentence. Our syntactic analysis component uses Hazm parser¹, and generates a dependency syntactic graph for each sentence. The erroneous syntactic analyzing of a sentence degrades the performance of later components of EP system. However, we alleviate this problem by enriching the syntactic dependencies with semantic information generated at WSD stage.

3.2.4 Semantic Enrichment

The aim of this stage is to augment the SRL frame and the dependency graph G_d with semantic information obtained in WSD phase. To enrich SRL frame elements with semantic information, we replace each SRL element with its corresponding disambiguated sense. Figure 3(c) shows the semantic enrichment result for the SRL frame given in Figure 3(b). To enrich dependency graph G_d with word senses, and create a syntactic-semantic graph G_{sd} , we start from the dependency graph G_d of sentence s , and a set of disambiguated senses for that sentence. If a disambiguated sense is a single token and covers a single node in G_d , it simply attach to the corresponding dependency node. If a disambiguated sense is a multi-word expression and covers more than one node in G_d , we merge the sub-graph referring to the same concept or entity to a single semantic node. Figure 3(e) shows the enrichment result for the graph G_d given in Figure 3(d).

3.3 Attribute Extraction

The attribute extraction (AE) component takes as input the query entity e , and a vocabulary of attributes A , and extracts filler values for the attributes in A . We focus on six kinds of attributes include: 'تاریخ تولد/ tarikhe tavallo/ date of birth', 'محل تولد/ mahale tavallo/ birth place', 'بستگان/ bastegan/ relatives', 'ملیت/ meliat/ nationality', 'شغل/ shogh/ occupation', and 'مدرک/ madrak/ degree'. These attributes are those extensively studied in IE tasks including slot/template filling, and knowledge base population tasks [37], [38]. We observe that the filler values for these attributes are from noun-based constructions, or sentences having a verb, which serves as an indicator for different attribute classes. Hence, we use two types of AE rules to extract attribute filler values: verb-based rules and noun-based rules. We applied verb-based rules on semantic augmented SRL

frames and noun-based rules on semantic boosted dependency graphs. In the following, we give more detail about these methods.

3.3.1 Verb-based Attribute Extraction

The procedure of extraction filler values for the given attributes from semantic boosted SRL frames includes two stages: (i) frame marking, and (ii) frame element mapping. The frame marking is responsible for labelling the verbs or phrases in SRL frames as indicators of possible values for target attribute classes, but it does not specify which of the elements should be considered as a filler value for any given attribute. The frame mapping looks at the elements of marked SRL frames, and decides which element corresponds to which attribute of the person in question.

3.3.1.1 Frame Marking

The frame marking identifies a set of potential SRL frames containing possible values for a given attribute class. We mark SRL frames by selecting an event verb for each attribute class of interest, and tagging frames for that target class. The main idea in using event verbs as attribute class indicator is that event verbs typically convey the main idea of a sentence. Let $X = \{a_1, \dots, a_k\}$, $X \in A$ be the list of attributes on question that their values can be extracted using verb-based AE rules. To mark SRL frames for a given attribute class $a_i \in X$, we define a seed event verb v specific to a_i . We supplement the event verb v with its synonymous verbs using Farsi version of WordNet [39], FarsNet [40] and form synonym vector, $S = \{s_1, \dots, s_m\}$, where $s_i \in S$ is the i th synonymous verb of v , and m indicates the number of synonymous verbs in S . Using synonymous verbs for SRL frame marking solves the problem of syntactic variation, and prevents inducing several patterns for extracting values for the same attribute class. The decision to using FarsNet comes from the fact that it has a flexible and well-defined lexicon schema, which is publicly known and accepted. We notice that an SRL frame argument may have multiple instances of a given attribute class, and could be considered as candidate value for multiple attributes.

Given attribute class recognizers, a semantic frame F is considered as a potential candidate for attribute class $a_i \in X$, if the predicate p in frame F matches up with one of the seed verbs defined for the attribute class a_i . For example, in the SRL frame given in Figure 3(d), the SRL frame marking have found that the verb predicate 'زاده شدن/ zAde shodan/ born' is an indicator for the attribute class of 'محل تولد/ mahale tavallo/ birth place'. Frame marking is important, since the tagged SRL frames form the pool of candidates for attributes of interest in the following stages.

3.3.1.2 Frame Mapping

Our procedure for frame mapping takes as input the SRL frame F that has been processed by frame marking, and maps the elements of frame F to corresponding attributes. The final representation we attempt to create

¹ <http://www.sobhe.ir/hazm/>

from SRL frames is similar to the frames in FrameNet [41]. In other words, we map the universal and verb-specific roles in SRL frames to template-specific roles. In English, there are resources to direct mapping the elements of SRL frame to FrameNet frame elements such as SemLink [42], but there are not still such resource in Farsi. To map the elements of SRL frame to attributes of interest, one can use different machine learning and data mining approaches proposed for slot/ template filling tasks [37], [38]. Since we have not sufficient training data, and the vocabulary of given attributes is a small closed-class set, here, we use a rule-based method to map the SRL frame elements to corresponding attributes. The overall strategy of our approach is similar to the rule-based methods taken by Angeli et al. [43], and Soderland et al. [20] in English, but our way of defining trigger words and extracting attributes' filler values is different.

For each target attribute class, we create manually a number of rules, based on error analysis over the SRL frames obtained from Wikipedia articles written in Farsi. Each rule is expressed as a number of regular expression patterns containing attribute-specific semantic and lexical constraints. These constraints ensure that the candidate SRL frame element to be a valid filler value for corresponding attribute. Each rule is run over given SRL frame and extracts a value for the attribute of interest when all constraints are met. A sample rule is shown in Figure 4, which determines the filler value for attribute class 'محل تولد/ mahale tavallod/ birth place'. In Figure 4, p refers to verb predicate of SRL frame, and s and g respectively refers for semantic role label and argument value in SRL frame element. It should be noticed that each predicate p in an SRL frame F may correspond with several syntactic frames in verb valency lexicon, which depends on its sense given in the sentence. Because there is not appropriate verb lexicon representing information about verb senses in Farsi, in our implementation, we assume that syntactic alternations belong to only one sense. However, this assumption makes some errors in the AE phase.

Extracting filler value for attributes using SRL frames has two important advantages: (i) since diverse expression forms of sentences with the same meaning are reduced into a single SRL frame, extracting attribute values from SRL frames is much simpler than those relying on syntactic information contained in sentences; (ii) because verb-based AE are easy to understand, one can extend and revise the initial AE rules with high quality rules.

Terms in rule	Value
Trigger seed verb	< p : ?birth place>
Query entity in	< s : Agent, g : ?person entity>
Entity type	<Person>
Attribute value in	< s :AM-LOC, g : ?location named-entity >
Attribute value type	<Location name>

Fig. 4. A sample rule designed for the attribute class 'محل تولد/ mahale tavallod/birth place'. This simple rule specifies the target attribute class and filler values for its arguments.

3.3.2 Noun-based Attribute Extraction

The filler values for attributes that their values contained in noun-based constructions cannot be extracted by verb-based AE rules. For example, in the sentence given in Figure 3, the verb-based AE cannot include that the phrase 'رئیس جمهور/raeis jomhour/President' is a filler value for the attribute of "occupation" for the person 'حسن روحانی/Hassan Rouhani'. To extract the attribute values contained in noun-based constructions, we use a rule-based approach, which exploits the syntactic information produced by dependency parser, and the lexical information in the form of pre-compiled keywords and named entities. We collect the list of keywords from online information resources such as Wikipedia¹, DBpedia² and FreeBase³ ontology, and tables found on the Web. For example, to find candidate values for the attribute of 'شغل/occupation', we collect a list of occupations from DBpedia and Wikipedia, and form keyword set 'occupation'. The overall strategy of our AE approach is similar to the implicit relation extraction method developed by Soderland et al. [6]. However, the limitation to their approach are that (i) the attribute filler that is a multi-word expression and covers more than one word in dependency graph cannot be extracted, and (ii) it cannot be directly applied in Farsi. We modify and extend their approach to extract both single-word and multi-word attributes' fillers from dependency graph. Borrowing the idea from the work of Bovi et al. [27], we couple syntactic dependencies and fully disambiguated entity mentions and word senses to solve the problem of multi-word filler value extraction. Let $Y = \{a_1, \dots, a_l\}$, $Y \in A$ be the list of attributes on question that their values can be extracted using noun-based AE rules. The procedure for noun-based AE is summarized in Figure 5.

```

Input: query entity  $e$ , query attribute  $a \in A$ , keyword list  $I$  specific to attribute  $a \in A$ , and graph  $G_{sd}$ .
Output: values for attribute  $a$ ,  $V$ 
if ( $G_{sd}$  contains keyword  $k \in I$ )
     $P = \text{shortestPath}(G_{sd}, e, k)$ ; // shortest path between  $e$  and  $k$  in  $G_{sd}$ 
     $P' = \text{validatePath}(P)$  // filter by constraints
    if ( $P'$  is valid)
         $v := k$ ;
         $V := V \cup \{v\}$ ;
    end
end
Return  $V$ ;
    
```

Fig. 5. The AE algorithm for extracting noun-based attribute values

The input to the algorithm is the query entity e , syntactic-semantic graph G_{sd} generated for sentence s , and a set of pre-compiled keywords I specific to attributes Y . The algorithms then iterates on the graph G_{sd} , and looking for a co-occurrence of a keyword $k \in I$, and the query entity e . An entity e and a keyword $k \in I$ in the

¹ <https://www.wikipedia.org/>
² <http://wiki.dbpedia.org/>
³ <https://www.freebase.com>

graph G_{sd} is considered to be related, (i) if there is a dependency path between them in which every dependency edge on the path tagged with one of the following labels: *app*, *moz*, *npostmod*, *npremod*, *apostmod*, *apremod*, *nez*, *npp*, *nadv*, *mos*, *ncl*, *acl*, *posdep*, and *predep*; and (ii) if the keyword k and the focus entity co-refer, i.e., they refer to the same entity. This constraint filters the meaningless and erroneous extractions. Since dependency arcs in G_{sd} is directed, there is no guarantee in finding a path between named entities and concepts. Thus, we use an undirected version of the graph G_{sd} , and follow the the assumption of tokens' locality information [44] to find the path between entity pairs. Among all of the paths found between the entity e and the keyword k , we chose the path with the shortest length. The idea behind this constraint follows the shortest path hypothesis [22], which states that the most valuable information about a relation is contained on the shortest path between two relation's argument nodes in the graph. In the dependency graph G_{sd} given in Figure 3(e), the resulting shortest path between entity 'حسن روحانی/Hassan Rouhani' and the keyword 'جمهوری/raies jomhour/President' contains dependency tag 'بدل/badal/app'. This path is a valid path and meets the constraint, thus the keyword 'جمهوری/raeis jomhour/President' is a valid filler value for the attribute 'شغل/occupation'.

4. Experiments and Results

In this section, we first describe benchmark datasets and performance metrics, and then give the results obtained by our approach and its counterparts.

4.1 Dataset

A key challenge to evaluate our EP approach is the lack of Farsi dataset suited for EP problem. Thus, we created a small Farsi corpus for evaluating our approach. All evaluations were carried out based on manual assessment. We first chose 30 typical person names and then queried Wikipedia for these names. The reason to using Wikipedia articles as benchmark comes from the fact that Wikipedia articles are rich source of knowledge on the Web and they frequently accessed by millions of users. We create our dataset by selecting a sample of 100 sentences from collected Wikipedia articles. Each sentence in the sample dataset contains at least a candidate filler value for one of the six attributes: 'تاریخ تولد/date of birth', 'محل تولد/birth place', 'بستگان/relatives', 'ملیت/nationality', 'شغل/occupation', and 'مدرک/degree'. In order to create ground truth for evaluation, two human annotators independently examined the sampled sentences to identify the relevant attributes, with an inter-annotator agreement. This type of evaluation follows previous work in the field of information extraction [25], [27], [45]. The annotators reached an agreement score of $\kappa = 70\%$ measured by Cohen's kappa coefficient, which considered to be within the substantial agreement boundaries [46].

The number of resulting <attribute, value> pairs in ground truth is 160.

However this dataset is small to evaluate the scalability of EP approach, but it have the desired characteristics that enables us to study the effectiveness of our EP approach in extracting entity-centric information. To the best of our knowledge, we are the first to investigate the EP in Farsi, thus our dataset to study EP is unique.

4.2 Performance Measures

In the experiments, we conducted evaluations using three criteria: (i) precision, (ii) recall, and (iii) F1 measure. For more detail about these metrics refer to [47]. The quality of the results is evaluated by comparing the profile <attribute, value> pairs obtained by the system and those attribute values in ground truth annotated by annotators. Formally, precision (P), recall (R), and F1 measure ($F1$) is defined as follows:

$$P = \frac{|S \cap G|}{|S|} \quad (3)$$

$$R = \frac{|S \cap G|}{|G|} \quad (4)$$

$$F1 = 2 \times \frac{P \times R}{P + R} \quad (5)$$

where S is the set of <attribute, value> pairs generated by the system, and G is the set of <attribute, value> pairs in the annotated gold standard set.

4.3 Numerical Results and Discussion

Table 1 shows the performances obtained by our AE approaches. In Table 1, we give the average performances of the pure verb-based AE (VAE) method, pure noun-based AE (NAE) method, and the combination AE (EP^+) method. In Table 1, we observe that the performance of AE method is increasing when incorporating both VAE and NAE, while either the VAE or the NAE cannot achieve good performance. The EP^+ approach achieves the best scores. However, the performances are far from ideal. This shows that profile extraction in Farsi text resources is a big challenge, and justifies that more effort is needed in this field.

Table 1. Performances of our AE approaches on the benchmark dataset

Method	P(%)	R(%)	F1(%)
VAE	32.86	20.51	25.26
NAE	43.64	21.14	28.48
EP^+ (VAE + NAE)	43.69	32.38	37.19

Table 2 shows the detailed performance for the six individual attributes obtained by our EP^+ system on benchmark dataset. As shown in Table 2, the attributes of 'تاریخ تولد/date of birth' and 'محل تولد/birth place' have achieved good performance, because their instances often expressed by easily predictable patterns in formal-style format. On the other hand, for the attributes of 'شغل/occupation', and 'بستگان/relatives', the approach cannot achieve good performance. The low score for these attributes is partially due to the fact that the set of values, which such attributes can take are often expressed with various forms in syntactic structure and vocabulary.

For the attributes ‘ملیت/nationality’ and ‘مدرک/degree’, the approach reports moderate result. Our approach achieved around 18-56% precision, 10-50% recall, and 13-53% F1 score for the given profile attributes.

Table 2. Detailed performance of the six individual attributes obtained by our approach (EP⁺) on benchmark dataset

Attribute class	P (%)	R (%)	F1 (%)
Birth place	56.25	50	52.94
Date of birth	53.85	43.75	48.28
Degree	40	33.33	36.36
Nationality	53.33	34.78	42.11
Occupation	40.54	22.39	28.85
Relatives	18.1	10	12.9
Overall	43.69	32.38	36.91

We implemented five AE methods as our baseline methods. These baseline methods include: (i) SRL-based AE, (ii) IMPLIE system [6], (iii) UvA_2 system [48], (iv) PolyUHK [49], and (v) our recent work in [30]. SRL-based AE is appropriate for the extraction of verb-based attributes. This baseline uses a set of hand-crafted mapping rules to map SRL frame elements to attributes in question. The overall strategy of SRL-based AE method is similar to those presented in [23] and [24]. IMPLIE is developed by the University of Washington team for TAC-KBP 2015 track. IMPLIE uses a set of syntactic rules to extract implicit noun-based relations from dependency graph. UvA_2 is developed by the university of Amsterdam team at WePS 2009 sharetask [50]. This method uses lists of pre-compiled keywords and Web-specific patterns for the personal AE. The overall strategy of UvA_2 is similar to AE methods presented in [7], [20], [48], [49], [51]. PolyUHK is a rule-based AE approach, which achieved the best performance in the WePS 2009 sharetask. For each attribute, PolyUHK first identifies a set of keywords and named entities. It then looks for a co-occurrence of one of the keywords regarding the focus attribute and the target person in a sentence. If a co-occurrence is found then the candidate keyword would be considered as a filler value for the focus attribute. Our previous work [30] is a simple pattern-matching method relying on pre-compiled keywords and hand-crafted rules. To fair comparison, we compare our EP⁺ approach with UvA_2 [48], PolyUHK [49] and our previous work [30]; our noun-based AE (NAE) with IMPLIE [6]; and our verb-based AE (VAE) with SRL-based AE method.

Table 3 summarizes the results obtained by baseline methods and our EP⁺ on the benchmark dataset. Comparing to baseline methods, our method outperforms the baseline methods. Our method achieves higher overall F1 score, 10.12% better than UvA_2, 8.34% better than [30], and 3.11% better than PolyUHK. This indicates that incorporating both verb-based AE and noun-based AE, and considering semantic enrichment is effective in increasing performance of the attribute extraction approach.

Table 3. Comparison of results obtained by baselines and our method on benchmark dataset

Method	P (%)	R (%)	F1 (%)
UvA_2 [48]	37.64	21.14	27.07
PolyUHK [49]	42.61	28.4	34.08
Emami et al. [30]	38.15	23.2	28.85
Our method	43.69	32.38	37.19

Table 4 shows the results obtained by our VAE method and SRL-based AE. From Table 4, we notice that VAE is outperformed pure SRL-based AE method. VAE method achieves a F1 score of 25.26% providing an improvement of about 1.16 F1 score points. This clearly shows the effect of semantic enrichment in the extraction of verb-based relations.

Table 4. Comparison of results obtained by our VAE method SRL-based AE on benchmark dataset

Method	P (%)	R (%)	F1 (%)
SRL-based AE	31.05	19.7	24.1
VAE	32.86	20.51	25.26

Table 5 shows the results obtained by our NAE method and IMPLIE system. The results clearly show that NAE outperforms IMPLIE, and achieved higher scores. The main reason to the low score of IMPLIE is that it cannot correctly extract the multi-word attributes, while some noun-based attributes are multiple-word mentions.

Table 5. Comparison of results obtained by our NAE and IMPLIE on benchmark dataset

Method	P (%)	R (%)	F1 (%)
IMPLIE [6]	42.55	19.38	26.63
NAE	43.64	21.14	28.48

Our method still suffers from several challenges that need to be addressed. Our manual investigation over incorrect extractions indicates that the performance scores for profile attributes can be raised if the following conditions are hold.

- *Creating more precise AE rules:* overall, our profiling approach reports low F1 score for some attributes on question. This fact indicates that EP in Farsi is still a big challenge. Obviously, the more precise the AE rules are, the higher the performance scores are. Therefore, if we spend more time in the development of more robust AE rules, the system performance will pick up.
- *Improving the performance of pre-processing components:* our manual investigation reveals that almost half of the incorrect extractions were because of the inefficiency of pre-processing and semantic analysis stages, and not because of the inefficiency of our AE method. Errors in pre-processing and semantic analysis stages are propagated to AE step and cause wrong extractions. Thus the low performance of pre-requisite stages is a bottleneck for efficient EP. However, improving pre-processing and semantic analysis is orthogonal to our problem and therefore out of the scope of this paper. Nonetheless, to alleviate the errors in semantic analysis stage, we enrich the analyzed text with semantic information extracted from a distant ontology. The effect of

semantic enrichment is shown in Table 6. In the table, EP^+ shows the scenario in which semantic enrichment is considered, and $-EP$ shows the scenario when the semantic enrichment is completely disregarded. We observe that semantic enrichment improve the result of EP, and the best results are obtained by EP^+ . The errors in semantic analysis stage leads to degradation of EP^- performance from 36.91 to 30.57% in terms of F1 score. This proves the importance of semantic enrichment in EP. We manually correct the errors in the output of pre-processing and semantic analysis, and give correct input to the AE stage. We observe that the results are improved, and the overall F1 score is raised to over 62%. This shows that errors in extractions were not completely because of the inefficiency of AE method.

- *Enriching discourse profile*: in this paper, we focused on the extraction of attributes on question only from the content provided in given dataset. One of the promising solutions to improve the result and alleviate the problem of data sparseness is to enrich local discourse profile with semantic information inferred from distant knowledge bases. This task is considered as our future work.

Table 6. The results obtained by EP^+ and EP^- in terms of F1 score

Attribute class	EP^-	EP^+
Birth place	44.4	52.94
Date of birth	45.16	48.28
Degree	25	36.36
Nationality	39.02	42.11
Occupation	22.6	28.85
Relatives	7.15	12.9
Overall	30.57	36.91

5. Conclusion

Entity profiling (EP) in poor-resource languages like Farsi is suffering from several challenges regarding the tools of language processing and annotated data. As an element of EP research to address these challenges, in this paper, we have investigated a specific variant of the general EP problem, namely the person profile extraction

from Farsi Web documents. Our approach identifies the persons in question from the text, and extracts their profile information. Our approach first parses each sentence of the text syntactically and semantically, and augments the local information with global semantic information derived from a distant knowledge base. It uses a semantic rule-based method to extract the attributes of persons, and form their discourse profile. We evaluated our EP approach with a small corpus collected from Farsi Wikipedia articles. Experimental results indicate that our approach is capable to extract the entity-centric information with a high performance.

On the whole, our EP approach can be considered as a foundation for more robust approaches to EP. There remain several important points to improve our research. First, we plan to automate the induction of attribute extraction rules which might to improve the performance and decrease manual engineering effort. Second work is to design a generic EP system to cover more entities and accurately extract their profiles. As the final results of EP system depend on the performance of three subtasks including pre-processing, semantic analysis, and attribute extraction, therefore, third interesting future work is to improve the pre-requisites' performance, which eventually can improve the overall quality of EP system. Since the problem of EP is far from being solved, our fourth future work is to integrate different information extraction and machine learning methods to complement the shortcomings of AE approach, and further improve the overall performance. We chose not to tackle cross-document EP, and instead spent our energy on document-level EP. Our EP approach identifies the entity-centric information only within a document, which is not enough for Web data as some information occur across documents. Our fifth future work is to work on algorithms for cross-document EP, which aims to gather the information about an entity distributed on multiple documents. In present study, we only focused on EP in formal-style text, while most of the entity-centric information in Web data is expressed in informal-style text. Finally, we would like to investigate EP in informal-style text.

References

- [1] P. Saeedi, H. Faili, and A. Shakery, "Semantic role induction in Persian: An unsupervised approach by using probabilistic models," *Lit. Linguist. Comput.*, 2014.
- [2] M. Shamsfard, "Challenges and open problems in Persian text processing," in *Proceedings of 5th Language & Technology Conference (LTC)*, Poznań, Poland, 2011, pp. 65–69.
- [3] H. Fadaei and M. Shamsfard, "Extracting conceptual relations from Persian resources," in *ITNG2010 - 7th International Conference on Information Technology: New Generations*, Las Vegas, Nevada, USA, 2010, pp. 244–248.
- [4] M. Moradi, B. Vazirnezhad, and M. Bahrani, "Commonsense Knowledge Extraction for Persian Language: A Combinatory Approach," *Iran. J. Inf. Process. Manag.*, vol. 31, no. 1, pp. 109–124, 2015.
- [5] M. Shamsfard, "Lexico-syntactic and Semantic Patterns for Extracting Knowledge from Persian Texts," *Int. J. Comput. Sci. Eng.*, vol. 2, no. 6, pp. 2190–2196, 2010.
- [6] S. Soderland, N. Hawkins, G. L. Kim, and D. S. Weld, "University of Washington System for 2015 KBP Cold Start Slot Filling," in *Proceedings of TAC-KBP 2015*, Maryland, USA, 2015.
- [7] W. Li, R. Srihari, C. Niu, and X. Li, "Entity profile extraction from large corpora," in *Pacific Association for Computational Linguistics Conference (PACLING-2003)*, Harifax, Canada, 2003.
- [8] X. YU and W. LAM, "An Integrated Probabilistic and Logic Approach to Encyclopedia Relation Extraction with Multiple Features," in *Proceedings of the 22nd*

- International Conference on Computational Linguistics (Coling 2008), Manchester, UK, 2008, pp. 1065–1072.
- [9] T. Lee, Z. Wang, H. Wang, and S. Hwang, "Attribute extraction and scoring: A probabilistic approach," in ICDE 2013, Brisbane, Australia, 2013, pp. 194–205.
- [10] F. M. Suchanek, G. Ifrim, and G. Weikum, "Combining Linguistic and Statistical Analysis to Extract Relations from Web Documents," in Proceedings of KDD, Philadelphia, Pennsylvania, USA, 2006, pp. 712–717.
- [11] J. Zhu, Z. Nie, X. Liu, B. Zhang, and J.-R. Wen, "StatSnowball : a Statistical Approach to Extracting Entity," in Proceedings of the 18th international conference on World wide web, Madrid, Spain, 2009, pp. 101–110.
- [12] N. Bach and S. Badaskar, "A review of relation extraction," *Lit. Rev. Lang. Stat. II*, 2007.
- [13] A. Sun, R. Grishman, and S. Sekine, "Semi-supervised relation extraction with large-scale word clustering," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 521–529.
- [14] F. Xu, "Bootstrapping Relation Extraction from Semantic Seeds," Saarland University, Saarbrücken, Germany, 2007.
- [15] F. Wu and D. S.Weld, "Autonomously Semantifying Wikipedia," in Proceedings of CIKM' 07, Lisboa, Portugal, 2007, pp. 41–50.
- [16] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, Suntec , Singapore, 2009, pp. 1003–1011.
- [17] K. Eichler, H. Hemsén, and G. Neumann, "Unsupervised relation extraction from Web documents," in Proceeding of the 6th International Conference on Language Resources and Evaluation, Marrakech, Morocco, 2008, pp. 1674–1679.
- [18] M. Banko, M. Cafarella, and S. Soderland, "Open information extraction from the web," in International Joint Conferences on Artificial Intelligence, Hyderabad, India, 2007, pp. 2670–2676.
- [19] S. Soderland, B. Roof, B. Qin, and S. Xu, "Adapting Open Information Extraction to Domain-Specific Relations," *AI Mag.*, vol. 31, no. 3, pp. 93–102, 2010.
- [20] S. Soderland, J. Gilmer, R. Bart, O. Etzioni, and D. Weld, "Open Information Extraction to KBP Relations in 3 Hours," in Proceedings of TAC-KBP 2013, Maryland, USA, 2013.
- [21] M. Yahya, S. E. Whang, R. Gupta, and A. Halevy, "ReNoun: Fact Extraction for Nominal Attributes," in Proceedings of EMNLP 2014, Doha, Qatar, 2014, pp. 325–335.
- [22] R. C. Bunescu and R. J. Mooney, "A shortest path dependency kernel for relation extraction," in Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, Canada, 2005, pp. 724–731.
- [23] M. Surdeanu, S. Harabagiu, J. Williams, and P. Aarseth, "Using predicate-argument structures for information extraction," in Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - ACL '03, Morristown, NJ, USA, 2003, pp. 8–15.
- [24] M. Gregory, L. Mcgrath, E. Bell, K. O. Hara, and K. Domico, "Domain Independent Knowledge Base Population From Structured and Unstructured Data Sources," in Twenty-Fourth International FLAIRS Conference, Palm Beach, Florida, USA, 2011, pp. 251–256.
- [25] P. Exner and P. Nugues, "Using semantic role labeling to extract events from Wikipedia," in CEUR Workshop Proceedings, Bonn, Germany, 2011, pp. 38–47.
- [26] A. Moro and R. Navigli, "Integrating Syntactic and Semantic Analysis into the Open Information Extraction Paradigm," in Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, Beijing, China, 2013, pp. 2148–2154.
- [27] C. Delli Bovi, L. Telesca, and R. Navigli, "Large-Scale Information Extraction from Textual Definitions through Deep Syntactic and Semantic Analysis," *Trans. Assoc. Comput. Linguist.*, vol. 3, pp. 529–543, 2015.
- [28] A. Moro, A. Raganato, and R. Navigli, "Entity Linking meets Word Sense Disambiguation : a Unified Approach," *Trans. Assoc. Comput. Linguist.*, vol. 2, pp. 231–244, 2014.
- [29] M. A. Heart, "Automatic Acquisition of Hyponyms from Large Text Corpora Lexico-Syntactic for Hyponymy Patterns," in Proceedings of the 14th conference on Computational linguistics, Stroudsburg, PA, USA, 1992, pp. 539–545.
- [30] H. Emami, H. Shirazi, A. A. Barforoush, and M. Hourali, "A Pattern-Matching Method for Extracting Personal Information in Farsi Content," *U.P.B. Sci. Bull., Ser. C*, vol. 78, no. 1, pp. 125–138, 2016.
- [31] R. Al-Rfou, V. Kulkarni, B. Perozzi, and S. Skiena, "Polyglot-NER: Massive Multilingual Named Entity Recognition," in Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, British Columbia, Canada, 2015, pp. 586–594.
- [32] E. Minkov, R. C. Wang, and W. W. Cohen, "Extracting Personal Names from Email : Applying Named Entity Recognition to Informal Text," *Comput. Linguist.*, pp. 443–450, 2005.
- [33] Y. Chen, S. Y. Mei Lee, and C. R. Huang, "A robust web personal name information extraction system," *Expert Syst. Appl.*, vol. 39, no. 3, pp. 2690–2699, 2012.
- [34] Z. M. Arani and A. Abdollahzadeh Barforoush, "Semantic Role Labeling using Syntactic Dependency Analysis and Noun Semantic Category," in 20th Annual Conference of Computer Society of Iran, Mashhad, Iran (In Farsi), 2015, pp. 619–624.
- [35] K. Kipper, A. Korhonen, N. Ryant, and M. Palmer, "A large-scale classification of English verbs," *Lang. Resour. Eval.*, vol. 42, no. 1, pp. 21–40, 2008.
- [36] H. Mohagheghiyani, "Comparison of Persian Syntactic Dependency Parsers," *Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran*, 2015.
- [37] M. Surdeanu, "Overview of the TAC2013 Knowledge Base Population Evaluation: English Slot Filling and Temporal Slot Filling," in Proceedings of the Sixth Text Analysis Conference (TAC 2013), Maryland, USA, 2013.
- [38] M. Surdeanu and H. Ji., "Overview of the English Slot Filling Track at the TAC2014 Knowledge Base Population Evaluation," in Proceedings of Text Analysis Conference (TAC2014), Maryland, USA, 2014.
- [39] C. Fellbaum, "WordNet: An Electronic Lexical Database," MIT Press, 1998.
- [40] M. Shamsfard, A. Hesabi, H. Fadaei, N. Mansoory, A. Famian, S. Bagherbeigi, E. Fekri, M. Monshizadeh, and S. M. Assi, "Semi Automatic Development Of FarsNet: The Persian Wordnet," in Proceedings of 5th Global WordNet Conference, Mumbai, India, 2010.
- [41] C. J. Fillmore, C. R. Johnson, and M. R. L. Petruck, "Background to FrameNet," *Int. J. Lexicogr.*, vol. 16, no. 3, pp. 1–28, 2002.

- [42] C. Bonial, K. Stowe, and M. Palmer, "Renewing and revising SemLink," in *The GenLex Workshop on Linked Data in Linguistics*, Pisa, Italy, 2013, pp. 9–17.
- [43] G. Angeli, A. Chaganty, A. Chang, K. Reschke, J. Tibshirani, J. Y. Wu, O. Bastani, K. Siilats, and C. D. Manning, "Stanford's 2013 KBP System," in *Proceedings of the Sixth Text Analysis Conference (TAC2013)*, Maryland, USA, 2013.
- [44] J. Christensen, S. Soderland, and O. Etzioni, "Semantic Role Labeling for Open Information Extraction," in *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, Los Angeles, California, 2010, pp. 52–60.
- [45] A. Fader, S. Soderland, and O. Etzioni, "Identifying relations for open information extraction," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Edinburgh, UK, 2011, pp. 1535–1545.
- [46] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.
- [47] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.
- [48] K. Balog, J. He, C. Monz, M. Tsagkias, K. Hofmann, V. Jijkoun, W. Weerkamp, and M. De Rijke, "The University of Amsterdam at WePS2," in *2nd Web People Search Evaluation Workshop (WePS 2009)*, 18th WWW Conference, Madrid, Spain, 2009.
- [49] Y. Chen, S. Lee, and C. Huang, "Polyuhk: A robust information extraction system for web personal names," *2nd Web People Search Eval. Work. (WePS 2009)*, 18th WWW Conf. Madrid, Spain, 2009.
- [50] J. Ariles, J. Gonzalo, and S. Sekine, "Weps 2 evaluation campaign: overview of the web people search clustering task," in *2nd Web People Search Evaluation Workshop (WePS 2009)*, 18th WWW Conference, Madrid, Spain, 2009.
- [51] I. Nagy, "Person Attribute Extraction from the Textual Parts of Web Pages," *Acta Cybern.*, vol. 20, no. 3, pp. 419–440, 2012.

Hojjat Emami received his BSc degree in Software Engineering from University of Tabriz, Iran. He received his MSc degree in Artificial Intelligence from University of Tabriz, Iran. He is currently a Ph.D student at the Department of Information and Communication Technology (ICT), Malek-Ashtar University of Technology, Tehran, Iran.

Hossein Shirazi received his BSc from Mashhad University, Iran. He received his MSc and Ph.D in Artificial Intelligence from the University of New South Wales, Australia. He is currently an associate professor at the Malek-Ashtar University of Technology, Iran.

Ahmad Abdollahzadeh Barforoush is a professor in Computer Engineering and IT Department of Amir Kabir University of Technology, Iran. He is the author of books entitled "Introduction to Distributed Artificial Intelligence" and "Software Quality Assurance Methodology". His research areas are: data quality, artificial intelligence, agent-based systems, automated negotiation, expert systems, natural language processing, decision support systems, business intelligence, data mining, data warehouse and software engineering.

An Effective Risk Computation Metric for Android Malware Detection

Mahmood Deypir*

Faculty of Computer and Information Technology, Shahid Sattari University of Science and Technology, Tehran, Iran
mdeypir@ssau.ac.ir

Ehsan Sharifi

Faculty of Computer and Information Technology, Shahid Sattari University of Science and Technology, Tehran, Iran
sharifi@ssau.ac.ir

Received: 15/May/2016

Revised: 05/Sep/2016

Accepted: 07/Sep/2016

Abstract

Android has been targeted by malware developers since it has emerged as widest used operating system for smartphones and mobile devices. Android security mainly relies on user decisions regarding to installing applications (apps) by approving their requested permissions. Therefore, a systematic user assistance mechanism for making appropriate decisions can significantly improve the security of Android based devices by preventing malicious apps installation. However, the criticality of permissions and the security risk values of apps are not well determined for users in order to make correct decisions. In this study, a new metric is introduced for effective risk computation of untrusted apps based on their required permissions. The metric leverages both frequency of permission usage in malwares and rarity of them in normal apps. Based on the proposed metric, an algorithm is developed and implemented for identifying critical permissions and effective risk computation. The proposed solution can be directly used by the mobile owners to make better decisions or by Android markets to filter out suspicious apps for further examination. Empirical evaluations on real malicious and normal app samples show that the proposed metric has high malware detection rate and is superior to recently proposed risk score measurements. Moreover, it has good performance on unseen apps in term of security risk computation.

Keywords: Mobile Device Security; Risk Computation; Android Malwares; Critical Permissions; Security Metric.

1. Introduction

Android becomes the most popular operating system for smartphones and tablets which made its users the largest target group for security threats. This operating system security architecture reduces the attack surface by restricting applications using permissions and sandboxing. Therefore, in order to perform malicious activities, e.g., stealing user's data, sending premium messages and making phone call, an attacker must deceive users to install a malicious app since other ways of intrusion are almost closed in Android. For installing an app, Android requires the user to grant privileges through the requested permissions. There are large number of applications (Apps) developed for this operating system which requires various permissions based on their functionalities. For an application, these permissions are displayed in the first screen of the installation program. The end user of an Android based mobile device must approve these permissions or discard to install the application. The privileges are remain unchanged until they are revoked from the app when the user issues the app removal process. Although, this security mechanism is very simple and straight forward for users, it causes many challenges. First, users usually does not spend much time for studying the permissions and think about their effects. Therefore, they tend to go forward and to complete the installation process. Moreover, an ordinary user does not have

technical skills about the Android permissions and their impacts. Therefore, this security model is not effective regarding to security and privacy of end users in order to preserve their personal information from disclosure or to prevent monetary resource abuse by various type of potential malwares. Consequently, an Android malware e.g., spyware, Trojan, Adware, can deceive the users by introducing itself as a useful app and stole their personal or business data as well as using their mobile phone credit and monetary. There exists some research regarding to enhance the Android security model and its security risk communication mechanism. Using better and intuitive titles for permissions, categorization of permissions based on their effects, reducing the number of permissions by merging similar ones, utilizing user reviews about apps, using visual security indicators for risky apps, and etc. are some samples of these efforts [1-6]. Additionally, a number of statistical and mining models have so far been presented in order to measure the security risk of Android apps. The number of critical permissions and the number of critical permissions combinations requested by an app are simple examples of the statistical measures of security risk for apps [2]. Based on an effective security measure, it can be possible to compute the security risk of an app and fire a warning signal to the user if the computed risk exceeds a predetermined threshold. Moreover, the users can compare similar functionality apps in term of their risk scores. Furthermore, Android markets require an

* Corresponding Author

effective risk computation metric to identify suspicious apps among vast number of newly submitted apps by developers for further examination. The reason is detailed analysis and deterministic malware detection for each app is a very time consuming process and systematic filtering of low risk apps is an important requirement. However, our evaluations show that current measures and models of Android risk computation do not have acceptable performance. That is, they don't compute relative high risk values for known malwares and low risk quantities for benign apps to well recognize malicious apps from non-malicious ones. In this paper, a new security risk score measurement has been proposed which has better performance with respect to previously proposed ones. This risk score benefits from statistics of permission usages in known malicious and clean apps. However, it can be simply extended to other features of Android apps including static and dynamic ones. Moreover, we have attempted to give better definition of permission criticality to aim users for making the best decision for new apps installation. We have shown effectiveness of the proposed metric through extensive experiments on large number of real Android app samples including both malwares and goodwares. The paper is organized as follows. In the next section, some previous research works regarding to Android security and malware detection are reviewed. The problem statement is presented in Section 3. In Section 4, the new security risk score metric is introduced. In this section, our algorithm for risk computation by the proposed metric is also described. Extensive experimental evaluations of the proposed measure with respect to previously proposed ones are presented and illustrated in Section 5. These experiments have been performed using known malwares in the Android world and ordinary useful apps belong to Google App store. Finally, Section 6 concludes the paper.

2. Related Works

The user of a mobile phone participates in their device security by approving requested permissions of an app or decline the permissions which is equal to cancel the installation process. Research findings show that, most users discard checking permissions requested by an Android app. There are researchers trying to overcome this problem and thus enhance the Android security architecture [3-6]. However, Android security architecture requires a simple and straightforward for risk computation of new untrusted applications. Felt et al [3] proposed solutions like changing the categorizations of the Android permissions, emphasizing on the security risk instead of permissions, and a method of approving permissions. In [7] it is suggested that a high level critical information access regarding to the user privacy including personal data, location information, and contact list are displayed instead of the permission names in the first installation page. However, similar to permission list, this high level information might be bypassed by end users. In order to

reduce the required space for displaying permissions and assisting the user for fast and effective decision making, in [1] visualizing summary risk and safety scores are suggested. These scores are quantities which can be computed based on various permissions requested by an app. It is shown that, for most users, displaying a summary of risk or safety scores by graphical indicators are more effective than textual information of the permissions in term of user notification. However, metric of risk or safety value computation for untrusted apps is not the main concern in [1]. Peng et al [8] introduce statistical measures and mining models to compute security risk scores and ranking apps based on the requested permissions. The approach can rank the applications in an Android app store like Google play based on their security risk values. Such a ranking aims the users to select more secure apps where there exist a number of apps with the same functionality and different security risk values. Moreover, similar definitions were introduced for the concept of security risk regarding to the list of permissions requested by apps. In [2], the work in reference [8] has been extended and a number of statistical and probabilistic generative risk scores for Android apps using permission usage patterns have been precisely described. All of the measures are defined based on the concept of critical permission which is defined as a permission which can access sensitive software and hardware mobile resources and its usage pattern in malicious apps. An Android malware usually abuses critical permissions and corresponding API functions within its code to perform a malicious activity. The proposed risk scores in [2] and [8] are generative and are mainly computed using benign apps permission usage information. However, for improving the performance, authors increase the impact of some critical permissions on the resulting risk score values. They manually selected nine critical permissions that can be misused by malwares but details of the approach for critical permission selection was not described. In fact, a systematic approach for recognizing critical permissions using information contained in previously known malicious and non-malicious apps is required. An automated system called RiskRanker was introduced in [9] to examine whether a particular app is risky in term of having dangerous behavior. While a mobile antivirus rely on known malware signatures in a reactive manner, RiskRanker system can proactively spot zero day Android malwares. Since deterministic detection of zero day malwares requires further analysis, the system can be used as a preprocessing step to sift through a large number of apps from an Android market by producing a prioritized list of suspicious apps based on their computed security risk. However, for risk computation of untrusted apps RiskRanker only relies on analysis of known malwares and does not take the information of known benign apps into account. Enck et al. [10] developed a system named Kirin which examines combinations of risky permissions to determine whether the permissions requested by an app satisfy a certain global safety policy.

In this system, permission combinations e.g., WRITE_SMS and SEND_SMS are manually specified. These combinations could be used in a malicious apps and therefore, are used to identify malwares. However, a systematic approach for identifying risky permissions or combination of them is required.

A number of approaches have been proposed in the literature to classify Android apps into malwares and benign apps [11-13]. The aim is to construct a mining model like naïve bayes, based on labeled apps augmented by some information regarding to static and dynamic behavior of malwares and clean apps in order to classify future malwares. However, in this context classification models usually suffer from significant misclassification error on unseen data since there is not a crisp boundary between malicious and non-malicious apps. Therefore, measuring amount of risk for newly unseen apps is preferable for decision making compared to deterministic malware detection by classification models. There is other category of researches which use static code analysis of decompiled apps to analysis malicious activates and behaviors within malwares. In this approach, permission to function mapping is performed as a preprocessing step to recognize which function calls are used and what is their ordering. For example, accessing contact list or storage and then sending a SMS is a malicious behavior used in some malwares. In this way, the extracted knowledges and patterns are used to distinguish malicious apps from ordinary applications [14-17]. Malware detection and risk score computation based on static source code analysis can be regarded as complementary method for permission analysis. However, it faces some challenges like code obfuscation and code writing techniques exploited by malware writers which prevent to extract suitable features for risk computation. Dynamic behavior analysis of the running Android apps is another method to detect malwares [18-21]. In this approach, an app is running in a testing environment to identify when and how a part of code is executed and which resources are misused. Both static and dynamic analysis are time consuming processes. Ordinary users and Android markets require fast approach of risk computation.

Permission based security analysis and malware detection are considered by a large number of researches. This is due to its simplicity, explainability, effectiveness, and faster analysis. Moreover, it can be augmented by static and dynamic analysis. A main drawback of this approach is unused permissions of apps since an app can request a permission without actually using it, i.e., over privileged Android apps. This offers opportunities to malware developers to gain access to otherwise inaccessible resources. However, this shortcoming can be overcome by static and dynamic analysis of source code and technique like the function to permission mapping in order to confirm permission usage and remove unused permissions. In [22], a certification technique, is proposed to identify over privileged application in the direction of better risk management assessment. In this technique both

runtime information and static analysis are combined to profile mobile applications and identify if they are over privileged or follow the least privilege principle. Coarse grain nature of permission is another problem since granting a permission for an app is equal to allow it to call a couple of API functions. Fortunately, almost all security measures, analysis, and classification based on permissions can also be extended to work using function calls in order to obtain more detailed evaluations. Other challenges and arising issues regarding to Android based security analysis including, incompetent permission administration, insufficient permission documentation, over claim of permissions, permission escalation attack, and TOCTOU (Time of Check to Time of Use) attack were reviewed in [23] and existing countermeasures were addressed. These findings are useful for better risk estimation using requested permissions. Barbara et al [24] proposed an approach to evaluate security models based on permissions by using the self-organizing maps (SOM). They apply the approach on thousands of apps in order to analysis permission distributions. They showed that, how requesting permissions by apps is related to applications categorization. Analyzing decompiled source code of an Android app was used in [25] in order to detect data leak within the app. In [26] a security tool named MAST has been developed to identify high probable malware apps using static code and permission usage analysis. PScout [27] is another Android security tool developed for source code analysis to extract permission to function mapping. Applying this tool on the Android source code reveals that its permission system has a little redundancy and this property remains stable within newer versions of the operating system. DREBIN is a system which works based on detailed set of static features of apps including function call, permission list and hardware usage to recognize malware by an SVM based classifier [28]. Androgaurd is a reverse engineering tool to disassemble and to decompile Android apps. It is designed to analyze malicious and non-malicious Android apps [29]. Some malicious apps repackaged malicious codes into benign apps and spread the resulting malwares for easily deceiving end users. Although, this method can be prevented by verifying digital signature of the original apps, some end user might be deceived. In [30], a mechanism named SCSdroid (System Call Sequence Droid) is devised which adopts the thread-grained system call sequences used by apps to extract the truly malicious common subsequences from the system call sequences to identify repackaged malicious apps without requiring the original benign applications. Static dataflow analysis of malwares and goodwares have been utilized in [31] to construct a k-nearest neighbor based classifier. In this classifier, dataflow related API-level features of malicious and non-malicious apps have been used as training samples for future malwares detection. Feizollah et. al in [32] review various types of features including static features, dynamic features, hybrid features and applications metadata which are used in the literature for

Android malware detection. Deterministic recognition of Android malwares encounters some challenges since the boundary between malwares and goodwares is not crisp. Therefore, it is preferred to compute security risk scores of apps instead of binary warning signal regarding to being malwares or goodwares. However, it requires to have effective risk score measurements for precise estimation of the risk value. In this study, we have proposed a new risk score measurement based on a decision making architecture to aim user and systems for making better decisions related to potential Android malwares.

3. Problem Statement

As mentioned previously, users require a convenient method to detect malicious application and make a correct decisions. However, at any time, all malwares and their signature are not fixed and known, i.e., zero day malwares. Therefore, in order to fire a warning signal about using a suspicious application a risk score measurement is desirable. This measure can be exploited in a security tool or embedded in Android to warn a user about malicious apps. It can utilize different aspect of an app to compute its security risk value. These aspect include, permissions, function calls, static or dynamic behavior and etc. Android permissions show what might be called or used in an app. In order to perform malicious activities, a malware requires using critical permissions. Critical permissions are those that can give an app access to sensitive resources and information. Here, permission list of apps are utilized in order to compute their security risk. We assume that there is a set P containing $|P|$ permissions in a mobile operating system: $P = \{p_1, p_2, p_3, \dots, p_n\}$. A mobile application A can request a subset of P to perform its activities. We use a binary variable named x_{Ap} to represent the status of permission x_p in application A . In the other words, x_{Ap} can be set when the permission x_p is requested by an application A . Otherwise it is unset. The problem is to measure the security risk of an input application A using its requested permissions. This measurement requires a formulation and a model which can well exploit historical statistics about previously known malwares and useful apps. For example consider the Table (1) which contains information regarding to permissions requested by a number of known apps including both malwares (+) and goodwares(-).

Table 1. Information about some malwares(+) and useful apps(-)

ID	Permissions	Malware
1	INTERNET, READ_PROFILE	-
2	BATTERY_STATS, BLUETOOTH	-
3	BROADCAST_SMS, WRITE_SMS	+
4	INTERNET, INSTALL_PACKAGE, READ_SMS	+
5	READ_SMS, WRITE_EXTERNAL_STORAGE	-
6	BATTERY_STATS, INTERNET	-
7	INSTALL_PACKAGE, READ_PROFILE	-
8	INTERNET, READ_SMS, BLUETOOTH	-

In this table, the second column shows the list of permissions really used in each app. For each app, the status or label of being malicious or useful are depicted in third column. A risk score of an unlabeled app is a value which can be computed based on the list of its permissions. The criticality of each permission is not pre-determined and changes over time. Since the permissions have different criticalities based on their historical usage or misuse, their contribution in computing risk score might be different from each other. A permission's criticality value can be related to its nature and amount of its usage in the previously known malware and goodwares.

An effective security risk score must compute higher values for malware samples than benign apps instances. The more relative risk score value for an untrusted app, the more potentiality of being malware is. In this study, the aim is to propose an effective, simple, and explainable security risk measurement. This measure of security can be used for user warning signal when they are going to install or use a suspicious application. Moreover, it can be used for apps prioritization based on their security risk or safety. Therefore, our aim is not to classify Android apps into malwares and goodwares but we are going to propose a security risk metric which is meaningful for both malicious and useful apps and can well distinguish malwares from goodwares by assigning higher risk values to malicious apps. Therefore, effectiveness of a risk measurement means having high detection rate for malwares within a set of unlabeled apps. Figure (1) illustrates the overall process of our decision making architecture based on the risk computation.

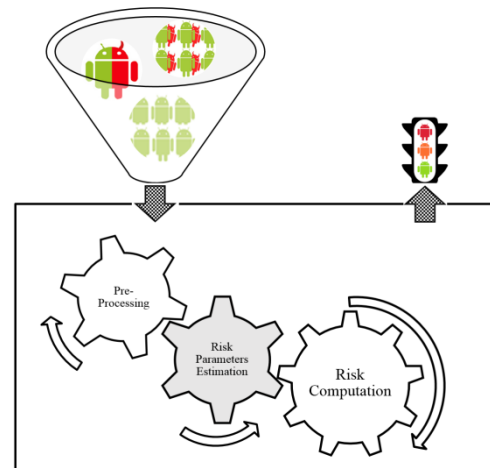


Fig. 1. Overall decision making architecture.

This process computes the risk of untrusted apps using analyzed previously known malicious and clean app samples. As shown in this figure, labeled malwares and benign apps are used to construct the model which consists of three main stage including data pre-processing, risk parameter estimation, and risk computation. The constructed model uses an effective measurement to compute risk of future input apps. In fact, the risk of an untrusted app or set of apps can be computed by the model. The computed risks can be seen as a guiding light

for selecting low risk apps for usage or selecting high risk apps, i.e., potential malwares for further analysis which leads to identify Android malicious apps.

4. The Proposed Method

Our evaluation shows that previously proposed criterions for risk measurements of Android apps do not have good performance because they operate based on imprecise definition of the criticality for permissions. We require a simple risk score which precisely benefits from the underlying statistics of known malwares and benign apps and exploit their discrimination power of permissions for identifying new malwares. In order to analysis statistical properties of permissions in apps and defining an effective risk score measure, we have used thousands of normal and malicious Android apps. For each app sample, requested permissions can be extracted using the Android Manifest.xml file exist inside the *apk* package file. Before, gathering statistics, a preprocessing can be performed to remove duplicate apps, i.e., several different versions of the same app and removing useless permissions by permissions to function mapping within each app. We have numbered permissions based on their alphabetical order from 1 to $|P|$ where P is the set of permissions in Android operating system. In order to obtain a better risk score metric based on permissions, 808 malwares and 71331 benign apps are analyzed. In this study, we have proposed a risk score measurement for effective risk computations of Android apps. As mentioned in [2], a good risk measurement has two main properties, high detection rate and high explainability.

In the devised measurement, we have designate a new formulation to assign higher risk values to permissions which have higher usage in malwares and very lower usage in benign apps. The idea is quite simple but produces interesting results. That is, the security risk of a permission is directly related to its usage in malware and inversely proportional to its usage in non-malicious apps. Given estimated risk values of permissions, one can compute risk of an Android app based on its permission list. We name our risk metric as *RF* (Rarity and Frequency based risk metric). Since the proposed measurement computes the risk values of permissions according to simple statistics of known malwares and useful Android apps, it has good explainability. Users can be effectively informed regarding to danger about approving risky permissions. They can made reasonable decision based on total risk score of an app which can be simply computed using security risks of its requested permissions. Moreover, Android markets can use the devised metric to handle the large number of daily submitted apps for security analysis by filtering out top most risky apps and examining them using time consuming and deterministic malware detection methods.

A. RF Metric

As mentioned previously, we require a simple risk score which precisely benefits from the underlying statistics of known malwares and benign apps. We leverage permission statistics of both malwares and goodwares to devise an effective risk metric. Permissions which are used frequently in malicious apps and rarely required by normal apps must have more impact on risk score measurement. For each permission, frequency of usage in malwares or rarity of usage in goodwares are not solely symptoms of having high risk. The reason is, it might also have high usage in both malwares and goodwares. On the other hand, requesting a permission might be rarely occurred in both normal and malicious apps. Therefore, an effective risk metric must take both of rarity in normal apps and frequency in malwares into account. In the proposed metric, for each Android permission, its frequency in both benign apps and malwares are considered. Based on this idea we have designed *RF* metric for computing security risk of apps according to the following equation:

$$s \quad RF(x_i) = \sum_{p=1}^{|P|} x_{ip} \cdot \left(\frac{N}{C_{pb} + \varepsilon} \right) \cdot \left(\frac{C_{pm}}{M} \right). \quad (1)$$

In the above equation, $|P|$ and x_{ip} are total number of permissions and status of p th permission in app x_i , respectively. Moreover, C_{pm} and M are usage count of p th permission in available malicious apps and total number of malwares, respectively. Finally, N and C_{pb} are total number of training benign app samples and the count of permission usage in the set of these samples, respectively. ε is a very smaller value used to prevent infinite or undefined numbers where the permission is not used by any analyzed normal apps. In this formulation, C_{pb} is computed as follow:

$$C_{pb} = \sum_{i=1}^N x_{ip}. \quad (2)$$

In the above equation, $x_{ip} = 1$ if i th app x_i uses p th permission and $x_{ip} = 0$ otherwise. Similarly C_{pm} is computed according to equation (3):

$$C_{pm} = \sum_{i=1}^M x_{ip}. \quad (3)$$

In formulation (1), the higher the score, the more risky the application is. In fact, for an app x_i , formulation (1) is the summation of risks for used permissions in the app. Therefore, *RF* metric can be also defined for each permission x_p as:

$$RF(x_p) = \left(\frac{N}{C_{pb} + \varepsilon} \right) \cdot \left(\frac{C_{pm}}{M} \right). \quad (4)$$

The symbols are defined similar to the previous equations. We named this risk score measurement, Rarity and Frequency based risk score measurement (*RF*) since for risk score computation, it takes the impact of both rarity of permissions in benign apps, i.e., the first component of the equation and frequency of them in malicious apps, i.e., the second component of the

formulation, into account. As can be inferred from formulation (4), a permission which is used more frequently in normal apps, its impact on risk computation of apps is reduced. On the other hand, a critical permission is frequently requested by malwares. For a permission, as the frequency in malwares and rarity in clean apps is large it is more critical and more risky.

For 20 top most obtained risky permissions, Table (2) represents their rank based on RF metric, their weights in malwares and goodwares, and their RF risk values. The permissions are sorted in descending order of their RF weights. RF values are not normalized and are computed according to equation (4).

Table 2. Information of top most critical permissions based on RF weights

Rank Based on RF	Permission Name	Usage in malwares	Usage in benign apps	RF Metric
1	WRITE_APN_SETTINGS	0.324	0.003	108
2	INSTALL_PACKAGES	0.218	0.003	72.667
3	DELETE_PACKAGES	0.062	0.001	62
4	WRITE_SMS	0.562	0.016	35.125
5	READ_SMS	0.679	0.023	29.522
6	DISABLE_KEYGUARD	0.29	0.014	20.714
7	READ_LOGS	0.269	0.013	20.692
8	RESTART_PACKAGES	0.348	0.022	15.818
9	WRITE_CONTACTS	0.417	0.031	13.452
10	MOUNT_UNMOUNT_FILESYSTEMS	0.104	0.008	13
11	RECEIVE_SMS	0.46	0.036	12.778
12	CHANGE_WIFI_STATE	0.262	0.021	12.476
13	SEND_SMS	0.489	0.043	11.372
14	RECEIVE_BOOT_COMPLETED	0.566	0.059	9.5932
15	ACCESS_WIFI_STATE	0.671	0.076	8.8289
16	ACCESS_LOCATION_EXTRA_COMMANDS	0.126	0.016	7.875
17	CALL_PHONE	0.415	0.069	6.0145
18	READ_CONTACTS	0.392	0.085	4.6118
19	READ_PHONE_STATE	0.931	0.222	4.1937
20	ACCESS_NETWORK_STATE	0.808	0.2941	2.7474

As would be seen in the experimental section, the relative values of estimated risks are considered to compute and to compare the risks of apps. As shown in Table (2), a permission with a relative high usage weight in malwares, might have a lower RF weight and thus low rank with respect to the other permissions. For example in this list `READ_PHONE_STATE`, has most usage in malwares but it has nineteenth rank regarding to RF risk value. On the other hand, for a permission, the rarity of usage in benign apps solely does not determine amount of risk value since it might be also rarely requested by malwares. For instance, in Table (2), `DELETE_PACKAGES` is rarer than `WRITE_APN_SETTINGS` and `INSTALL_PACKAGES` but it is less risky than these permissions. Based on the proposed risk score measurement we have re-defined the criticality concept of permissions in Android platform.

Criticality of a permission: It is a relative and variable property which directly proportional to its usage in the current malware samples, and inversely proportional to its normal usage in benign apps.

There are some important points regarding the above definition. First, the criticality is a relative property. That is, we cannot categorize permission into two separate sets, i.e., critical and not critical. In the other words, the permission can be compared together based on their criticality or risk value. This value can be estimated using a metric like RF in equation (4). The second point is regarding to the variable nature of the criticality. That is, based on permission usage pattern for current malwares and useful apps development, the criticality of permissions and number of critical permissions might be changed over time. It is obvious that the amount of risk for a permission is not fixed and must be periodically recomputed or updated due to developing new malwares and thus new permissions usage patterns. Finally, the last point is about approach for accessing critical resources and sensitive data through permissions by malicious apps. Malware developers are not interested in using some permissions to perform malicious activities due to some reasons despite critical resources and private data access through the permissions. For example, based on our analyses which is partly shown in Table (2), permissions related to using Bluetooth capabilities, i.e., `BLUETOOTH` and `BLUETOOTH_ADMIN` are not used frequently in malicious apps and have RF values very close to zero. This might be due to restrictions of using such capabilities.

B. The Algorithm

In this section, the pseudo code of algorithms for computing risk of permissions and apps based on the proposed RF metric are described. In these algorithms, it is supposed that preprocessing is performed on all app samples including malwares, benign apps, and untrusted input apps. The preprocessing consist of permission extraction, removing unused permissions, and removing duplicate apps, i.e., various versions of distinct apps, and etc. Figure (2) depict pseudo code of the algorithm for computing RF metric for the permissions based on training normal and malicious app samples. Algorithm for prioritizing a set of apps based on their security risk value is shown in Figure (3). This algorithm gets three parameters named SP , SB , and SM which are the set of Android permissions, set of benign app samples, and set of malwares, respectively. In line 1, the number of normal and malicious apps are obtained. In lines 2 through 14 for all permissions, the RF metric is computed. For each permission, in lines 3 through 7 counts of the permission usage in normal apps is accumulated in C_{pb} variable. Similarly, using lines 8 through 12 similar counting and accumulation is performed for malwares using C_{pm} variable. According to equation (4), in line 13, for each permission x_p , C_{pb} and C_{pm} are used to compute risk value of the permission based on the rarity value of the permission in normal apps and its frequency value in malicious apps, respectively. Finally, in line 15 a list containing computed risk values of all permissions is returned. These values are used for computing risk values of input apps which is described by the next algorithm.

Algorithm RFCompute(SP, SB, SM)	
Begin	
1.	$N = SB ; M = SM ;$
2.	for each permission $x_p \in SP$ do
3.	for each sample $s \in SB$ do
4.	If x_p is requested by s then
5.	$C_{pb} = C_{pb} + 1;$
6.	end if;
7.	end for;
8.	for each sample $s \in SM$ do
9.	If x_p is requested by s then
10.	$C_{pm} = C_{pm} + 1;$
11.	End if;
12.	End for;
13.	$RF(x_p) = (\frac{N}{C_{pb} + \epsilon}) \times (\frac{C_{pm}}{M})$
14.	End for;
15.	return $RF;$
End;	

Fig. 2. Risk computation for set of permissions

The overall structure of *RFCompute* algorithm consists of an outer loop and two inner loops. The number of rounds for outer loop is equal to $|P|$ which is the number of permissions in *SP* set. First and second inner loops have $|SB|$ and $|SM|$ numbers of iterations, respectively which are the sizes of benign set and malware set, respectively. Usually the number of analyzed benign apps are greater than the number of malicious apps as you can see in our analysis and experimentation. Therefore, the complexity of *RFCompute* algorithm is calculated as follow:

$$O(RFCompute) = O(|P| \times (|SB| + |SM|)) = O(|P| \times |SB| + |P| \times |SM|) = O(|P| \times |SB|) \quad (5)$$

Therefore, as the number of analyzed app is increased, a more time is required for risk computation of Android permission. However, the process of risk computation is performed once and risk value of each future app can be computed using obtained risk of permissions.

Risk computation can be performed for an individual apps. However, risk values of Android apps are also meaningful where untrusted apps are compared together based on their risk or where high risk apps must be identified. For example, when a user wants to compare some same functionalities untrusted apps to select lowest one or when top most risky apps must be selected for further examination to identify zero day malwares in an Android market or in a user device. In this situations, a prioritized list of available untrusted apps is desirable. Figure (3) briefly describes risk computations based on *RF* metric for a list of preprocessed input apps in order to prioritize them according to their risks. In this algorithm, *SP*, *SA*, and *RF*, respectively, are set of permissions, set of untrusted input apps, and the list of computed *RF* risk values of the permissions as illustrated in the previous algorithm. In lines 1 through 3, risk of each input app is computed. In line 4 apps are sorted based their risk and finally the sorted list of apps as well as their risk values are returned in line 5. The sorting order can be either in

descending or ascending order based on the application of risk computation.

Algorithm RiskPrioritization(SP, SA, RF)	
Begin	
1.	for each app $x_i \in SA$ do
2.	$RF(x_i) = \sum_{x_p \in SP} (x_p \times RF(x_p));$
3.	end for;
4.	$SA = \text{Sort}(SA, RF);$
	// Sort input apps based on their <i>RF</i> risks in descending order
5.	return $SA;$
End;	

Fig. 3. Apps prioritization based on RF metric

C. An Example

For better describing the overall process based on the proposed metric, after preprocessing of malicious and clean app samples, consider the following example. In this toy example which is designed similar to a real situation, our approach for computing risk of the permissions and any app *A* is explained.

Example: Suppose that, there is a set of labeled apps including both malwares and useful apps according to Table (1). Here, it is not important how these apps were labeled.

In order to compute security risk score of unknown apps, the risk values of all permissions must be computed. For all Android permissions, statistics regarding to their rarities in goodwares and their frequencies in malwares must be computed to obtain risk score of future apps. Suppose that based on the above example, we have an unlabeled Android app *A* which requires *INSTALL_PACKAGES*, *INTERNET*, *READ_SMS*, and *BLUETOOTH* permissions. For these permissions, the value of rarity and frequency are computed. For the first permission, it is requested by one benign and one malicious apps. These values for the second permission are 3 and 1, respectively. *READ_SMS* is requested by 2 benign and one malicious apps, respectively. Finally, *BLUETOOTH* is requested only by two normal apps. Based on obtained values of rarity in benign apps and frequency in malwares, the security risk of this app according to equation (1) is estimated as:

$$RF(A) = RF(INSTALL_PACKAGES) + RF(INTERNET) + RF(READ_SMS) + RF(BLUETOOTH) = (6/1 \times 1/2) + (6/3 \times 1/2) + (6/2 \times 1/2) + (6/2 \times 0/2) = 3 + 1 + 1.5 + 0 = 5.5 \quad (6)$$

As can be inferred from the above computation, in this example, *BLUETOOTH* permissions don't have any contributions in the resulting value since it was not used by any malware. The computed risk value can be used to prioritize several apps based on their risks. For an app, having a security risk essentially is not a reason for being malicious but it is a warning signal for the user or can be used as a pre-processing step for more detailed analysis. Risk scores of apps are also relative values and can aid users to select low risk apps. That is, having more than one app with the same functionality and various security risk scores, selecting lowest risk app is a more preferable decision.

5. Experimental Evaluation

In order to evaluate the proposed risk score measurements, required codes are developed using Matlab 2013. We have obtained publically available preprocessed malwares and goodwares datasets as well as source codes of some previous approaches belong to authors of reference [2] from the web¹. For useful ordinary apps, Market 2011 and Market 2012 are used which we named them as Benign 2011 and Benign 2012, respectively, since they contains non-malicious apps of Google app store at year 2011 and 2012 A.D. These dataset contain permission information of 71331 and 136534 useful apps, respectively. Both malwares and benign apps datasets have 122 columns which are alphabetically ordered permissions of apps in recent versions of Android operating system. Table (3) summarizes characteristics of the used datasets for our evaluations.

Table 3. Android apps datasets specifications for evaluation and comparisons

Dataset Name	Number of Apps	Brief Description
Benign2011	71331	Useful apps of Google App store in 2011 A.D
Benign2012	136534	Useful apps of Google App store in 2012 A.D
Malwares	808	A number of known Android malwares

In order to compare the proposed measurement against previously proposed ones, our proposed *RF* and couple of previously proposed risk score measurements have been evaluated. Table (4) summarizes all of these metrics. Some of them are statistical and others are probabilistic mining models. The interested readers are referred to [2], [8], and [23] for more details.

Table (4): Summarization of previous risk scores

Risk Metric	Meaning
RCP	Rare Critical Permission
RPCP	Rare Pairs of Critical Permissions
RS	Rarity based risk Score
RSS	Rarity based risk Score with Scaling
BNB	Basic Naive Bayes model
PNB	Naive Bayes with informative Priors
MNB	Mixture of Naive Bayes models
HMNB	Hierarchical Mixture of Naive Bayes models
Kirin	Certain combinations of dangerous permissions

In the experimentation, the main concern is detection rate. That is, detecting malwares by assigning relative higher risk to them. Using Benign 2011 and Malware datasets, the detection rates are computed with respect to a range of warning rates from 0 to 1.

A. ROC Curves

Figure (4) shows resulting ROC curves of all metrics where horizontal and vertical axes are warning rate and detection rate, respectively. The only exception is Kirin method which contains fixed rules and does not require warning rate parameter. For this method, instead of ROC curve, its fixed detection rate value is depicted by a single point. In order to evaluate a risk metric, we have placed all malwares and ordinary apps in the same list and sort

them in descending order of computed security risk values of the metric. The more malwares placed in the top of sorted list, the stronger security risk score is. For evaluation, we use 10-fold cross validation approach. For this purpose, Benign 2011 and malwares are placed in the same list and at each fold, both models are made using 90 percent of the list. Using each model separately, the ordered remaining 10 percent of the list is obtained. For various percentage values, top most security risk score apps are selected from the ordered list. Subsequently, for each model's ordered list, it is determined that what percent of malwares are contained in the selected apps. In this setting, the percentage of selection from each ordered list and determined percentage of malwares are named warning rate and detection rate, respectively. In the other words, number of false positives and true positives are directly proportional to warning and detection rates, respectively. Although, the ranges of computed risk scores of the compared measurements are different, based on this approach, they can be fairly compared together since there is not any absolute warning rate threshold. It is obvious that, as a risk measurement is stronger, a larger number of malwares are resided on the top of the ordered list and thus the measurement has more detection rate. Additionally, a stronger risk score measurement has high detection rate in smaller warning rate e.g., 1%, 5% since in this experimental setting smaller warning rate is equal to smaller fraction of top most risk score apps. In the other words, the end user expects that the top high risk score apps are malicious not normal.

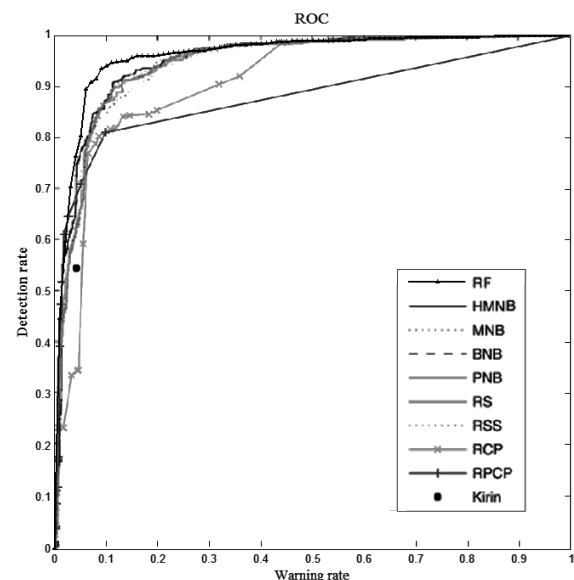


Fig. 4. Detection rate for various warning rates.

As can be seen from Figure (4), the proposed metric is superior to the other approaches especially for smaller warning rates. As warning rate increases, the performance gaps are reduced and all metrics converge to full detection rate. However, smaller warning rate is more desirable for user where number of false positives are smaller. Moreover, area under the curve for *RF* metric is close to one which shows the effectiveness of the proposed risk

1. <https://github.com/hao-peng/AppRiskPred>

score measurement. Therefore, obtained results confirms the superiority of *RF* in term of assigning relative higher risk values to malicious apps than non-malicious ones. The reason is, *RF* considers both rarity of permissions in normal apps and frequency of misused ones in malicious apps.

B. Area Under The Curves (AUC)

For better illustration of this experiment, Area Under Curve (*AUC*) of *ROC* curves are computed for various risk scores metrics since some *ROC* curves are very close to each other especially in larger warning rate values. The *AUC* is computed up for small warning rate values of *ROC* curves. The results is plotted in Figure (5).a and Figure (5).b for 1 percent and 5 percent of warning rates, respectively. Similar results are obtained for other values. As shown in this figure, the proposed *RF* measure has better performance than other metrics. In fact, *RF* is significantly better than other metrics especially for small warning rates where users are interested in. The reason is the better distinguishing power of *RF* which can better differentiate malwares from goodwares. Moreover, the proposed *RF* metric utilizes permission usages statistics of both malwares and goodwares together while the other risk scores mainly focuses on malware or goodware permission usage patterns or manually take the impact of malware statistics into account.

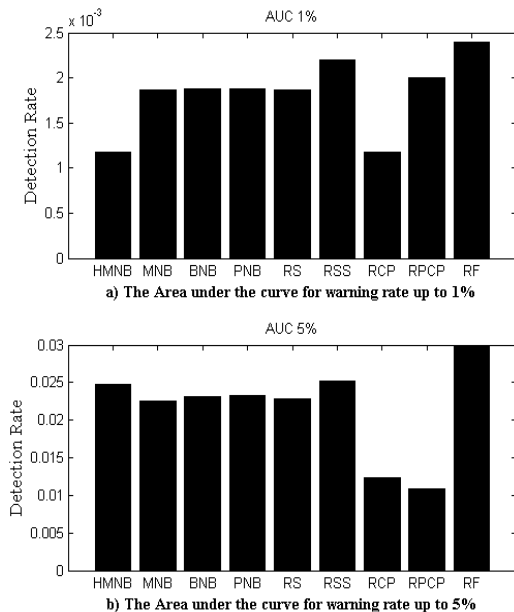


Fig. 5. Comparison of Area Under Curve (*AUC*) up to 1% and 5% warning rates.

For example, *RSS* which has closest detection rate to our proposed *RF* metric, considers only the rarity of permissions in benign apps augmented by scaling factors to increase the impact of some manually selected critical permissions. This measurement takes the weights of rare permissions into account and exploit it to compute estimated value of risk. However, a permission may be rarely used in both malicious and non-malicious apps.

C. Performance on Unseen Data

In order to evaluate generalization of the proposed risk measurement, we must apply it on unseen apps. For this purpose, we repeat the above experiment using Benign 2012 and malwares. That is, we obtain usage statistics of permission using Benign 2011 and test it on Benign 2012. In the other words, we use whole set of Benign 2011 for training and whole set of Benign 2012 for test. In training and testing phases, *RF* values of permissions are computed according to usage statistics of the permissions in Benign 2011. Subsequently, detection rates of various warning rates for Benign 2012 and Benign 2011 are computed and resulting *ROC* curves are obtained and shown in Figure (6). As can be seen from this figure, the metric has high performance for seen and unseen apps. However, for unseen apps, detection rate is slightly degrades. This is due to change in permission usage patterns in newly developed apps which leads to change in risk of permissions and apps. Therefore, in order to obtain better estimation of security risk, usage statistics of permissions must be periodically updated since the criticality values of permissions are not fixed.

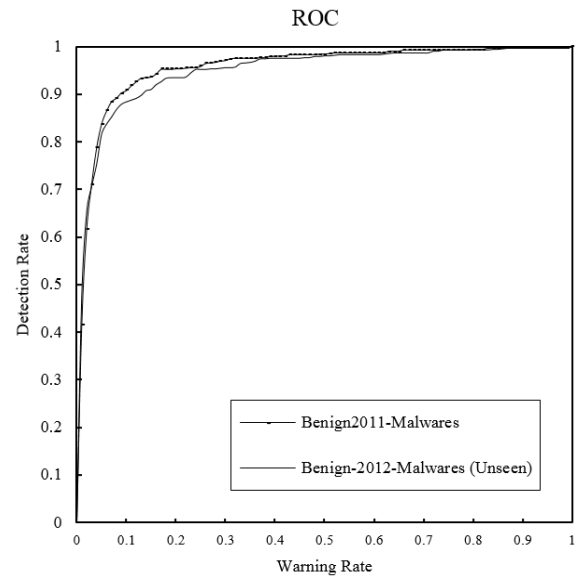


Fig 6. Performance of the proposed risk metric for seen and unseen apps

In fact, permission usage pattern of Android apps is changed over time since new apps with various services and capabilities and thus new permission requirements are introduced in the world. On the other hand, malware developers use new techniques to entice users for malicious apps installation which also leads to change in permission usage pattern.

6. Discussion and Conclusion

In this study, a new risk score metric namely *RF* is devised which has better detection rate with respect to other measurements due to precise identification of the critical permissions. Empirical evaluations on real

Android apps show that *RF* computes relative high risk values for known malwares rather than ordinary apps since it can well differentiate between permissions in term of their usage in malwares and clean apps. As a result, *RF* has high detection rate in comparison to previous risk score measurement. Moreover, the proposed measurement is highly explainable since it can be computed for an app by simply summation of the risk values of critical permissions requested by that app. Risk values of the permissions can be pre-computed using available known malwares and goodwares. An overview on top most critical permissions listed in Table (2) obtained by the proposed metric shows that these permissions are examples of those ones that an app can perform malicious activities by granting a subset of them. In this study, all analyzed malicious apps are categorized into the same

category named malwares. However, by using larger and categorized malware datasets we can compute risk scores more precisely. In the other words, exploiting prior knowledge of malware types including Trojan, Adware, Spyware and etc. could enhances the obtained performance since various malware types have different impacts and thus various security risk values. For example, an Adware can be less dangerous than a spyware. Computing *RF* for pair of permissions can further improve the performance of devised approach and thus obtaining better estimation of security risk values. Although the proposed approach is based on permission analysis it can be extended to or completed using other features like Android function calls and dynamic running flow analysis which contain more detailed information.

References

- [1] Gates, C. S., Chen, J., Li, N., & Proctor, R. W. (2014). Effective risk communication for android apps. Dependable and Secure Computing, IEEE Transactions on, 11(3), 252-265.
- [2] Gates, C. S., Li, N., Peng, H., Sarma, B., Qi, Y., Potharaju, R., & Molloy, I. (2014). Generating summary risk scores for mobile applications. Dependable and Secure Computing, IEEE Transactions on, 11(3), 238-251.
- [3] Chin, E., Felt, A. P., Sekar, V., & Wagner, D. (2012, July). Measuring user confidence in smartphone security and privacy. In Proceedings of the Eighth Symposium on Usable Privacy and Security (p. 1). ACM.
- [4] Felt, A. P., Greenwood, K., & Wagner, D. (2011, June). The effectiveness of application permissions. In Proceedings of the 2nd USENIX conference on Web application development (pp. 7-7).
- [5] Felt, A. P., Ha, E., Egelman, S., Haney, A., Chin, E., & Wagner, D. (2012). Android permissions: User attention, comprehension, and behavior. Tech. Rep. UCB/EECS-2012-26, UC Berkeley.
- [6] Kelley, P. G., Consolvo, S., Cranor, L. F., Jung, J., Sadeh, N., & Wetherall, D. (2012). A conundrum of permissions: installing applications on an android smartphone. In Financial Cryptography and Data Security (pp. 68-79). Springer Berlin Heidelberg.
- [7] Kelley, P. G., Cranor, L. F., & Sadeh, N. (2013, April). Privacy as part of the app decision-making process. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 3393-3402). ACM.
- [8] Peng, H., Gates, C., Sarma, B., Li, N., Qi, Y., Potharaju, R., & Molloy, I. (2012, October). Using probabilistic generative models for ranking risks of android apps. In Proceedings of the 2012 ACM conference on Computer and communications security (pp. 241-252). ACM.
- [9] Grace, M., Zhou, Y., Zhang, Q., Zou, S., & Jiang, X. (2012, June). Riskranker: scalable and accurate zero-day android malware detection. In Proceedings of the 10th international conference on Mobile systems, applications, and services (pp. 281-294). ACM.
- [10] Enck, W., Ongtang, M., & McDaniel, P. (2009, November). On lightweight mobile phone application certification. In Proceedings of the 16th ACM conference on Computer and communications security (pp. 235-245). ACM.
- [11] Jang, J. W., Kang, H., Woo, J., Mohaisen, A., & Kim, H. K. (2016). Andro-dumpsys: anti-malware system based on the similarity of malware creator and malware centric information. Computers & Security.
- [12] Sarma, B. P., Li, N., Gates, C., Potharaju, R., Nita-Rotaru, C., & Molloy, I. (2012, June). Android permissions: a perspective combining risks and benefits. In Proceedings of the 17th ACM symposium on Access Control Models and Technologies (pp. 13-22). ACM.
- [13] Cen, L., Gates, C., Si, L., & Li, N. (2015). A probabilistic discriminative model for android malware detection with decompiled source code, in Dependable and Secure Computing, IEEE Transactions on, vol.12, no.4, (pp.400-412). IEEE.
- [14] Desnos, A. (2012, January). Android: Static analysis using similarity distance. In System Science (HICSS), 2012 45th Hawaii International Conference on (pp. 5394-5403). IEEE.
- [15] Schmidt, A. D., Bye, R., Schmidt, H. G., Clausen, J., Kiraz, O., Yüksel, K., & Albayrak, S. (2009, June). Static analysis of executables for collaborative malware detection on android. In Communications, 2009. ICC'09. IEEE International Conference on (pp. 1-5). IEEE.
- [16] Zhou, Y., Wang, Z., Zhou, W., & Jiang, X. (2012, February). Hey, You, Get Off of My Market: Detecting Malicious Apps in Official and Alternative Android Markets. In NDSS.
- [17] Aafer, Y., Du, W., & Yin, H. (2013). DroidAPIMiner: Mining API-level features for robust malware detection in android. In Security and Privacy in Communication Networks (pp. 86-103). Springer International Publishing.
- [18] Christodorescu, M., Jha, S., & Kruegel, C. (2008, February). Mining specifications of malicious behavior. In Proceedings of the 1st India software engineering conference (pp. 5-14). ACM.
- [19] Rieck, K., Holz, T., Willems, C., Düssel, P., & Laskov, P. (2008). Learning and classification of malware behavior. In Detection of Intrusions and Malware, and Vulnerability Assessment (pp. 108-125). Springer Berlin Heidelberg. [17]
- [20] Shabtai, A., & Elovici, Y. (2010). Applying behavioral detection on android-based devices. In Mobile Wireless Middleware, Operating Systems, and Applications (pp. 235-249). Springer Berlin Heidelberg.

- [21] Burguera, I., Zurutuza, U., & Nadjm-Tehrani, S. (2011, October). Crowdroid: behavior-based malware detection system for android. In Proceedings of the 1st ACM workshop on Security and privacy in smartphones and mobile devices (pp. 15-26). ACM.
- [22] Geneiatakis, D., Fovino, I. N., Kounelis, I., & Stirparo, P. (2015). A Permission verification approach for android mobile applications. *Computers & Security*, 49, 192-205.
- [23] Fang, Z., Han, W., & Li, Y. (2014). Permission based android security: Issues and countermeasures. *computers & security*, 43, 205-218.
- [24] Barrera, D., Kayacik, H. G., van Oorschot, P. C., & Somayaji, A. (2010, October). A methodology for empirical analysis of permission-based security models and its application to android. In Proceedings of the 17th ACM conference on Computer and communications security (pp. 73-84). ACM.
- [25] Enck, W., Ocateau, D., McDaniel, P., & Chaudhuri, S. (2011, August). A Study of Android Application Security. In *USENIX security symposium* (Vol. 2, p. 2).
- [26] Chakradeo, S., Reaves, B., Traynor, P., & Enck, W. (2013, April). Mast: triage for market-scale mobile malware analysis. In Proceedings of the sixth ACM conference on Security and privacy in wireless and mobile networks (pp. 13-24). ACM.
- [27] Au, K. W. Y., Zhou, Y. F., Huang, Z., & Lie, D. (2012, October). Pscout: analyzing the android permission specification. In Proceedings of the 2012 ACM conference on Computer and communications security (pp. 217-228). ACM.
- [28] Arp, D., Spreitzenbarth, M., Hubner, M., Gascon, H., & Rieck, K. (2014, February). DREBIN: Effective and Explainable Detection of Android Malware in Your Pocket. In *NDSS*.
- [29] Desnos, A. (2013). Androguard-Reverse engineering, Malware and goodware analysis of Android applications. URL code. google.com/p/androguard.
- [30] Lin, Y. D., Lai, Y. C., Chen, C. H., & Tsai, H. C. (2013). Identifying android malicious repackaged applications by thread-grained system call sequences. *computers & security*, 39, 340-350.
- [31] Wu, S., Wang, P., Li, X., & Zhang, Y. (2016). Effective detection of android malware based on the usage of data flow APIs and machine learning. *Information and Software Technology*, 75, 17-25.
- [32] Feizollah, A., Anuar, N. B., Salleh, R., & Wahab, A. W. A. (2015). A review on feature selection in mobile malware detection. *Digital Investigation*, 13, 22-37.

Mahmood Deypir received his Ph.D. in 2011 and M.Sc. in 2006, both from Shiraz University. He is currently assistant professor in the Computer and Information Technology department at Shahid Sattari University of Science and Technology. His research interests include Data Mining and Cyberspace Security. He has published a number of papers in ISI journals and international conferences.

Ehsan Sharifi received the B.Sc. degree with honors in software engineering from the Shahid Sattari University in 2003 and the M.Sc. degree in software engineering from the PNU University of Tehran, in 2012. He is currently a PhD candidate in software engineering at the Amirkabir University of Technology of Tehran. In 2004, he joined the Department of Computer Engineering and Information Technology of Shahid Sattari University. His current research interests include the Software Quality, Software Modeling, Ontology Engineering, Network Security, and Fuzzy Systems. He has published numerous papers in leading academic journals and conference.

A New Node Density Based k -edge Connected Topology Control Method: A Desirable QoS Tolerance Approach

Mohsen Heydarian*

Information Technology and Computer Engineering Collage, Azarbijan Shahid Madani University, Tabriz, Iran
Heydarian@Aazaruniv.ac.ir

Received: 13/Dec/2015

Revised: 13/Jul/2016

Accepted: 13/Aug/2016

Abstract

This research is an ongoing work for achieving consistency between topology control and QoS guarantee in MANET. Desirable topology and Quality of Service (QoS) control are two important challenges in wireless communication networks such as MANETs. In a Mobile Ad hoc Network, MANET, nodes move in the network area; therefore, the network topology is randomly and unpredictably changed. If the network topology is not controlled properly, the energy consumption is increased and also network topology probably becomes disconnected. To prevent from this situation, it is necessary to use desirable dynamic topology control algorithms such as k -edge connectivity methods. This paper tries to improve the three following parameters according to the k -edge connectivity concepts: (1) network performance, (2) reduce energy consumption, and (3) maintain the network connectivity. To achieve these goals, as a new method, we enhance k -edge connectivity methods using an improved definition of node density. The new method is called as: Node Density Based k -edge connected Topology Control (NDB^kTC) algorithm. For the first time the node density definition is dynamically used. The new method, computes the node density based on a new equation which consists of the following factors: the relative velocity of nodes, distance between nodes, the number of nodes and the transmission range of nodes. The results show that our new method improves the network performance compared with the existing methods. Also we will show that the new method can hold QoS in a desirable tolerance range.

Keywords: Local topology Control; k -edge Connectivity; Node Density; MANET; Optimized energy Consumption; QoS.

1. Introduction

Topology Control and QoS are two opposite functional services in the MANET because increasing in QoS will increase the number of links whereas topology control will decrease the number of links. This paper tries to present a new method which can hold these two functional services in the MANETs

1.1 Topology Control

A number of existing topology control algorithms including Local Minimum Spanning Tree (LMST) [5], Relative Neighborhood Graph (RNG) [6] and the Local Shortest Path Tree (LSPT) [7], guarantee 1-edge connectivity; meaning that, with removal of just one link, the network may lose its connectivity. Therefore, these algorithms are not practical for MANETs, due to changeable topology of them. For MANETs, many reliable topology control algorithms are introduced, including Fault-tolerant Local Spanning Sub graph (FLSS) [8] and Local Tree based Reliable Topology (LTRT) [9]. The mentioned algorithms can guarantee k -edge connectivity which means that with removal of $(k-1)$ arbitrary edges, the network doesn't lose its connectivity [2]. Using same value of k for whole network in order to make redundancy is disadvantage of the algorithms. Because due to different moving speeds of the network nodes, an unnecessary redundancy maybe made in the

network; in other words, a large value of k maybe not necessary in the parts of the network in which the average moving speed of nodes is low. The greater value of k for a node implies more directly connected neighbors, high energy consumption and high interference. Therefore, the value of k must be as small as possible.

In order to overcome the above issue, H. Nishiyama et al. [10] proposed a dynamic method, namely DLTRT which tries to compute the optimal value of k . The authors use the moving speed of nodes and probabilities to dynamically compute the appropriate value of k for each section of the network. The method supposes that all nodes in a part of the network move with the maximum speed existing in that part. This paper believes that always it is not necessary to consider the worst cases, and the real situation is the best case must be used efficiently. Our idea is to use the concept of node density and introduce a new equation to compute it. The new proposed equation improves the previous definitions of node density in the context of mobile networks. Also, it is the first time that a new topology control algorithm uses the node density dynamically.

1.2 QoS Control

Against the best effort, on the Internet and in other networks, QoS (Quality of Service) is the idea that transmission rates, error rates, bandwidth, delay and jitter can be measured, improved, and, to some extent, guaranteed in advance. It is important that a topology

* Corresponding Author

control algorithm mustn't reduce QoS conditions. This research shows that the new algorithm, NDB^kTC, does not reduce the QoS stability [25, 26].

The remainder of this paper is organized as follows: at first the paper presents some of the well known existing topology control algorithms in MANETs and points their deficiencies, in section 2. Section 3 describe DLTRT algorithm, briefly. Section 4 explains our new node density based method. The experimental results and comparisons are depicted in section 5. Section 6 concludes the paper and finally the future works will be presented in section 7.

2. Related Works

Many of topology control algorithms in wireless ad hoc networks are based on minimum spanning tree, MST, and RNG [6] approaches [10]. The main goal in MST is to find a tree in the given graph so that includes all nodes of the graph while the total weight of edges is minimal. Li et al. [5] introduced an MST based algorithm referred to as LMST, which is the local version of MST. In LMST, each node sends a "hello" message contains its ID and current location, using the maximum transmission range. Note that the nodes inform their own location using GPS technology. Each node constructs its local graph afterwards receiving the same information and computes the minimum spanning tree using Prim's algorithm [11]. The vertices of this tree which are directly (one hop away) connected to the node, are remained as neighbors of the node. Li et al. [12] introduced an algorithm called k-localized minimum spanning tree (LMST_k) and claimed that the nodes degree is up to six; this can decrease the contention and interference in the MAC level.

The main idea of RNG [6] is to delete the redundant edges. An edge (u, v) is redundant if there is a node w so that the weights of both (w, u) and (w, v) are less than (u, v) . Cartigny et al. [13] proved that in a given graph G , the resulting topology of LMST is a sub graph of RNG. Li et al. [12] proposed a lower weighted structure called Incident MST and RNG Graph (IMRG) which utilizes both MST and RNG.

As another corporation of MST and RNG, two algorithms are introduced, namely RNG based Broadcast Oriented Protocol (RBOP), and LMST based Broadcast Oriented Protocol (LBOP) [14]. In these algorithms, the broadcast is initiated at the source, and is propagated following the rules of neighbor elimination, on the topology derived from RNG and LMST approaches.

Another spanning tree algorithm for topology control is the shortest path tree (SPT) [21]. R. Meng [22] presented an algorithm based on SPT, called LSPT. According to LSPT, an edge (u, v) is redundant if there is a two hop path such $[u, w, v]$ between u and v such that weight $\{(u, w) + (w, v)\} < \text{weight}\{(u, v)\}$. Li and Halpern [7] extended the algorithm to k -hop path using Dijkstra algorithm [16].

Although the above algorithms are simple and practical in wireless ad hoc networks, their resulting topology is 1-edge connected. That means the network may lose its connectivity with removal of just one link. Therefore, the next studies are focused on fault-tolerant strategies. A fault is the removal of some links in the network which can destroy the network connectivity.

k -edge connectivity is utilized to add fault-tolerance to topology control algorithms. The purpose of k -edge connectivity is to guarantee the network connectivity while some links of the network may be destroyed. Bahramgiri et al. [17] introduced CBTC (α) algorithm to guarantee k -edge connectivity which is based on cone based topology control algorithm [18]. In this algorithm, the transmission power of node u so determined that there is at least one node in each cone of degree α in the coverage area of node u . They demonstrate that if $\alpha < 2\pi/3k$, the algorithm guarantees k -edge connectivity. Li and Hou [8] proposed FLSS algorithm that its resulting topology is k -edge connected. Its main idea is to add an edge with the smallest weight into the set of edges until k -edge connectivity is guaranteed. The disadvantage of FLSS is its high complexity (see Table 1).

LTRT [9] is local version of tree based reliable topology, TRT, [19]. The complexity of LTRT is $O(k(m + n \log n))$ and is better than FLSS, practically. LTRT uses the same value of k in whole network to guarantee k -edge connectivity that is a disadvantage. This results in unnecessary energy consumption in the parts of the network with low moving speed of nodes.

In order to eliminate this weakness of LTRT, recently Nishiyama et al. [10] proposed DLTRT algorithm that is a dynamic version of LTRT. In DLTRT, an appropriate value of k is determined for a certain section of the network. This computation is based on the nodes moving speeds and the probability that a node moves out of coverage area of another node. R. Azzeddine et al. [12] introduced a k -edge connected algorithm, called SFL. In SFL, each node uses two broadcasts to determine its transmission range instead of $(k+1)$ times broadcasting. Due to the few number of broadcasting, the complexity of SFL is lower than LTRT (see Table 1). It can be seen that the number of nodes (m) and links (n) can affect on the time complexity of algorithms.

Table 1. Time Complexity of Algorithms

Algorithm	Complexity
LMST	$O(m+n \log n)$
RNG	$O(n \log n)$
LSPT	$O(m+n \log n)$
FLSS	$O(m(m+n))$
LTRT	$O(k(m+n \log n))$
SFL	$O(2(m+n \log n))$

3. DLTRT Algorithm

H. Nishiyama et al. [10] proposed an algorithm namely dynamic LTRT, DLTRT. The basis of DLTRT is to consider the concept of k -edge connectivity and the fact of different moving speeds in the network. They

compute three probabilities: the probability that a node moves out of coverage area of another node, ρ , the probability that the network be disconnected, ρ_{global} , and the probability that a node loss all its direct links, ρ_{local} . A network is disconnected if and only if there is a node that loses all links to its neighbors. The following equation is used to compute the value of k :

$$\rho^k \leq \rho_{\text{local}} \quad (1)$$

The smallest value of k satisfying equation (1) is the appropriate k .

In the beginning, each node periodically broadcasts a "hello" message, contains its ID and current location and speed, to the maximum transmission range. Subsequently, each node u receives the same messages from its neighbors and constructs its local graph. It is assumed that all neighbors of u move with the maximum speed of the neighborhood. The maximum distance r that a node v can move through the slot time Δt , can be obtained using $r = 2V_{\text{max}} \Delta t$ (see Fig. 1). If R represents the transmission radius of node u , according to [5] the probability that node v moves out of coverage area of node u after Δt seconds, ρ , is computed using one of the following three equations:

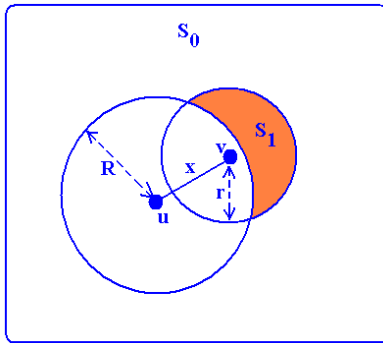


Fig. 1. Calculation of the probability that node v moves out of coverage area of node u [10].

if $0 < r < R$:

$$\rho = \int_{R-r}^R \frac{2xS_1}{S_0 r^2} dx, \quad (2)$$

if $R < r < 2R$:

$$\rho = \frac{\pi(r+R)}{S_0 r^2} (r-R)^3 - \int_{r-R}^R \frac{2xS_1}{S_0 r^2} dx, \quad (3)$$

if $r \geq R$:

$$\rho = \frac{\pi(r^2 - R^2)R^2}{S_0 r^2}. \quad (4)$$

4. Proposed New Method Based on Node Density

In the previous section, DLTRT algorithm was explained briefly. Although this algorithm resolves the

weakness of previous algorithms, it has another weakness. Authors of [10] use the worst situation in their computations; this means that, they suppose that all nodes in a certain part of the network move with the maximum speed existing in that part. In such a case, the probability that the neighbors move out of the coverage area of node u , see Fig. 1, is increased. Therefore, node u has to increase its transmission range more and as a result, the energy consumption in node u is increased. This research believes that the new algorithm can use more real situation of the environment and increase the network performance.

To understand the worst and real cases of the environment around the nodes and their impact on the performance, consider Fig. 2. As it can be seen, Fig. 2(a) shows a node u , in time t , which three neighbor nodes v_1 , v_2 and v_3 are in its transmission range, R_u . Suppose that the moving speed of neighbors is as $V_{v_1} > V_{v_2} > V_{v_3}$. Now the new location of the neighbor nodes after Δt seconds must be known and determined the transmission of node u such that it does not lose all links to the neighbors. If the worst situation is considered, i.e. it can be supposed that all neighbors move with the maximum speed, V_{v_1} , and in opposite direction of node u , the new situation of nodes is similar to Fig. 2(b). As mentioned in the previous section, the distance r that the nodes can get away from node u is equal to $r = 2 * V_1 * \Delta t$. Therefore, with a high probability, all neighbor nodes would move out of transmission range of node u after Δt seconds. Therefore, node u must increase its transmission range to R'_u , that here is equal to distance between u and v_3 , in order to maintain some of its connections with neighbors.

Now consider the real situations, i.e. each neighbor has its own moving speed and direction. As the moving speed for every neighbor is lower than the previous case, the probability that they move out of the transmission range of node u becomes less. So, the new locations of the neighbor nodes will be similar to Fig. 2(c). Similar to the previous case, node u increase its transmission range to R'_u . The difference is that in this case, real situation, the amount of increasing is low. Therefore, the energy consumption and interference are less than the worst case has.

In order to utilize the real situations and improve the network performance, the concept of node density is used in the network. Our new method use arguments of [10] to compute the value of k while the difference is the addition of computation and broadcasting of k to the proposed algorithm. In fact, using of node density is our main idea and innovation. The main idea behind utilizing the node density concept is that in the dense sections of the network, the topology is more stable. As a result, the nodes can decrease energy consumption by decreasing their transmission range. The new method introduces a new equation to compute node density which presented in the following subsection. Furthermore, SFL [20] is used as k -edge connected algorithm, because its complexity is lower than LTRT (see Table 1).

4.1 Node Density: Identification and Application

Before description of proposed algorithm, we investigate the concept of node density and the method of its calculation. The studies which use the concept of node density, such as [21] and [22], identify node density as $\frac{n}{N}$

or the network density as $\frac{n\pi R^2}{A}$. In these identifications, n depicts the number of neighbor nodes, N is the total number of nodes in the network, R is the average transmission range, and A stands for the network area. There are two major disadvantages in these identifications:

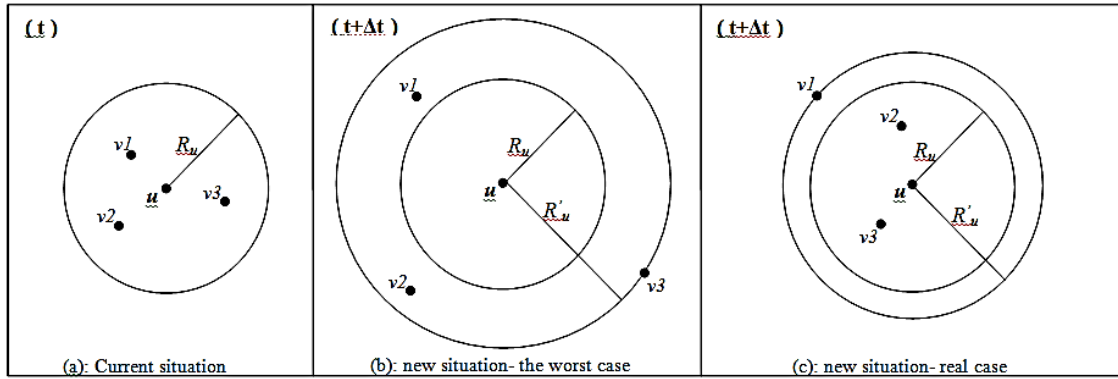


Fig. 2. The worst and real situations of environment around node u

The methods which use above identifications for computation of node density consider density as a constant value and use it globally. However, similar to moving speed, the node density is not equal in different sections of the network.

In these identifications, the distance and relative velocity of nodes are not considered, while these two factors have a significant effect on stability of the local topology.

In this paper, for the first time, the new method uses the concept of node density dynamically in local topology control problem. Also the above disadvantages will be addressed. First, the local density is dynamically computed for each node. This means that in each topology update, the node density is computed regarding the new situation and available information. In addition, we introduce an equation to compute node density which considers all items affecting the node density and stability of topology.

Our proposed equation uses the following factors in computation of node density: the number of neighbors of the node, the distance of neighbors from the node, the relative velocity of the neighbors in the rest frame of the node, and the transmission range of the node. There is a direct relationship between node density and the number of neighbors; in other words, as the number of neighbors becomes larger, the density value will become greater too.

The node density has an opposite relationship with distance, the relative velocity of neighbors and the transmission range of the node; that means, if the node covers its neighbors by a smaller transmission radius, and the distance and relative velocity of the neighbors are smaller, the node density will be increased.

Based on the above analysis, the new following equation is proposed to compute the node density:

$$D_u = \frac{n_u^2}{(N-1) \sum_{i=1}^{n_u} (x_{iu} + v_{iu})} \cdot \left(\frac{R_{max} - R_u}{R_{max}} \right)^2 \tag{5}$$

where n_u is the number of neighbors of node u , N is the total number of nodes in the network, x_{iu} is the distance between nodes i and u , and v_{iu} is the relative velocity of node i in the rest frame of node u . Also, R_u is the transmission range of node u and R_{max} is the maximum transmission range. As it can be seen in equation (5), the node density will be increased if the number of nodes is increased; the distance of neighbors from the node, relative velocity of the neighbors in the rest frame of the node, and transmission range of node are decreased. Therefore, the density of node u is maximized when all nodes are located in the transmission range of u , the nodes relative velocities in the rest frame of u are zero, they have the minimum distances from u , and transmission range of the u is small as far as possible.

Now suppose that there are two nodes A and B in the network, as shown in Fig. 3. As it can be seen, the number of neighbors is identical for A and B . Suppose that the transmission radius of A and B are equal and the average speed of the neighbors for node B is greater than for node A . If the node density be computed regarding the number of neighbors, the density of A and B are equal, while this is not logically true; because it is clear that the local topology of node B is more unstable than node A , due to the greater distances and speeds of the neighbors of node B .

Therefore, this is necessary that the new method changes and improves the identification of node density in the mobile networks. The procedure of the new proposed method and its algorithm are investigated in the following subsection.

4.2 The Algorithm and Steps of the New Method

Procedure 1 shows the algorithm of our proposed method which is explained it as the following.

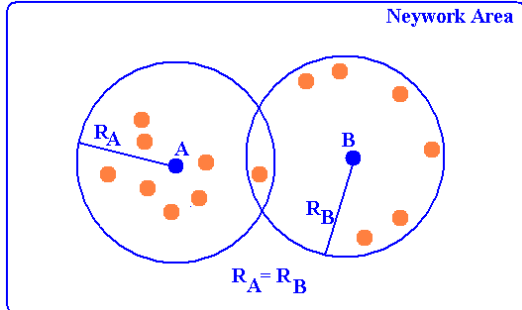


Fig. 3. Two Nodes with same transmission range and number of neighbors and different density.

Steps 1-6: At the beginning, each node broadcasts a “hello” message to the maximum transmission range. This message contains ID, current location, current speed and moving direction of the node. When a node u receives the same.

Procedure 1: Density based topology control in each node

- 1: **loop**
- 2: Calculate the current moving speed and direction.
- 3: Broadcast a “hello” message.
- 4: Build the local graph similar to Fig.4.
- 5: Calculate ρ based on the local graph by using one of Eqs. (2) - (4).
- 6: Determine the optimal value of k by using Eq. (1) withcalculated value of ρ .
- 7: Calculate node density using Eq. (5).
- 8: Broadcast the density and k in a message.
- 9: select the nearest neighbor in terms of density.
- 10: if the value of k related to the selected neighbor is smaller than the calculated k , replace it as new k .
- 11: Run a k -edge connected algorithm with the optimalvalue of k .
- 12: Keep the determined transmission range during theperiod of topology update.
- 13: **end loop**

messages from its neighbors, it constructs the local graph as Fig. 4. The vertices of this graph are node u and the neighbor nodes which their messages received by node u . There is an edge between every two vertices of this graph while its weight equals to the Euclidian distance between its vertices. The distance can be calculated using the current location of the nodes.

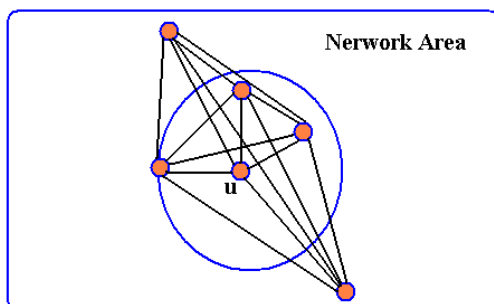


Fig. 4. Local graph of node u.

In computation of the minimum spanning trees for a given graph, if the weights of some edges are equal, the resulting tree may not be unique. This can affect performance of the algorithm. To prevent this, the following weight function is used in order to compute the weights of the network links. Here, w is the weight function and d is the Euclidian distance function. It is supposed that the nodes know their own position using GPS technology.

$$w(u_1, v_1) > w(u_2, v_2)$$

$$\iff d(u_1, v_1) > d(u_2, v_2)$$

$$\text{or } (d(u_1, v_1) = d(u_2, v_2))$$

$$\&\& \max\{id(u_1), id(v_1)\} > \max\{id(u_2), id(v_2)\}$$

$$\text{or } (d(u_1, v_1) = d(u_2, v_2))$$

$$\&\& \max\{id(u_1), id(v_1)\} = \max\{id(u_2), id(v_2)\}$$

$$\&\& \min\{id(u_1), id(v_1)\} > \min\{id(u_2), id(v_2)\}$$

The probability ρ and the value of k are calculated using the method explained in section 3.

Steps 7-10: Each node computes its density using Eq. (5) and broadcasts it along with the calculated k using maximum transmission range. When node u receives the same information from its neighbors, compares its own density with them and selects the nearest neighbor in terms of density. If the value of k related to the selected neighbor is smaller than the calculated k , node u replaces it as new k . Now, a k -edge connected algorithm is run using the new k in order to determine the new transmission radius. In the proposed procedure, every k -edge connected algorithm can be used. Therefore SFL algorithm [20] is used because its complexity is low (see Table 1). After running the algorithm, each node uses the determined transmission radius until the next topology update procedure is started. The topology update is completed when the Procedure 1 is performed once.

5. Compare: Simulation Calculations and Numerical Results

In this section, the new method, NDB^kTC , according to Procedure 1 introduced in the previous section will be simulated. Afterwards, the new method will be compared with DLTRT [10]. The following metrics are employed for the comparison: connectivity rate, average transmission range and the node degree.

Simulation computations are done according to the parameters of Table 2, in Matlab R2012a on a Core i5-4200M-2.5 GHz laptop with 4 GB of RAM and in windows 8 environment. According to [10] for the beginning, it is supposed that $\rho_{local}=0.0022$, as the expected connectivity rate is considered equal to 80%. 100 nodes are placed uniformly in an square area of $1000*1000$ (m^2). The maximum transmission range is set to 250 m. Each node can move with speed of 0-25 m/s in random direction. The expected connectivity rate is 80%. The topology update interval is set to 10s for all nodes.

Random way point [23] is used as mobility model. Simulation results are presented in the following.

Table 2. Simulation parameters and related values.

Simulation Parameter	Value
Simulation area	1000*1000 m
Maximum transmission range	250 m
Number of nodes	100
Topology update interval	10 s
Average moving speed	0~25 m/s
Expected connectivity rate	80%

5.1 Comparison of Connectivity Rate

Two nodes are connected if there is at least one path between them. A network is connected if and only if every two nodes of it are connected [2]. According to this identification, the following equation is used to compute the connectivity rate [10]; assuming that N is the total number of nodes in the network.

$$C = \frac{\sum_{u,v \in N} c_{uv}}{|N|(|N|-1)} \tag{6}$$

where,

$$c_{uv} = \begin{cases} 1, & \text{if } u \neq v \text{ and } (u, v) \text{ is connected,} \\ 0, & \text{otherwise.} \end{cases}$$

Suppose that graph $G(N, E)$ is as Fig. 5-left [10]. As it be seen, the graph G is connected; so according to Eq. (6) the connectivity rate of G in this case is 100%. If one of the links from G is deleted, as shown in Fig. 5-right, the connectivity rate of G will be 40%.

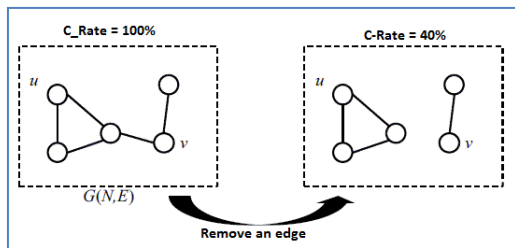


Fig. 5. Change The connectivity rate with removal of one edge.

Fig. 6 shows the connectivity rate in both our new method and DLTRT algorithm. It is clear that the connectivity of new method is lower than DLTRT algorithm in some cases. But, it still satisfies the expected connectivity rate, 80%. The reason of low connectivity rate of our new method is decreasing the value of k . With decreasing the value of k in the node, the number of times k-edge connected algorithm is run, step 11 in procedure 1, is decreased. So, the determined transmission range to the node is decreased. This results in decreasing the number of edges and reducing the connectivity rate.

5.2 Comparison of Average Transmission Range

As previously mentioned, decreasing the value of k reduces the determined transmission range. Fig. 7 shows

the average transmission range in both DLTRT and NDB^kTC. It can be seen that with increasing the average moving speed, the difference between DLTRT and NDB^kTC is increased. This is because, in the lower moving speeds, the difference of value of k between the neighbors is low and as a result the value of k has not an impressive change. But, when the average moving speed is increased, the difference of value of k is increased; so, the change of k in node density based method will be greater. According to [24] the amount of energy consumption in the network can be obtained from the following equation:

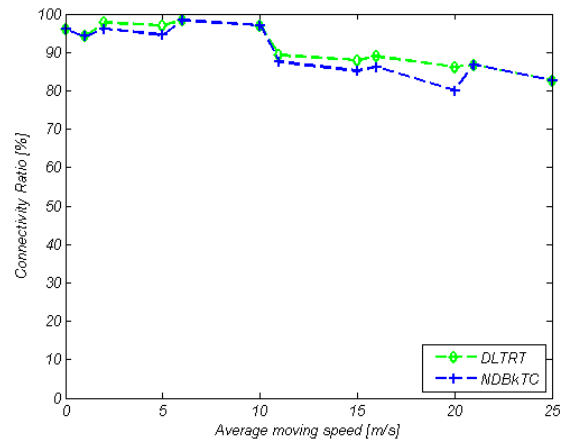


Fig. 6. Connectivity rate in new method compared with DLTRT.

$$E = \sum_{u \in N} (R_u)^\alpha \tag{7}$$

where R_u is the node u transmission radius and α is distance-power gradient or the path loss element. In an environment without obstacles, α is considered equal to two [24]. According to equation (7) when transmission radius is decreased, the amount of energy consumption is decreased at least with power of 2, and vice versa. Therefore, our new method decreases energy consumption by reducing the transmission range.

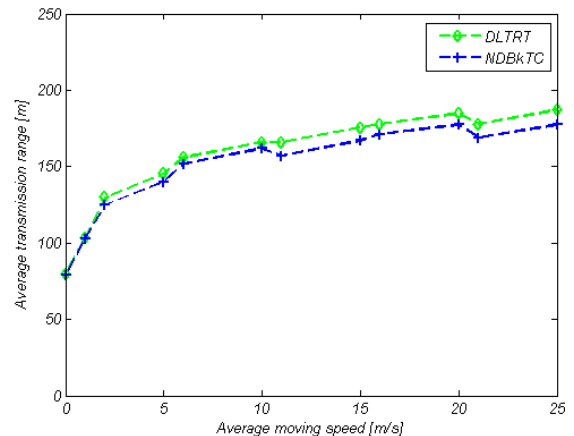


Fig. 7. Average transmission range in new method compared with DLTRT.

5.3 Comparison of Node Degree

Node degree is the number of direct neighbors of the node; in other words, the neighbors which are connected to the node with one direct link. The node degree can be studied as logical node degree and physical node degree [1]. The logical node degree equals to the number of nodes determined by the algorithm; in fact, the number of logical neighbors. The physical node degree is the number of neighbors in the transmission range of the node. Interferences are decreased with decreasing in the node degree. Fig. 8 shows the logical node degree in both DLTRT and NDB^kTC.

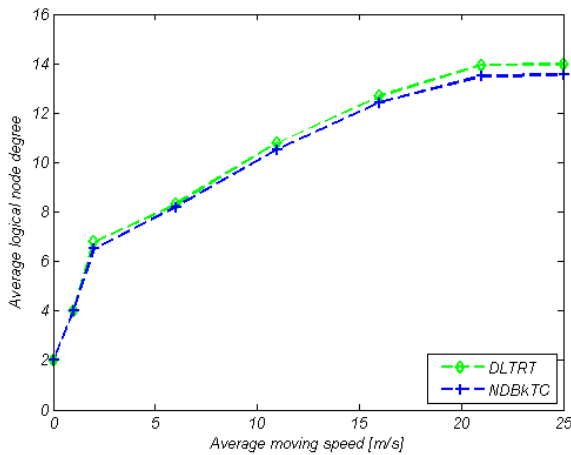


Fig. 8. Logical node degree in new method compared with DLTRT.

As it can be seen from Fig. 8, the logical node degree in our new method is lower than in DLTRT. This difference is increased when the average moving speed is increased. This is because in this case due to increasing the difference of moving speed related to the nodes which are located in a same section, the difference of computed k for the nodes is increased. Therefore, it is more probable to change (decrease) k values computed to the nodes, based on the analogy of their densities. As a result, the times to repeat the k -edge connected algorithm and the number of logical neighbors is decreased.

The simulation results show that the new method decreases physical node degree up to 1.3 nodes compared with DLTRT. In fact, in NDB^kTC due to decreasing the transmission range, physical node degree is decreased.

5.4 QoS Analysis

Figure 8 shows some mechanism for controlling QoS which we consider 3 parameters delay, bandwidth and lost rate in this paper.

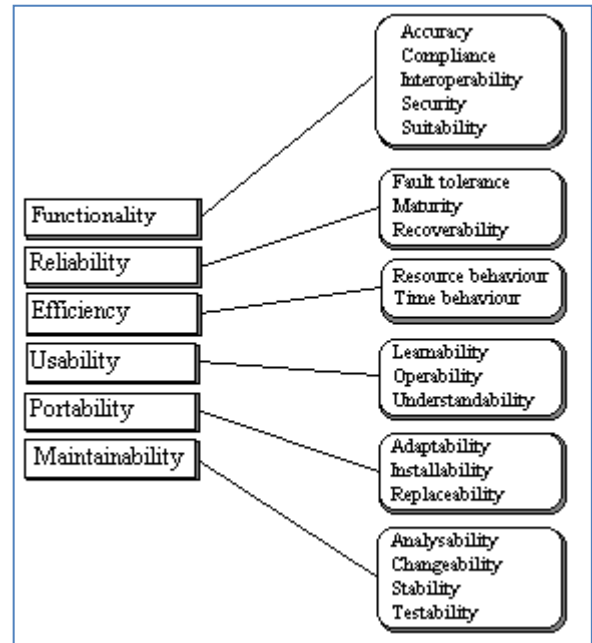


Fig. 9. Some mechanism for controlling and improving QoS in networks [25, 29].

Also in Figure 9 we can see some new and modern mechanism for controlling QoS in networks such as Differential Services Networks [25, 27].

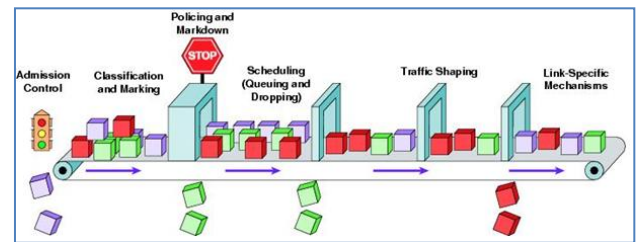


Fig. 10. some new and modern mechanism for controlling QoS in networks such as Differential Services Networks [25, 28].

For analyzing the situation of QoS in NDB^kTC and DLTRT, we consider the following heuristic formulation:

$$F_{QoS} = \frac{\text{Used Band}}{\text{Available Band}} * \left(1 - \frac{\text{Lost Rate}}{\text{Total packets}}\right) * \left(1 - \frac{\text{delay}}{\text{total time}}\right) \tag{8}$$

If F_{QoS} vanishes, this means that lost rate or delay has been increased and so QoS will be failed, whereas, if F_{QoS} limits to 1, this means that QoS has a desirable situation. $F_{QoS} = 0$ shows that all available bandwidth has been used and delay and lost rate are equal to 0. Figure 10 shows values of F_{QoS} before and after applying NDB^kTC to the network. This figure shows that however NDB^kTC reduces the number of links and bandwidth, it maintains F_{QoS} in a desirable range and does not fall QoS conditions. Before applying NDB^kTC, DLTRT transmits and delivers packets.

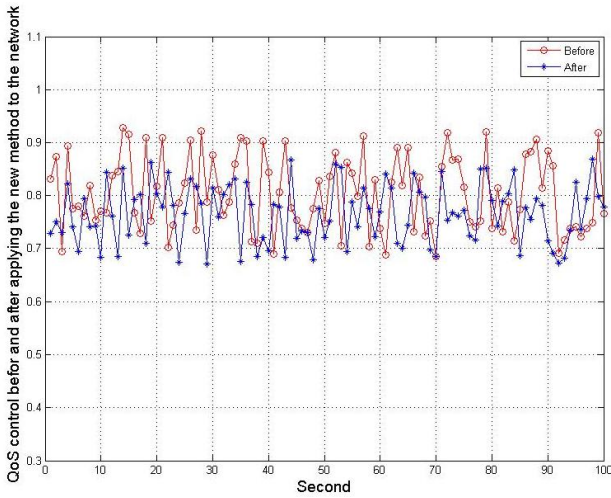


Fig. 11. NDBkTC decreases the number of links but can holds QoS conditions (F_{QoS}) in a desirable ranges.

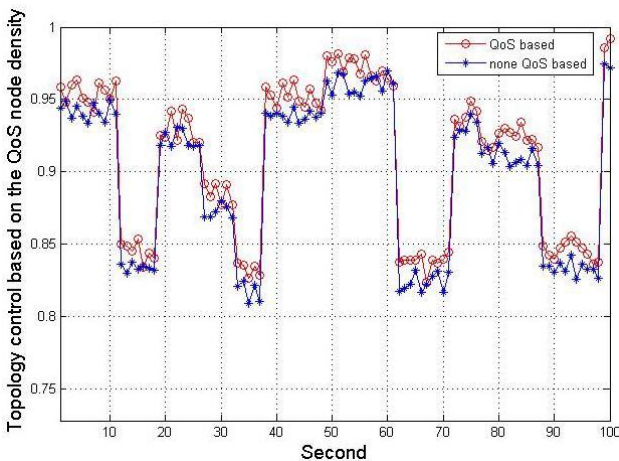
6. New QoS k-edge Connected Algorithm

In this paper in the subsection 4.1 a new node density definition has been presented and used. Now this definition is improved using F_{QoS} in order to increase the QoS guarantee of the new method NDB^kTC which has been presented in the section 4. We call this new node density formulation as QoS Node Density factor which is as the following:

$$D_u^{QoS} = F_{QoS} \times D_u \tag{9}$$

As we see formulation (9) shows that F_{QoS} can affect on the D_u . In fact if two nodes have the same D_u , D_u^{QoS} selects that node that has the better QoS conditions. Based on the formulation (9) we compare results of NDB^kTC in two states: using D_u^{QoS} and D_u . For achieving these results we re-run the simulation shown in Table 2. Figure 11 shows that D_u^{QoS} as compared to D_u increases stability and consistency of QoS in

NDB^kTC. D_u^{QoS} as compared to D_u increases the average of QoS parameter as %5.17.



7. Conclusion

This paper has investigated the effect of the nodes movement over the mobile ad hoc networks connectivity, network QoS, and analyzed the algorithms which guarantee k-edge connectivity. We outlined disadvantages of the algorithms and proposed a new node density based method in order to overcome the disadvantages. In this method, for the first time, the concept of node density was used in local topology control, dynamically. In order to improve existing identifications of node density which was used as a constant in the previous studies, a new equation was proposed to compute it. In this new formula, all factors affecting the node density and stability of topology have been used. In fact, our purpose is to efficiently utilize available information in the computations, instead of using the worst cases.

The new method has been compared with DLTRT. The criteria used for comparison are as follow: connectivity rate of the network, average transmission range and the logical and physical node degree. The results show that our new method with maintaining the expected connectivity rate, improves the performance factors.

Simulation results show that however NDB^kTC reduces the number of links and bandwidth, it maintains F_{QoS} in a desirable range and does not falls QoS conditions. Before applying NDB^kTC, DLTRT transmits and delivers packets.

8. Future Works

This paper used the concept of node density in order to improve the performance factors in MANETs. Although the new proposed method is better than the existing ones, but it can be improved by using other techniques. We will try to improve our new method in the future studies. Some of these techniques include the following:

1. Combination of node density with clustering: The new proposed method uses the density of the neighbor nodes to determine the optimal value of k related to k-edge connectivity. Density based clustering is one of the clustering techniques. In density based clustering, a node can construct a new cluster or can be a member of another existing cluster. The condition required to construct a new cluster by a node is that the node has a certain number of neighbors in its coverage area. This procedure runs in each node and obtained clusters are merged as far as possible. The local versions of density based clustering are introduced which we can use them to improve our new method. The new idea can be using the information of clusters in order to determine the optimal value of k .
2. Neural networks: As shown in section 5, the new proposed method in this paper improves the factors of the network performance compared with the existing methods. Therefore its determined

transmission range for every node of the network is better than the previous methods. The neural network technique can be used as the following.

In order to use neural networks a data set must be provided for training the network. To construct this data set, the following steps must be done: In the beginning, we simulate a MANET according to this paper and run it using the proposed new algorithm. When a node computes its density, a feature vector is constructed for it. This vector consists of the following elements: the computed density for the node, current transmission range, the average distance of the neighbors from the node, and the number of the neighbor. Afterwards, the remaining of the algorithm is run and the transmission range is determined for the node. This determined transmission range must be saved in the target vector, as the related element of the node. This procedure is repeated for all nodes until termination of the simulation.

Suppose that the number of the network nodes is N , the simulation time is T , and topology update interval is Δt . The number of the existing samples in data set in the end of each simulation will be equal to $(T / \Delta t) * N$. So, if we repeat the simulation for M times, the number of samples will be $M * (T / \Delta t) * N$ samples. We can train the

neural network using 80% of the data set samples as training set. The remaining of samples is used to test the correction of the trained network.

Now new simulation must be done using the trained neural network. If the results have a performance as well as the proposed method in this paper, the algorithm can reduce the computation overhead impressively, in topology control.

3. Inserting new factors in the node density equation: the new proposed equation in this paper for node density, uses four various factors: the moving speed of nodes, the distance between nodes, the number of nodes, and transmission range of nodes. We proposed this new equation to use in MANETs and for topology control problem. In the other types of networks such as Mobile IP and VANETs, or in the other problems, e.g. routing, it is maybe required to use other factors in the equation. Therefore, we can allocate a part of the future studies to use the concept of node density in other types of network and other problems. It is clear that a new equation must be presented to compute the node density.

References

- [1] Paolo Santi, *Topology control in ad hoc and sensor networks*, John Wiley & Sons, 2005, pp 16-100.
- [2] B. Bollobas, "Modern graph theory", *Graduate Texts in Mathematics*, vol. 189, no. 10, pp. 56-99, Jan, 1998.
- [3] J. Pan, L. Cai, Y. Hou, Y. Shi, and X. Shen, "Optimal base-station locations in two-tiered wireless sensor networks", *IEEE Trans. Mobile Comput.*, vol. 4, no. 5, pp. 458-473, Sep. 2005.
- [4] J. Zhang, J. Chen, and Y. Sun, "Transmission power adjustment of wireless sensor networks using fuzzy control algorithm", *Wireless Communication and Mobile Computing (Wiley)*, vol. 9, no. 6, pp. 805-818, June 2009.
- [5] N. Li, J.C. Hou, and L. Sha, "Design and analysis of an MST-based topology control algorithm", *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 259-270, May 2005.
- [6] G. Toussaint, "The relative neighborhood graph of finite planar set", *Pattern Recognition*, vol. 12, no. 4, pp. 261-268, 1980.
- [7] L. Li and J. Y. Halpern, "Minimum energy mobile wireless networks revisited", in *Proc. 2001 IEEE ICC*, 2001, pp. 278-283.
- [8] N. Li and J. C. Hou, "Localized fault-tolerant topology control in wire-less ad hoc networks", *IEEE Trans. Parallel and Distributed Systems*, vol. 17, no. 4, pp. 307-320, Apr. 2006.
- [9] K. Miyao, H. Nakayama, N. Ansari, and N. Kato, "LRT: an efficient and reliable topology control algorithm for ad-hoc networks", *IEEE Trans. Wireless Commun.*, vol. 8, no. 12, pp. 6050-6058, Dec. 2009.
- [10] H. Nishiyama, T. Ngo, N. Ansari, N. Kato, "On Minimizing the Impact of Mobility on Topology Control in Mobile Ad Hoc Networks", *IEEE Transaction On Wireless Communications*, vol. 11, no. 3, pp. 1158-1166, March. 2012.
- [11] Prim, R. C., "Shortest connection networks and some generalizations", *Bell System Technical Journal*, vol. 36, no. 6, pp. 1389-1401, November 1957.
- [12] X. Li, Y. Wang, and W. Song, "Applications of k-local MST for topology control and broadcasting in wireless ad hoc networks", *IEEE Trans. Parallel and Distrib. Syst.*, vol. 15, no. 12, pp 1057-1069, Dec. 2004.
- [13] J. Cartigny, D. Simplot, and I. Stojmenovic, "Localized minimum-energy broadcasting in ad-hoc networks", in *Proc. 2003 IEEE INFO-COM*, 2003, pp. 2210-2217.
- [14] J. Cartigny, F. Ingelrest, D. Simplot-Ryl, and I. Stojmenovic, "Localized LMST and RNG based minimum-energy broadcast protocols in ad hoc networks", *Ad Hoc Networks*, vol. 3, no. 1, pp. 1-16, 2005.
- [15] V. Rodoplu and T. H. Meng, "Minimum energy mobile wireless networks", *IEEE J. Sel. Areas Commun.*, vol. 17, no. 8, pp. 1333-1344, Aug. 1999.
- [16] E. W. Dijkstra, "A note on two problems in connexion with graphs", *Numerische Mathematik*, vol. 1, no. 1, pp. 269-271, Dec. 1959.
- [17] M. Bahramgiri, M. T. Hajiaghayi, and V. S. Mirrokni, "Fault-tolerant and 3-dimensional distributed topology control algorithms in wireless multi-hop networks", *Wireless Network*, vol. 12, no. 2, pp. 179-188, Mar. 2006.
- [18] L. Li, J. Y. Halpern, P. Bahl, Y. Wang, and R. Wattenhofer, "Analysis of a cone-based distributed topology control algorithm for wireless multi-hop networks", in *Proc. 2001 ACM PODC*, pp. 264-273, 2001.
- [19] N. Ansari, G. Cheng, and R. Krishna n, "Efficient and reliable link state information dissemination", *IEEE Commun. Lett.*, vol. 8, no. 5, pp. 317-319, May 2004.
- [20] R. Azzeddine and W. Dong, "A simple fault-tolerant local topology control algorithm for sensor networks", presented at *IEEE conference on Industrial electronics and applications (ICIEA)*, Japan, 2012.

- [21] Ch. Bettstetter, "On the minimum node degree and connectivity of a wireless multi-hop network", presented at the MOBIHIC'02 Int. Conf, EPF Lausanna, Switzerland, 2002.
- [22] N. Bulusu, D. Estrin, L. Girod, J. Heidemann, "Scalable coordination of wireless sensor networks: self-configuring localization systems", in Proc. Of the international symposium communication theory and applications (ISCTA'01), UK, 2001, pp. 1-6.
- [23] M. Shiwen, Fundamentals of communication networks, Cognitive Radio Communications and Networks, New Yourk, Elsevier, 2010, pp. 201-234.
- [24] S. Wolf, Optimization problems in self-organization networks, Berlin, Logos Verlag Berlin GmbH, 2010, pp. 160-161.
- [25] A. Isazadeh, M. Heydarian, "Optimal multicast multichannel routing in computer networks", Computer Communications, vol. 31, pp. 4149-4161, June, 2008.
- [26] M. Heydarian, "A new high performance approach: merging optimal multicast sessions for supporting multisource routing", Supercomputing, vol. 63, pp. 871-896, May, 2013.
- [27] Mrunal Gavhale, Pranay D. Saraf, "Survey on Algorithms for Efficient Cluster Formation and Cluster Head Selection in MANET", Procedia Computer Science, Vol. 78, pp. 477-482, 2016.
- [28] Hasan Abdulwahid, BinDai, BenxiongHuang, Zijing Chen, "Scheduled-links multicast routing protocol in MANETs", Journal of Network and Computer Applications, vol. 63, pp. 56-67, May, 2016.
- [29] S. Gundry, J. Zou, M. Umit Uyar, C. S. Sahin, J. Kusyk, "Differential evolution-based autonomous and disruption tolerant vehicular self-organization in MANETs", Ad Hoc Networks, Vol. 25, Part B, pp. 454-471, Jan. 2015.

Mohsen Heydarian received the B.Sc. degree in Applied Mathematics from University of Tabriz in 1999, the M.Sc. degree in Applied Mathematics (Numerical Analysis) from Tabriz University in 2002, and the Ph.D. degree in Applied Mathematics and Computer Science from Tabriz University in 2010. He is currently an Assistance Professor in the Department of Information Technology and Computer Engineering at Azarbaijan Shahid Madani University and has been a founding member of this department since 2010. His current research interests include Communication Technologies, Mathematical Optimization, and Information Technology.

High-Resolution Fringe Pattern Phase Extraction, Placing a Focus on Real-Time 3D Imaging

Amir Hooshang Mazinan*

Department of Control Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran
ahmazinan@gmail.com

Ali Esmaeili

Department of Electronics Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran
Esmaeili63@gmail.com

Received: 05/Aug/2015

Revised: 05/Jun/2016

Accepted: 05/Aug/2016

Abstract

The idea behind the research is to deal with real-time 3D imaging that may extensively be referred to the fields of medical science and engineering in general. It is to note that most effective non-contact measurement techniques can include the structured light patterns, provided in the surface of object for the purpose of acquiring its 3D depth. The traditional structured light pattern can now be known as the fringe pattern. In this study, the conventional approaches, realized in the fringe pattern analysis with applications to 3D imaging such as wavelet and Fourier transform are efficiently investigated. In addition to the frequency estimation algorithm in most of these approaches, additional unwrapping algorithm is needed to extract the phase, coherently. Considering problems regarding phase unwrapping of fringe algorithm surveyed in the literatures, a state-of-the-art approach is here organized to be proposed. In the aforementioned proposed approach, the key characteristics of the same conventional algorithms such as the frequency estimation and the Itoh algorithm are synchronously realized. At the end, the results carried out through the simulation programs have revealed that the proposed approach is able to extract image phase of simulated fringe patterns and correspondingly realistic patterns with high quality. Another advantage of this investigated approach is considered as its real-time application, while a significant part of operations might be executed in parallel.

Keywords: High-Resolution Fringe Patterns; 3D Imaging; Itoh Algorithm; Wavelet Transformation.

1. Introduction

Due to the fact that optical measurement approaches have the merit of being the precise outcomes in the areas of medical science and engineering, it is coherent to realize it first with respect to other related ones via more reliable image processing tools. There are high-resolution optical noncontacts measuring approaches that can be effective techniques to calculate the exact depth of objects, as long as their surface is not taken into real consideration. In one such case, the most effective noncontact measuring techniques can include the structured light patterns that are provided for an object to derive its depth. In fact, the most common type of structured light pattern can now be recognized as the fringe pattern. It is to note that non-contact measuring approaches may be exploited, in order to obtain the depth distribution of such an object. It is carried out via fringe pattern through a number of sequential steps. Now, the fringes acquired from interferometer or reticulated projector is provided for the surface of the object to be considered, while the image acquired from an axis except projection axis is to be used. Hereinafter, the projected fringe pattern transforms as its phase is modulated through the distribution of object depth. Therefore, image of transformed fringe pattern can be demodulated or analyzed via a potential fringe analysis approach to be able to extract the phase distribution of the

pattern. In conclusion, the depth distribution of the object is to obtain though the extracted demodulated phase distribution, coherently.

The fringe pattern analysis approaches including Fourier fringe analysis can generally be divided into two key steps that can be listed as the extraction of wrapped phase and phase unwrapping, as well. It should be noted that the first step of the phase of the fringe pattern to be extracted is to use the Fourier or wavelet transform. There are some necessary filters to be applied to the frequency domain, while phase that is generated in the aforementioned step is wrapped. These are to be eliminated by using the phase unwrapping approach, which is the second phase of the fringe pattern analysis approach. It can be shown that the processing time is to be considerable because of the existence of noise, in wrapped phase mapping. With this goal, optimizing the approaches is prominent so that noise is removed and, in turn, the processing time is decreased.

It is reasonable to note that despite unwrapped phase that is directly extracted by these approaches, most of them suffer from drawbacks, which can now be addressed. Furthermore, the Fourier transform approach is efficient in surfaces that have the uniform phase variations. Due to the fact that the measured surface has severe and abrupt changes, the frequency of fringe pattern should be more than these variations. Therefore, considering the Fourier

* Corresponding Author

transform approach from such a pattern and extracting the desired frequency to be named as the phase though the inverse of the Fourier transform approach can be problematic. Now, as long as the feature of imaging speed is important to consider, the approach is applicable, while it merely utilizes one fringe pattern and Fourier analysis [1]-[3]. It is obvious that the applications of the same approaches are measuring surface vibration through high speed imaging. By executing this one, the phase can be extracted in wrapped form and subsequently the phase unwrapping approach can be employed to interpret the validated phase information.

It is highly likely that another fringe analysis approach in this area can be taken into consideration as the wavelet transformation, which is the efficient choice for the non-uniform surfaces. In sequel, the aforementioned wavelet transformation may be carried out, whilst one the fringe pattern and the noise effect on the extracted phase that is less than Fourier approach can be resulted [4]-[5].

The lengthy processing time is known as the disadvantage of this approach that is a consequence of wavelet transformation. Besides, derived outputs in edges of image have errors which might be problematic in some applications. Among approaches concerning the wavelet transform is to be utilized, the phase estimation and the frequency estimation ones might be addressed. The phase extracted by the phase estimation approach is wrapped and needs the phase unwrapping algorithm. The phase estimation approach is more precise than the Fourier transform one in the fringe pattern analysis, while it is consuming more execution time. Yet, both cases need the phase unwrapping algorithm as the extracted phase can be wrapped.

Cusack et al and Karout et al have all proposed their potential algorithms [6]-[7]. It should be noted that the normal image may include thousands of phase discontinuities. Some of them are intrinsic, while some others are consequence of noise or phase extraction algorithm. Distinguishing intrinsic discontinuities and those resulted from noise is challenging which makes phase unwrapping more complicated. Moreover, accumulator nature of this procedure makes it even more difficult. In cases such as Frequency Estimation method there is no need for phase unwrapping as phase is directly extracted unwrapped [8]-[10].

Right now researchers in medical and engineering fields complain about inappropriate efficiency of current phase unwrapping techniques [11]. The researchers have even tried to collect a group of existing phase unwrapping algorithms to separately exploit their advantages for a set of applications [12]-[13]. So far, nobody knows why phase unwrapping algorithms (even powerful ones) demonstrate low performance in some cases [14]. The most evident and essential drawback of recent phase unwrapping algorithms is lack of generality. In other words, each phase unwrapping algorithm cannot be employed for unwrapping of all types of wrapped phases. In this study, firstly, a novel and powerful algorithm is proposed to analyze various types of fringe algorithms

which is applicable to inclined surface objects. Afterwards, performance of proposed algorithm is compared to conventional algorithms.

2. The Proposed Approach

The behavior of proposed approach is determined using a method different from other fringe pattern analysis methods and does not need complicated calculations. In this algorithm a composition of characteristics of conventional methods is exploited. The following subsections elaborate on proposed method.

2.1 The Main Idea

As mentioned before, most fringe pattern analysis methods extract the phase in wrapped form and require a phase unwrapping algorithm; however, they may face various problems in unwrapping step which are mentioned in [14]. The proposed method to address this problem is supervising phase unwrapping operation through comparing main wrapped phase pattern with a proper pattern. The unwrapping procedure continues similar to other algorithms; whereas, in case of confronting any discontinuities, corresponding pixels are compared to similar ones in reference pattern. If the same discontinuity exists in reference pattern, it will be considered in unwrapping operation; otherwise it will be considered as fake discontinuity and would be ignored and the algorithm continues. Fake discontinuities are neglected as they might be noise and may threaten unwrapping operation of whole image.

Using valid reference pattern means using another method in parallel which necessarily meet following requirements:

- Small processing time.
- Avoiding destruction of output image as a result of noise in input image (in some methods the output image might be completely destroyed by an unwanted noise).
- Unimportance of output precision.

In almost all current methods if the noise variance exceeds a certain value, the output image will be completely lost. Even in conditions where destructive effect of noise is reduced, the processing time increases. On the other hand the output of this method is merely used for comparison in case of suspicious pixels; so, there is no need for a perfect output with high precision.

The traditional idea that is realized to deal with frequency estimation method or phase gradient method has a number of problems that are all solved through the proposed approach, which is entirely illustrated in Fig. 1, listed as: this approach is only potentially suitable for generating reference output [15]. Its output does not have sufficient precision in most cases and additionally, it does not need phase unwrapping algorithm after phase extraction. As a result the noise does not significantly affect the output, though it has improper output. Ignoring its large processing time, this method is a good candidate

for reference pattern. Thus, the estimation frequency might be reduced which considerably reduces processing time. Although this attempt destroys output image, experiments have demonstrated that the information of this image is sufficient for a reference pattern. It is still necessary to have tradeoff between number of frequencies and generated output.

2.2 The Basic Flowchart of the Proposed Algorithm

The basic flowchart of the proposed algorithm is given in the form of Fig. 1. In this procedure, fringe pattern projected on the object is simultaneously processed by two methods. In the first one, the image of projected fringe pattern is unwrapped using a conventional method with low computational load. In fact, this takes place based on the core of phase gradient method and Itoh algorithm. Furthermore, in the second method, the phase of the same image is derived via frequency estimation method and then it is wrapped manually and intentionally (it is done using atan function in MATLAB). The image obtained by this method is called reference pattern whose information is utilized in the first method. In this flowchart DIFF1 denotes the difference between two adjacent pixels in projected fringe pattern after filtering operation; while, DIFF2 is the difference between the same pixels in reference pattern. As mentioned, when first method is being executed and it faces discontinuity, the similar pixels in reference pattern are checked. The observed discontinuity is valid if there is a discontinuity in its corresponding pixels in reference pattern.

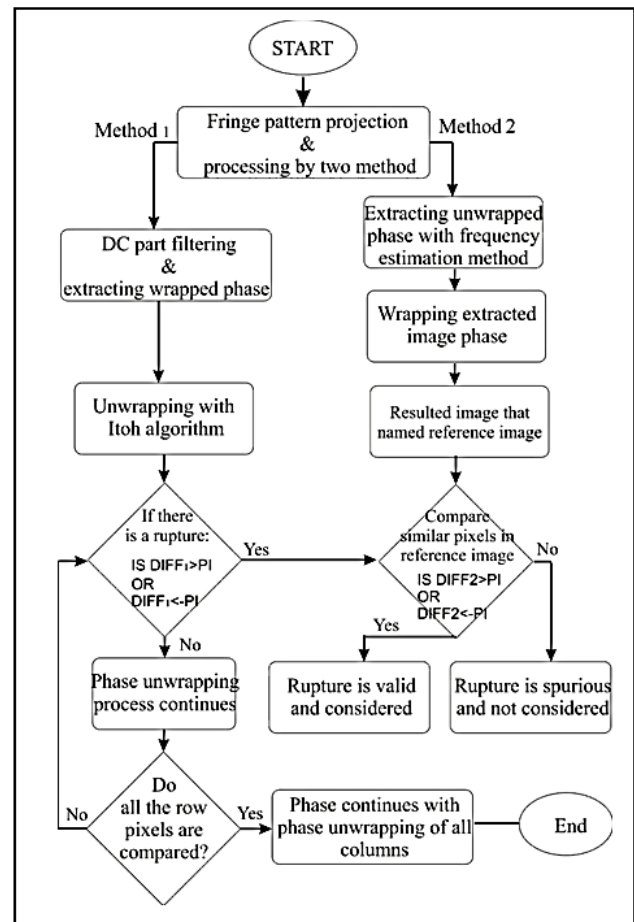


Fig. 1. The schematic diagram of the proposed algorithm

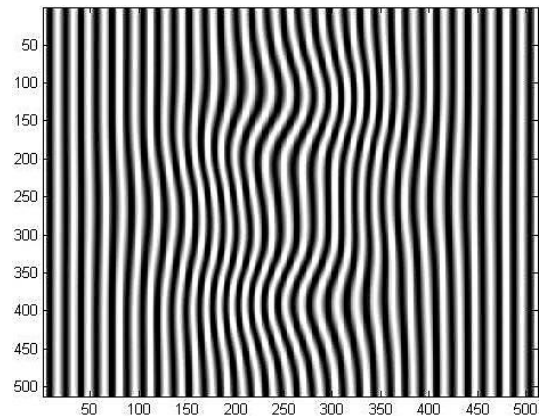
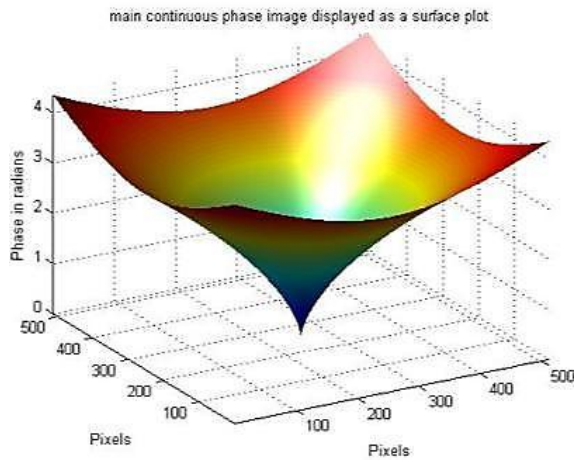


Fig. 2. The fringe pattern of the main image without the noise.

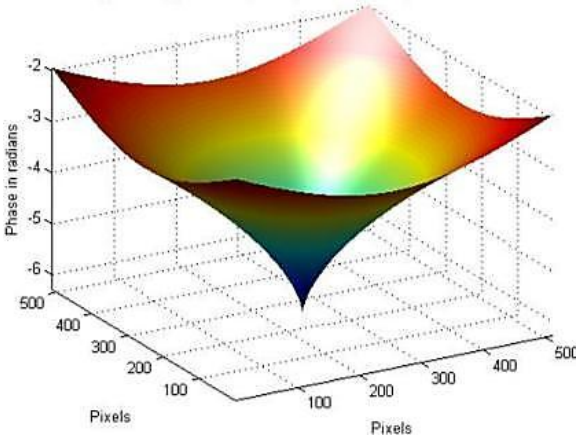
3. The Simulation Results

The proposed approach is now carried out through a simulation program, as long as the first condition is given by including the noise variance to be equaled to 0.5 in the fringe pattern. For this purpose, an image with the undergoing phase relation is provided and its fringe pattern is simulated via the Gaussian distribution. The simulated fringe pattern is now illustrated in Fig. 2, where after processing of this one via the noisy image by the proposed approach, an image that is similar to the input image and without any deficiencies is also provided. Now, in order to illustrate the applicability of the proposed approach, the same image of Fig. 2 with similar noise variance is processed under the some conventional algorithms including the phase estimation, the frequency estimation and finally the first and the second Itoh methods. The proposed approach is in fact carried out, where the phase

Variation of the image is considerable. The main image, which has a relatively large slope and also the image provided by the proposed approach, is illustrated in Fig. 3.



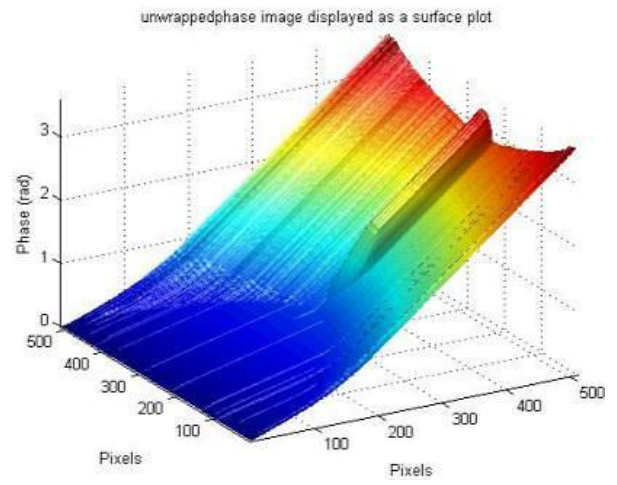
(a)



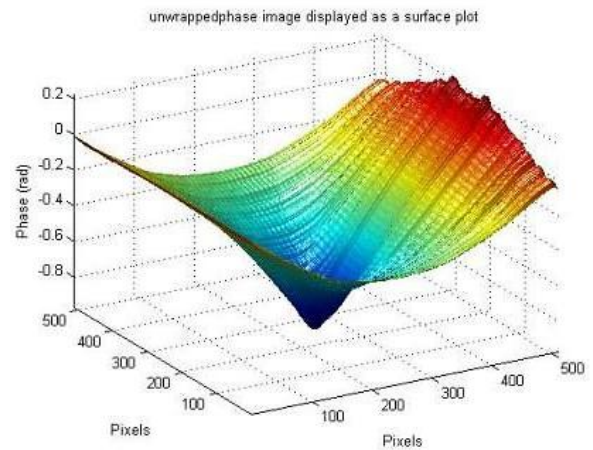
(b)

Fig. 3. a) The main image, which has the relatively large slope, b) the image provided by the proposed algorithm.

Figures 4 and 5 depict the results, as long as the frequency estimation algorithm is executed in two conditions in Fig. 4 including (a) with the small number of frequency samples and (b) with the large number of frequency samples. As it can be seen in Fig. 4(a), the output image is completely distorted and cannot be interpreted while in Fig. 4(b), it is somehow interpretable. Experiments have revealed that through the proposed method, both images would be acceptable and similar and desired results might be achieved; nevertheless, condition (a) is selected in the proposed approach owing to its small execution time. Figure 5 illustrates the results of image given in Fig. 3 with the phase estimation approach including Fig. 5(a) indicates the output image and Fig. 5(b) indicates the difference between main image and output image.

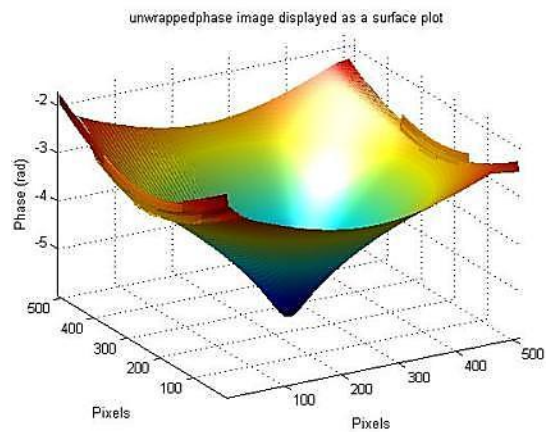


(a)



(b)

Fig. 4. The results of image illustrated in Fig. 3 with the frequency estimation approach a) with the small number of frequency samples, b) with the large number of frequency samples.



(a)

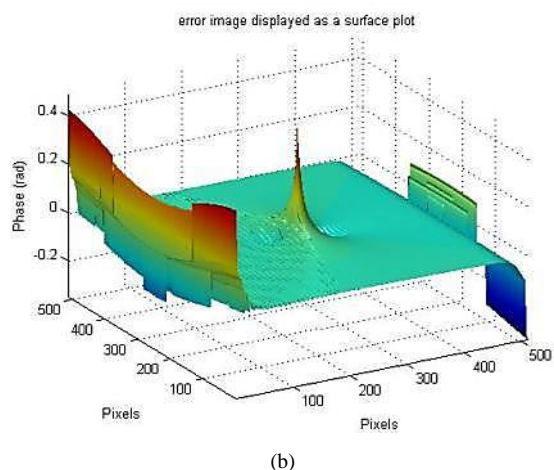


Fig. 5. The results of image illustrated in Fig. 3 with the phase estimation approach a) output image, b) difference between main image and output image.

References

- [1] A. Abdulbasit, Zaid Ahmed. "Fringe Pattern Analysis". Liverpool John Moores University, 2008.
- [2] X. Su, W. Chen, Z. Qc and Y. Chao "Dynamic 3-D Shape Measurement Method Based on Ftp". Optics and Lasers in Engineering, vol. 36, pp. 46-64, 2001.
- [3] M. Takeda, H. Ina and A. Kobayashi, "Fourier-Transform Method of Fringe Pattern Analysis for Computer-Based Topography and Interferometry". Journal of the Optical Society of America, pp. 156-160, 1982.
- [4] N. Karpinsky, S. Zhang, "High-resolution, real-time 3D imaging with fringe analysis". Journal of Real-Time Image Processing, vol.7, issue 1, pp. 55-66, March 2012.
- [5] G. Munther, "Mother Wavelets". Liverpool John Moores University.2009.
- [6] R. Cusack, J.M. Huntley and H. T. Goldrein. "Improved noise immune phase-unwrapping algorithm", Applied Optics, pp. 781-789, 1995.
- [7] S. A. Karout, M. A. Gdeisat, D. R. Burton and M. J. Lalor, "Two-dimensional phase unwrapping using a hybrid genetic algorithm". Applid Optics, vol. 46, pp. 730-743, 2007.
- [8] P. Sandoz, "Wavelet Transform as a Processing Tool in White-Light Interferometry". Optics Letters, vol. 22, pp. 1065-7, 1997.
- [9] M. Cherbuliez, P. Jacquot and X. Colonna de Lega, "Wavelet Processing of Interferometric Signals and Fringe Patterns". SPIE-Int. Soc. Opt. Eng., 1999.
- [10] A. Dursun, S. Ozder and F. N. Ecevit, "Continuous Wavelet Transform Analysis of Projected Fringe Patterns". Measurement Science and Technology, vol. 15, pp. 1768-1772, 2004.
- [11] H. Pottle, "The Phase Unwrapping Problem". Liverpool John Moores University. 2012.
- [12] F. Qiangian, P. M. Meaney and K. D. Paulsen. "The multidimensional phase unwrapping Integral and applications to microwave tomographical image reconstruction". IEEE Transactions on Image Processing, no. 11, vol. 15, pp. 3311-24, 2006.
- [13] S. A. Karout, M. A. Gdeisat, D. R. Burton and M. J. Lalor. "Residue Vector, An Approach To Branch-Cut Placement In Phase Unwrapping: Theoretical Study". Applied Optics, vol. 46, no. 21, pp. 4712-4727, 2007.
- [14] R. Kulkarni, P. Rastogi, "Patch-wise denoising of phase fringe patterns based on matrix enhancement", Optics and Lasers in Engineering, in press, 2016.
- [15] A. Abid, M. Gdeisat, D. Burton and M.A. Lalor, "Comparison between Wavelet Fringe Analysis Algorithms". Journal of Photon, Manchester UK,2006.
- [16] B. Li, Y. Wang, J. Dai, W. Lohry, S. Zhang, "Some recent advances on superfast 3D shape measurement with digital binary defocusing techniques". Optics and Lasers in Engineering, vol.54, pp. 236-246, 2014.
- [17] Y. Xu, L. Ekstrand, J. Dai, S. Zhang, "Phase error compensation for three- dimensional shape measurement with projector defocusing". Applied Optics, pp. 2572-81, 2011.
- [18] S. Zhang. "Flexible 3-D shape measurement using projector defocusing: extended measurement range". Optics Letters, pp. 931-3, 2010.
- [19] W. Lohry, S. Zhang, "3D shape measurement with 2D area modulated binary patterns". Optics and Lasers in Engineering, vol. 50, issue 7, pp. 917-921, 2012.
- [20] J. Dai, B. Li, S. Zhangm, "Intensity-optimized dithering technique for three-dimensional shape measurement with projector defocusing". Optics and Lasers in Engineering, pp.79-85, 2014.
- [21] S. Zhang, "Three-dimensional range data compression using computer graphics rendering pipeline". US 20140063 024A1 2014.
- [22] A. Abdulbasit, "FPGA Implementation for Fringe Pattern Demodulation Using the One-Dimensional Modified Morlet Wavelet Transform", International Journal of Engineering and Innovative Technology, vol. 3, pp. 2277-3754, 2013.
- [23] A. Abdulbasit, "Fringe pattern demodulation using the one-dimensional continuous wavelet transform: field-programmable gate array implementation". Applied Optics, vol. 52, 2013.
- [24] S. Saadi, M. Touiza, F. Kharfi, A. Guessoum. "Dyadic wavelet for image coding implementation on a Xilinx MicroBlaze processor: application to neutron radiography". Applied Radiation and Isotopes, vol. 82, pp. 200-210, 2013.

4. Conclusions

An approach is proposed in this research to solve the problems concerning the unwrapping of phase extracted though the fringe pattern. The proposed approach is realized based on a number of traditional algorithms to be correspondingly taken into account including the frequency estimation approach and the corresponding Itoh approach. In fact, the key suggestions of the research are to present the new tools for processing of the fringe pattern; nonetheless, a number of other issues might be investigated in future research. The proposed one does not have an acceptable performance in case of images with higher than average noise. It might be improved by changing two utilized parallel methods. Furthermore, the performance of algorithm can increased by using other mother wavelets as the known Morlet mother wavelet is now exploited in the proposed approach.

Amir Hooshang Mazinan received the Ph.D. degree in 2009 in Control Engineering. Dr. Mazinan is the Associate Professor and also the Director of Control Engineering Department at Islamic Azad University, South Tehran Branch, Tehran, Iran, since 2009. He is now acting as the Associate Editor in Transactions of the Institute of Measurement and Control (Sage publisher) and the Guest Editor in Computers & Electrical Engineering Journal (Elsevier Publisher), as well. Moreover, he is a member of Editorial Board in three international journals and also a member of programming committee in four international conferences. Dr. Mazinan has more than 110

journal and conference papers in so many reputable publishers. His current research interests include intelligent systems, model based predictive control, over-actuated space systems modeling and control, time–frequency representation, filter banks, wavelet theory and image–video processing.

Ali Esmacili received the M.Sc. degree in 2014 in Electronics Engineering. His current research interests include signal, image and video processing.

Hybrid Task Scheduling Method for Cloud Computing by Genetic and PSO Algorithms

Amin Kamalinia

Department of Computer Engineering, Urmia Branch, Islamic Azad University, Urmia, Iran
Amin.kamalinia@gmail.com

Ali Ghaffari*

Department of Computer Engineering, Tabriz Branch, Islamic Azad University, Tabriz, Iran
a.ghaffari@iaut.ac.ir

Received: 11/Sep/2015

Revised: 17/Aug/2016

Accepted: 28/Sep/2016

Abstract

Cloud computing makes it possible for users to use different applications through the internet without having to install them. Cloud computing is considered to be a novel technology which is aimed at handling and providing online services. For enhancing efficiency in cloud computing, appropriate task scheduling techniques are needed. Due to the limitations and heterogeneity of resources, the issue of scheduling is highly complicated. Hence, it is believed that an appropriate scheduling method can have a significant impact on reducing makespans and enhancing resource efficiency. Inasmuch as task scheduling in cloud computing is regarded as an NP complete problem; traditional heuristic algorithms used in task scheduling do not have the required efficiency in this context. With regard to the shortcomings of the traditional heuristic algorithms used in job scheduling, recently, the majority of researchers have focused on hybrid meta-heuristic methods for task scheduling. With regard to this cutting edge research domain, we used HEFT (Heterogeneous Earliest Finish Time) algorithm to propose a hybrid meta-heuristic method in this paper where genetic algorithm (GA) and particle swarm optimization (PSO) algorithms were combined with each other. The experimental results of simulation are shown that the proposed algorithm optimizes the average makespans of the HEFT_UpRank, HEFT_DownRank, HEFT_LevelRank and MPQMA for 100 independent task graphs scheduling with 10, 50 and 100 tasks. Total optimization of makespans by the proposed algorithm against the other algorithms were 6.44, 10.41, 6.33 and 4.8 percent respectively.

Keywords: Cloud Computing; Task Scheduling; Genetic Algorithm; Particle Swarm Optimization Algorithm.

1. Introduction

In the recent years, with huge advancement in IT (information technology) based systems [1-3], cloud computing is considered as one of the most important trends [4]. Cloud computing is considered to be a novel scientific tool and asset for high-performance computing (HPC). It refers to a technology which uses internet and central distant service provision in order to maintain data and applications. Moreover, this technology can be used in high-performance computing to centralized storages, memory, processing and bandwidth. Cloud computing is used as a technology to supply the resources of information and communication technology (ICT) dynamically and scalably all over the internet. Also, cloud computing is a pattern which provides computational resources are delivered to users on demand over the Internet as a public service [5]. Furthermore, Task scheduling in software defined networks (SDNs) [6] based cloud computing is an important challenges for future work. Scheduling is regarded as a decision-making process which is regularly used in the majority of production and service-providing industries which is used to enhance efficiency optimization [7]. Indeed, scheduling refers to the allocation of limited resources to tasks throughout time [8]. It should be noted that unique

features and characteristics of resource management and service scheduling distinguish cloud computing from other computing methods. Whereas centralized scheduling in a clustered system is aimed at enhancing the efficiency of the entire system, distributed scheduling in a grid computing system is intended to enhance efficiency for a certain final user. When compared with other systems, scheduling in cloud computing is much more complicated; hence, a centralized scheduling is required [9]. Each cloud provider is obliged to provide services for the users. It should be noted that the cloud provider provides services without mentioning the location of host infrastructures and data centers. On the other hand, commercial features make it necessary for cloud computing to consider the users' needs and preferences with respect to the quality of services all over the world.

In cloud computing, there is a data center which includes interconnected equipment and machines where they have high-speed links and connections with each other. Such an environment is appropriate for processing a mass of diverse tasks and activities. Scheduling in distributed systems refers to the allocation of multiple tasks to multiple machines which intends to enhance optimization; hence, it is considered to be an NP-complete. It can be argued that heuristic algorithms are usually used as less than ideal and desirable algorithms to

* Corresponding Author

achieve relatively good solutions. Hence, in recent years, evolutionary algorithms are used to better optimize solutions. In this paper, a novel algorithm has been proposed where genetic and particle swarm optimization algorithms were combined and also the HEFT algorithm was used to schedule tasks in the context of cloud computing. Also, we simulated and evaluated proposed scheme and analyzed it statistically. The results of simulation and statistical analysis of proposed method indicate that the proposed algorithm is optimized the average makespans of the HEFT_UpRank, HEFT_DownRank, HEFT_LevelRank and MPQMA for 100 independent task graphs scheduling with 10, 50 and 100 tasks. Percent of the total optimization of makespans for the mentioned algorithms were 6.44, 10.41, 6.33 and 4.8 respectively.

The reminder of the paper is organized as follows: in Section 2, studies related to task scheduling are briefly reviewed. In Section 3, HEFT algorithm is described and discussed. In Section 4, genetic algorithm (GA) is described and reviewed. Section 5 is concerned with PSO algorithm. Section 6 describes the algorithm proposed in this paper. Section 7 describes experimental results of the proposed algorithm. Finally, Section 8 presents the conclusion and suggestions for further research.

2. Related Works

Most of studies on task scheduling issues has been studied in distributed high performance computing (HPC) environments such as clusters, Grid [10] and also cloud computing. Numerous research studies have been conducted on the issue of scheduling in cloud computing. Some of the related studies are reviewed in this section of the study. Researchers considered the virtualization features and commercial features in cloud computing to propose a task scheduling algorithm for the first time based on Berger model [11]. This algorithm maintains the dual fairness limitation in the process of task scheduling. The first categorization limitation selects the user's tasks based on service quality priorities by creating a public wait function. In selecting a user's tasks, task categories are taken into consideration to avoid the fairness of resources in the selection process. The second limitation is related to define resource justice function which is used to judge about the justice and fairness of resource allocation. The main motivation of researchers in [12] was to design and develop a cloud resource server for efficient handling of cloud resources and doing tasks for scientific programs with respect to the deadline determined by the user. The deadline was based on task scheduling. Task scheduling was combined and implemented with particle swarm optimization algorithm. This solution was intended to reduce task execution time and cost based on the defined fitness function. Researchers developed a new task scheduling algorithm for executing massive programs

and applications in cloud [13]. This economical and low-cost task scheduling algorithm operates based on two heuristic methods. The first strategy dynamically maps the tasks to the best virtual machines in terms of cost according to the Pareto dominance. The second strategy which complements the first strategy reduces financial costs from unimportant tasks.

Researchers in [14], proposed a novel memetic task scheduling algorithm on cloud environment using multiple priority queues which named MPQMA (multiple priority queues and a memetic algorithm). This algorithm employs a genetic algorithm with local search algorithm to solve scheduling problem in heterogeneous computing systems. The main goal of this algorithm is using advantage of MA to increase the convergence speed of the solutions. Experimental results of randomly generated graphs discovered that the MPQMA algorithm optimized the other four current algorithms in terms of makespan with fast convergence of solutions. In work [15], the researchers proposed a population based meta-heuristic algorithm based on particle swarm optimization (PSO) to schedule applications on cloud resources. This algorithm considers both computation cost and data transmission cost. Experiment results are gained with a workflow application by varying its computation and communication costs. The algorithm is compared with existing 'Best Resource Selection' (BRS) algorithm in terms of the cost savings. The results illustrated that PSO betters BRS in times cost savings and distribution of workload onto resources. In research [16], the concept of project scheduling with the workflow scheduling problem are integrated to formulate a mathematical model that aims to minimize the makespan. In order to solve the workflow scheduling optimization problem two Artificial Bee Colony algorithms are applied. This algorithm is compared with the optimal solutions obtained by Gurobi optimizer to evaluate performance of ABC on the different workflows. The experimental results depict that ABC can be utilized as a practical method for complex workflow scheduling problems in the cloud computing environment.

In [17], a task scheduling based on Ant Colony Optimization (ACO) for task scheduling problem is proposed to minimize the makespan of the tasks submitted on the cloud environment. In addition, the ACO is applied to improve the efficiency of Cloud computing system. Experimental results are achieved by Cloud simulator which called CloudSim. In this work, the various graph from 100 to 500 of tasks are used to evaluate the algorithm in different situations.

Using genetic algorithm and multiple priority queues called MPQGA, the researchers proposed a task scheduling method in heterogeneous computational systems [18]. The rationale behind this method was to benefit from both heuristic and evolutionary algorithms and make up their shortcomings. The algorithm proposed in [18] utilized the genetic algorithm for allocating task priority and made use of the EFT heuristic method for mapping and dedicating tasks to the processor. In MPQGA

method, crossover and mutation operators, and the appropriate fitness function were designed for the scenario of directed acyclic graph. The results of experiments indicated that the MPQGA algorithm performed better than the two non-evolutionary methods and the random search method with respect to scheduling quality.

3. HEFT Algorithm

In general task scheduling algorithms divided into static or dynamic [19]. The static task scheduling algorithm HEFT was first introduced in [20]. In this method, the scheduling algorithm was used for a limited number of heterogeneous processors. It was used for parallelizing the processors so as to enhance efficiency and fasten scheduling. Before discussing the HEFT algorithm, it is necessary to introduce the terms EFT (earliest finish time) and EST (earliest start time). EST and EFT refer to the earliest starting time and the earliest finishing time of the execution of the task n_i on the processor p_j . The value of EST for then try task is equal to zero which has been defined in Equation (1). For other tasks in the graph, the values of EST and EFT are defined recursively according to Equations (2) and (3). For measuring the EFT of task n_i , all the procedures of this task should be scheduled. In these equations, $pred(n_i)$ stands for the entire procedures of the task n_i and $avail\{j\}$ refers to the earliest time of the p_j processor which is ready to execute the task. If n_k is a recent task which is dedicated to the processor p_j , then, $avail\{j\}$ refers to the time at which the processor p_j has finished the execution of task n_k and it is ready to execute another task in case it has used a non-insertion-based scheduling method. Internal max in Equation (2) measures the time at which all the required data for n_i has arrived at p_j . After task n_m has been scheduled on the processor p_j , the earliest starting time and the earliest finishing time of task n_m on the processor p_j will be equal to AST (actual start time) and AFT (actual finishing time). It should be noted that, according to Equation (5), AFT will be equal to the smallest obtained EFT for that task. After the scheduling of all the graph tasks, the scheduling makespan will be equal to the AFT of the exit task. In case there are several output tasks or in case there are no pseudo-task, the makespan of the scheduling will be obtained through Equation (4). Moreover, $c_{m,i}$ refers to the communication costs from node m to node i . If the two tasks m and i are allocated to the same processor, $c_{m,i}$ will be equal to zero.

$$EST(n_{entry}, p_j) = 0 \quad (1)$$

$$EST(n_i, p_j) = \max \left\{ avail\{j\}, \max_{n_m \in pred(n_i)} (AFT(n_m) + c_{m,i}) \right\} \quad (2)$$

$$EFT(n_i, p_j) = \omega_{i,j} + EST(n_i, p_j) \quad (3)$$

$$makespan = \max\{AFT(n_{exit})\} \quad (4)$$

$$AFT(n_i, p_j) = \min_{1 \leq l \leq m} EFT(n_i, p_l) \quad (5)$$

In the HEFT algorithm, the priorities are determined recursively based on task upward rank according to Equation (7). In this equation, $succ(n_i)$ refers to a set of the successors of task n_i and $\bar{c}_{i,j}$ stands for the average cost of the communication edge (i,j) and $\bar{\omega}_i$ refers to the average computational cost of task n_i which is measured through Equation (8). As its name denotes, since rank starts from the output node and is measured recursively, hence, it is referred to as upward rank. The upward rank of the output node is measured through Equation (6). Basically, $rank_u(n_i)$ refers to the length of critical path from task n_i to the output task which also includes the computational cost of task n_i .

$$rank_u(n_{exit}) = \bar{\omega}_{exit} \quad (6)$$

$$rank_u(n_i) = \bar{\omega}_i + \max_{n_j \in succ(n_i)} (\bar{c}_{i,j} + rank_u(n_j)) \quad (7)$$

$$\bar{\omega}_i = \sum_{j=1}^q \omega_{i,j}/q \quad (8)$$

Similarly, downward rank is obtained recursively through Equation (9). $pred(n_i)$ refers to the set of procedures of the task n_i . As the name suggests, downward rank is obtained recursively through the downward graph movement of the task which starts from the input node of graph. The downward rank of the input node is equal to zero. In general, $rank_d(n_i)$ is the longest distance from the input node to the task n_i where the computational cost of the task is not considered.

$$rank_d(n_i) = \max_{n_j \in pred(n_i)} (rank_d(n_j) + \bar{\omega}_j + \bar{c}_{i,j}) \quad (9)$$

The HEFT algorithm has two phases. The first phase is concerned with prioritization of tasks so that the priorities of all tasks are measured. The second phase is concerned with the selection of the processor so that task are chosen based on their priorities and the scheduling of each selected task is allocated to the best processor which can minimize the finishing time of the task.

Task prioritization phase: in this phase, the priority of each task can be measured through different methods some of which are mentioned below. The priority of each task is determined through upward rank and downward rank according to the procedure reported in [20] which have been defined in Equations (7) and (9). Also, priorities can be measured by combining the two methods which have been described in [18] in which Equation (10) is first used to level the graph; then, the values of levels and the values of upward rank and downward rank are used to produce a new prioritization queue. As a result, those tasks which are at the same level are arranged in a descending order. After one of the mentioned methods is selected and their values for each task are calculated, a list of tasks is produced based on descending order of tasks. In case the value of the selected method is equal for several tasks, the tasks are randomly

selected. It should be noted that upward rank is based on average computation and communication cost. It is obvious that the descending order of the upward rank values create a topological order of tasks which is regarded as a linear order in which precedence limitations are preserved.

$$Level(T_i) = \begin{cases} 0, & \text{if } T_i = T_{entry}; \\ \max_{T_j \in pred(T_i)} (Level(T_j)) + 1, & \text{otherwise} \end{cases} \quad (10)$$

Processor selection phase: in the majority of task scheduling algorithms, the earliest time for the accessibility of the processor p_j for executing a task is when p_j has finished the previous task. Moreover, some algorithms have the insertion-based policy. As a case in point, HEFT is based on insertion-based policy which considers the probability of inserting a task in the idle time slot between two previous scheduled tasks. The length of the idle time slot of the processor is the distance between the starting execution time and the finishing time of two tasks which were consecutively executed on the same processor. At least, it should be able to execute the computational cost of the task. Furthermore, scheduling an idle time slot should consider the precedence limitations.

4. Genetic Algorithm

Genetic algorithm is deemed to be a search and optimization method which is based on the principles of genetics and natural selection [21]. Genetic algorithm is a type of evolutionary algorithms which has been inspired by the Darwin theory about evolution. This algorithm was developed by John Holland at the Michigan University during the 1960's and 1970's. Later, one of Holland's students named David Goldberg was able to propose a solution based on evolutionary algorithms to a challenging issue about the control of gas pipeline transmission [22,23]. The major contribution of Holland was published in a book entitled "Adaptation in Natural and Artificial Systems" [24]. Holland's theory was expanded and now it was developed into a powerful algorithm for solving the search and optimization problems. This algorithm has the following three operators: selection, crossover and mutation operators. The details about the implementation of these operators have been discussed later in this paper.

5. PSO Algorithm

Particle swarm optimization (PSO) algorithm is a population-based random optimization method which was proposed by Russell Eberhart and James Kennedy in 1999. The development of this algorithm was inspired from the swarm behavior of birds or fish [24,25]. This system begins with a population which has random solutions and it updates the generation to find an optimal solution. In contrast with genetic algorithm, none of the evolutionary operators such

as crossover and mutation are available in the PSO algorithm. Solutions in PSO algorithm are referred to as particles which move in the problem search space and follow the current optimal particle [26]. In this algorithm, each particle follows the particle which has a better fitness function among all the particles. However, it does not forget its own experience. Hence, it follows the condition and state in which it has the best fitness function. Thus, in each iteration of the algorithm, each particle determines its next position based on two values: first, the best position that the particle has ever had indicated by pbest and also the best position that all the particles have ever had indicated by gbest. In other words, gbest refers to the best pbest in the entire population. Conceptually, pbest for each particle refers to the memory which a particle has experienced about its best position. gbest represents the public knowledge of the population and when particles change their positions based on gbest, they try to keep up with the knowledge of the population. Conceptually, the best particle connects all the particles of the population with each other [26,27]. In this method, the next position for each particle is determined according to the following equation:

$$v_i(t+1) = w \cdot v_i(t) + c_1 r_1 (pbest(t) - x_i(t)) + c_2 r_2 (gbest(t) - x_i(t)) \quad (11)$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (12)$$

In Equation (11), $v_i(t)$ refers to the speed or velocity of particle i in the time unit of t . Also, w which is indicated by α refers to the coefficient or inertia weight for controlling exploitation and exploring the search space. C_1 and C_2 are the learning parameters. In other words, they are constant accelerators which change the speed changes of the particle towards pbest and gbest. Indeed, the value of these two variables are equal to 2. The values of r_1 and r_2 are two random variables which vary between 0 and 1. In Equation (12), $x_i(t)$ represents the position of particle i in the time unit of t .

6. The Proposed Algorithm

The input of the problem is a directed acyclic graph which is indicated by $G = (V, E)$. Each node is a member of V set which is a vertex of the graph and indicates one task from all the set of tasks; the weight of these nodes determine the execution time of the tasks. This graph also includes a set of edges; in other words, it includes E which indicates the prerequisite relations among the tasks. In case there exists an edge such as (t_i, t_j) , it means that task t_j cannot start until task t_i is finished. These edges are weighted and the weight of each edge indicates the communication cost of sending a message between two tasks. This cost exists when two related tasks are executed on different processors or machines and in case they are executed on the same processor or machine, the cost of communication between them will be zero.

Directed acyclic graph illustrated in Figure 1 includes the following tasks: $t_0, t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9, t_{10}$ which are the input of the proposed algorithm. The node t_0 is the entry task and t_{10} is the exit task. Table 1 indicates the costs of executing tasks on the m_0, m_1 and m_2 . Also, $\bar{\omega}$ indicates the average costs of executing tasks on the machines. As noted, each task is executed with a different cost on each machine which indicates the heterogeneity of the computational context of tasks.

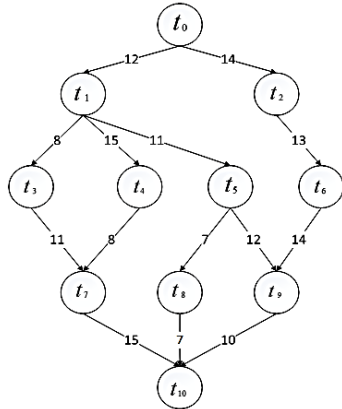


Fig. 1. DAG with 11 tasks

Table 1. Task execution costs on machines

Tasks	m_0	m_1	m_2	$\bar{\omega}$
t_0	7	9	8	8
t_1	10	9	14	11
t_2	5	7	6	6
t_3	6	8	7	7
t_4	10	8	6	8
t_5	11	13	15	13
t_6	12	15	18	15
t_7	10	13	7	10
t_8	8	9	10	9
t_9	15	11	13	13
t_{10}	8	9	10	9

One of the most important challenges in scheduling tasks in the cloud computing context is the selection of the best solutions for allocating resources to the tasks so that the cost and task finishing time are reduced. Inasmuch as there are a lot of tasks and there are different solutions for different tasks, hence, the selection of a solution is not a unique choice. That is, there is a set of choices and each choice is not preferred to the other choice. In the proposed hybrid method in this paper, a set of answers is produced by the genetic algorithm; then, these answers are considered as the initial population for the PSO algorithm and based on these answers, the next population for the genetic algorithm is produced with the help of PSO algorithm. At the end of this stage, based on the PSO algorithm, the whole produced answers are updated and the stages are repeated again. In each repetition, first, the particles find answers with respect to the operators of mutation and crossover. Then, PSO algorithm is used to produce children without moving entry and exit nodes. Hence, an optimal population is produced. It should be noted that if the children's priority is violated after the production of children, they will be sorted from left to right so that the priorities are

not violated. In the proposed algorithm, the solutions prevented premature convergence before achieving an absolute optimal solution. It should be noted that after the crossover and mutation operators are executed each time, the replacement process is carried out so that the produced children are compared with their parents. If the fitness function of the children are not better than their parents, then, they are eliminated. Otherwise, they will replace and eliminate their parents. Figure 2 illustrates the flowchart of the genetic and PSO algorithms proposed in this paper.

In the PSO algorithm, each particle includes a solution which cover the context of the problem. In each iteration, the fitness or cost function is measured for all the particles. Then, the memory of each particle (pbest) is compared with the obtained value and in case the value of particle cost function is smaller than the value of its memory, particle memory will be equal with the current state of that particle; if these conditions occur, then, in this way, the memory value will be compared with the gbest value. As a result, the minimum solution is obtained for the problem. The implementation of the PSO algorithm is considered to be computationally simple; in case appropriate values are used for its parameters, it is highly probable to find an optimal answer. To avoid local optimality, the PSO algorithm functions in a way that when it is placed in an optimality, the particles mutate to other parts of the search space. Then, in other parts, they search for optimal answers.

In the genetic algorithm, once the initial population is created, the appropriateness of the answers is measured by means of the fitness function value. For having an optimal answer in the proposed method, the proposed model should have a small value for the fitness function.

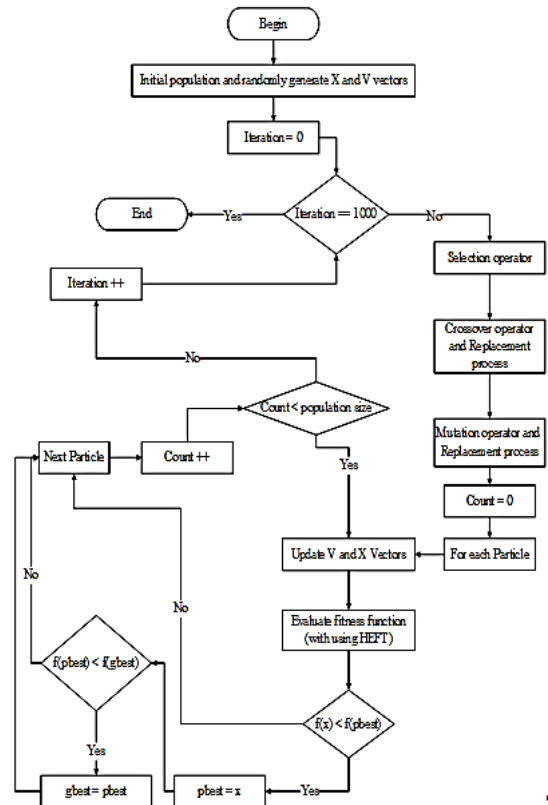


Fig. 2. Flowchart of the proposed algorithm in this paper

6.1 The Production of Initial Population

The initial population includes particles which are independent of each other where the sizes of chromosomes are fixed. In this paper, for having variety and appropriate initial values, three traditional heuristic methods were used to give initial values to the three particles. The three methods include upward rank, downward rank and a combination of these two methods based on their rank [18]. The initial values of the three particles were given according to the above-mentioned methods for the graph included in Figure 1. Indeed, multiple priority queues are produced which is shown in Table 2 for the directed acyclic graph. The remaining particles were randomly valued which is explained later in the paper. That is, the beginning and end of the chromosome which are the start and exit nodes are established in the chromosome. Those between these two nodes are randomly selected from left to right and are sorted provided that the priorities are not violated.

Table 2. Task Priorities

Tasks	$rank_u(t_i)$	$rank_d(t_i)$	level	$rank_u(t_i) + rank_d(t_i)$
t_0	102	0	0	102
t_1	79	20	1	99
t_2	80	22	1	102
t_3	52	39	2	91
t_4	50	46	2	96
t_5	57	42	2	99
t_6	61	41	2	102
t_7	34	62	3	96
t_8	25	62	3	87
t_9	32	70	3	102
t_{10}	9	93	4	102

6.2 Measuring Makespan for each Particle

For measuring makespan for each particle in this paper, tasks should be executed based on a processor or machine allocation method. This operation was conducted by means of the HEFT processor allocation method on each particle which is discussed below.

6.3 Fitness Function

The fitness value plays a significant role in deciding which particles should be used to produce the next generation. In this paper, makespan of a DAG is obtained from finishing time of exit task in an application which this makespan assumed the fitness of algorithm. In scheduling issue, the purpose of allocating task is to reduce the makespan without violating priorities. The makespan is obtained through Equation (4) and the fitness function is obtained through Equation (13).

$$fitness_i = makespan_i \tag{13}$$

6.4 Selection Operator

One of the significant parts of genetic algorithm is selection which has a remarkable impact on convergence. Indeed, a particle with a better fitness value is more likely to mate. One of the best implementation methods is the roulette wheel. This method assumes that the selection

probability is a ratio of particle fitness. Some of the particles will be reselected for the genetic operation based on their fitness. A particle with the highest fitness is highly probable to be selected. Particles are measured according to their fitness. The value of the fitness function is always greater than zero. p_i stands for the probability of each particle to be selected is measured through Equation (14). Algorithm 1 shows the selection pseudo code.

$$p_i = \frac{fitness_i}{\sum_{j=1}^{PopSize} fitness_j} \tag{14}$$

Algorithm 1. Roulette wheel pseudo code

```

1: Generate a random number  $R \in [0,1]$ 
2: For  $i=1$  to PopSize do
3:   If  $p_i > R$  then
4:     Select the chromosome;
5:     Return the chromosome;
6:   end if
7: end for.
    
```

Algorithm 2. Pseudo code of single-point combination operator

```

1: Choose randomly a suitable crossover point  $i$ ;
2: Cut the first parent's chromosome and the second parent's chromosome into left and right segments
3: Generate a new offspring, namely the child one;
4: Inherit the left segment of the first parent's chromosome to the left segment of the child one's chromosome;
5: Copy genes in second parent's chromosome that do not appear in the left segment of first parent's chromosome to the right segment of child one's chromosome;
6: Generate a new offspring, namely the child two;
7: Inherit the left segment of the second parent's chromosome to the left segment of the child two's chromosome;
8: Copy genes in first parent's chromosome that do not appear in the left segment of second parent's chromosome to the right segment of child two's chromosome;
9: if offspring's fitness values are better than their parents then replace them
10: if step 9 is true then compare fitness value of offspring with local best if offspring's fitness value is better then replace it.
11: if step 10 is true then compare fitness value of offspring with global best if offspring's fitness value is better then replace it.
    
```

6.5 Crossover Operator

The population of a genetic algorithm is evolved and completed by crossover and mutation. In the method used in this paper, the crossover operator is regarded as a significant operation. Crossover is a function of replacing some genes of one parent with genes of another parent. In the task scheduling issue, the crossover operator combines the two parents with each other so as to produce two valid children.

In this paper, single-point crossover was implemented according to the method mentioned in [18]. That is, firstly, a random point between 1 and n is selected and the crossover point takes the priority queue of both parents from left to right in case they are not identical. For example, consider the particles depicted in Figure 3. The crossover point which is equal to 6 produces the single point of two new children. Indeed, it uses crossover

operator to replace some genes. The left part of children inherit their parents' genes. Then, some selected genes are eliminated from the parent and the remaining genes are added to the child from left to right. Consequently, the child will also be valid [18]. Then, the value of fitness function will be measured for each child. The fitness values of children are compared with those of parents and in case the fitness values of children are better than those of parents, the children will replace parents. Then, the fitness value of each child will be compared with the memory of that particle (pbest). If the fitness value is better than pbest, it will replace the memory of that particle. Also, if the mentioned conditions occur, the fitness value of particles will be compared with the gbest value. In case the pbest value of particle is less than the gbest value, it will replace it. Algorithm 2 represents the pseudo code of this operator.

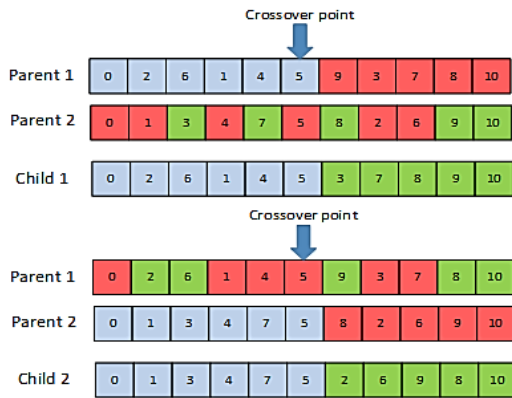


Fig. 3. Crossover operator

6.6 Mutation Operator

This operator replaces a gene with another one based on a certain probability. Mutation operator causes variety and diversity in the population. Accordingly, it expands the search space and prevents the algorithm from local optimization. Usually, this operator is done after the crossover operator and helps to gain a better solution. A new chromosome is obtained by exchanging two genes if the precedence constraint is not violated [18]. In this paper, the mutation operator is inspired from [18]. In other words, at first, a gene is randomly selected. Then, based on this method, the first successor for the task (t_j) from the mutation point to the end is obtained. If there is m^{th} gene which is a member of $[i + 1, j - 1]$ and the priorities of t_m are not in front of t_i , t_i and t_j can be replaced with each other which is illustrated in Figure 4. If these conditions do not occur, hence, the mutation operator will be executed from the beginning. After exchanging the genes, the fitness of the child is calculated by fitness function. The fitness value of the generated child will be compared with its parent. If the fitness results of the child is better, the child will replace with the parent. After that, the fitness value of the child will be compared with the memory of that particle (pbest). The fitness value of the particle is replaced with pbest if the obtained fitness betters the pbest. Moreover, if this condition is established, the fitness value of the particle will be compared with the gbest value and in

case the value of this particle is less than the gbest value, it will be replaced with the gbest. Algorithm 3 represents the pseudo code of the mutation operator.

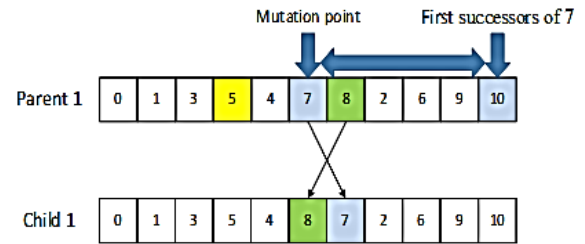


Fig. 4. Mutation operator

Algorithm 3. Pseudo code of two-point mutation operator

```

1: A randomly chosen chromosome.
2: Choose randomly a gene  $T_i$  in the selected chromosome;
3: Find the first successor  $T_j \in Succ(i)$ ;
4: Choose randomly a gene  $T_k$  in the interval  $[i + 1, j - 1]$ ;
5: if  $1 < i$  for all  $T_l \in Pred(k)$  then
6:   Generate a new offspring by interchanging gene  $T_i$  and
   gene  $T_k$ ;
7: return the new offspring;
8: else
9:   Go to Step 1;
10: if the fitness of the new offspring is better than its parent
then replace it with parent
11: if step 10 is true then compare fitness value of offspring
with local best if offspring's fitness value is better then
update local best.
12: if step 11 is true then compare fitness value of offspring
with global best if offspring's fitness value is better then
update global best.

```

6.7 Termination Condition

The genetic and PSO algorithms are regarded as random methods which can be executed for ever by means of a rule. In practice, a termination condition should be carried out. The usual methods operate by considering the fitness evaluations or the working times of the computer or by exploring the population diversity. In this paper, the termination condition is realized when the algorithm has been executed for 1000 times.

6.8 Complexity Analysis

The complexity of the proposed method is $O(\text{generators} \times n^2 \times e \times m)$, where *generators* is the number of iterations, n is the number of subtasks, e is the number of edges and m is the number of machines.

7. Experimental Results

Certain measurement criteria were used for evaluating efficiency which are mentioned later in the paper. It should be noted that the entire implementation procedure was conducted in Visual Studio 2013 and the C#.net programming language was used to implement the algorithm. There are some parameters in the combined algorithm which have a significant impact on the performance of the algorithm; these parameters are given in Table 3. In this table, the parameter *ini* determines the

number of population (particles) and the parameter w stands for the inertia weight which is aimed at balancing the speed of particles. The values of the parameters C_1 and C_2 help particles learn how to locate the optimal points. The parameters $srate$, $crate$, $mrates$ refer to the rates of the selection, crossover and mutation operators, respectively.

Table 3. Values of the parameters

parameters	Values
ini	80
C_1	1.5
C_2	1.5
r_1	Randomly
r_2	Randomly
W	0.4
srate	30
crate	80
mrates	20

7.1 Comparing Measurements

In this section the proposed algorithm is compared with other three heuristics and a GA algorithms in term of the makespan. To do this, some metrics such as SLR and CCR are used in comparison. Furthermore, random graph and statistical analysis are used in experimental comparisons.

7.1.1 Task Makespan

The makespan of the directed acyclic graph, as shown in Figure 1, was simulated on three prioritization methods by means of upward rank, downward rank and the combination these two methods and MPQMA algorithm. The simulation results for Figure 1 with the proposed and other algorithms are illustrated in Figure 6, Figure 7, Figure 8 and Figure 9 respectively. The obtained results as in mentioned figures were 76, 74, 76 and 70 for the four algorithm. When scheduling by the proposed algorithm is finished, the makespan for graph execution is equal to 68 which is illustrated in Figure 10. These gained results from algorithms are shown in Figure 5 by Bar chart which depicts that the makespan of scheduling for the proposed algorithm has better performance than the other four algorithms.

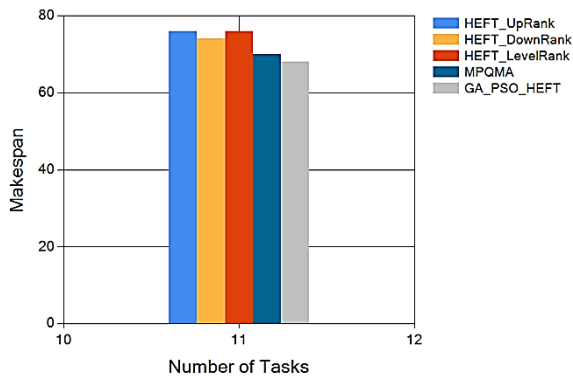


Fig. 5. Bar graph representing the makespan vs. number of tasks for the graph of Fig. 1

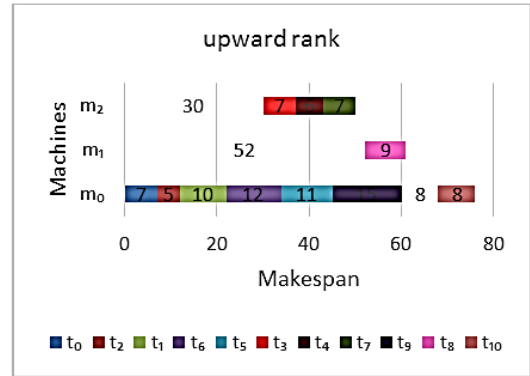


Fig. 6. Gantt chart for task scheduling with upward rank prioritization (Makespan = 76)

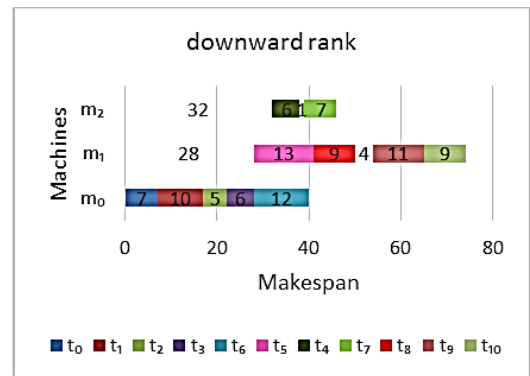


Fig. 7. Gantt chart for task scheduling with downward rank prioritization (Makespan = 74)

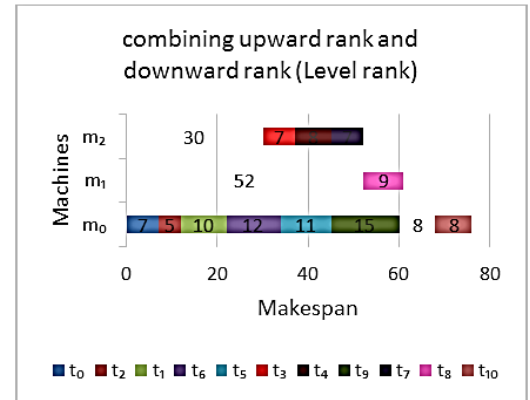


Fig. 8. Gantt chart for task scheduling with both upward and downward ranks (Makespan = 76)

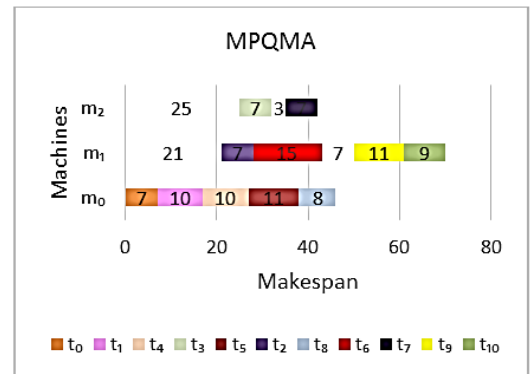


Fig. 9. Gantt chart for task scheduling with MPQMA (Makespan = 70)

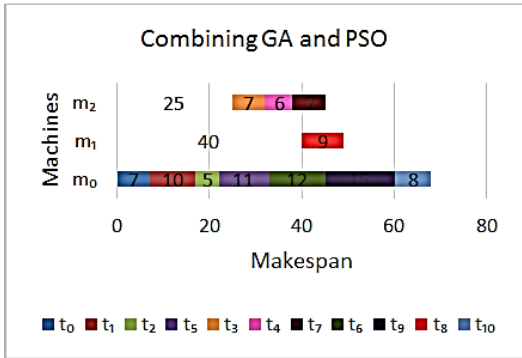


Fig. 10. Gantt chart for task scheduling with a combination of the prioritization of genetic and PSO algorithms (Makespan = 68)

7.1.2 Scheduling Length Ratio (SLR)

In this paper, for measuring SLR, the minimum critical path between the processors or machines should be obtained. Hence, the critical path method (CPM) was used [28].

The measurement of the main efficiency of the scheduling algorithm on the graph is the scheduling length. Since a large collection of task graphs with different features is used, the scheduling length to the low bound should be converted into a rule which is referred to as scheduling length ratio (SLR). The value of SLR on the graph is obtained through Equation (15).

The denominator of the ratio is the set of minimum computation costs of the tasks on CP_{min} . In an unscheduled directed acyclic graph, if the computation cost of each node n_i is adjusted with less value, then, the critical path will be based on minimum computation costs which are indicated by CP_{min} . The SLR of a graph cannot be less than one. In the task scheduling algorithm, the smallest SLR will have better efficiency [18].

$$SLR = \frac{makespan}{\sum_{T_i \in CP_{min}} \min_{P_k \in P} (\omega(T_i, P_k))} \tag{15}$$

7.1.3 Communication to Computation Ratio (CCR)

CCR indicates that the used DAGs in this paper is either communication-intensive or computation-intensive. For a graph, the value of CCR is obtained by measuring the mean communication cost (numerator of Equation (16)) divided by the mean computation cost (denominator of Equation (16)) in the computational system. Hence, CCR value is measured through the following equation [18].

$$CCR = \frac{\frac{1}{e} \sum_{edge(T_i, T_j) \in E} \overline{C(T_i, T_j)}}{\frac{1}{n} \sum_{T_i \in T} \overline{\omega(T_i)}} \tag{16}$$

7.2 Comparison of SLR vs. CCR for Graph Depicted in Fig. 1

The comparison of SLR vs. CCR calculations in the proposed method was done and simulated using upward and downward ranks and using the combination of priority methods. The results of comparisons are depicted in Figure 11. As shown in this figure, the method proposed in

this paper has less scheduling length rate which is attributed to the fact that the makespan in the proposed method is less than other methods. Consequently, this leads to the minimization of Equation (15).

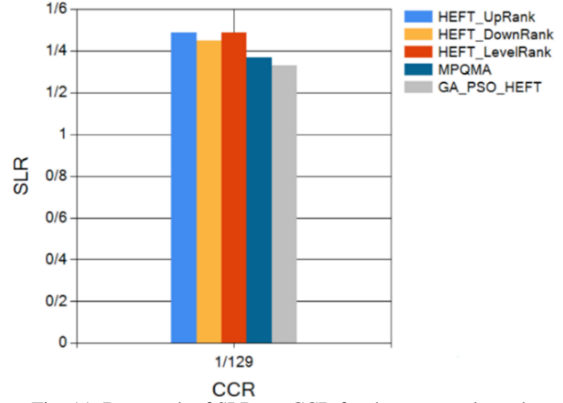


Fig. 11. Bar graph of SLR vs. CCR for the proposed graph

7.3 Comparing SLR vs. CCR for Random graph

In this paper, for a more extensive comparison and evaluation, randomly produced graphs were used. In this section, a random graph is examined. The directed acyclic graph which was randomly produced has 30 tasks; in total, it has 72 edges. This graph was executed on 9 machines. As illustrated in Figure 12, the proposed method has less SLR than the other methods. Furthermore, in the proposed method, the makespan on the random graph was 176. Accordingly, with respect to the results demonstrated in Figure 13, it can be maintained that the proposed method has better performance than the other methods.

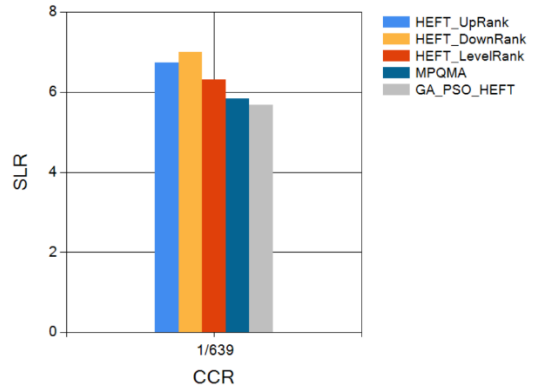


Fig. 12. Bar graph of SLR vs. CCR for the random graph

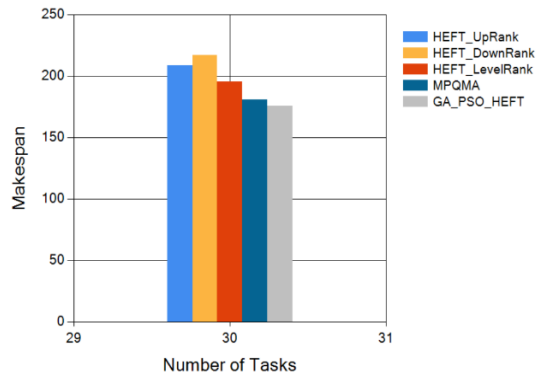


Fig. 13. Bar graph of the makespan vs. the number of tasks for the random graph

7.4 Evaluation of the Randomly Produced Graphs

For evaluating results on different graphs with 10, 50 and 100 tasks using 8 machines, 100 iterations of the three mentioned tasks were produced. Figure 14, Figure 15 and Figure 16 represent the results obtained from the experiments. In Figure 14, average makespans of 100 independent task graphs with 10 tasks for HEFT_UpRank, HEFT_DownRank, HEFT_LevelRank, MPQMA and the proposed algorithms are: 99.27, 101.03, 99.48, 92.64 and 92.18 respectively. With respect to the obtained results the proposed algorithm has better performance than the other methods.

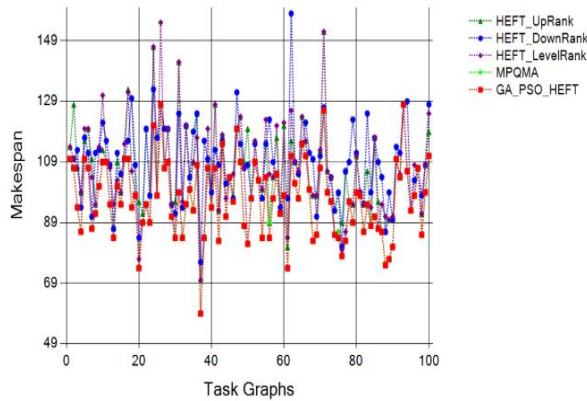


Fig. 14. Makespan of the produced 10-fold graph with 8 machines and 100 independent executions

According to the results in Figure 15, the average makespans of 100 independent task graphs scheduling with 50 tasks for the HEFT_UpRank, HEFT_DownRank, HEFT_LevelRank, MPQMA and the proposed algorithms are 469.29, 487.41, 469.69, 461.54 and 435.97 respectively. With respect to the obtained results the proposed algorithm better makespans of the other methods.

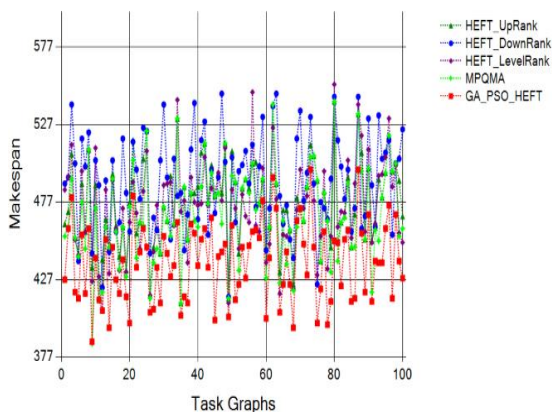


Fig. 15. Makespan of the produced 50-fold graph with 8 machines and 100 independent executions

Furthermore, the obtaining average makespans results of 100 independent task graphs with 100 tasks in Figure 16 for HEFT_UpRank, HEFT_DownRank, HEFT_LevelRank, MPQMA and the proposed algorithms are: 948.79, 985.41, 946.52, 939.73 and 897.3 respectively. According to the results the proposed algorithm optimizes the other methods in terms of makespan.

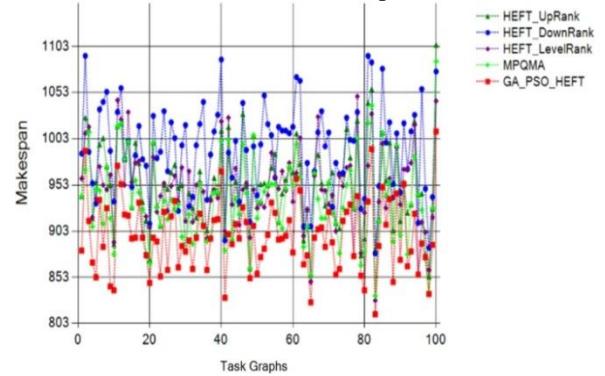


Fig. 16. Makespan of the produced 100-fold graph with 8 machines and 100 independent executions

8. Conclusion and Suggestions for Further Research

As discussed in the paper, task scheduling is considered to be one of critical challenges in cloud computing systems. In the past, numerous task scheduling methods have been used in cloud computing. In this paper, to enhance resource efficiency and minimize the total task execution time, the researchers used a novel cost function which was based on a combination of PSO and genetic algorithms. The cost function was used to measure task execution time on available resources in the context of cloud computing. The purpose of proposing the hybrid or combinatory model was to benefit from the capabilities meta-heuristic methods since they have high speed in finding optimal solutions. The new method introduced in the proposed algorithm was intended to reduce and shorten the length of the critical path and reduce the communication costs among the processors. Finally, the obtained results from the implementation of the proposed method indicated that it optimizes other mentioned current algorithms. In future, it is possible to design an appropriate scheduling for similar algorithms.

References

- [1] A. Ghaffari, "Real-time routing algorithm for mobile ad hoc networks using reinforcement learning and heuristic algorithms," *Wireless Networks*, pp. 1-12, 2016.
- [2] Z. Mottaghinia and A. Ghaffari, "A Unicast Tree-Based Data Gathering Protocol for Delay Tolerant Mobile Sensor Networks," *Information Systems & Telecommunication*, p. 59, 2016.
- [3] A. Ghaffari, "Congestion control mechanisms in wireless sensor networks: A survey," *Journal of Network and Computer Applications*, vol. 52, pp. 101-115, 6// 2015.
- [4] C.-S. Chen, W.-Y. Liang, and H.-Y. Hsu, "A cloud computing platform for ERP applications," *Applied Soft Computing*, vol. 27, pp. 127-136, 2// 2015.
- [5] Y.-D. Lin, M.-T. Thai, C.-C. Wang, and Y.-C. Lai, "Two-tier project and job scheduling for SaaS cloud service providers," *Journal of Network and Computer Applications*, vol. 52, pp. 26-36, 6// 2015.
- [6] R. Masoudi and A. Ghaffari, "Software defined networks: A survey," *Journal of Network and Computer Applications*, vol. 67, pp. 1-25, 5// 2016.
- [7] M. Pinedo, *Scheduling : theory, algorithms, and systems*, 4th ed. New York: Springer, 2012.
- [8] Y. Robert and F. d. r. Vivien, *Introduction to scheduling*. Boca Raton: CRC Press, 2010.
- [9] F. Magoulès, J. Pan, and F. Teng, *Cloud Computing : Data-Intensive Computing and Scheduling*. Boca Raton: CRC Press, 2012.
- [10] Y. Li and W. Cai, "Update schedules for improving consistency in multi-server distributed virtual environments," *Journal of Network and Computer Applications*, vol. 41, pp. 263-273, 5// 2014.
- [11] B. Xu, C. Zhao, E. Hu, and B. Hu, "Job scheduling algorithm based on Berger model in cloud environment," *Advances in Engineering Software*, vol. 42, pp. 419-425, 7// 2011.
- [12] T. S. Somasundaram and K. Govindarajan, "CLOUDRB: A framework for scheduling and managing High-Performance Computing (HPC) applications in science cloud," *Future Generation Computer Systems*, vol. 34, pp. 47-65, 5// 2014.
- [13] S. Su, J. Li, Q. Huang, X. Huang, K. Shuang, and J. Wang, "Cost-efficient task scheduling for executing large programs in the cloud," *Parallel Computing*, vol. 39, pp. 177-188, 4// 2013.
- [14] B. Keshanchi and N. J. Navimipour, "Priority-Based Task Scheduling in the Cloud Systems Using a Memetic Algorithm," *Journal of Circuits, Systems and Computers*, vol. 25, p. 1650119, 2016.
- [15] S. Pandey, L. Wu, S. M. Guru, and R. Buyya, "A Particle Swarm Optimization-Based Heuristic for Scheduling Workflow Applications in Cloud Computing Environments," in *2010 24th IEEE International Conference on Advanced Information Networking and Applications*, 2010, pp. 400-407.
- [16] Y. C. Liang, A. H. L. Chen, and Y. H. Nien, "Artificial Bee Colony for workflow scheduling," in *2014 IEEE Congress on Evolutionary Computation (CEC)*, 2014, pp. 558-564.
- [17] L. Wang and L. Ai, "Task Scheduling Policy Based on Ant Colony Optimization in Cloud Computing Environment," in *LISS 2012: Proceedings of 2nd International Conference on Logistics, Informatics and Service Science*, Z. Zhang, R. Zhang, and J. Zhang, Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 953-957.
- [18] Y. Xu, K. Li, J. Hu, and K. Li, "A genetic algorithm for task scheduling on heterogeneous computing systems using multiple priority queues," *Information Sciences*, vol. 270, pp. 255-287, 6/20/ 2014.
- [19] X. Kong, C. Lin, Y. Jiang, W. Yan, and X. Chu, "Efficient dynamic task scheduling in virtualized data centers with fuzzy prediction," *Journal of Network and Computer Applications*, vol. 34, pp. 1068-1077, 7// 2011.
- [20] H. Topcuoglu, S. Hariri, and W. Min-You, "Performance-effective and low-complexity task scheduling for heterogeneous computing," *Parallel and Distributed Systems*, *IEEE Transactions on*, vol. 13, pp. 260-274, 2002.
- [21] G. Giftson Samuel and C. Christofer Asir Rajan, "Hybrid: Particle Swarm Optimization–Genetic Algorithm and Particle Swarm Optimization–Shuffled Frog Leaping Algorithm for long-term generator maintenance scheduling," *International Journal of Electrical Power & Energy Systems*, vol. 65, pp. 432-442, 2// 2015.
- [22] R. L. Haupt and S. E. Haupt, *Practical genetic algorithms*, 2nd ed. Hoboken, N.J.: John Wiley, 2004.
- [23] F. T. Hecker, M. Stanke, T. Becker, and B. Hitzmann, "Application of a modified GA, ACO and a random search procedure to solve the production scheduling of a case study bakery," *Expert Systems with Applications*, vol. 41, pp. 5882-5891, 10/1/ 2014.
- [24] S. N. Sivanandam and S. N. Deepa, *Introduction to genetic algorithms*. Berlin ; New York: Springer, 2007.
- [25] A. Mahor and S. Rangnekar, "Short term generation scheduling of cascaded hydro electric system using novel self adaptive inertia weight PSO," *International Journal of Electrical Power & Energy Systems*, vol. 34, pp. 1-9, 1// 2012.
- [26] D. Y. Sha and H.-H. Lin, "A multi-objective PSO for job-shop scheduling problems," *Expert Systems with Applications*, vol. 37, pp. 1065-1070, 3// 2010.
- [27] A. P. Engelbrecht, *Computational intelligence : an introduction*, 2nd ed. Chichester, England ; Hoboken, NJ: John Wiley & Sons, 2007.
- [28] U. Defense Acquisition and Press, *Scheduling guide for program managers*. Fort Belvoir, VA; Washington, DC: Defense Acquisition University Press ; For sale by the U.S. G.P.O., Supt. of Docs., 2001.

Amin Kamalinia received his BS degree in Software Engineering from Bostan-Abad Branch, Islamic Azad University, Bostan-Abad, Iran, in 2011 and his MS degree in Software Engineering from Science and Research Branch, Islamic Azad University, Urmia, Iran in 2014. His research interests include Grid & Cloud computing, Task Scheduling and Programming.

Ali Ghaffari received his BSc, MSc and PhD degrees in computer engineering from the University of Tehran and IAUT (Islamic Azad University), TEHRAN, IRAN in 1994, 2002 and 2011 respectively. As an assistant professor of computer engineering at Islamic Azad University, Tabriz branch, IRAN, his research interests are mainly in the field of wired and wireless networks, Wireless Sensor Networks (WSNs), Mobile Ad Hoc Networks (MANETs), Vehicular Ad Hoc Networks (VANETs), networks security and Quality of Service (QoS). He has published more than 60 international conference and reviewed journal papers.