p-ISSN 2521-4209
e-ISSN 2521-4217

جامعة التنمية البشرية
UNIVERSITY OF HUMAN DEVELOPMENT

UHD Journal of Science and Technology

A Scientific periodical issued by University of Human Developement

Vol.1 No.(2) December 2017

2017 2717

www.jst.uhd.edu.iq

Kurdistan Regional Government
University of Human Development

UHD Journal of Science and Technology

A periodic scientific journal issued by University of Human Development

Introduction

UHD Journal of Science and Technology (UHDJST) is a semi-annual journal published by the University of Human Development, Sulaymaniyah, Kurdistan Region, Iraq. UHDJST member of ROAD, e-ISSN: 2521-4217, p-ISSN: 2521-4209 and a member of Crossref, DOI: 10.21928/issn.2521-4217. UHDJST publishes original research in all areas of Science, Engineering, and Technology. UHDJST is a Peer-Reviewed Open Access journal with Creative Commons Attribution Non-Commercial No Derivatives License 4.0 (CC BY-NC-ND 4.0). UHDJST provides immediate, worldwide, barrier-free access to the full text of research articles without requiring a subscription to the journal, and has article processing charge (APC). UHDJST applies the highest standards to everything it does and adopts APA citation/referencing style. UHDJST Section Policy includes three types of publications: Articles, Review Articles, and Letters.

By publishing with us, your research will get the coverage and attention it deserves. Open access and continuous online publication mean your work will be published swiftly, ready to be accessed by anyone, anywhere, at any time. Article Level Metrics allow you to follow the conversations your work has started.

UHDJST publishes works from extensive fields including, but not limited to:

- Pure Science
- Applied Science
- Medicine
- Engineering
- Technology

Scope and Focus

UHD Journal of Science and Technology (UHDJST) publishes original research in all areas of Science and Engineering. UHDJST is a semi-annual journal published by the University of Human Development, Sulaymaniyah, Kurdistan Region, Iraq. We believe that if your research is scientifically valid and technically sound then it deserves to be published and made accessible to the research community. UHDJST aims to provide a service to the international scientific community enhancing swap space to share, promote and disseminate the academic scientific production from research applied to Science, Engineering, and Technology.

SEARCHING FOR PLAGIARISM

We use plagiarism detection: detection; According to Oxford online dictionary, Plagiarism means: *The practice of taking someone else's work or ideas and passing them off as one's own.*

Section Policies

No.	Title	Peer Reviewed	Indexed	Open Submission
1	Articles: This is the main type of publication that UHDJST will produce	☑	☑	☑
2	Review Articles: Critical, constructive analysis of the literature in a specific field through summary, classification, analysis, comparison.	☑	☑	☑
3	Letters: Short reports of original research focused on an outstanding finding whose importance means that it will be of interest to scientists in other fields.	☑	☑	☑

PEER REVIEW POLICIES

At UHDJST we are committed to prompt quality scientific work with local and global impacts. To maintain a high-quality publication, all submissions undergo a rigorous review process. Characteristics of the peer review process are as follows:

- The journal peer review process is a "double-blind peer review".
- Simultaneous submissions of the same manuscript to different journals will not be tolerated.
- Manuscripts with contents outside the scope will not be considered for review.
- Papers will be refereed by at least 2 experts as suggested by the editorial board.
- In addition, Editors will have the option of seeking additional reviews when needed. Authors will be informed when Editors decide further review is required.
- All publication decisions are made by the journal's Editors-in-Chief on the basis of the referees' reports. Authors of papers that are not accepted are notified promptly.
- All submitted manuscripts are treated as confidential documents. We expect our Board of Reviewing Editors, Associate Editors and reviewers to treat manuscripts as confidential material as well.
- Editors, Associate Editors, and reviewers involved in the review process should disclose conflicts of interest resulting from direct competitive, collaborative, or other relationships with any of the authors, and remove oneself from cases in which such conflicts preclude an objective evaluation. Privileged information or ideas that are obtained through peer review must not be used for competitive gain.
- Our peer review process is confidential and the identities of reviewers cannot be revealed.

Note: UHDJST is a member of CrossRef and CrossRef services, e.g., CrossCheck. All manuscripts submitted will be checked for plagiarism (copying text or results from other sources) and self-plagiarism (duplicating substantial parts of authors' own published work without giving the appropriate references) using the CrossCheck database. Plagiarism is not tolerated.

For more information about CrossCheck/iThenticate, please visit
http://www.crossref.org/crosscheck.html.

OPEN ACCESS POLICY

This journal provides immediate open access to its content on the principle that making research freely available to the public supports a greater global exchange of knowledge. Open Access (OA) stands for unrestricted access and unrestricted reuse which means making research publications freely available online. It access ensures that your work reaches the widest possible audience and that your fellow researchers can use and share it easily. The mission of the UHDJST is to improve the culture of scientific publications by supporting bright minds in science and public engagement.

UHDJST's open access articles are published under a Creative Commons Attribution CC-BY-NC-ND 4.0 license. This license lets you retain copyright and others may not use the material for commercial purposes. Commercial use is one primarily intended for commercial advantage or monetary compensation. If others remix, transform or build upon the material, they may not distribute the modified material. The main output of research, in general, is new ideas and knowledge, which the UHDJST peer-review policy allows publishing as high-quality, peer-reviewed research articles. The UHDJST believes that maximizing the distribution of these publications - by providing free, online access - is the most effective way of ensuring that the research we fund can be accessed, read and built upon. In turn, this will foster a richer research culture and cultivate good research ethics as well. The UHDJST, therefore, supports unrestricted access to the published materials on its main website as a fundamental part of its mission and a global academic community benefit to be encouraged wherever possible.

Specifically:

- The University of Human Development supports the principles and objectives of Open Access and Open Science
- UHDJST expects authors of research papers, and manuscripts to maximize the opportunities to make their results available for free access on its final peer-reviewed paper
- All manuscript will be made open access online soon after final stage peer-review finalized.
- This policy will be effective from 17th May 2017 and will be reviewed during the first year of operation.
- Open Access route is available at http://journals.uhd.edu.iq/index.php/uhdjst for publishing and archiving all accepted papers,
- Specific details of how authors of research articles are required to comply with this policy can be found in the Guide to Authors.

ARCHIVING

This journal utilizes the LOCKSS and CLOCKSS systems to create a distributed archiving system among participating libraries and permits those libraries to create permanent archives of the journal for purposes of preservation and restoration.

LOCKSS: Open Journal Systems supports the LOCKSS (Lots of Copies Keep Stuff Safe) system to ensure a secure and permanent archive for the journal. LOCKSS is open source software developed at Stanford University Library that enables libraries to preserve selected web journals by regularly polling registered journal websites for newly published content and archiving it. Each archive is continually validated against other library caches, and if the content is found to be corrupted or lost, the other caches or the journal is used to restore it.

CLOCKSS: Open Journal Systems also supports the CLOCKSS (Controlled Lots of Copies Keep Stuff Safe) system to ensure a secure and permanent archive for the journal. CLOCKSS is based upon the open-source LOCKSS software developed at Stanford University Library that enables libraries to preserve selected web journals by regularly polling registered journal websites for newly published content and archiving it. Each archive is continually validated against other library caches, and if the content is found to be corrupted or lost, the other caches or the journal is used to restore it.

PUBLICATION ETHICS

Publication Ethics and Publication Malpractice Statement

The publication of an article in the peer-reviewed journal UHDJST is to support the standard and respected knowledge transfer network. Our publication ethics and publication malpractice statement is mainly based on the Code of Conduct and Best-Practice Guidelines for Journal Editors (Committee on Publication Ethics, 2011) that includes;

- General duties and responsibilities of editors.
- Relations with readers.
- Relations with the authors.
- Relations with editors.
- Relations with editorial board members.
- Relations with journal owners and publishers.
- Editorial and peer review processes.
- Protecting individual data.
- Encouraging ethical research (e.g. research involving humans or animals).
- Dealing with possible misconduct.
- Ensuring the integrity of the academic record.
- Intellectual property.
- Encouraging debate.
- Complaints.
- Conflicts of interest.

ANIMAL RESEARCHES

- For research conducted on regulated animals (which includes all live vertebrates and/or higher invertebrates), appropriate approval must have been obtained according to either international or local laws and regulations. Before conducting the research, approval must have been obtained from the relevant body (in most cases an Institutional Review Board, or Ethics Committee). The authors must provide an ethics statement as part of their Methods section detailing full information as to their approval (including the name of the granting organization, and the approval reference numbers). If an approval reference number is not provided, written approval must be provided as a confidential supplemental information file. Research on non-human primates is subject to specific guidelines from the Weather all (2006) report (The Use of Non-Human Primates in Research).
- For research conducted on non-regulated animals, a statement should be made as to why ethical approval was not required.
- Experimental animals should have been handled according to the highest standards dictated by the author's institution.
- We strongly encourage all authors to comply with the '*Animal Research: Reporting In Vivo Experiments*' (ARRIVE) guidelines, developed by NC3Rs.
- Articles should be specific in descriptions of the organism(s) used in the study. The description should indicate strain names when known.

ARTICLE PROCESSING CHARGES

UHDJST is an Open Access Journal (OAJ) and has article processing charges (APCs). The published articles can be downloaded freely without a barrier of admission.

Address

University of Human Development, Sulaymaniyah-Kurdistan Region/Iraq
PO Box: Sulaymaniyah 6/0778

Contact

Principal Contact

Dr. Aso Darwesh

Editor-in-Chief

University of Human Development – Sulaymaniyah, Iraq

Phone: +964 770 148 5879

Email: jst@uhd.edu.iq

Support Contact

UHD Technical Support

Phone: +964 770 158 4888

Email: jst@uhd.edu.iq

Contents

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Study the Characteristics of Finite Impulse Response Filter Based on Modified Kaiser Window

Muzhir Shaban Al-Ani

Department of Computer Science, College of Science and Technology, University of Human Development, Sulaimani, Kurdistan Region - Iraq

ABSTRACT

Finite impulse response (FIR) plays an important part between all other types of filters. There are many types windows used to design of FIR filters. Most important types are as follows: Hanning, hamming, rectangular, triangular, Blackman, Kaiser, etc. The characteristics of these filters depend on the number of generated coefficients in addition to the side lobes of the filter spectrum. The aim of this work is to study and evaluate Kaiser Window type depends on the variation of its factors applied for resizing the impulse response to reach a suitable size the filter. Kaiser Window is an important filter window that can be used to get many types of windows depending on their parameters. The proposed filter approach is designed and implemented through mixing of many filter factors. The filter characteristics are achieved using different values of filter size and attenuation. The implementation of the proposed Kaiser filter window provides an adequate and easy way to measure the window coefficients and maximum side lobe levels. The benefit of Kaiser Window that you can generate many types of window depending on the parameters change.

Index Terms: Filter Coefficients, Finite Impulse Response Filter, Kaiser Filter, Sidelobes

1. INTRODUCTION

There are two main categories of filters; infinite impulse response (IIR) and finite impulse response (FIR) [1]. These two categories of filters can be implemented for one-dimensional signals and two-dimensional (2D) signals, these filters can be implemented through software and hardware algorithms [2]. In addition, these can be implemented through time domain or frequency domain. Many of filter algorithms are implemented by hybridizing both software and hardware to achieve flexibility and real-time processing [3]. According to the characteristics of the impulse response, filters can be divided into four types; low-pass filter (LPF) that passes low frequencies, high-pass filter that passes high frequencies, band-pass filter that passes a band of frequencies, and band-stop filter that stop a band of frequencies [4]. The implementation of filter is realized through the convolution process between the input signal and the impulse response. To generate a FIR filter, the coefficients of time domain filter must be limited in number by multiplying by a window function of a finite width [5]. The basic idea of applying windows through the design of filters is to truncate the sequence of the filter to be limited to certain values. Various types of windows are used for resizing the signal into limited values [6].

Many algorithms are implemented considering the factors of FIR filters to frequency domain specifications [7]. A narrow band LPF with good filter gain is introduced [8].

Access this article online

DOI: 10.21928/uhdjst.v1n2y2017.pp1-6

E-ISSN: 2521-4217
P-ISSN: 2521-4209

Corresponding author's e-mail: muzhir.al-ani@uhd.edu.iq

Received: 14-05-2017 Accepted: 19-06-2017 Published: 29-08-2017

A hardware digital filter is implemented based on minimizing of computational expenses [9]. An adaptive analog filter is introduced based on echo cancelation [10]. An efficient pipelined filter is implemented using many structures [11]. A wideband efficient linear phase filter is implemented [12]. A high quality frequency domain applied in analog speech scrambler is reconstructed [13]. An efficient filter design based on FIR structure is implemented [14]. An optimized filter design technique is achieved [15]. A filter bank structure based on low complexity is implemented [16]. A programmable architecture applied programmable logic device (PLD) and FFT [17], then introduced PLD and digital signal processor [18]. Fast algorithm of filter design based on LUT is performed [19], then for 2D applications [20]. A real time digital filter processing based on hardware processor is realized [21].

2. RELATED WORK

Several researches have touched on this subject, the following are number of published works:

Mohindru *et al.* designed a new mathematical model to find the transfer function of LPF FIR based on Fourier transform and rotation angle. In contrast with the variation of the angle of rotation from 0 to $\pi/2$, then it is possible to find the width of the transition band and the attenuation of the band. By raising the length of the filter, you can find the frequency response of the digital filter. The response characteristics of FIR filter can be organized through the operation, by controlling only one parameter and remaining other coefficients fixed [22].

Kumar *et al.* designed a novel approach for the structure of filters with variable band of finite linear impulse response band using different polynomials. In this work, the transfer function of a variable bandwidth filter, which is a linear combination of the linear phase filters with fixed coefficients and previous polynomials, is operated separately as control parameters to control the bandwidth of the filter. The proposed method also provides a better performance of tunable BPFs with larger filter captures compared to previously published results [23].

Pak *et al.* proposed a new approach to manage the horizon size of the nonlinear FIR filter. The proposed approach is to carry out an estimation of the state through a FIR filter bank. In this approach, the state estimate is obtained by weighting the average of several estimates of a bank of impulse responses using different horizon sizes. The filter size used for this approach is selected to maximize the likelihood function. The simulations of this approach gives good results compared with the conventional approach [24].

Leighton P. Barnes, George C. Verghese, studied the relationship between the Wiener filter and the coherence function, then defines the causal relationship between the wide-sense stationary (WSS) process. This causal consistency is interpreted in a modeling context and is used to show that a measure depends on the frequency of causality it can and cannot represent. In addition, the Wiener causal convergence of FIR filters with the Wiener causality of IIR filter is studied because the length of the filter passes the infinity. The main results show Lp convergence frequency responses under certain conditions of continuity support in the power spectra, and the upper asymptotic limits for the convergence error. Then, under the same conditions, the uniform convergence of AR approximations shows power spectra as the order model tends to infinity [25].

Huang *et al.* presented a filter design in the closed form based on the compensation of transfer characteristics. In this approach, a new filter design based on a convolution of window is presented and the relationship between window spectrum and filter performance is developed. Then, a schematic design of three step filter is designed these are designing an irregular filter, designing a compensation filter, and the sum of filter. This scheme can be simplified into a closed form characterized by two analytical formulas by merging the intermediate steps [26].

Pak *et al.* proposed an efficient nonlinear FIR. The proposed least square extended FIR filter is derived using a least squares criterion and one through the property. This approach is a special filter designed for the constant speed motion model and does not require noise information such as Gaussian noise covariance. If the noise information is very uncertain, this approach can provide a constant performance, whereas the non-linear estimators of the existing state, such as the extended Kalman filter and particle filter, degradation of performance often in the same condition. The simulations results indicated the robustness of this approach against the uncertainty of the noise model [27].

Boukharouba implemented a new technique for the FIR filters, where the desired frequency response is a smooth rectangular function. In this approach, go directly to smooth out the ideal desired response in the frequency domain not in time domain. The impulse response of the filter is a sampled

version of the inverse Fourier transform of the frequency response. This approach achieved the best performance results of filter specifications compared with the results of other works [28].

3. KAISER WINDOW DESIGN

Kaiser Window have many parameters that affect the overall filter design. The ripple parameter α enabling the designer to trade-off the transition and ripple. It is defined in the interval – M ≤ n ≤ M, and otherwise it is zero. The equations of the Kaiser Window and their parameters are shown below:

Measuring the ripple (δ) factor and then measure the attenuation factor (A) as the following:

$$A = -20 \log_{10} \delta \quad (1)$$

Measure the value of α according to the value of attenuation (A) factor as the following:

$$\propto = 0.1102 (A-8.7) \; A > 50$$

$$\alpha = 0.5842 (A-21)^{0.4} + 0.07886 (A-21) \; 21 \le A \le 50 \quad (2)$$

$$\alpha = 0 \; A < 21$$

Find the window size (M) according to the calculated attenuation value (A) and the transition width (TW) as below:

$$M \ge \frac{(A-7.95)}{28.72*\mathrm{TW}} \quad (3)$$

Kaiser Window is calculated using the following equation:

$$w(n) = \frac{I_0(\alpha\sqrt{1-(n/M)^2}}{I_0(\alpha)} \quad -M \le n \le M \quad (4)$$

And the zero order Bessel function I_0 (α) is measured as below:

$$I_0(x) = 1 + \sum_{n=1}^{\infty}\left[\left(\frac{x}{2}\right)^n \frac{1}{n!}\right] \quad (5)$$

4. PROPOSED FIR FILTER ALGORITHM

There are many parameters affected the filter design, these parameters are ripple factor, attenuation factor, TW, number of coefficients, and window type in addition to the required accuracy.

The length of the filter concentrated on the number of values of impulse response samples in the FIR filter, then the impulse response is varied from $n = 0$ to $n = M-1$, where M is the filter length.

The proposed filter approach is demonstrated in Fig. 1 and is shown below:
Step 1: Ripple (δ) measurement, considering pass band ripple and stop band ripple.
Step 2: Attenuation measurement depending on the ripple value.
Step 3: TW measurement depending on the number of coefficients.
Step 4: Filter length measurement based on the TW.
Step 5: Bessel function $I_0(x)$ measurement depending on the calculated parameters.
Step 6: Kaiser Window elements measurement.
Step 7: Filter coefficients measurement depending on the given cutoff frequency.
Step 8: Filter coefficients truncation based on the required Kaiser Window.
Step 9: Realize the filter according to the measured parameters.

5. RESULTS AND DISCUSSIONS

To realize and study the FIR filter characteristics, it is better to generate different types of windows. These windows are; Barthann, Bartlett, Blackman, Blackman-Harris, Bohman, Chebyshev, Flattop, Gaussian, Hamming, Hanning, Nuttall, Parzen, Rectangular, Taylor, Triangular, Tukey, and Kaiser. To study the shape characteristics of these windows, these are generated using different length number of window. Fig. 2 shows these windows in time domain and frequency domain generated using 20 values. Fig. 3 shows these windows in time domain and frequency domain generated using 50 values. These two figures illustrated that the time domain shape varying from rectangular shape to different type of Sinc function shape or hat shape. In addition, these two

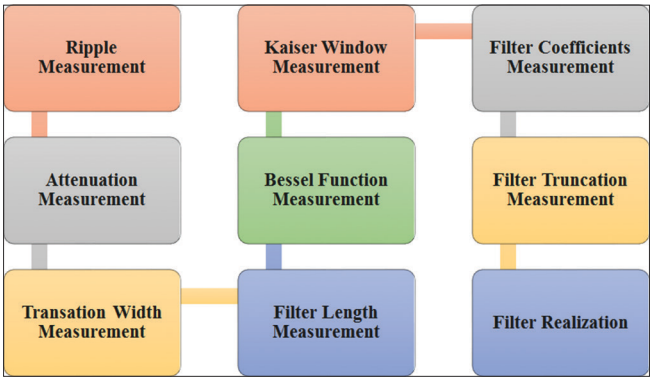

Fig. 1. Proposed finite impulse response filter algorithm

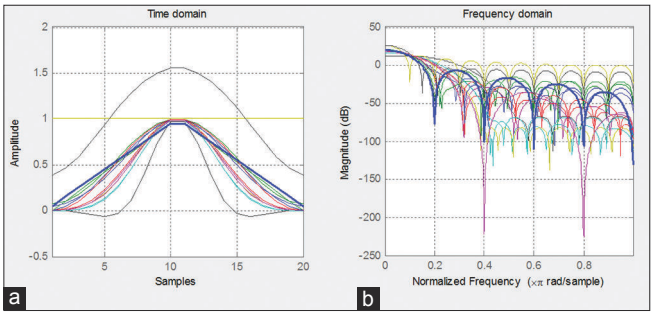

Fig. 2. (a and b) Time domain and frequency domain representation of 20 values

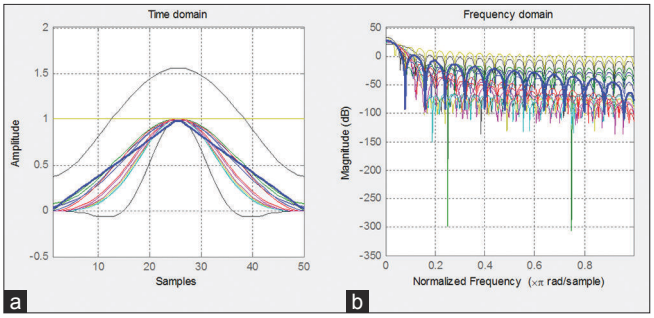

Fig. 3. (a and b) Time domain and frequency domain representation of 50 values

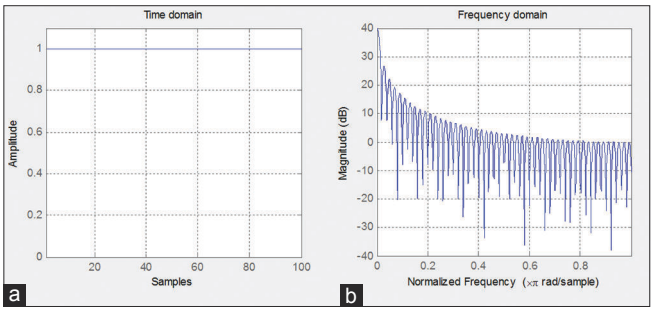

Fig. 4. (a and b) Kaiser characteristics beta = 0

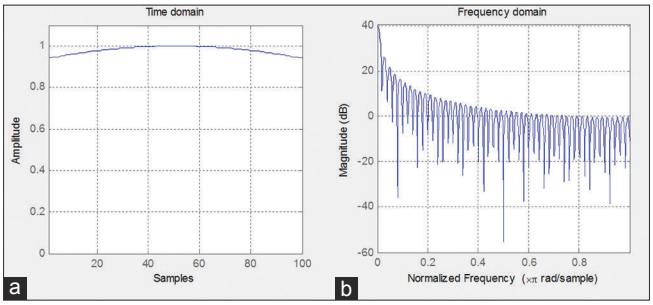

Fig. 5. (a and b) Kaiser characteristics beta = 0.5

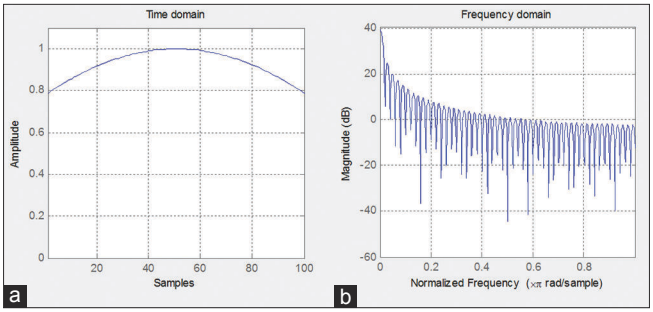

Fig. 6. (a and b) Kaiser characteristics beta = 1

TABLE I
Kaiser Parameters of Low-beta Values

Beta value	Leakage factor (%)	Sidelobe attenuation dB	Main lobe width −3 dB
0	9.26	−13.3	0.017578
0.1	9.22	−13.3	0.017578
0.2	9.13	−13.3	0.017578
0.3	8.97	−13.4	0.017578
0.4	8.75	−13.5	0.017578
0.5	8.48	−13.6	0.017578
0.6	8.16	−13.8	0.017578
0.7	7.80	−14.0	0.017578
0.8	7.40	−14.2	0.017578
0.9	6.98	−14.4	0.017578
1.0	6.53	−14.7	0.017578

figures in frequency domain show that the relative sidelobe concentrated on −26.8 dB where the latest on appears on about −75 dB.

Kaiser Window is an important filter window and it have weighting factor parameter. The value of beta is the parameter of Kaiser Window that affects the sidelobe attenuation in the frequency domain of the window. Fig. 4 indicates the Kaiser characteristics when beta equal to zero in which indicated that the leakage factor equal to 9.26% and the sidelobe attenuation equal to −13.3 dB. Fig. 5 indicates the Kaiser characteristics when beta equal to 0.5 in which indicated that the leakage factor equal to 8.48% and the sidelobe attenuation equal to −13.6 dB. Table I demonstrates the Kaiser parameters of low beta values in which shows the variation of beta value from 0 to 1 with the increment of 0.1 in which the leakage factor varying from 9.26% to 6.53%. In addition, the values of sidelobe attenuation is varied from −13.3 to −14.7 dB, with main lobe width at −3dB equal to 0.017578 for all values of beta.

Fig. 6 indicates the Kaiser characteristics when beta equal to one in which indicated that the leakage factor equal to 6.53% and the sidelobe attenuation equal to −14.7 dB. Fig. 7 indicates the 10 in which indicated that the leakage factor equal to 0% and the sidelobe attenuation equal to −74.1 dB,

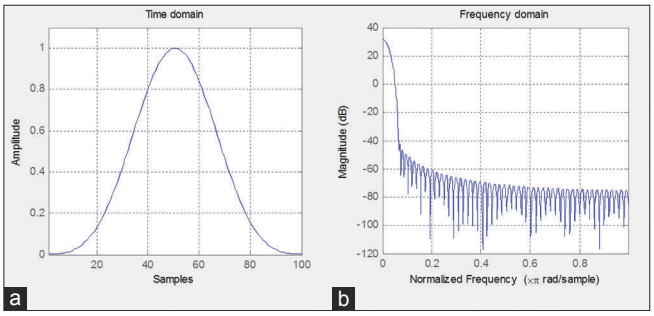

Fig. 7. (a and b) Kaiser characteristics beta = 10

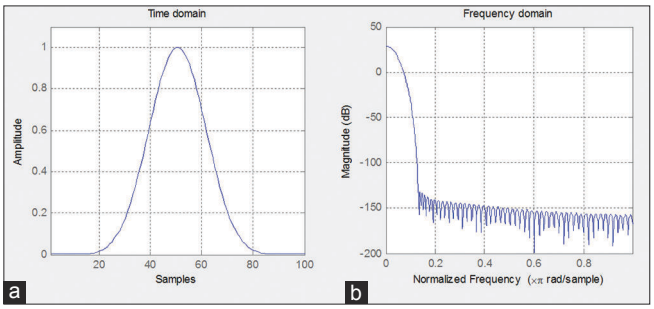

Fig. 8. (a and b) Kaiser characteristics beta = 20

in this case, the shape of the window is approach to a hat. Fig. 8 indicates the Kaiser characteristics when beta equal to 20 in which indicated that the leakage factor equal to 0% and the sidelobe attenuation equal to −154.9 dB, in this case, the shape of the window is stretch hat. Table II demonstrates the Kaiser parameters of high-beta values in which shows the variation of beta value from 1 to 20 with the increment of 1, in this case, the leakage factor varying from 6.53% to 0%, considering that the zero value of leakage factor starting when beta equal to 6. In addition, the values of sidelobe attenuation is varied from −14.7 to −154.9 dB, with main lobe width at −3 dB varying from 0.017578 to 0.048828 corresponding to the variation of the value of beta.

6. CONCLUSION

In this work, an adequate algorithm of Kaiser parameters is measured to be ready for the filter implementation. The framework to study the characteristics of FIR filter based on modified Kaiser Window is implemented through two scenarios. Two scenarios are implemented; First scenario used low Beta values and second scenario used high Beta values. The obtained results indicated that the implemented Kaiser Window filter has good characteristics and stable. The comparison of sidelobe attenuation and the values of beta it is possible to select an adequate and optimal

TABLE II
Kaiser Parameters of High-beta Values

Beta value	Leakage factor (%)	Sidelobe attenuation dB	Main lobe width −3 dB
1	6.53	−14.7	0.017578
2	2.47	−18.6	0.019531
3	0.61	−24.0	0.021484
4	0.12	−30.4	0.023438
5	0.02	−37.1	0.025391
6	0.00	−44.0	0.027344
7	0.00	−51.0	0.029297
8	0.00	−58.4	0.031250
9	0.00	−66.1	0.033203
10	0.00	−74.1	0.035156
11	0.00	−82.3	0.035156
12	0.00	−90.3	0.037109
13	0.00	−98.1	0.039063
14	0.00	−105.8	0.041016
15	0.00	−113.8	0.041016
16	0.00	−122.3	0.042969
17	0.00	−130.4	0.044922
18	0.00	−138.7	0.044922
19	0.00	−147.1	0.046875
20	0.00	−154.9	0.048828

values to be used for the filter design. The obtained results indicated that an improvement of the filter characteristics during the attenuation value will increase. In addition, an improvement happens on the filter performances when filter length will increase, that is increase the number of coefficients. The implemented Kaiser Window indicated that is easy to generate many types of windows through simple modification. The sidelobe attenuation of Kaiser Window is about −14 dB when beta value is 1 and this value will increase vastly up to −154.9 when beta reach to 20.

REFERENCES

[1] M. S. M. Al-Ani and S. Lorenzo. "System implementation of LUT FFT." 6th Mediterranean Electrotechnical Conference (MELECON'91), Yugoslavia, 22-24 May. 1991.

[2] L. Nozal, S. Lorenzo, B. Rui and M. S. M. Al-Ani. "Real time and low cost image processing architecture based on programmable logic device (PLD)." Intelligent for Mechanical System, Proceedings IROS '91, Osaka, Japan, pp. 279-284, 3-5 Nov. 1991.

[3] M. S. M. Al-Ani, S. Lorenzo, L. Nozal and B. Rui. "FFT LUT for image processing." The International Society for Optical Engineering (SPIE'92), Section of Algorithms Technique and Active Vision, USA, pp. 121-129, 15-20 Nov. 1992.

[4] G. I. Raho, A. J. Dawood and M. S. Al-Ani. "Real time fast algorithm of 2D DWT based DSP technology." *International Journal of Application or Innovation in Engineering and Management*, vol. 2, no. 10, pp. 1-5, Oct. 2013.

[5] L. Nozal, S. Lorenzo, B. Rui and M. S. M. Al-Ani. "Hardware Structure + digital signal processing on real time." International Conference on Industrial Electronics, Control, Instrumentation and

Automation, San Diego, California, USA, pp. 1397-1402, 9-13 Nov. 1992.

[6] Q. Al-Shayea and M. S. Al-Ani. "Efficient window approach of FIR filter design (MSK2)." *IJCSNS International Journal of Computer Science and Network Security*, vol. 16, no. 2, pp. 63-68. Feb. 2016.

[7] Oliveira, H. Jr., H. Petraglia and A. Petraglia. "Frequency domain FIR filter design using fuzzy adapting simulated annealing." *Circuits, Systems, Signal Process*, vol. 28, pp. 899-911, 2009.

[8] A. A. Lanne, T. V. Merkucheva and A. I. Solonina. "Calculation of narrowband low pass filters with finite impulse response." *Radioelectronic and Communication Systems*, vol. 52, no. 6, pp. 311-316, 2009.

[9] A. A. Lanne and T. V. Merkucheva. "Filter with double symmetry." *Radioelectronic and Communication Systems*, vol. 52, no. 5, pp. 256-260, 2009.

[10] H. S. Yazdi and M. Rezaei. "The wheatstone bridge-based analog adaptive filter with application in echo cancellation." Analog Integral Circuits and Signal Processing, 16 Sep. 2009.

[11] A. Blad and O. Gustafsson. "Integer linear programming-based bit-level optimization for high speed FIR decimation filter architectures." *Circuits, Systems, Signal Process*, vol. 29, pp. 81-101, 2010.

[12] R. Lehto, T. Saramäki and O. Vainio. "Synthesis od wideband linear-phase FIR filters with a piecewise-polynomial-sinusoidal impulse response." *Circuits, Systems, Signal Process*, vol. 29, pp. 25-50, 2010.

[13] Y. C. Lim, J. W. Lee and S. W. Foo. "Quality analog scramblers using frequency response masking filter banks." *Circuits, Systems, Signal Process*, vol. 29, pp. 135-154, 2010.

[14] Y. Wei and Y. Lian. "Frequency response masking filters based on serial masking schemes." *Circuits, Systems, Signal Process*, vol. 29, pp. 7-24, 2010.

[15] Y. J. Yu and Y. C. Lim. "Optimization of linear phase FIR filters in dynamically expanding subexpression space." *Circuits, Systems, Signal Process*, vol. 29, pp. 65-80, 2010.

[16] L. Rosenbaum, P. Löwenborg and H. Johansson. "Two classes of cosine modulated IIR/IIR and IIR/FIR NPR filter banks." *Circuits, Systems, Signal Process*, vol. 29, pp. 103-133, 2010.

[17] L. Nozal, S. Lorenzo, B. Rui and M. S. M. Al-Ani. "Pipe-line programmable logic device (PLD) a new solution for image processing." SICE'91 Conference, Yonezawa, Japan, pp. 1097-1100, 17-19 Jul. 1991.

[18] L. Nozal, S. Lorenzo, B. Rui and M. S. M. Al-Ani. "A new vision system programable logic devices digital signal processor architecture (PLD+DSP)." International Conference on Industrial Electronics Control and Instrumentation (IECON'91), Japan, pp. 2014-2018, 28 Oct. 1 Nov. 1991.

[19] M. S. M. Al-Ani, S. Lorenzo and L. Nozal. "Fast 2D convolution filter based on LUT FFT." IEEE International Symposium on Industrial Electronics, Chain, pp. 446-449, 25-27 May. 1992.

[20] M. S. M. Al-Ani, S. Lorenzo, L. Nozal and B. Rui. "Fast image filtering implementation." Second International Conference on Document Analysis and Recognition (ICDAR'93), In: Cooperation with the IEEE Computer Society and IGS, Tsukuba Science City, Japan. 20-22 Oct. 1993.

[21] B. Rui, S. Lorenzo, L. Nozal and M. S. M. Al-Ani. "Digital signal processor accelerator board for image processing on VME bus based system." Machine Vision Application, Architectures, and System Integration, Boston, Manssachusetts, USA, pp. 85-89, 17-18 Nov. 1992.

[22] P. Mohindru, R. Khanna and S. S. Bhatia. "New tuning model for rectangular windowed FIR filter using fractional Fourier transform", London: SIViP, Springer-Verlag, 2013.

[23] A. Kumar, S. Sumana and G. K. Singh. "A new closed form method for design of variable bandwidth linearphase FIR filter using different polynomials." *AEU-International Journal of Electronics and Communications*, vol. 68, pp. 351-360, 2014.

[24] J. M. Pak, S. Y. Yoo, M. T. Lim and M. K. Song. "Weighted average extended FIR filter bank to manage the horizon size in nonlinear FIR filtering." *International Journal of Control, Automation, and Systems*, vol. 13, no. 1, pp. 1-8, 2015.

[25] P. L. Barnes and G. C. Verghese. "Uniform FIR approximation of causal wiener filters, with applications to causal coherence." *Signal Processing*, vol. 122, pp. 129-137, 2016.

[26] X. Huang, Y. Wang, Z. Yan, H. Xian and M. Liu. "Closed-form FIR filter design with accurately controllable cut-off frequency, circuits syst signal process," New York: Springer Science and Business Media, 2016. Available: https://link.springer.com/article/10.1007/s00034-016-0330-7. [Apr. 27, 2016].

[27] J. M. Pak, P. S. Kim, S. H. You, S. S. Lee and M. K. Song. "Extended least square unbiased FIR filter for target tracking using the constant velocity motion model. *International Journal of Control, Automation and Systems*, vol. 15, no. X, pp. 1-5, 2017.

[28] A. Boukharouba. "Smoothed rectangular function-based FIR filter design, circuits system signal process," New York: Springer Science and Business Media, 2017. Available: https://link.springer.com/article/10.1007/s00034-017-0529-2. [Feb. 21, 2017].

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Transmission Control Protocol Global Synchronization Problem in Wide Area Monitoring and Control Systems

Yahya Ahmed Yahya[1] and Ahmad T. Al-Hammouri[2]

[1]Department of Information Technology, Zakho Technical Institute, Duhok Polytechnic University Zakho, Duhok, Kurdistan Region, Iraq,
[2]Department of Network Engineering and Security, Jordan University of Science and Technology, Irbid, Jordan

ABSTRACT

The electrical power network is a significant element of the critical infrastructure in modern society. Nowadays, wide area monitoring and control systems (WAMC) are becoming increasingly an important topic that motivates several researchers to improve, develop, and find the problems that hinder progress toward WAMC systems. WAMC is used to monitor and control the power network so the power network can be adapt to failures in automatic way. In this work, verification of the extent found a problem in transmission control protocol (TCP) which is called global synchronization and its impact on utilizing the buffer of the routers. A simulation models had been belt of WAMC system using OMNeT++ to study the performance of TCP in two queuing algorithms for measuring transmission of phasor measurement units and to test if global synchronization problem occurs. Three scenarios were used to test the survival of this problem on the system. It is found that the problem of global synchronization occurred in two scenarios which in turn causes low utilization for a buffer of routers.

Index Terms: Global Synchronization, Phasor Measurement Units, Power Network, Transmission Control Protocol, Wide Area Monitoring and Control Systems

1. INTRODUCTION

Electric power grid is one of the most important topics in the modern societies. The sharp increase of demand on electricity, the electricity trade between the neighboring countries, and the long distances to transport electricity motivates the researchers and industries to propose and improve systems to monitor and control the electric power grid over a wide area. Hence, wide area monitoring and control systems (WAMC) became an important topic.

Access this article online

DOI: 10.21928/uhdjst.v1n2y2017.pp7-12

E-ISSN: 2521-4217
P-ISSN: 2521-4209

The communication network must transmit measurements with low latency and with high accuracy [1]. There are many factors affecting these two requirements, for example, the bandwidth, the type of medium, and the protocol that are used for transmission.

Phasor measurement unit (PMU) measurements can be transmitted over different types of transmission media such as wired or wireless. However, the best medium that can be used is the fiber-optic cable [1], [2]. The reason to choose fiber optic is the advantages including high data transfer rates, immunity to electromagnetic interference, and very large channel capacity [3].

There are many protocols used with WAMC system including the user datagram protocol (UDP), multiprotocol label switching (MPLS), resource reservation protocol (RSVP),

Corresponding author's e-mail: yahya.ahmed@dpu.edu.krd

Received: 10-03-2017 Accepted: 25-03-2017 Published: 29-08-2017

and synchronous digital hierarchy. These protocols can be used individually or in combination (more than one of these protocols work with each other as one protocol), for example, using UDP with MPLS and RSVP as a one main protocol furnishes quality of services features.

Surveying the literature indicates that many protocols were used with WAMC systems except TCP. However, the NASPInet standard [4], [5] mentions that TCP can be used as the transport layer protocol to deliver PMU measurements. In general, architecture that is used in most WAMC systems is shown in Fig. 1.

The aim of this work is to study the effect of the global synchronization problem when we are using TCP with WAMC systems.

2. TCP GLOBAL SYNCHRONIZATION PROBLEM

TCP is one of more than a few transport protocols [6], [7]. It is reliable and stream oriented. In addition, TCP is connection-oriented, meaning it establishes a connection when data need to be transferred. The data are sent in packets and in an ordered manner at transport layer. It supports flow and congestion control [7].

Researchers after a deep study found a problem in TCP called global synchronization that can be defined as the pattern of each sender decreasing and increasing transmission rates at the same time as other senders [6], [7], [8].

As shown in Fig. 2, we built a topology to study the problem of global synchronization using OMNeT++. The topology consists of the two senders, two receivers, and two routers. Each sender connected to one of two routers by bandwidth link equal to 100 Mbps and propagation delay equal to 3 ms. In the other hand, each receiver connects to the other router by bandwidth link equal to 100 Mbps and propagation delay equal to 3 ms. The two routers were connected by a bandwidth link equal to 10 Mbps and propagation delay equal to 1 ms. The buffer in the two routers was drop tail with 65-packet capacity.

After running the experiment, the result of congestion window (cwnd) is as shown in Fig. 3. The synchronization in both cwnd flows appeared clearly because both cwnds have to deal with the queue reaching its limits at the same time.

The low utilization of buffer seen can be clearer in Fig. 4. The queue length is switching between full (65 packets) to nearly (5 packets). This oscillating load is caused by the synchronization of cwnd. This problem motivates the researchers, and they proposed many methods to solve this problem. Some researchers suggest adjusting router functionalities or modification in the transmission protocol parameters, and others suggested increasing the size of the router buffer capacity. However, these solutions will cause another unexpected problems because when increasing the size of router buffer, it is likely to increase queuing delays. The most adopted algorithm to reduce the global synchronization is random early drop queuing policy.

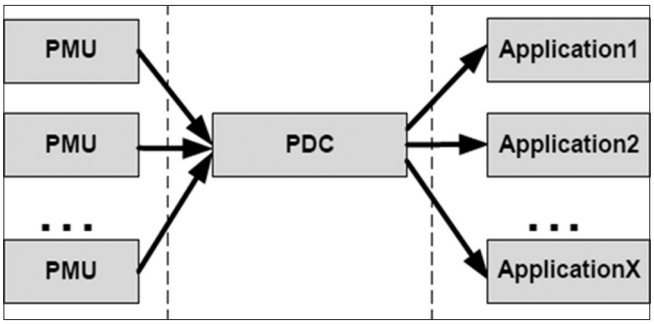

Fig. 1. General wide area monitoring and control Architecture [4]

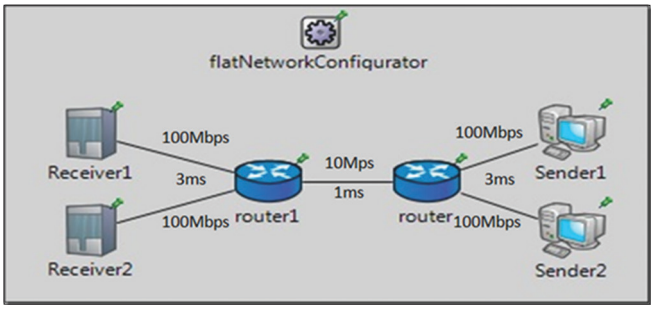

Fig. 2. Simple topology to test global synchronization

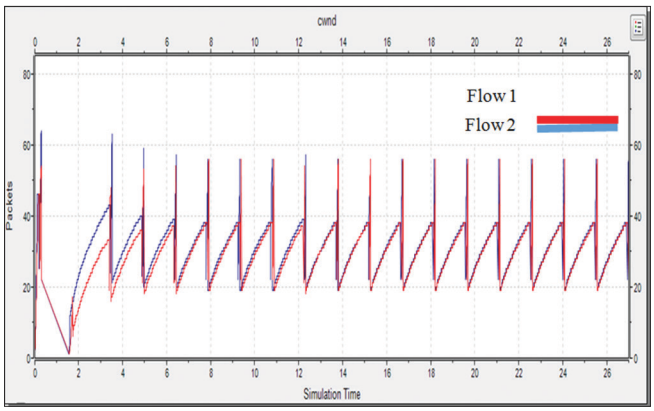

Fig. 3. Congestion windows of two senders

It is liable in WAMC that several senders or PMUs can start sending in the same time. Hence, this study is to show how the global synchronization problem effects on WAMC.

3. DESIGN OF A WAMC SYSTEM IN OMNET++

A. PMU Building

This was to construct the first component of WAMC which is the PMU using C++ language. The PMU generates packets in a certain period of time, and each packet contains three information's:

1. Sequence number of the packet: Which can be arranged systematically by giving a specific number to each packet.
2. Source ID: It is the number that helps to know from which PMU the packet had been generated.
3. Time stamp: It is the time of when the packet is generated until it reaches to PDC.

B. PDC Level 1 (L1) Building

This was to construct the second component of WAMC, which is the PDC using the C++ language. The PDC is similar to the server of the computer, as its function is to collect all the measurements sent to it from the PMUs that are connected with the PDC.

The PDC function is to assemble all the packets that were produced from PMUs in one packet and to send it to PDC level 2.

For more details about PDC L1 flowchart, PDC L1 checks each packet after arriving from PMU "if this sequence number sent before that time," when "Yes" PDC will delete this packet, and if "No" PDC check whether "if this sequence number is a first time seen." When "Yes," PDC will add a time-out to this sequence number of the packet.

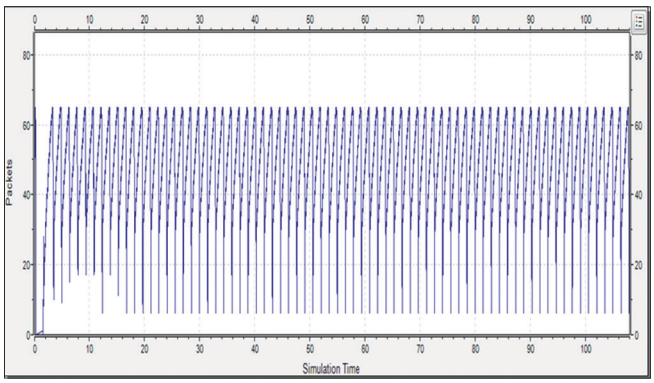

Fig. 4. The queue buffer of router

The benefit of this time-out is that when the timer fires PDC does not receive all packets from all PMUs, and then PDC will neglect and send only the arrived packets. After adding time-out to a packet with a sequence number, PDC will store the packet in the buffer and wait for packets that have the same sequence number. When the answer is "No," "if this sequence number is first time seen," PDC will store this packet in the buffer and checks whether that "all packets for one sequence number received or their time-out finish." When "No," wait until it is Yes.

When "Yes," PDC encapsulates the packets that have this sequence number in one packet that has a new sequence number, new source ID, and a new time stamp, then it will be sent to PDC L2.

C. PDC Level 2 (L2) Building

This was to construct the third component of WAMC, which is the PDC using the C++ language. This PDC will receive the above mentioned single packet (produced in part II), then it will calculate the time of end-to-end latency and packet delivery ratio.

4. SIMULATIONS AND RESULTS

In this section, three scenarios were created as follows:

1. Twenty PMUs Scenario (high traffic rate)
2. Nineteen PMUs scenario (medium traffic rate)
3. Eighteen PMUs scenario (low traffic rate).

Through the study of the three scenarios, it was noticed in some cases of using the TCP protocol, the global synchronization problem occurred. Therefore, the study focuses on the buffer of the router, which may cause low utilization to the link bandwidth.

A. Twenty PMUs Scenario (High Traffic Rate)

The WAMC systems topology in this scenario has 20 PMUs, one PDC L1, and one PDC L2. Fig. 5 shows the topology of this scenario. There were two routers in the topology. The size of buffer capacity in each router is 100 packets. The links parameters that were used in all scenarios were listed in Table I. The packet size that sends from PMU is 512B, and the rate of date packet that generates from PMUs is 120 p/s. A time-out of PDC is 0.024s. Table I shows all scenarios parameters.

The buffer result of router when using TCP with the FIFO queue algorithm is shown in Fig. 6.

TABLE I
Links Parameters of All Scenarios

Link between	Data rate (bps)	Propagation delay (ms)
PMUs and router	150 M	3-6
Router and PDC L1	10 M	1
PDC L1 and router1	150 M	3
Router1 and PDC L2	150 M	3

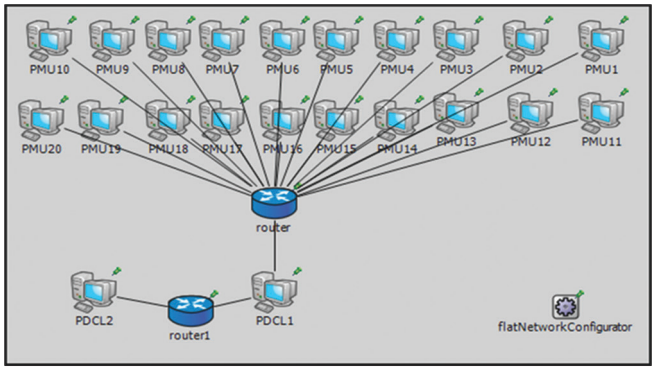

Fig. 5. The topology in 20 phasor measurement units scenario

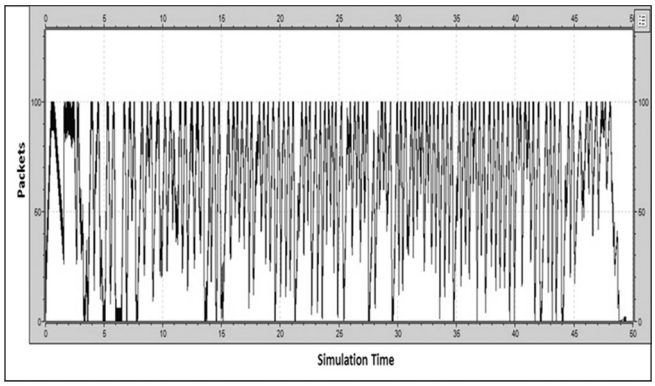

Fig. 6. The buffer of router when using transmission control protocol with the FIFO queue algorithm

The buffer exhibits oscillation between 0 and 100, which means that global synchronization had occurred in this scenario. When we used TCP with RED queue algorithm, the results are shown in Fig. 7.

Now, the buffer exhibits oscillation, but this oscillation did not to exceed 50 packets as before using TCP with FIFO.

B. Nineteen PMUs Scenario (Medium Traffic Rate)

The parameter of this scenario is as mentioned in 20 PMUs scenario. The only difference between the scenarios is the number of PMUs. Fig. 8 shows topology of 19 PMUs scenario.

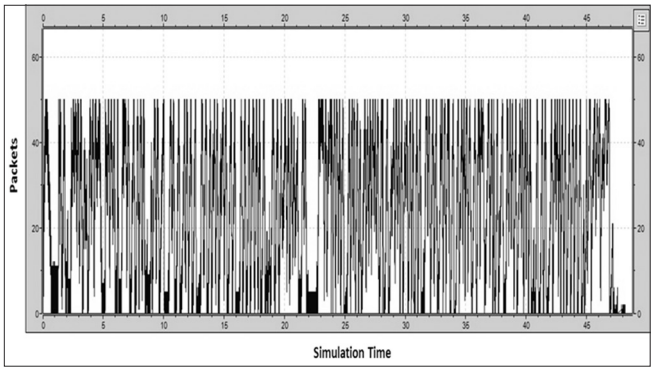

Fig. 7. The buffer of router when using transmission control protocol with the RED queue algorithm

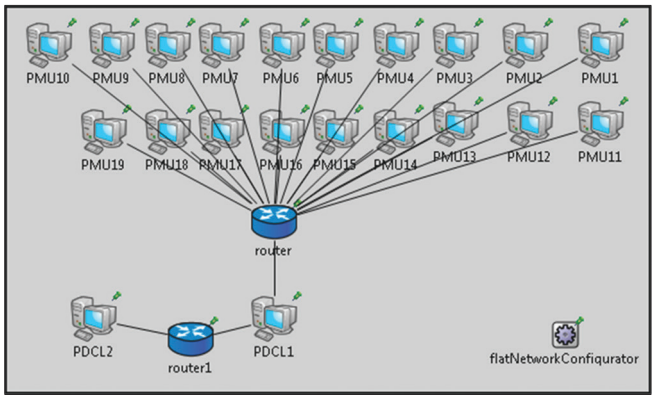

Fig. 8. The topology in 19 phasor measurement units scenario

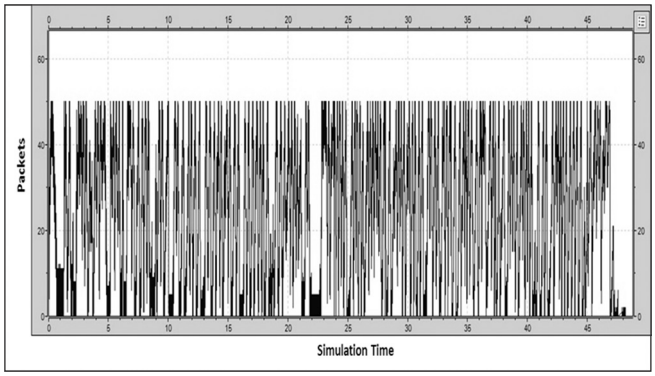

Fig. 9. The buffer of router when using transmission control protocol with the FIFO queue algorithm

For 19 PMUs scenario (TCP protocol and FIFO), the result of buffer is shown in Fig. 9.

The buffer exhibits oscillation between 0 and 100, which means that global synchronization had occurred in this scenario. However, the exhibits oscillation is lower than the scenario of 20 PMUs. When we used TCP with RED, the results are shown in Fig. 10.

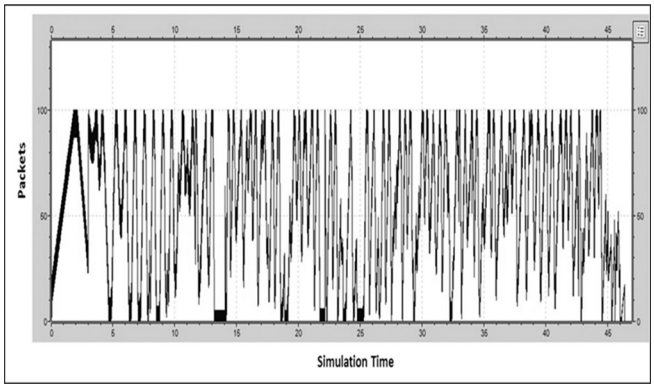

Fig. 10. The buffer of router when using transmission control protocol with the RED queue algorithm

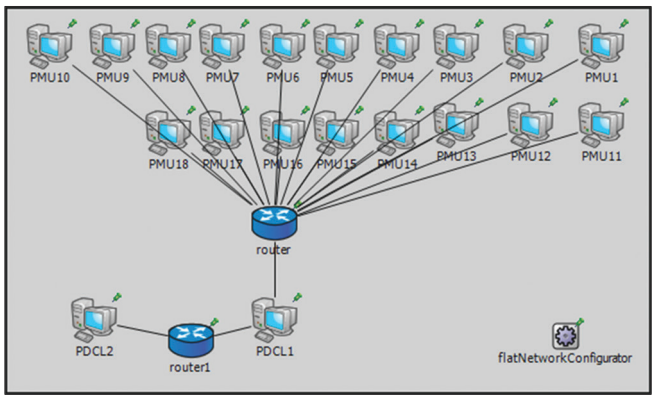

Fig. 11. The topology in 18 phasor measurement units scenario

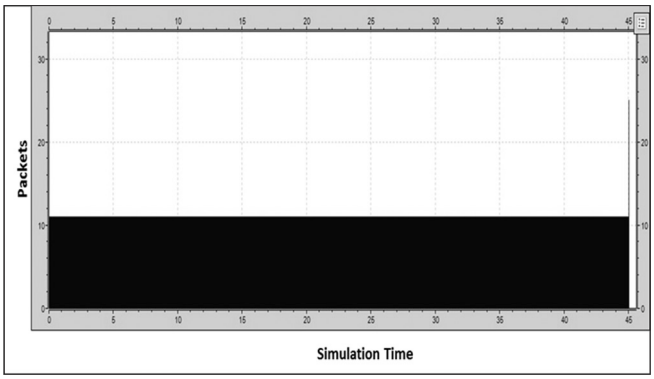

Fig. 12. The buffer of router when we used transmission control protocol with the FIFO queue algorithm.

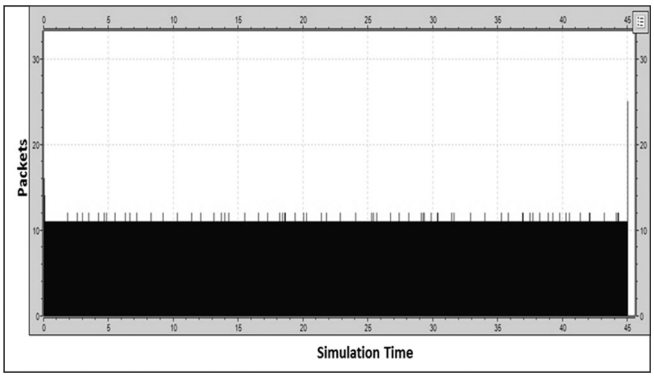

Fig. 13. The buffer of router when we used transmission control protocol with the RED queue algorithm

The buffer exhibits oscillation, but this oscillation did not exceed 50 packets as before using TCP with FIFO. The oscillation in this scenario is the same of scenario 20 PMUs.

C. Eighteen PMUs Scenario (Low Traffic Rate)

The parameter of this scenario is mentioned in 20 and 19 PMUs scenario. The only difference between the scenarios is the number of PMUs. Fig. 11 shows topology of 18 PMUs scenario.

In 18 PMUs scenario, there is no oscillation of packets which means that there is no global synchronization in this scenario. The buffer exhibits oscillation between 0 to 10 or 11 while the buffer size is 100 packets. Figs. 12 and 13 show the buffer of routers when we used TCP with FIFO and RED queue algorithm of router was used.

6. CONCLUSION

In this study, WAMC system using OMNeT++ was created and applied TCP protocol (FIFO and RED) queue algorithm that was used to deliver the measurements and control information over the networks.

The global synchronization problem occurred with the TCP protocol when the PMUs sent the measurements in the synchrony manner in two scenarios (20 and 19 PMUs) in other words when the traffic rate is high and medium. Whereas in the scenario of 18 PMUs when traffic rate is low, the global synchronization was not occurred. According to this study, it is recommended not to used TCP protocol in WAMC systems to do not use TCP protocol in WAMC system during have high rate traffic.

REFERENCES

[1] C. Moustafa and L. Nordström. "Investigation of communication delays and data incompleteness in multi-pmu wide area monitoring and control systems." Electric Power and Energy Conversion Systems, EPECS'09. International Conference on IEEE, 2009.

[2] H. Erich, H. Khurana and T. Yardley. "Exploring convergence for SCADA networks." Innovative Smart Grid Technologies (ISGT),

IEEE PES, IEEE, 2011.

[3] C. Moustafa, K. Zhu and L. Nordstrom. Survey on Priorities and Communication Requirements for PMU-Based Applications in the Nordic Region. PowerTech, 2009.

[4] C. Moustafa, A. Layd and L. Jordan. "PMU traffic shaping in IP-based wide area communication." Critical Infrastructure (CRIS), 2010 5th International Conference on IEEE, 2010.

[5] Y. Yorozu, M. Hirano, K. Oka and Y. Tagawa. "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," *IEEE Translation Journal on Magnetics in Japan*, vol. 2, pp. 740-741, Aug. 1987. (Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982).

[6] Z. Lixia and C. David. "Oscillating behavior of network traffic: A case study simulation". *Internetworking: Research and Experience,* vol. 1, pp. 101-112, 1990.

[7] S. Chakchai. "Loss synchronization of TCP connections at a shared bottleneck link." *Department of Computer Science and Engineering*, St. Louis: Washington University, 2006.

[8] H. Sofiane and R. David "loss synchronization, router buffer sizing and high-speed TCP versions: Adding RED to the mix". IEEE 34th Conference on Local Computer Networks, 2009.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

An Improved Parallel Multiple Sequence Alignment Algorithm on Multi-core System

Mohammed W. Al-Neama[1], Salwa M. Ali[2] and Kasim A. Al-Salem[3]

[1]*Department of Computer Science, Education College for Girls, Mosul University, Mosul, Iraq*
[2]*Department of Mathematics, Faculty of Science, Ain Shams University, Cairo, Egypt*
[3]*Department of Computer Science, University of Cihan/Sulaimanya, Kurdistan Region, Iraq*

ABSTRACT

In this paper, we introduce an improved parallel algorithm for computing the number of exact matches nid (*S*,*T*) in the local alignment of two biological sequences *S* and *T*. This number is used in the first stage of progressive alignment to compute the distance between two sequences. The distance computations are usually its most computationally intensive part. Therefore, this work concentrates on improving an algorithm for this stage using vectorizing technique and running on multi-core. Our program is able to compute *nid* (*S*,*T*) between very long sequences, up to 34 k residues by C++ with OpenMP library on an Intel Core-i7-3770 quad-core processor of 3.40 GHz and main memory of 8 GB. It outperforms ClustalW-MPI 0.13 with 2.9-fold speedup, and the efficiency reached 0.35. Furthermore, a higher speedup with improved efficiency can be accomplished. Its performance figures vary from a low of 0.438 GCUPS to a high of 3.66 GCUPS as the lengths of the query sequences decrease from 34,500 to 9200.

Index Terms: Bioinformatics, Distance Computation, Multi-cores, Multiple Sequence Alignment, Parallel Programming

1. INTRODUCTION

Sequence alignment refers to the search of similarity regions within biological sequences of DNA, RNA, or proteins. Besides the biological significance and interpretation of the results, the main problem of sequence alignment from the computer science point of view is the very large number of residue-to-residue comparisons that are needed when searching for similarities.

The biological definition of the problem makes it time-consuming taking in consideration the fact that in practice, a single genome may contain in the order of 10^9 residues.

Access this article online

DOI: 10.21928/uhdjst.v1n2y2017.pp13-24

E-ISSN: 2521-4217
P-ISSN: 2521-4209

Thus, a computer program can therefore intensively search for regions of similarity between sequences to detect such relationships.

Alignment could be global or local. Global alignment is to find the best match between the entire sequences, whereas local alignment must find the best match between certain regions of the sequences.

Global alignment algorithms such as Needleman–Wunsch (NW) [1], which attempt to align every residue in every sequence, are most useful when the sequences in the query set are similar and of roughly equal size. Local alignments instead, like in Smith–Waterman (SW) [2], are more often used for dissimilar sequences that are suspected to contain regions of similarity within their larger sequence context.

Fig. 1 shows a global and a local alignment of the same two sequences. It shows that if sequences are not sufficiently similar, a global alignment tends to spread gaps that hide possible regions of similarity. Residue mismatches are called

Corresponding author's e-mail: mwneama@uomosul.edu.iq

Received: 10-03-2017 Accepted: 25-03-2017 Published: 29-08-2017

```
    --T--CC-C-AGT--TATGT-CAGGGGACACG--A-GCATGCAGA-GAC
      |  || |  ||  | | | |||     || |  | |  | ||||   |
    AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG--T-CAGAT--C

                        tccCAGTTATGTCAGgggacacgagcatgcagagac
                           ||||||||||||
aattgccgccgtcgttttcagCAGTTATGTCAGatc
```

Fig. 1. A global and a local alignment of the same two sequences

substitutions (mutations in genetic terminology), and the character "-" denotes gaps that may be insertions or deletions depending on the point of view.

A. Substitution Matrix

Biologists have learned that some mutations are more likely than others (e.g., a DNA mutation from cytosine to adenine is more common than cytosine to guanine) and they need alignment algorithms to reflect this property. For this purpose, substitution matrices have been built using statistical data from known sequences and mutations. Fig. 2 shows the BLOSUM62 [3], a common substitution matrix for amino acids alignments.

Definition 1: A substitution matrix (*sbt*) on an alphabet $\Sigma = \{a_1, a_2..., a_n\}$ has $n \times n$ entries, where each entry (i, j) assigns a score for a substitution of the letter a_i by the letter a_j in an alignment.

The elements on the main diagonal have the highest values to encourage matching of identical residues in alignment algorithms. Amino acids are divided into colored groups according to the chemistry of the side group.

B. Gap Penalty

Besides having substitution matrices for mutations, it is also desirable to score insertions and deletions gaps differently. First, to avoid having gaps all over the alignment, gaps have to be given penalties, just like unmatching amino acids. This penalty cannot be derived from the database alignments used to create the substitution matrices such as BLOSUM since these matrices were derived from ungapped alignments. The score given to an insertion/deletion is commonly called a gap penalty.

There are two schemes used, namely, linear gap (g) penalty and affine gap penalty. Having only one score for any gap inserted is called a linear. However, insertions and deletions often involve a longer stretch of sequence in a single event. For this reason, two different gap penalties are usually included in the alignment algorithms: One penalty for having a gap at all (gap opening penalty (go)) and another smaller penalty for

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Fig. 2. The BLOSUM62 substitution matrix

extending already opened gaps (ge). This is called an affine and is actually a compromise between the assumptions that the insertion or deletion is created by one or more events [4].

Definition 2: Given a sequence S over the alphabet Σ of length L, a sequence S^g of length L^g over Σ U "-" is called a gapped sequence of S if $L^g \geq L$ and there exist a transformation $T(S) = S^g$ such that:

1. $\forall\ 1 \leq i \leq L, \exists\ 1 \leq j \leq L^g; S(i) = S^g(j)$
2. If $S^g(p) = S(i)$, and $S^g(q) = S(j)$; and $i < j$ then $p < q$.

2. PAIR-WISE ALIGNMENT

Pair-wise sequence alignment is defined as an alignment of two sequences to determine how similar they are. In most sequence similarity calculations, a similarity score is inferred from the alignment. Gap insertions are allowed until the resulting sequences are of the same size, and the alignment must obey the restriction that gaps cannot appear in the same position in both sequences. This score is determined based on a substitution matrix and specific penalties for the insertions and deletions gaps.

In the following, the two most common pair-wise alignment algorithms used to compute the similarity matrix H for each a pair of sequences S and T with their length L_s and L_t, respectively, are explained in detail.

Definition 3: Given a pair of sequences S and T over the alphabet Σ with their lengths L_s and L_t, respectively. Let S^g and T^g be two-gapped sequences with lengths L_s^g and L_t^g, respectively. A pair-wise sequence alignment of S and T is defined to be a matrix M of size $(2 \times n)$ with $n = \max(L_s^g, L_s^g, L_t^g)$ with the following properties: $\forall\ 1 \leq i \leq n$.

1. $M(s, i) = S^s(i)$ and $M(t, i) = T^s(i)$
2. If $M(s, i)$ = "-" then $M(t, i) \neq$ "-" and vice versa.

Definition 4: Given two sequences S and T on an alphabet Σ. A similarity score function of an alignment.

Score: $M \rightarrow R$

Score $(M) \in R$

Assigns a similarity score to each pair of characters in M.

Definition 5: A score of alignment is a real value function (score) that associate for each alignment $M(S,T)$ a real value function Score $(M(S,T))$.

The problem of pair-wise alignment S and T is to find an alignment $M_o(S,T)$ with optimal score $(M_o(S,T))$, that is,

Score $(M(S,T)) \leq$ Score $(M_o(S,T))$ for all alignments $M(S,T)$.

A. NW Algorithm

The NW algorithm [1], introduced in 1970, as the first dynamic programming tool to compute a global alignment for any pair of biological sequences. The NW algorithm achieves its goal by going through the following three phases:

- The initialization phase: Initiates the $H(0,0)$ matrix element by 0. The first row and column are initialized with the costs of gaps of lengths s and t.
  i.e., $H(s,0) = g.s$ and $H(0,t) = g.t \; \forall \; 1 \leq s \leq L_s, 1 \leq t \leq L_t$; g is a gap penalty.
- The score computation phase: Computes all other values of matrix $H(s,t)$ using one of the following recursive formula:

$$H_L(s,t) = max\begin{Bmatrix} H_L(s-1,t-1) + sbt(S(s),T(t)), \\ H_L(s-1,t) + g, \\ H_L(s,t-1) + g \end{Bmatrix}$$

or

$$H_A(s,t) = max\begin{Bmatrix} H_A(s-1,t-1) + sbt(S(s),T(t)), \\ E(s,t) + g, \\ F(s,t) + g \end{Bmatrix}$$

$$E(s,t) = max\begin{Bmatrix} H_A(s,t-1) + go, \\ E(s,t-1) + ge \end{Bmatrix}$$

$$F(s,t) = max\begin{Bmatrix} H_A(s,t-1) + go, \\ E(s,t-1) + ge \end{Bmatrix}$$

Where, sbt is a substitution matrix and (g, go, and ge) are the gaps penalty.

- The trace back phase: Recovers the alignment by tracing back the path starting from the last element $H(L_s+1; L_t+1)$.

A. SW Algorithm

From the evolution perspective, two-related sequences could evolve independently with many independent mutations lowering the similarity between the sequences. Aligning the sequences with noised information often fails to produce a biologically meaningful alignment. In these cases, the local alignment, proposed by Smith and Waterman [2], identifies the longest segment pair that yields the best alignment score is more preferable. In the SW's algorithm, the longest segment pair between two aligning sequences that yield the optimal alignment is identified by comparing all possible segments of all lengths between the two sequences through dynamic programming technique.

The main difference between this technique and NW's is that negative scores are set to zeroes. This modification produces an alignment score matrix with positive scores. Thus, the backtracking procedure of the algorithm starts at the highest positive score cell and proceeds until it encounters a cell with zero score. The longest segment pair identified in these backtracking steps is the optimal scored local alignment of the two sequences.

The similarity matrix score H is filled using one of the following recurrence formula:

$$H_L(s,t) = max\begin{Bmatrix} 0, \\ H_L(s-1,t-1) + sbt(S(s),T(t)), \\ H_L(s-1,t) + g, \\ H_L(s,t-1) + g \end{Bmatrix}$$

or

$$H_A(s,t) = max\begin{Bmatrix} 0, \\ H_A(s-1,t-1) + sbt(S(s),T(t)), \\ E(s,t) + g, \\ F(s,t) + g \end{Bmatrix}$$

$$E(s,t)=max\begin{Bmatrix} H_A(s,t-1)+go, \\ E(s,t-1)+ge \end{Bmatrix}$$

$$F(s,t)=max\begin{Bmatrix} H_A(s-1,t)+go, \\ F(s-1,t)+ge \end{Bmatrix}$$

Where, *sbt* is a substitution matrix and (*g*, *go*, and *ge*) are the gaps penalty.

Fig. 3 shows the example of calculating the pair-wise local alignment $S = \{ATCTCGTATGAT\}$ and $T = \{GTCTATCAC\}$ using the SW algorithm.

The similarity matrix H is shown for:

$g = -1$,

$$sbt(S(s),T(t))=\begin{cases} 2 & if\ S(s)=T(t) \\ -1 & \text{Otherwise} \end{cases}$$

From the highest score ($H(8,11) = 10$), a procedure of trace-back carries out the corresponding alignment as shown in Fig. 3.

3. ALIGNMENT FOR MULTIPLE SEQUENCE

Multiple sequence alignment (MSA) refers to the alignment of more than two biological sequences (DNA, RNA, or protein). It is considered as an extension of pair-wise sequence alignment as discussed in the previous section. It helps in many criteria such as identifying diagnostic patterns or motif to characterize protein families, demonstrating homology between new sequences and existing families of sequences.

	\	A	T	C	T	C	G	T	A	T	G	A	T
\	0	-2	-4	-6	-8	-10	-12	-14	-16	-18	-20	-22	-24
G	-2	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12
T	-4	-2	1	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
C	-6	-3	0	3	2	1	0	-1	-2	-3	-4	-5	-6
T	-8	-4	-1	2	5	4	3	2	1	0	-1	-2	-3
A	-10	-5	-2	1	4	4	3	2	4	3	2	1	0
T	-12	-6	-3	0	3	3	3	5	4	6	5	4	3
C	-14	-7	-4	-1	2	5	4	4	4	5	5	4	3
A	-16	-8	-5	-2	1	4	4	3	6	5	4	7	6
C	-18	-9	-6	-3	0	3	3	3	5	5	4	6	6

The best global alignment would be: ATCTCGTATGAT

where "|" = match ; "-" = gap G--TC-TATCAC

Fig. 3. Local alignment using Smith–Waterman

MSA is NP-complete problem [5] since the computational cost grows exponentially with the expansion of biological datasets. This leads to the development of many algorithms aiming at reaching the most accurate and efficient alignment. Most commonly used algorithms are classified into two categories, progressive and iterative.

Recent studies have shown significant progress in enhancing the quality, accuracy, and speed of MSA tools. However, the big dataset of sequences of biologically relevant length can be difficult and time-consuming to align. Thus, many MSA tools have been proposed. In this section, only important MSA paradigms are introduced. They are used as a base for various MSA distinguished tools [6]. A progressive method most widely used MSA tools utilize the progressive method that was first introduced in 1987 [7]. For aligning N sequences, it goes through the following three main stages (Fig. 4):

- Stage 1: Generates all possible $N(N-1)/2$ pair-wise sequence alignment to construct a distance matrix computing the similarity between of each two sequences
- Stage 2: Creates a guide-tree using all the pair-wise distances using a clustering method such as unweighted pair-group method with arithmetic [8] or neighbor-joining [9]
- Stage 3: Builds-up the final multiple alignments by the progressive inclusion of the N sequences alignment based on the range given by the guide tree.

Stage 1 is calculated using the match between the residues of the two sequences found by the local alignment. The number of exact match is computed by counting the identical residues appearing in the same column in the local alignment excluding gaps, using the equation [10]:

$$Dist(S,T)=1-\frac{nid(S,T)}{min\{L_s,L_t\}} \quad (2)$$

Where, *nid(S,T)* indicates the number of exact matches using SW algorithm to align S and T. For instance, the *nid*-value in Fig. 3 is 6. Actually, this method will run-out storing the similarity matrix H, which is not practical for the datasets with long sequences.

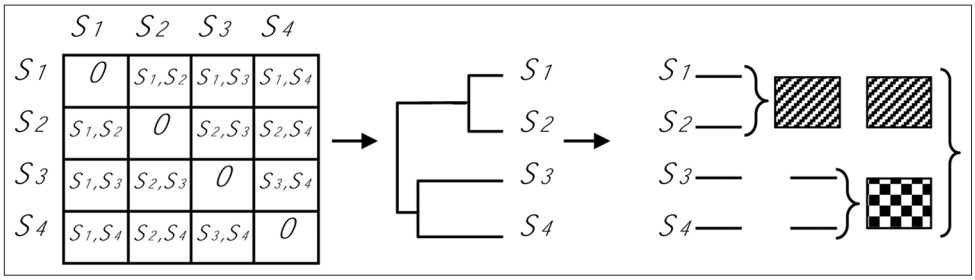

Fig. 4. Multiple sequence alignment produced with ClustalW

Liu *et al.* [11] presented new recurrence relations equations (3) and (4) for the *nid*-value calculation that is compatible for parallel systems such as GPUs. They facilitate nid-computations without calculation of the actual trace-back, to reduce the storage space using the matrices $N_L(s,t)$ and $N_A(s,t)$ for linear and affine gap penalty, respectively. The computations of these matrices are given by following recurrences:

$$N_L(s,t)=\begin{cases} 0 & if\ H_L(s,t)=0 \\ N_L(s-1,t-1) & if\ H_L(s,t)=H_L(s-1,t-1) \\ +m(s,t) & +sbt(S(s),T(t)) \\ N_L(s,t-1) & if\ H_L(s,t)=H_L(s,t-1)+g \\ N_L(s-1,t) & if\ H_L(s,t)=H_L(s-1,t)+g \end{cases} \quad (3)$$

Where,

$$m(s,t)=\begin{cases} 1 & \text{if } S(s)=T(t) \\ 0 & \text{otherwise} \end{cases}$$

For linear gap penalty, and for affine gap penalty, we use:

$$N_A(s,t)=\begin{cases} 0 & if\ H_A(s,t)=0 \\ N_A(s-1,t-1) & if\ H_A(s,t)=H_L(s-1,t-1) \\ +m(s,t) & +sbt(S(s),T(t)) \\ N_E(s,t-1) & if\ H_A(s,t)=E(s,t) \\ N_F(s-1,t) & if\ H_A(s,t)=F(s,t) \end{cases} \quad (4)$$

Where,

$$m(s,t)=\begin{cases} 1 & \text{if } S(s)=T(t) \\ 0 & \text{otherwise} \end{cases}$$

$$N_E(s,t)=\begin{cases} 0 & if\ t=1 \\ N_A(s,t-1) & if\ E(s,t)=H_A(s,t-1)+go \\ N_E(s,t-1) & if\ E(s,t)=E(s,t-1)+ge \end{cases}$$

$$N_F(s,t)=\begin{cases} 0 & if\ t=1 \\ N_A(s,t-1) & if\ E(s,t)=H_A(s-1,t)+go \\ N_F(s,t-1) & if\ E(s,t)=E(s-1,t)+ge \end{cases}$$

4. PROPOSED METHOD

This section propose vectorized and parallelized algorithm for computing the sequence alignment. The aim of this work is switching of all the previous matrices to vectors and computing it by parallel. The main goal for it is to reduce used storage and speed-up runtime, for aligning long biological sequences fast.

The proposed algorithm will implement on the optimal local alignment (SW) algorithm because of the following features:

- It has been trusted by biologists for almost two decades with quality that is still comparable to more recent algorithms
- Its alignment results are similar to biologists expectations
- It is relatively fast, simple, understandable, and provides fairly good alignments across a diverse range of sequence types
- It has highly cited aligner, especially for big dataset of sequences as mentioned in recent studies [12]-[15].

SW algorithm [2] compares two sequences by computing a distance that represents the minimal cost of transforming one segment into another, with respect to the given scoring system.

As previously mentioned, there are two approaches to compute SW algorithm, based-on-gap penalty, linear, and affine gap penalties. Gap penalties prefer more continuous

gaps to opening new gaps. Therefore, it encourages that gaps occur in loop regions instead of in highly structured regions.

The background biological meaning for this is that biologically divergence is often less likely in highly structured regions, which are commonly very important to the fold and function of a protein. In this paper, the affine gap penalty will be used.

A. Vectorization Approach

Vectorizing all matrices presented in the previous section is the main contribution of the proposed method. It was used when computing the aligning sequences of long lengths, aiming at accelerating computations, and used less space. This approach was based on the calculation of each elements of antidiagonal D in the similarity matrix (H) with affine gap penalty is based only on the computed elements of the four antidiagonals; two from (H) matrix with one from both E and F matrices.

This postulate that just one vector D for current antidiagonal, with six buffers previously computed D_1, D_2, D_E, D_{E1}, D_F, and D_{F1} are enough to calculate the elements of D. After the newly D is computed, D_1 is replaced by D_2, D_2 is replaced by D, D_E is replaced by D_{E1}, and D_F is replaced by D_{F1}.

In the subsequent iteration, this cyclic method is used to replace the six buffers D_1, D_2, D_E, D_{E1}, D_F, and D_{F1}. The values of all cells in D are computed in terms of its diagonal neighbor stored at D_1, with its left and upper neighbors stored at D_2 in addition the cells in D_E and D_F, and the maximum value is selected indicating the highest score.

Fig. 5 shows the vectorization approach when aligning pair of sequences $S = \{GCTACTCAC\}$ and $T = \{GCTAGG\ TATGAT\}$ with their lengths are 9 and 12, respectively. It illustrates the calculation of the elements of H_A, using affine gap ($go = 7$, $ge = -1$) and a substitution cost of (2) if the characters are match and (−1) otherwise, and how it is replaced by D, using D_1, D_2, D_E, D_{E1}, D_F, and D_{F1}.

This figure shows the dependence relationship of the elements of the matrices, which is visualized by considering its antidiagonals D_1, D_2, D_E, D_{E1}, D_F, and D_{F1} dependencies. It is clear that antidiagonal D in iteration $i = 9$ computations depend on the four previously computed antidiagonals D^7, D^8, $\mathrm{D}_{\mathrm{E}}^{8}$, and D^8_F. Therefore, all other computed antidiagonals can be neglected.

Furthermore, it shows the dependence relationship of cell $D(5)$ with its left neighbor $D_E(5) = -9$, upper neighbor $D_F(4) = -8$, and upper left neighbor $D_1(4) = 5$. Where $D_2(5) + go$ is a maximum value of ($D_{E1}(5) + ge$ and $\mathrm{D}_{\mathrm{E}}^{1}(5) + go$); $D_2(4) + go$ is a maximum value of ($D_{F1}(4) + ge$ and $D_F(4) + go$). Using this way, all elements along vector D are computed in parallel from all elements in vectors D_1, D_2, D_E, and D_F.

To verify these postulates, authors have proposed the following new recurrence and theorem.

Theorem: The pair-wise local alignment of the sequences S and T, with an affine gap penalty (go for opining gap and ge for extending gap), substitution matrix (sbt), in iteration (i), and with element (j), the equation:

$$D^i(j) = H_A(j, i-j+1) \tag{5}$$

Gives a vectorization D^i of the matrix H_A and the following relations hold:

$$N^i(j) = N_A(j, i-j+1) \tag{6}$$

$$nid(S,T) = \max_i N^i(i_{max})$$

Where $max(2, s-L_s) \le i \le min((s+1), (L_s+1))$ and (i_{max}) indicates the position of the maximum value in the vectors D^i.

Proof: From Equation (1), we get:

$$H_A(j, i-j+1) = max\begin{cases} 0, \\ H_A(j-1, i-j+stb(S(j), \\ T(i-j+1)), \\ E(j, i-j+1), \\ F(j, i-j+1) \end{cases}$$

$$E(j, i-j+1) = max\begin{cases} H_A(j, i-j)+go, \\ E(j, i-j)+ge \end{cases}$$

$$F(j, i-j+1) = max\begin{cases} H_A(j-1, i-j)+go, \\ F(j-1, i-j+1)+ge \end{cases}$$

and from Equation (5), we get:

$$D^i(j) = max\begin{cases} 0, \\ D^{i-2}(j-1)+stb(S(j), T(i-j+1)), \\ D_E^i(j), \\ D_F^i(j) \end{cases}$$

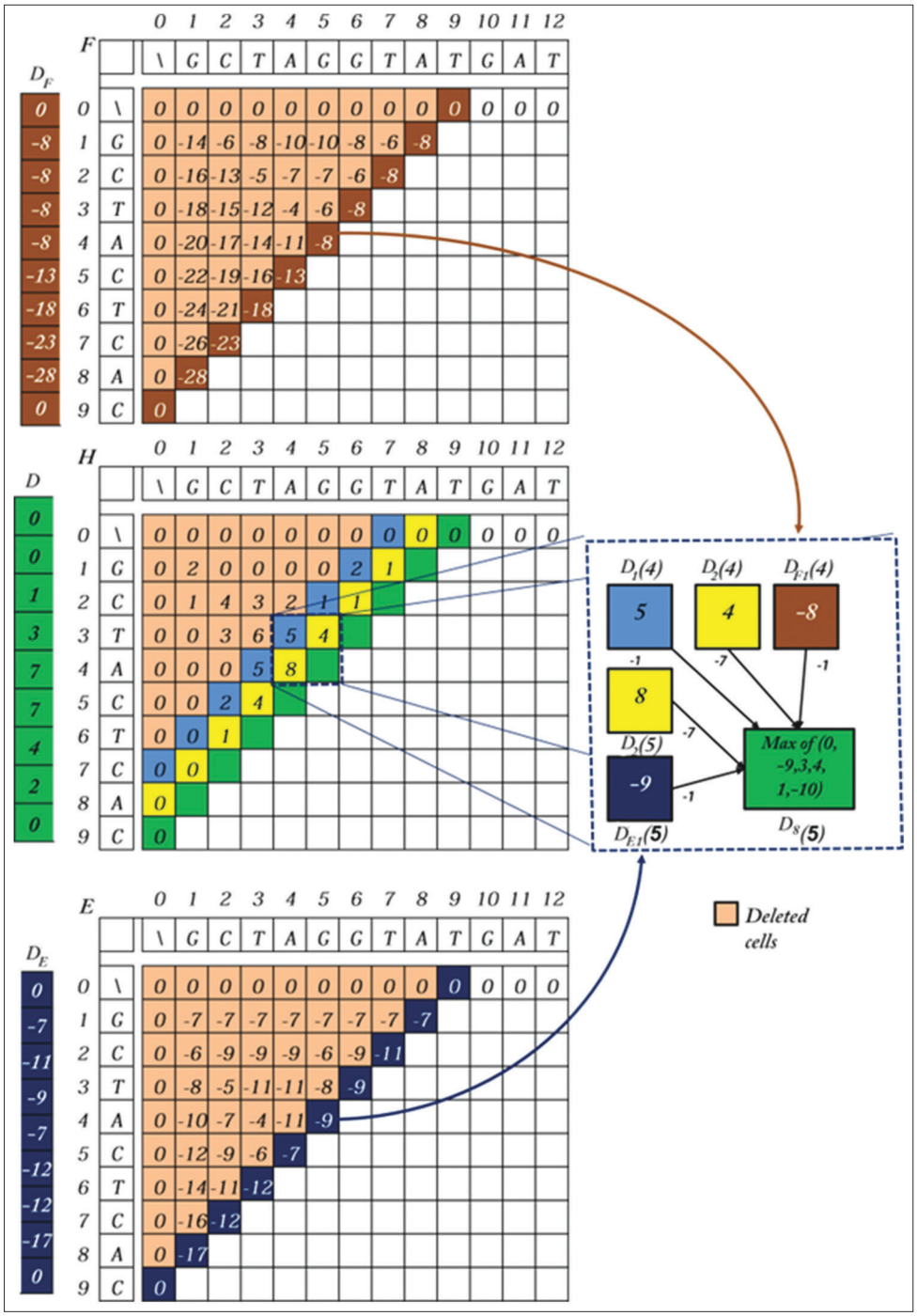

Fig. 5. Relation between *H*, *D*, D_1, D_2, D_E, and D_F

$$D_E^i(j) = max\begin{cases} D^{i-1}(j-1) + go, \\ D_E^{i-1}(j) + ge \end{cases}$$

$$D_F^i(j) = max\begin{cases} D^{i-1}(j) + go \\ D_F^{i-1}(j-1) + ge \end{cases}$$

Since D^{i-1} is computed in the previous iteration of D^i, D^i_E, and D^i_F, that is, in one iteration before D^i, D^i_E, and D^i_F, hence D^i_2 is computed in the second iterations before D^i, let us denote D^{i-1}, D^{i-2}, D^{i-1}_E, and D^{i-1}_F by D^{i-1}, D^{i-2}, D^i_{E1}, and D^i_{F1}, respectively. Then, we get:

$$D^i(j) = max\begin{cases}0,\\ D^i_2(j-1) + stb(S(j),T(i-j+1)),\\ D^i{}_E(j),\\ D^i{}_F(j)\end{cases}$$

$$D^i_E(j) = max\left\{D^i_1(j-1) + go, D_{E1}(j) + ge\right.$$

$$D^i_F(j) = max\left\{D^i_1(j) + go, D_{F1}(j-1) + ge\right. \quad (7)$$

To proof of $N^i_D(j)$ let the Equation (6), gives the vectorization $N_A(s,t)$:

$$NA(j,i-j+1) =$$
$$\begin{cases}0 & if\ H(j,i-j+1)=0\\ N_A(j-1,i-j) & if\ H(j,i-j)=H(j-1,i-j)\\ +m(j,i-j+1) & +sbt(S(j),T(i-j+1))\\ N_E(j,i-j+1) & if\ H_A(j,i-j+1)=E(j,i-j+1)\\ N_F(j,i-j+1) & if H_A(j,i-j+1)=F(j,i-j+1)\end{cases}$$

$$N_E(j,i-j+1) =$$
$$\begin{cases}0 & if\ i-j=0\\ N_A(j,i-j) & if\ H_A(j,i-j+1)=H_A(j,i-j)+go\\ N_E(j,i-j) & if\ H_A(j,i-j+1)=E(j,i-j)+ge\end{cases}$$

$$N_F(j,i-j+1) =$$
$$\begin{cases}0 & if\ k=1\\ N_A(j-1,i-j+1) & if\ H_A(j,i-j+1)=H_A(j-1,i-j+1)+go\\ N_F(j-1,i-j+1) & if\ H_A(j,i-j+1)=F(j-1,i-j+1)+ge\end{cases}$$

After vectorizing the matrix N_A as shown in Fig. 5.

$$N^i_D(j) = \begin{cases}0 & if\ D^i(j)=0\\ N^{i-2}_D(j-1) & if\ D^i(j)=D^i_1(j-1)+\\ +m(j) & +sbt(S(j),T(i-j+1))\\ N^i_E(j) & if\ D^i(j)=D^i_E(j)+g\\ N^i_F(j-1) & if\ D^i(j)=D^i_F(j)+g\end{cases}$$

$$N^i_E(j) = \begin{cases}0 & if\ j=1\\ N^i_A(j) & if\ D(j)=D^i(j)+go\\ N^i_E(j) & if\ D(j)=D^{i-1}_E(j)+ge\end{cases}$$

$$N^i_F(j) = \begin{cases}0 & if\ j=1\\ N^i_A(j-1) & if\ D^i(j)=D^i(j-1)+go\\ N^i_F(j-1) & if\ D^i(j)=D^{i-1}_F(j-1)+ge\end{cases}$$

Then, we get:

$$N^i_D(j) = \begin{cases}0 & if\ D^i(j)=0\\ N^i_1(j-1) & if\ D^i(j)=D^i_1(j-1)\\ +m(j) & +sbt(S(j),T(i-j+1))\\ N^i_2(j) & if\ D^i(j)=D^{i-1}_E(j)+g\\ N^i_2(j-1) & if\ D^i(j)=D^i_F(j)+g\end{cases} \quad (8)$$

Where,

$$m(j) = \begin{cases}1 & S(j)=T(i-j+1)\\ 0 & \text{otherwise}\end{cases}$$

$$N^i_E(j) = \begin{cases}0 & if\ i-j=0\\ N^i_A(j) & if\ E^i(j)=D_i(j)+go\\ N^i_E(j) & if\ E^i(j)=D^i_{E1}(j)+ge\end{cases}$$

$$N^i_F(j) = \begin{cases}0 & if\ j=1\\ N^i_A(j-1) & if\ D^i(j)=D(j-1)+go\\ N^i_F(j-1) & if\ D^i(j)=D^i_{F1}(j-1)+ge\end{cases}$$

We now show that for a given *i*, $N_A(j)$ is equal to the number of exact matches in the optimal (*i*) suffix alignment.

Case 1: *D(j) = 0*. The alignment is empty. Hence, $N_A(j)=0$.

Case 2: *D(j) = Dⁱ(j−1)+sbt(S(j),T(i−j+1))*. The alignment ends with *S(j)* aligned to *T(i−j+1)*, which contributes *m(S(j), T(i−j+1))* to the *nid*-value. The residual number is than equal to the *nid*-value got in the optimal *j−1* suffix alignment. Hence,

$N_A(j) = N_1(j-1) + m(S(j), T(i-j+1))$

Case 3: *D(j) = D²(j−1) + go*. The alignment ends with *S(j)* aligned to a gap, which contributes zero exact matches. The residual number is than equal to the number got in the optimal *D²(j−1)* suffix alignment. Hence,

$N_A(j) = N_2(j-1)$

Case 4: $D(j) = D^2(j) + ge$. The alignment ends with $T(i-j+1)$ aligned to a gap, which contributes to zero exact matches. The remaining number is equal to the number found in the optimal $D^2(j)$ suffix alignment. Hence,

$N_A(j) = N_2(j)$

The increase in N_D^i occurs only at the vectors D_i which has matching in its elements. Hence,

$N_A(x_{max}, y_{max}) = \text{nid}(S,T)$ is obtained by maximization $N_D^i(i_{max})$.

B. Parallelization Approach

Another critical challenge that must be dealt with is the gigantic explosion in the amount of molecular data which makes the ability to align a huge number of long sequences becoming even more essential.

For example, the Ribosomal Database Project Release 10 [16] consists of more than million sequences. This leads to the massive number of calculations. Even if the sequences are short, and pair-wise calculations can be done relatively quickly, say at a rate of 5000^{-1} s, then their alignment still requires almost 12 days of CPU time. Another difficulty is how to store the similarity matrix elements, as it will take up to 40 GB of memory. This leads to the need of new approaches to parallelize the calculations using sort of sophisticated parallel and distributed systems such as multi-cores.

Recent studies refer some attempts have made to accelerate computation of similarity matrix. Ying *et al.*, uses GPU's in [17]. They show speedup comparing with the serial CPU program. Wirawan *et al.*, [18] introduced a parallel algorithm on the cell broadband engine multi-core system for the calculation by taking benefit of the 128-bit SIMD vectorization registers of each SPEs and used half word values (16 bits) for the computation. Their results show a good speedup comparing with sequential ClustalW program.

Recently, Al-Neama *et al.* [19] proposed a new parallel algorithm of distance matrix computations of ClustalW is based on OpenMP system. It achieves speedup of about 2.39 on 50 sequences of the average length of 9200 nucleotide; tested on Core-i7 Intel Xeon 2.83GHz of the processor.

The second contribution of the proposed algorithm is parallelizing the computations to align the long-sequences dataset. The multithreads technique is used to apply the parallelism that reduces runtime necessary for repeated tasking synchronization and exchanging data. In addition, it makes efficient the scheduling through a task allocation policy that prefers the distribution according to the location of data. Each core (P) in the processor has a thread that its responsibility is calculation the maximum value of D's and all threads runs in parallel. Calculations of the vector D are distributed over the total number of available cores (P). The elements of all vectors D_1, D_2, D_E, D_{E1}, D_F, D_{F1}, and D are accessible through the core's shared memory.

The maximum value for each elements of D using Equation (7) and N_D^i using Equation (8) are calculated in parallel. The value of each cell is evaluated in terms of its diagonal neighbor stored at D^1, with its left and upper neighbors stored at D^2, with D_E and D_F, and then the maximum value is selected indicating the highest score.

Fig. 6 shows the parallelization approach. It displays the scheduling of calculations of the elements of D and distributed them on the available cores labeled $P0$, $P1$, $P2$, and $P3$. They run in parallel on each four sequent elements of D, then they sequentially run on the second sequent 4 rows, and so on.

5. PERFORMANCE EVALUATION

The performance of the conceived parallel implementation of the proposed algorithm was extensively evaluated using different processing parameters. The evaluation methodology that was followed to correctly study the results obtained by the described solution is presented. The analysis goes from pair-wise alignment methods (SW). The improvements achieved due to introduced optimizations for multi-core system.

In this section, all needed information about the experimental setup for performance measurements is illustrated. It includes specifications of used platforms, details of experimented biological datasets, and characteristics of other programs used during comparisons.

A. Platform

As obviously clear from the previous section, the presented algorithm was designed to parallelize computations on a multi-core-based environment. Thus, to correctly evaluate the performance of the both original and proposed methods, the platform was considered as specified: An Intel quad-core-i7-3770, with 3.40 GHz processor and main memory of 8 GB; implementing on 64-bit Linux

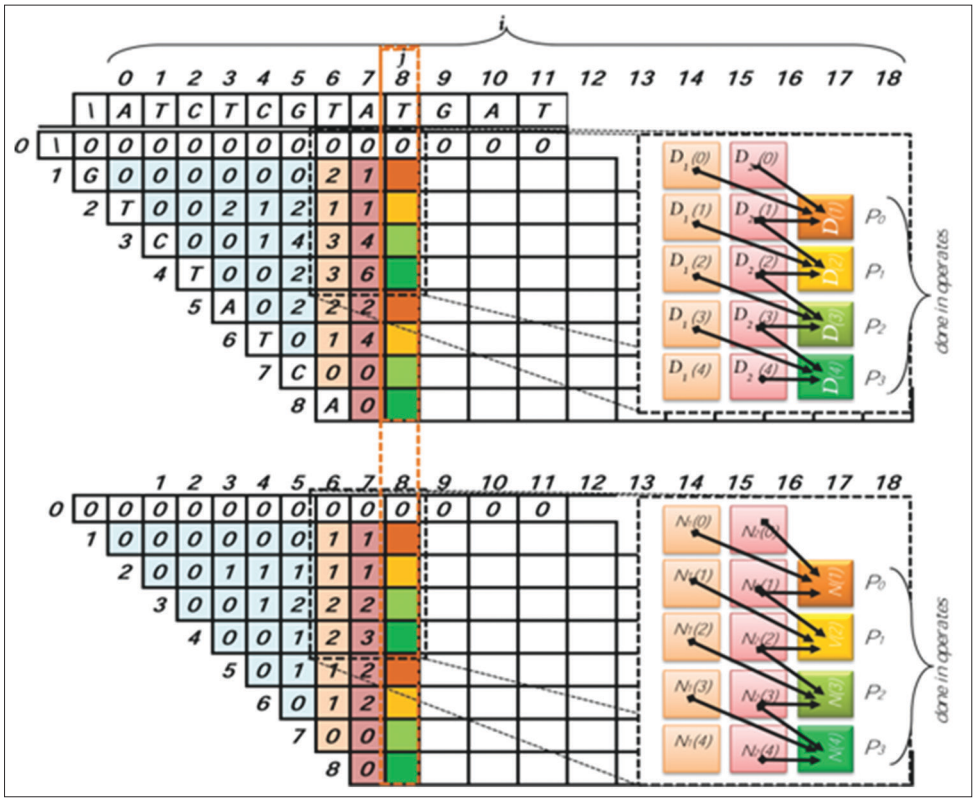

Fig. 6. Scheduling *D*'s computations on 4 cores

TABLE I
Used Benchmark Dataset Specifications

Sequences' No.	Sequences' length	Standard deviation
50	34,500	5.307
100	19,700	1.417
500	9200	514
400	856	7
1,000	858	8
400	408	3
2,000	266	2
4,000	247	2

OS (Ubuntu) and running using C + + with OpenMP library.

B. Datasets

The tests have been conducted using a variety of data sets. These data sets sequences including long, medium-length, and short sequences. These lengths are ranging from 400 to 34,500 residues to study the solution's overall performance against multiple different sizes. The data sets consist of sequences selected from NCBI [20].

TABLE II
Sequential Performance Measurements of Our Algorithm versus ClustalW-MPI

Sequences' No.	Sequences' length	Our algorithm	CW-MPI	Speedup
50	34,500	79,027	164,613.66	2.083
100	19,700	51,203	96,876.08	1.892
500	9,200	47,140	80,892.24	1.716
1000	858	2572	3569.94	1.388
400	856	382	626.86	1.641
400	408	163	187.12	1.148
2000	266	903	1,008.65	1.117
4000	247	2,631	2767.81	1.052

Table I shows the used data set with the number of sequences and average their length with a numerical measure of the scatter of a data set (standard deviation).

C. Programs

Overall, tests have been conducted on the specific platforms using various groups of data sets. To evaluate the implementation of our algorithm, it was tested in

TABLE III
Parallel Performance Measurements of Our Algorithm versus ClustalW-MPI

Sequences' No.	Sequences' Length	Our Alg.	CW-MPI	Speedup
50	34,500	33,270	93,157.03	2.92
100	19,700	25,602	67,075.93	2.62
500	9,200	28,803	67,109.92	2.33
1000	858	240.66	517.42	2.15
400	856	115.73	223.59	1.93
400	408	1,955	3,67.98	1.72
2000	266	677.25	819.47	1.21
4000	247	2026	2329.75	1.15

TABLE IV
Efficiency Comparisons Using 8 Cores

Sequences' No.	Sequences' length	CW-MPI
50	34,500	0.35
100	19,700	0.33
500	9200	0.29
1000	858	0.27
400	856	0.24
400	408	0.22
2000	266	0.15
4000	247	0.14

TABLE V
Performance Comparison (in GCUPS) for Scanning the Datasets

Sequences' No.	Sequences' length	Our algorithm	CW-MPI
50	34,500	0.438	0.157
100	19,700	0.750	0.286
500	9200	3.667	1.574
1000	858	2.435	1.133
400	856	1.153	0.597
400	408	1.886	1.094
2000	266	2.104	1.739
4000	247	2.428	2.111

comparison to popular and efficient MSA program named ClustalW-MPI. This program is available online at: http://www.mybiosoftware.com/alignment/3052.

The runtime and speedup are considered most common performance measurements. Runtime is the elapsed time for all calculation, including all additions, comparisons, and maximum values. Speedup is the ratio between the runtime of the two involved programs.

Table II gives the runtime (in sec) and speedup of the two sequential of the proposed program against the ClustalW-MPI program computing the distance computation

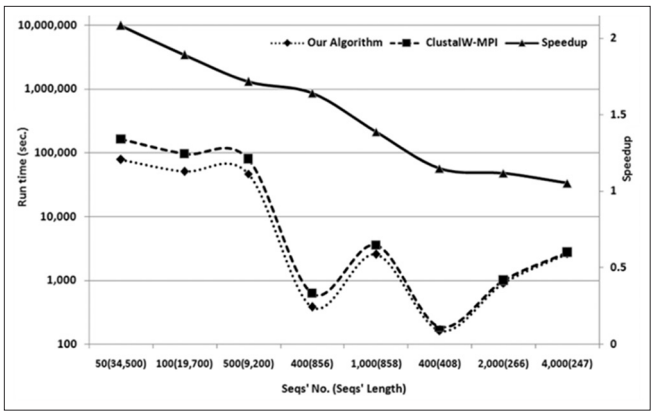

Fig. 7. Performance comparison between sequential our algorithm, ClustalW-MPI

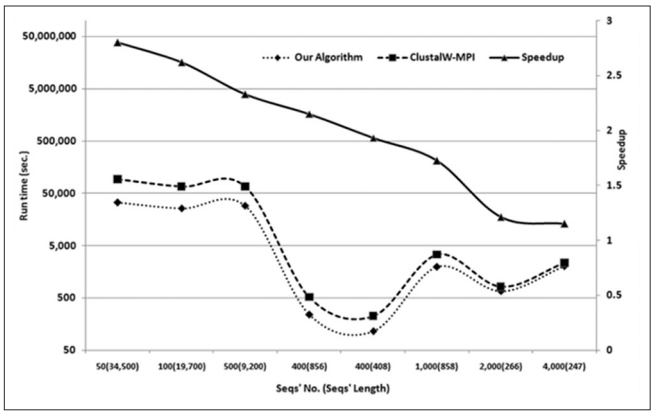

Fig. 8. Performance comparison between parallel our algorithm, ClustalW-MPI

to illustrate how vectorization approach accelerates calculations.

Fig. 7 illustrates that the longer sequences have the more acceleration of our algorithm calculation. Furthermore, our program achieves a speedup up to 2-fold over ClustalW-MPI.

Furthermore, parallel runtime of both is shown in Table III. Fig. 8 shows the execution time and speedup of proposed parallel program on the above-mentioned datasets. Our parallel program achieved reducing the runtime of aligning sequences of length 34,500 from 164,613 s using ClustalW-MPI to 79,027 s using the proposed algorithm using 8 cores.

The speed-up of our parallel program achieved significant speedup of almost 3 for aligning sequences of longest length of sequence. Obviously, the sequences with the short length, the decrease the overall performance

To conform that the proposed program is most efficient than other executed programs, the parallel efficiency of the available cores was evaluated. It is the ratio of the speedup (S) with respect to the number of processors (P) [21]. It is given by the following equation:

$$E = \frac{S}{P} \tag{8}$$

Results are shown in Table IV. It is clear that the efficiency of our algorithm is exponentially increasing as the length of sequences increases. In addition, our program supreme efficiency was up to 0.35 with respect to the ClustalW-MPI for the longest sequence length up to (34 k).

There is another performance measurement used in computational biology called billion cell updates per second (GCUPS). A GCUPS represents the time for a complete computation of one entry in the similarity matrix, including all comparisons, additions, and maximum operations. Table V shows the performance comparison for the datasets.

6. CONCLUSIONS

This paper presented a new parallel algorithm for computing the *nid*-value in the pair-wise local alignment. The results of the proposed algorithm were used in the first stage of MSA. Since the new algorithm was implemented using vectorizing technique, we have got a significant improvement in the performance.

The program was able to calculate *nid*-value for sequences with length up to (34 k) residues. It surpasses ClustalW-MPI 0.13 with 2.9 speedup and the efficiency reached 0.35. A better performance can be achieved if more cores are provided. Furthermore, it can be accomplished a higher speedup with improved efficiency.

Program's performance figures vary from a low of 0.438 GCUPS to a high of 3.66 GCUPS as the lengths of the query sequences decrease from 34,500 to 9200.

REFERENCES

[1] S. B. Needleman and C. D. Wunsch. "A general method applicable to the search for similarities in the amino acid sequence of two proteins." *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443-453, 1970.

[2] T. F. Smith and M. S. Waterman. "Identification of common molecular subsequences." *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195-197, 1981.

[3] S. Henikoff, J. G. Henikoff and S. Pietrokovski. "Blocks+: A non-redundant database of protein alignment blocks derived from multiple compilations." *Bioinformatics*, vol. 15, no. 6, pp. 471-479, 1999.

[4] B. Schmidt. *Bioinformatics: High Performance Parallel Computer Architectures*. Florida: CRC Press, 2010.

[5] L. Wang and T. Jiang. "On the complexity of multiple sequence alignment." *Journal of Computational Biology*, vol. 1, no. 4, pp. 337-348, 1994.

[6] R. C. Edgar and S. Batzoglou. "Multiple sequence alignment." *Current Opinion in Structural Biology*, vol. 16, no. 3, pp. 368-373, 2006.

[7] D. F. Feng and R. F. Doolittle. "Progressive sequence alignment as a prerequisitetto correct phylogenetic trees." *Journal of Molecular Evolution*, vol. 25, no. 4, pp. 351-360, 1987.

[8] P. Sneath and R. Sokal. "Unweighted pair group method with arithmetic mean," in *Numerical Taxonomy*, San Francisco: W.H. Freeman and Company, pp. 230-234, 1973.

[9] N. Saitou and M. Nei. "The neighbor-joining method: A new method for reconstructing phylogenetic trees." *Molecular Biology and Evolution*, vol. 4, no. 4, pp. 406-425, 1987.

[10] K. Chaichoompu and S. Kittitornkun. "Multithreaded clustalw with im-proved optimization for intel multi-core processor," in *Communications and Information Technologies, 2006. ISCIT'06. International Symposium on. IEEE*, 2006, pp. 590-594.

[11] W. Liu, B. Schmidt, G. Voss and W. Muller-Wittig. "Streaming algorithms for biological sequence alignment on gpus." *IEEE Transactions on Parallel and Distributed Systems*, vol. 18, no. 9, pp. 1270-1281, 2007.

[12] F. F. Ghaleb, N. M. Reda and M. W. Al-Neama. "An overview of multiple sequence alignment parallel tools." in *Proc. CSCCA '13,* Dubrovnik, Croatia, 2013, pp. 91-96.

[13] Y. Liu, B. Schmidt and D. L. Maskell. "Msaprobs: Multiple sequence alignment based on pair hidden markov models and partition function posterior probabilities." *Bioinformatics*, vol. 26, no. 16, pp. 1958-1964, 2010.

[14] J. D. Thompson, B. Linard, O. Lecompte and O. Poch. "A comprehensive benchmark study of multiple sequence alignment methods: Current challenges and future perspectives." *Plosone*, vol. 6, no. 3, pp. 2101-2113, 2011.

[15] J. Daugelaite, A. O. Driscoll and R. D. Sleator. "An overview of multiple sequence alignments and cloud computing in bioinformatics," ISRN Biomathematics, 2013.

[16] J. R. Cole, Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, T. Marsh, G. M. Garrity and J. M. Tiedje. "The ribosomal database project: improved alignments and new tools for rRNA analysis." *Nucleic Acids Research*, vol. 37, no. 1, pp. D141-D145, 2009.

[17]. Z. Ying., X. Lin., S. C. W. See and M. Li. "GPU-accelerated DNA distance matrix computation," in *Proc. ChinaGrid 2011*, Dalian, Liaoning, China, 2011, pp. 42-47.

[18]. A. Wirawan, C. K. Kwoh and B. Schmidt. "Multi threaded vectorized distance matrix computation on the Cell/BE and x86/SSE2 architectures." *Bioinformatics Advance*, vol. 26, no. 10, pp. 1368-1369, 2010.

[19]. M. W. Al-Neama, N. M. Reda and F. F. Ghaleb. "Multiple sequence alignment on multi-cores." *International Journal of Biomathematics*, vol. 8, no. 6, p. 1550084, 2015.

[20]. National Center for Biotechnology Information (NCBI); 2017. Available: https://www.ncbi.nlm.nih.gov/. [Last Accessed on 2017 Aug 24].

[21]. G. Hager and G. Wellein. *Introduction to High Performance Computing for Scientists and Engineers*. Boca Raton, FL: CRC Press, 2010.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Model Dependent Controller for Underwater Vehicle

Wesam M. Jasim

Department of Information Technology, College of Computer Science and Information Technology, University of Anbar, Ramadi, Iraq

ABSTRACT

In this work, a model dependent control design method based feedback scheme was investigated for autonomous underwater vehicle control. The controller was designed with the nonlinear terms - inertia, hydrodynamic damping, and gravitational, and buoyancy - of the vehicle dynamical model consideration. Then, the model independent controller (PD) was also investigated, with no nonlinear terms consideration. The stability analysis of the proposed model dependent feedback controller was obtained based on a Lyapunov function. The simulation results of the proposed controller were compared with that of PD controller. The comparison shows the validation of the proposed controller.

Index Terms: Model Dependent Controller, PD Controller, Underwater Vehicle

1. INTRODUCTION

An AUV is a robot able to work in six degrees of freedom with actuators and sensors diving autonomously under the water to perform tasks. The dynamics of the underwater vehicle are nonlinear and acted to disturbances. However, to perform its tasks quickly and accurately, two important problems faced the researchers. They are the problem of identifying the accurate model and designing the suitable control techniques. Therefore, researchers in the robotic control field have been attracted to build their algorithms to solve these problems. In this paper, the problem of building the control algorithm of the underwater vehicle is addressed.

Recently, different control techniques were presented, the H_∞ approach for linear parameter varying polytopic systems was addressed to guarantee the performance of the vehicle [1]. An optimal control with game theory was presented for position control problem of the underwater vehicle in Patel *et al.* [2].

Several controllers were discussed in Ferreira *et al.* [3], some of them were based on Lyapunov control theory, and the others were gathered with linear or nonlinear control theory for underwater vehicle horizontal and vertical motions. A feedback control algorithm was presented in Vervoort [4] to stabilize the underwater vehicle with the linearized model.

A linear quadratic regulator (LQR) algorithm was implemented in Prasad and Swarup [5] for underwater vehicle stabilization combined with Model Predictive control (MPC) for position and velocity control. The simulation results showed a stable response compared with a sliding mode controller performance. While a LQR controller was presented for depth control problem of an underwater vehicle in Joo and Zu [6]. The simulation results show the success of the proposed algorithm.

Authors in Mohd-Mokhtar *et al.*'s [7] study present a PID controller for underwater vehicle identified model. The simulation results show that the controller performs accurately when the identified model error was 98% compared with that when the error was 70%.

Access this article online

DOI: 10.21928/uhdjst.v1n2y2017.pp25-30 **E-ISSN:** 2521-4217 **P-ISSN:** 2521-4209

Corresponding author's e-mail: wmj_r@yahoo.com

Received: 10-03-2017 Accepted: 25-03-2017 Published: 29-08-2017

Nonlinear control laws were developed in Elnashar [8] for the six degrees of freedom of an underwater vehicle in several motion strategies. The stability of the system was analyzed based on phase plane analysis. An adaptive signal for the unknown forces compensation gathering with a following controller in a limited space for an autonomous vehicle was proposed in Mukheriee *et al.* [9] without state velocity measurement.

A MPC controller was presented for an AUV low speed tracking control in Steenson *et al.* [10]. The controller was tested in simulation, and the results were verified by testing the strategy in a tank at zero speed. Authors in Rathore and Kumer [11] study proposed a PID controller for an underwater vehicle steering control. The PID controller parameters were optimized based on genetic algorithm and harmonic search method. The simulation results show the robustness of the proposed controller.

The sliding mode control strategy was proposed for an underwater vehicle position control in Tabar *et al.* [12]. The controller was applied to overcome the effect of the disturbances. Zhou *et al.* [13] proposed a state feedback sliding mode controller for a nonlinear dynamic system of an underwater vehicle with disturbance consideration. The simulation results show a good performance.

In this paper, a model dependent controller for autonomous underwater vehicle is proposed. The controller was developed to include the nonlinear dynamics of the vehicle. Then, a model independent controller was presented, and it is results were compared with that of the former controller.

In the following, section II presents the underwater vehicle dynamical model. Section III provides the description of the designed nonlinear feedback control algorithm. Section IV provides simulation results. Our conclusion and future work are given in section V.

2. AUV MODELLING

The nonlinear dynamical model of a 6DOF underwater vehicle can be described based on two reference frames; fixed reference frame (inertial reference frame) I and the body frame (motion reference frame) B, shown in Fig. 1. The dynamics and kinematics of the vehicle are expressed as follows [14]:

$$\begin{aligned} M\dot{v} + C(v)v + D(v)v + g(\eta) &= \tau \\ \dot{\eta} &= J(\eta)v \end{aligned} \tag{1}$$

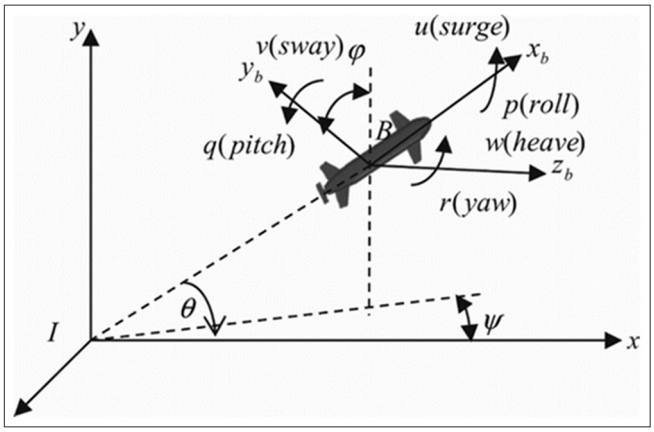

Fig. 1. Underwater vehicle frames

Where, $M=M^T$ is a positive $R^{6\times6}$ inertia matrix with the added masses, $C(v)=-C(v)^T$ is $R^{6\times6}$ Coriolis and centripetal matrix, $D(v)$ is a positive $R^{6\times6}$ hydrodynamic damping matrix, $g(\eta)$ is $R^{6\times1}$ gravitational and buoyancy vector, $\tau=[\tau_X, \tau_Y, \tau_Z, \tau_K, \tau_M, \tau_N]^T$ is $R^{6\times1}$ forces and torque input vector, $v=[u,v,w,p,q,r]^T$ is the linear and angular velocity vector, $\eta=[x,y,z,\varphi,\theta,\psi]^T$ is $R^{6\times1}$ motion vector in surge, sway, heave, roll, pitch, and yaw, respectively, and $J(\eta)$ is $R^{6\times6}$ body frame to inertial frame transformation matrix.

$$J(\eta) = \begin{bmatrix} c\psi c\theta & -s\psi c\phi + c\psi s\theta s\phi & c\psi s\theta c\phi + s\psi s\phi \\ s\psi c\theta & s\psi s\theta s\phi + c\psi s\phi & s\psi s\theta c\phi - c\psi s\phi \\ -s\theta & c\theta s\phi & c\theta s\phi \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & s\phi t\theta & c\phi t\theta \\ 0 & c\phi & -s\phi \\ 0 & \frac{s\phi}{c\theta} & \frac{c\phi}{c\theta} \end{bmatrix}$$

Where, *s*, *c*, and *t* are sine, cosine, and tan.

Assuming that the vehicle is symmetry about the three planes, the vehicle is operate in low speed, roll, and pitch movement is neglected, the body frame is considered to be at same position of the center of gravity, no disturbance is considered, and all the dynamic states can be decoupled, and the dynamical system Eq. (1) can be rewritten as follows:

$$M\dot{v} + D(v)v + g(\eta) = \tau$$
$$\dot{\eta} = J(\eta)v \quad (2)$$

Where,

$$M = \begin{bmatrix} m + X_{\dot{u}} & 0 & 0 & 0 & 0 & 0 \\ 0 & m + Y_{\dot{v}} & 0 & 0 & 0 & 0 \\ 0 & 0 & m + Z_{\dot{w}} & 0 & 0 & 0 \\ 0 & 0 & 0 & I_x + K_{\dot{p}} & 0 & 0 \\ 0 & 0 & 0 & 0 & I_y + M_{\dot{q}} & 0 \\ 0 & 0 & 0 & 0 & 0 & I_z + N_{\dot{r}} \end{bmatrix}$$

and

$$D = \begin{bmatrix} X_u + X_{u|u|}|u| & 0 & 0 & 0 & 0 & 0 \\ 0 & Y_v + Y_{v|v|}|v| & 0 & 0 & 0 & 0 \\ 0 & 0 & Z_w + Z_{w|w|}|w| & 0 & 0 & 0 \\ 0 & 0 & 0 & K_p + K_{p|p|}|p| & 0 & 0 \\ 0 & 0 & 0 & 0 & K_q + K_{q|q|}|q| & 0 \\ 0 & 0 & 0 & 0 & 0 & N_r + N_{r|r|}|r| \end{bmatrix}$$

and only four degrees of freedom are considered to control the vehicle, i.e., control (x,y,z) and ψ states.

3. CONTROLLER DESIGN

In this section, the aim is to design a feedback control algorithm for the path following problem of the underwater vehicle. To this end, the first equation of Eq.2 will be used. Our control scheme consists of two approaches. The first approach is to design a model dependent controller to find the desired control vector τ. The second approach is to design a model independent control algorithm to obtain the desired control vector τ. The controllers' stability analyses are guaranteed based on Lyapunov function as exponential and asymptotic stability, respectively.

The main task is to derive the underwater vehicle toward the desired position η_d from the initial position to satisfy the following equilibrium point.

$$\lim_{t\to\infty} \tilde{\eta} = \lim_{t\to\infty} (\eta_d - \eta) = 0 \quad (3)$$

Now, the following theorem can be addressed:

Theorem 1: Considering the dynamics of Eq.2 under the feedback control law of the form:

$$\tau = M\dot{v} + D(v)v + g(\eta) + \dot{\tilde{\eta}} + \dot{\tilde{v}} \quad (4)$$

Where, $\dot{\tilde{\eta}} = -k_p\tilde{\eta}$, $\dot{\tilde{v}} = -k_d\tilde{v}$, k_p and k_d are diagonal matrices, $\tilde{\eta} = \eta_d - \eta$ is the motion vector error, and is the linear and angular velocity error. Then, the closed loop system of Eq.2 and Eq.4 is exponentially stable.

Proof: Let us suggest the following Lyapunov candidate:

$$V = \frac{1}{2}\tilde{\eta}^T\tilde{\eta} + \frac{1}{2}\tilde{v}^T\tilde{v} \quad (5)$$

Calculating the time derivative of the proposed Lyapunov function we obtain:

$$\dot{V} = \tilde{\eta}^T\dot{\tilde{\eta}} + \tilde{v}^T\dot{\tilde{v}} \quad (6)$$

Substituting the value of $\dot{\tilde{\eta}}$ and $\dot{\tilde{v}}$ into Eq.6 we get:

$$\dot{V} = -\tilde{\eta}^T k_p\tilde{\eta} - \tilde{v}^T k_d\tilde{v} \le 0 \quad (7)$$

Then, Eq.7 is less than zero leaded k_p and k_d are positive definite diagonal matrices, and it can be concluded based on the Barbalat's lemma [15] that the closed loop system Eq.2 and the control law Eq.4 is globally asymptotically stable, which meets the condition of Eq.3.

Now, the model independent feedback control law, i.e., it is a PD controller without the effect of the hydrodynamic

damping, gravitational and buoyancy forces, and the inertia terms are:

$$\tau = -\Gamma_1 \tilde{\eta} - \Gamma_2 \tilde{v} \quad (8)$$

Where, Γ_1 and Γ_2 are positive diagonal matrices.

4. SIMULATIONS

The model dependent controller Eq.4 has been applied for the four state to be controlled of the autonomous underwater vehicle, which presented in Singh and Chowdhury [16]. A vehicle MATLAB simulator was implemented with the following matrices:

$$M = \begin{bmatrix} 99 & 0 & 0 & 0 & 0 & 0 \\ 0 & 108.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 126 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.05 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1.002 & 0 \\ 0 & 0 & 0 & 0 & 0 & 29.1 \end{bmatrix}$$

$$D = \begin{bmatrix} d_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & d_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & d_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & d_5 & 0 \\ 0 & 0 & 0 & 0 & 0 & d_6 \end{bmatrix}$$

With

$$d_1 = 10 + 227.18|u|$$
$$d_2 = 405.41[v]$$
$$d_3 = 10 + 227.18|w|$$
$$d_4 = 0.05 + 5.21|p|$$
$$d_5 = 0.025 + 3.22|q|$$
$$d_6 = 1.603 + 12.937|r|$$

$$g(\eta) = \begin{bmatrix} 0 \\ 0 \\ -19.6 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

To validate the proposed model dependent control performance, two paths are considered. Then, the vehicle retested under the model independent PD controller Eq.8 and its results are compared with that of the model dependent controller Eq.4.

First, the following desired path was tested:

$$\begin{cases} x_d = 20\sin(t/10) \\ y_d = 20\cos(t/10) \\ z_d = 10 \\ \psi_d = \pi/2 \end{cases}$$

Then, the following desired path was tested:

$$\begin{cases} x_d = 5t \\ y_d = 3t \\ z_d = 20 \\ \psi_d = \pi/4 \end{cases}$$

The vehicle was started from zero initial condition in both cases.

Figs. 2-5 show the typical results when the vehicle was commanded to follow the first desired path under the proposed controller compared with that of the PD controller. While Figs. 6-9 present the results obtained when the second path was used.

Figs. 2 and 6 present the motion of the vehicle toward *x*-axis under the two controllers in the first and the second

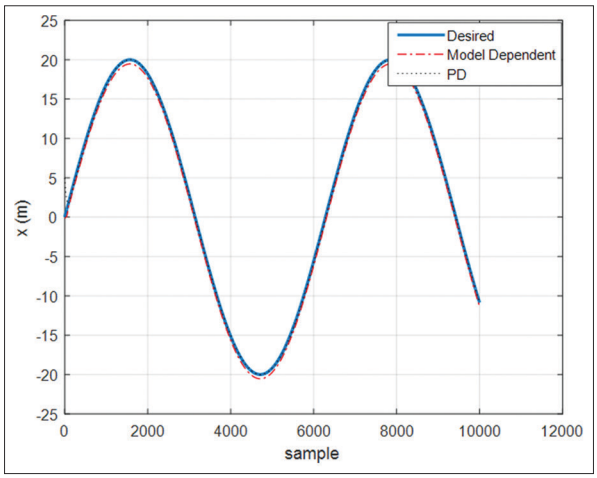

Fig. 2. Motion of the vehicle toward x-direction in first path

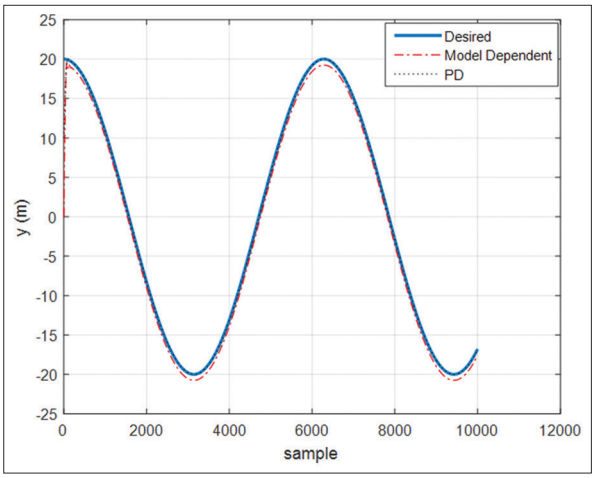

Fig. 3. Motion of the vehicle toward y-direction in first path

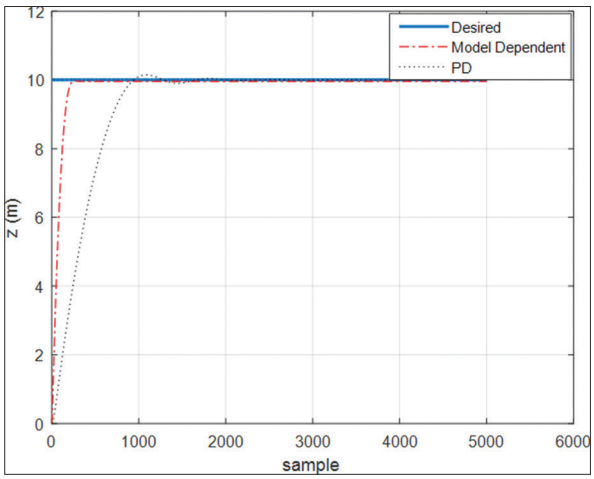

Fig. 4. Motion of the vehicle toward z-direction in first path

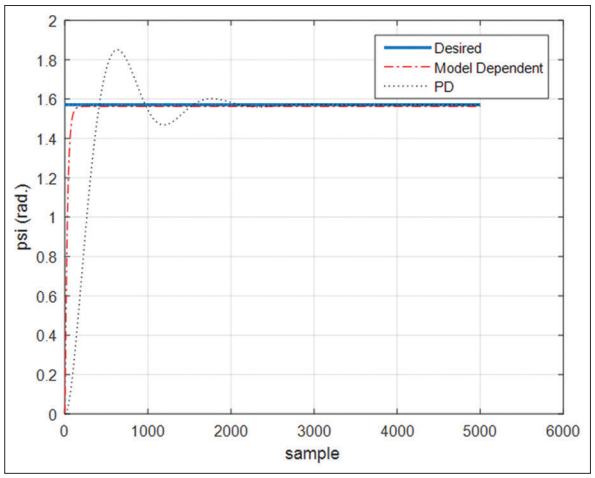

Fig. 5. Rotating of the vehicle along z-axis in first path

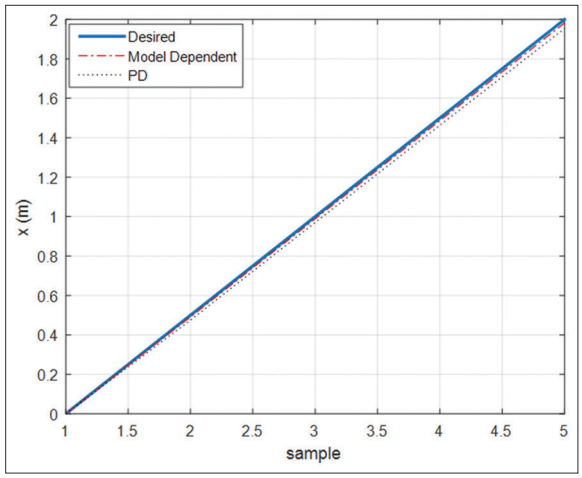

Fig. 6. Motion of the vehicle toward x-direction in second path

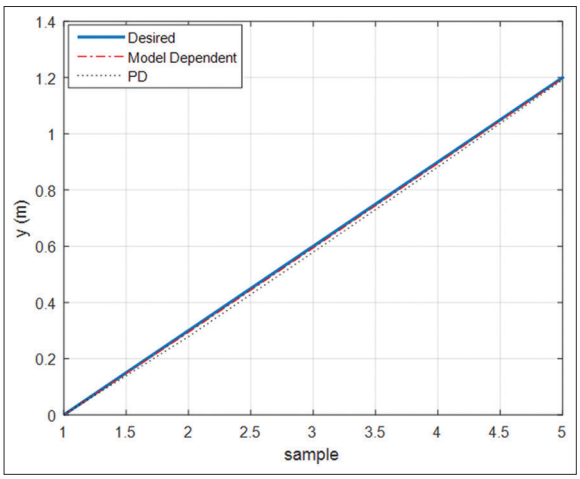

Fig. 7. Motion of the vehicle toward y-direction in second path

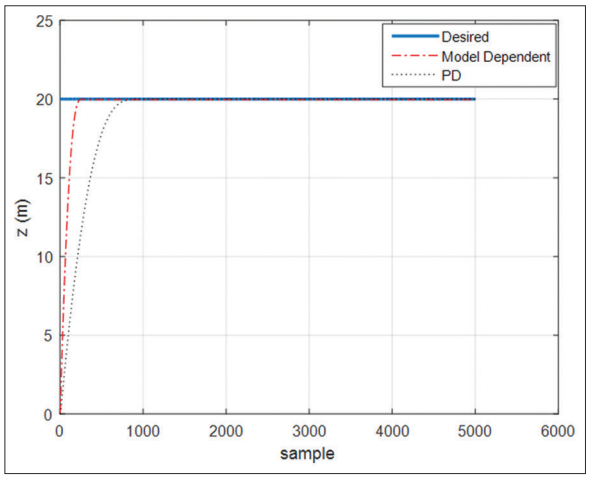

Fig. 8. Motion of the vehicle toward z-direction in second path

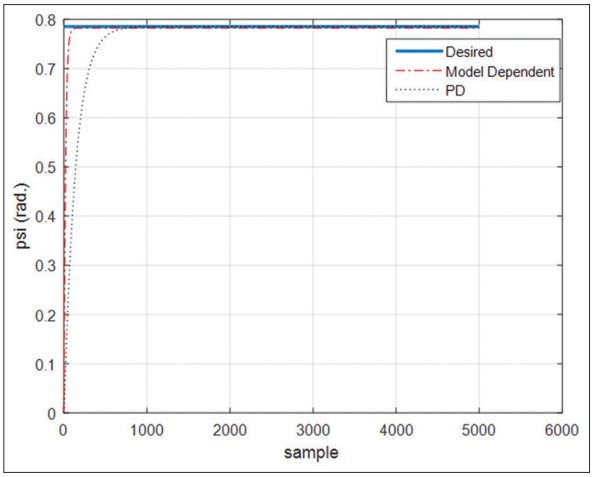

Fig. 9. Rotation of the vehicle along z-axis in second path

cases, respectively. It can be seen that the performance of the vehicle under the proposed controller is faster than that of PD controller to catch the desired path with smaller oscillation.

From Figs. 3, 4, 7, and 8, one can conclude that the vehicle moved toward y and z-axes with very small error when the proposed controller was used compared with some over shoot when the PD controller was used.

The performance of the rotation angle ψ was obtained in Figs. 5 and 9 for the first and the second cases, respectively. In these figures, no oscillation was appearing in both cases, but the proposed controller performs faster than the PD one.

It is quite obvious that the performance of the vehicle path following under the proposed controller is much better than the performance of the vehicle under the PD controller.

6. CONCLUSIONS

The design of a model dependent controller of an underwater vehicle has been addressed in this paper. The proposed controller includes the vehicle nonlinear dynamic terms. The control system stability was guaranteed through Lyapunov theory. The simulation results of using the proposed controller show better performance compared with that of PD controller. Our work toward this subject is to apply the proposed controller for a swarm of underwater vehicle control problem.

REFERENCES

[1] E. Roche, O. Sename and D. Simon. "LPV/H_∞ control of an autonomous underwater vehicle AUV," Procedings of the European Control Conference, 2009.

[2] N. M. Patel, S. E. Gano and J. E. Renaud. "Simulation model of an autonomous underwater vehicle for design optimization," 45th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference, pp. 1-15, Apr. 2004.

[3] B. Ferreira, M. Pinto, A. Matos and N. Cruz. "Control of the MARES autonomous underwater vehicle," in OCEANS, pp. 1-10. Oct. 2009.

[4] J. H. A. Vervoort. "Modeling and Control of an Unmanned Underwater Vehicle." master thesis, University of Technology Eindhoven, 2008.

[5] M. P. R. Prasad and A. Swarup." Position and Velocity Control of Remotely Operated Underwater Vehicle using Model Predictive Control." *Indian Journal of Geo-Marine Sciences*, vol. 44, no. 12, pp. 1920-1927, 2015.

[6] M. G. Joo and Z. Qu." An autonomous underwater vehicle as an underwater glider and its depth control." *International Journal of Control, Automation, and Systems*, vol. 13, no. 5, pp. 1212-1220, 2015.

[7] R. Mohd-Mokhtar, M. H. R. Aziz, M. R. Arshad, A. B. Husaini and M. M. Noh. "Model identification and control analysis for underwater thruster system." *Indian Journal of Geo-Marine Sciences*, vol. 42, no. 8, pp. 992-998, 2013.

[8] G. A. Elnashar. "Dynamics modelling, performance evaluation and stability analysis of an autonomous underwater vehicle." *International Journal of Modelling, Identification and Control*, vol. 21, no. 3, pp. 306-320, 2014.

[9] K. Mukheriee, I. N. Kar and R. K. P. Bhatt. "Adaptive gravity compensation and region tracking control of an AUV without velocity measurement." *International Journal of Modelling, Identification and Control*, vol. 25, no. 2, pp. 154-163, 2016.

[10] L. V. Steenson, S. R. Turnock, A. B. Phillips, C. Harris, M. E. Furlong, E. Rogers and L. Wang. "Model predictive control of a hybrid autonomous underwater vehicle with experimental verification," in *Proc. Of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the MaritimeEnvironment*, 2013.

[11] A. Rathore and M. Kumer. "Robust steering control of autonomous underwater vehicle: Based on PID tuning evolutionary optimization technique." *International Journal of Computer Applications*, vol. 117, no. 18, pp. 1-6, 2015.

[12] A. F. Tabar, M. Azadi and A. Alesaadi. "Sliding mode control of autonomous underwater vehicles." *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 8, no. 3, pp. 546-549, 2014.

[13] H. Y. Zhou, K. Z. Liu and X. S. Feng. "State feedback sliding mode control without chattering by constructing hurwitz matrix for AUV movement." *International Journal of Automation and Computing*, vol. 8, no. 2, pp. 262-268, 2011.

[14] T. I. Fossen, *Guidance and Control of Ocean Vehicles*. New York: John Wiley & Sons, 1994.

[15] J. J. Slotine, and Li, W. *Applied Nonlinear Control*. New Jersey: Prentice Hall, 1991.

[16] M. P. Singh and B. Chowdhury. *Control of Autonomous Underwater Vehicles*. Rourkela: Bachelor of Technology in Electrical Engineering, National Institute of Technology, 2011.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

A Hybrid Simulated Annealing and Back-propagation Algorithm for Feed-forward Neural Network to Detect Credit Card Fraud

Ardalan Husin Awlla

Ministry of Education, Sulaimani 46001, Iraq

ABSTRACT

Due to the ascent and fast development of E-commerce, utilization of credit cards for online buys has significantly expanded, and it brought about a blast in the credit card fraud. As credit card turns into the most prevalent method of installment for both online and also normal buy, cases of fraud associated with it are additionally rising. In actuality, false exchanges are scattered with veritable exchanges, and basic example for coordinating procedures is not frequently adequate to identify those frauds accurately. Usage of effective fraud recognition frameworks has in this manner gotten to be basic for all credit card distributing banks to decrease their losses. Many current systems based on artificial intelligence, Fuzzy logic, machine learning, data mining, sequence alignment, genetic programming, and so on have advanced in distinguishing different credit card fake transactions. A reasonable seeing on all these methodologies will absolutely lead to an efficient credit card fraud detection framework. This paper suggested an anomaly detection model based on a hybrid simulated annealing (SA) and back-propagation algorithm for feed-forward neural network (FFNN), which joined the significant global searching capability of SA with the precise local searching element of back-propagation FFNNs to improve the initial weights of a neural network toward getting a better result for detection fraud.

Index Terms: Artificial Neural Network, Back-propagation, Back-propagation Feed-forward Neural Network, Feed-forward Neural Network, Simulated Annealing, Simulated Annealing-back-propagation Feed-forward Neural Network

1. INTRODUCTION

The convenience of credit cards is common in modern day community. Credit card utilization has expanded among the clients since credit card installment is key one and it is helpful to pay the amount. It is utilized either online or conventional shopping. Due to the expansion and fast advancement in the fields such as E-commerce, the utilization of credit card is also expanded radically [1]. As the use of credit card is development, the credit card fraud is additionally increments. The fraud is characterized as a restricted movement by a client for whom the record was not anticipated [2]. The clients who are utilizing the credit card not having the associations with the cardholder and has no goal of making the repayments for the obtain they done. At present, commercial fraud is turning into a serious issue, and successful identification of credit card is a troublesome effort for the experts [3].

Identifying credit card fraud is a tough effort when applying traditional methods; therefore, the growth of the credit card fraud discovery model has matured off significance, either in the educational or trades the society recently.

Credit card fraud detection belongs to the classification and identification problem with a large number of non-linear

Access this article online

DOI: 10.21928/uhdjst.v1n2y2017.pp31-36

E-ISSN: 2521-4217
P-ISSN: 2521-4209

Corresponding author's e-mail: ardalan.husin@gmail.com

Received: 10-03-2017 Accepted: 25-03-2017 Published: 29-08-2017

situations, which cause it significant to consider non-linear integrated ways to explaining the problem [4].

Artificial neural network (ANN) is a mathematical description of the network of neurons in the mind and share relationships functionalities, such as accepting inputs, processing it, and then produces output [5]. It follows a combined graph of nodes, which are twisted by the weighted links related to the biological neurons. There are different models ANN, for example, feed-forward neural network (FFNN), multiple-layered perceptron, and Kohonen network. Adaptive resonance network and the initial two nets work as a classifier, i.e. these can learn from patterns, and the knowledge can be immediately supervised. Although the other nets learn from attention and later update the network weights, through serve unsupervised learning system seen in a case of clustering. In this paper, a FFNN has been improved for classification intention. FFNN allows the information to pass from the input to output layer in a feed-forward path through the hidden layer(s) [6]. All FFNNs, as stated, possibly trained in a supervised way so that it can learn the feature pattern accessible within the data. To achieve the wanted accuracy in class prediction, fit training is compulsory. While training, the purpose is to catch the network learning feature as the best, which is mirrored by reducing the squared error (i.e., the squared variation between the calculated and the wanted output).

There are various algorithms to optimize such learning method. Backpropagation (BP) is one of the standard traditional ANN training algorithms for supervising learning. The weights are adjusted and updated with a statement delta rule to minimize the prediction error during iterations. The weight improvement methodology covers BP the errors from output layer into hidden layer, so obtaining the optimal set of weights [7].

Simulated annealing (SA) is a probabilistic meta-algorithm for global optimization [8]. It is parallel to the physical method where a solid is casually begin cooled till it is construction is in a cold state, which occurs at a minimum energy form [9]. Similarly, BP algorithm, in SA, the weight has to go into some configuration on the rule till it leads the global minimum [10]. There are besides various other optimization methods such as evolutionary algorithm, for example, genetic algorithm, practical swarm optimization, genetic programming (GP), and so on, there are behind scope of this paper.

The principal purpose of this paper is work to experiment the achievement hybrid of SA and BP compare with BP in the FFNN structure for detection credit card fraud.

2. FEATURE SELECTION

The essential step in developing credit card fraud detection is how to extract the key features. They will influence in recognition rate and improved false alarms. By flattering feature, the data reservation will also be enhanced, so the training and time for data set will be more able for classification that runs under constant environment.

The example dataset that we are running was obtained from a data mining blog. This dataset includes the rundown of the transactions of 20,000 dynamic credit card holders recent months. The input fields incorporate credit card ID, authentication type, current balance, average bank balance, book balance, total number credit card used, and 8 distinctive cardholder classifications such as overdraft, average overdraft, number of location usage, and so on. The data set essentially gives the analysis of the cardholders' exchanges without expressing whether the exchanges were legal or fraudulent. Concerning a given cardholder the dataset based on the following critical values, we can identify which exchange is legal or fraud:

1. Based on credit card usage frequency: Frequency can be found as total number card used/credit cardholder age, if the result <0.2, it implies this property is not relevant for fraud.
2. Based on a number of location credit card usage: Number of locations credit card used per day so far achieved from the dataset, if location is <5, it means this property is not relevant for fraud.
3. Based on credit card average overdraft: With respect to card used happened so far considers, the average overdraft can be found as number of overdraft/total number of card used, if overdraft with respect to card used is <0.02, it means this property is not relevant for fraud.
4. Based on credit card book balance: Regular book balance can be found as current book balance/average book balance, if book balance is equal or <0.25, it implies *that* this property is not relevant for fraud (Table I).

3. FFNN STRUCTURE

According to chosen features from the dataset, we created different networks. The number of hidden layers of every network is restricted to one for active and manageable calculation. The amount of neurons in the hidden layer is changed to test the results [5]. Fig. 1 illustrates the last proper structure of the network achieved it among them.

TABLE I
Sample of Dataset

Transaction No.	1	2	3	4	5
Credit card ID	11111	11112	11113	11114	11115
Authentication type	111	112	113	114	115
Current balance	20000	25000	15000	100000	15000
Average bank balance	80000	55000	70000	60000	61000
Book balance	0.25	0.4545	0.214	1.6666	0.245
Total number card used	13	40	21	90	85
Overdraft	4	20	3	29	17
Average overdraft	0.3076	0.5	0.142	0.3222	0.2
Number of location usage	3	4	2	11	3
Amount of transaction	9000	15000	8500	12000	19000
Card holder age	25	64	50	21	43
Average daily balance	2666	1833	2333	2000	2033
Card frequency	0.52	0.625	0.16	4.2857	0.18
Card holder marital status	0	1	0	1	1

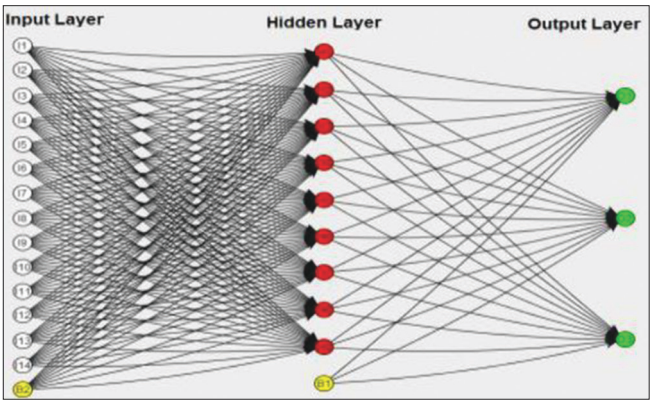

Fig. 1. Description of the FFNN developed

The log sigmoid function in equation 1 is applied as the transfer function connected by the neurons in hidden and output layers to achieve the outputs.

$$f(x)=(1+e^{1-x})^{-1} \quad (1)$$

4. ALGORITHMS

A. Back-propagation Algorithm

BP is a common method for training ANN. The algorithm operates in two forms. First, a training input pattern is given to the input layer, which is forward to the hidden layer then into output layer to produce the network output. Mean square error (MSE) is next calculated by analyzing the estimated output and the target output for all inputs as we explained in equation 3, where "N" indicates the number of entire instances.

$$MS\frac{1}{N}\sum_{1=0}^{N}(T0-E0)^2 \quad (2)$$

In the next step, besides the MSE, the network information back propagates from the output layer to the input layer, and specific connector weights are updated utilizing a "generalized Δ rule" that is held of learning rate (η) and momentum constant (α) [11]. Equations 3 and 4 display the rule of weight updating. In particular equations, the characters "w" means the weights between the connectors "I" and "j" and "t" is the position iteration. An excellent style manual for science writers is [7].

$$\Delta w_{i.j}^{t}=-\eta.\frac{\partial MSE}{\partial w_{i.j}^{t}}+\alpha.w_{i.j}^{t} \quad (3)$$

$$w_{i.j}^{t+1}=\Delta w_{i.j}^{t}+w_{i.j}^{t} \quad (4)$$

For achieving the average η which gives the least MSE, rigorous parametric research has been conducted in this research [11]. The η at which the error is minimal is determined for the association with the SA algorithm. In this research, the momentum constant (α) is fixed equal to 0.9 for all states to speed up the learning method. The epoch size is fixed to as 1500.

B. Simulated Annulling Algorithm

The critical parameter for SA is a temperature (T) which is the similarity of the T in physical system. Beginning at a high T, the algorithm ends the minimum T with continuous decrease with attaining of a thermal equilibrium status at each T [8]. At any T, the weights are randomized. A recent set of weights is accepted as the new optimized set if the MSE with this set is under than the prior set or with a possibility that the present set of weights will reach to the global minimum. As estimated in BP, cost function and transfer functions are utilized. This

research is expected if the number of adjustment in the weight set is more than 10 either the number of iterations is more than 1500 then the equilibrium state at a critical T is supposed to be done. The primary T is determined as 10°C, and ultimate T is 1°C randomly. In addition, the T is reduced with by a determinant of 0.95 random because it is challenging to obtain the accurate values of initial, ultimate, and more the threatening of T. The implementation algorithm is as follows (E(s)) is an actual function.

SA Algorithm

1) *set initial solution in S*
2) *Set initial solution T*
3) *While not terminate do*
4) *Repeat k times*
5) *Chose S′ a randomly element from N(s)*
6) $\Delta E = E(S') - E(S)$
7) *If* $(\Delta E \leq 0)$ *then*
8) S_{i+1} *S′*
9) *else if* S_{i+1} *S′ with probability* $e^{(-\Delta E/T)}$
10) S_{i+1} *S′*
11) *end if*
12) *end repeat*
13) *decreased T*
14) *end do*

C. Hybrid Algorithms

To defeat the local minimum issue of BP because of initial random weight parameters of the network, various optimization algorithms have been attempted by numerous researchers, which enhance the execution of the classification at the cost of more impalement time. In this paper I, hybridized two algorithms, joining global search SA algorithm, and local search gradient algorithm that defeats the local minimum issue with high speculation and quick union speed.

The hybrid SA-BP is a training algorithm joining the SA algorithm with the BP algorithm. SA is global optimization algorithm, which has a powerful capability to investigate the whole search space. This algorithm has a drawback that the search over the global optimum solution is slow. In opposite, the BP has exact and quick local searching capacity to investigate locally the optimum result, but it gets stuck to discover global optimum result in complex pursuit space.

By joining the SA and the gradient-based BP algorithm, another algorithm alluded to as hybrid SA-BP algorithm as shown in Fig. 2. The suggested hybrid algorithm has two stages: Initial one a global search stage, the FFNN is trained utilizing the SA algorithm for few pre-characterized temperature or training error is less than some predefined value, then training mechanism changed to the second stage for searching locally utilizing a deterministic technique the BP algorithm. In this paper, it achieved SA-BP hybrid training algorithm as a strong option way to deal with BP algorithm.

Following steps is the pseudo code for the hybrid SA-BP algorithm:

1. Randomly initialize the weights of the FFNN system appeared in Fig. 1
2. Evaluate weights using SA used in the neural network follow a temperature annealing schedule with the algorithm
3. While first temperature value is under or equal to minimum error then select the best solution for MLP then go to step 7
4. Select a moving method with some probability
5. Try a new solution
6. Evaluate
7. Select the best solution
8. Initialize parameters of BP learning algorithm
9. Initialize weights of the MLP utilizing best solution of SA
10. While new epoch is under or equal to maximum epoch or error converges to minimum error do
11. Using BP update weights to minimize error with training data
12. End while
13. Assess execution of classification with test data
14. End while

5. EXPERIMENTAL STUDY

The network packets that are obtained are separated into two sections. The first section about eight hundred records is used to train SA and BP neural network module. The second section is about two hundred records applied to test the credit card fraud detection. The efficiency of the neural network relies on the number, type, and amount of features and learning algorithm applied to train the neural network. Hence, as to evaluate the execution of a credit card fraud recognition strategy; we have to display a quantitative estimate. In our credit card fraud detection system, we mostly classify the network traffic into two categories, which they are normal and abnormal network traffic. Hence, we need to realize the true positive, true negative, false positive, and finally false negative to define true-positive rate (TPR) and false-negative rate (FNR). TPR and FNR) can be calculated using the following mathematical equations [5], [6].

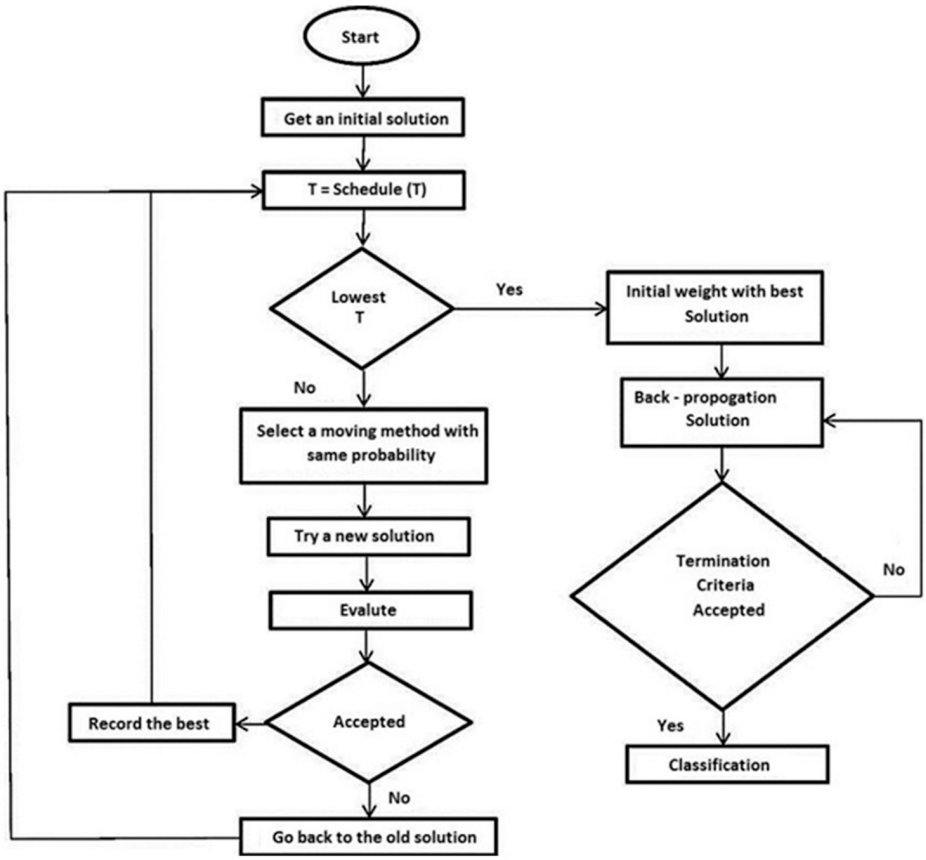

Fig. 2. Structure of hybrid algorithm for classification

$$TPR=TP/(TP+FN) \quad (5)$$

The TPR measures the performance of credit card fraud detection technique concerning the possibility of a suspect data reported correctly as abnormal data. Then, again the FPR measures the performance of credit card fraud detection technique as far as the possibility of a normal traffic reported as abnormal data.

As introduced, the length of parameter temperature has been taken from 10°C to 1°C. The balanced state for it includes of each 10 changes made in set of weights or 1500 iterations, the momentum is 0.9, learning rate is 0.7, maximum error to reaches is 0.01, and weight and threshold values are randomly initialized before training. Consequently, Figs. 2 and 3 show the result of training and test case for the detection rate and FPR of BP and SA-BP (Table II).

Experimental result in Fig. 3. clearly show that SA-BPFFNN more secure in detection credit card farud in comparison to BPFFNN. Furthermore, from the Fig. 4. SA-BPFFNN

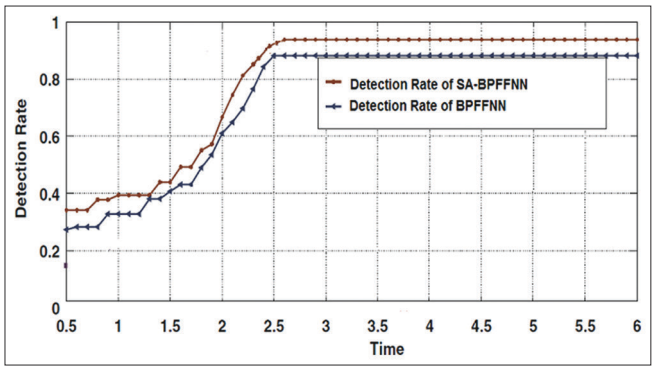

Fig. 3. Detection Rate of SA-BPFFNN and BPFFNN

significantly reduce the false-positive rate compare to BPFFNN.

6. CONCLUSION

As utilization of credit cards turn out to be increasingly regular in each field of the everyday life, master card or

TABLE II
Summaries the Result During Training and Testing

Desired output SA-BP	Yes	0	1	0	1	0
	No	0	0	1	0	1
	Maybe	1	0	0	0	0
Actual output SA_BP	Yes	0.033	0.9723	0.023	0.9865	0.0146
	No	0.3687	0.0173	0.969	0.0075	0.867
	Maybe	0.5983	0.0104	0.008	0.006	0.1184
Desired output BP	Yes	1	0	1	0	0
	No	0	1	0	1	0
	Maybe	0	0	0	0	1
Actual output BP	Yes	0.0035	0.9148	0.0253	0.9517	0.1277
	No	0.4019	0.073	0.8770	0.0303	0.870
	Maybe	0.5946	0.0122	0.0977	0.0180	0.0023

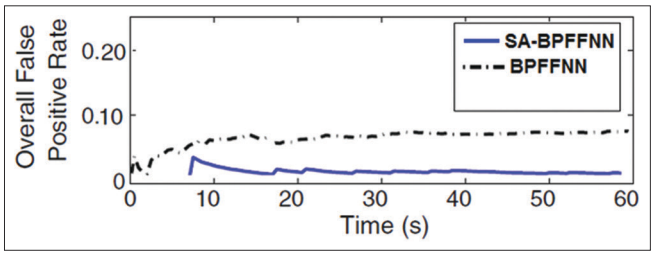

Fig. 4. False positive rate of SA-BPFFNN and BPFFNN

credit card fraud has turned out to be much more rampant. To enhance security of the financial transaction frameworks in an automated and successful way, constructing an accurate and effective credit card fraud detection framework is one of the key efforts for the financial institutions. Credit card fraud detection refers to the classification and recognition issues. The paper hybrids the SA algorithm with BPFFNN for fraud detection where the simulated neural network can learn knowledge from a large number of a dataset for training and examining the result of detection. The analysis result illustrated that the using BP is a simple local minimum algorithm, and SA is a good global search algorithm or optimization algorithm based on the analysis, the experimental results indicate that the accuracy of BPFFNN is under than applied SA to BPFFNN algorithm.

REFERENCES

[1] N. S. Halvaiee and M. K. Akbari. "A novel model for credit card fraud detection using artificial immune systems." *Applied Soft Computing*, vol. 24, pp. 40-49, Nov. 2014.

[2] C. Yin, A. H. Awlla, Z. Yin and J. Wang. "Botnet detection based on genetic neural network." *International Journal of Security and Its Applications*, vol. 9, pp. 97-104, Nov. 2015.

[3] V. Van Vlasselaer, C. Bravo, O. Caelen and B. Baesens. "A novel approach for automated credit card transaction fraud detection using network-based extensions." *Decision Support Systems*, vol. 75, pp. 38-48, Jul. 2015.

[4] D. Sanchez, M. A. Vila, L. Cerda and J. M. Serrano. "Association rules applied to credit card fraud detection." *Expert Systems with Applications*,vol. 36, pp. 3630-3640, 2009.

[5] S. Suganya and N. Kamalraj. "A survey on credit card fraud detection." *International Journal of Computer Science and Mobile Computing*, vol. 4, pp. 241-244, Nov. 2015.

[6] J. Bernal and J. Torres-Jimenez. "SAGRAD: A program for neural network training with simulated annealing and the conjugate gradient method." *Journal of Research of the National Institute of Standards and Technology*, vol. 120, pp. 113-128, 2015.

[7] S. J. Subavathi and T. Kathirvalavakumar, "Adaptive modified backpropagation algorithm based on differential errors." *International Journal of Computer Science, Engineering and Applications*,vol. 1, no. 5, pp. 21-33, Oct. 2011.

[8] A. T. Kalai. "Simulated annealing for convex optimization." *Mathematics of Operations Research*, vol. 31, pp. 253-266, 2006.

[9] C. M. Tan, Ed. *Simulated Annealing*. Vienna, Austria: In-Teh is Croatian Branch of I-Tech Education and Publishing KG, Sep. 2008.

[10] S. H. Zhan, J. Lin, Z. J. Zhang and Y. W. Zhong. "List-based simulated annealing algorithm for traveling salesman problem." *Computational Intelligence and Neuroscience*, vol. 2016, pp. 12, Mar. 2016.

[11] N. A. Hamid, N. M. Nawi, R. Ghazali and M. N. M. Salleh. "Solving local minima problem in back propagation algorithm using adaptive gain, adaptive momentum and adaptive learning rate on classification problems," International Conference Mathematical and Computational Biology. Malacca, Malaysia, pp. 448-455, Apr. 2011.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

A Simple Software Rejuvenation Framework Based on Model Driven Development

Hoger Mahmud

Department of Computer Science, College of Science and Technology, University of Human Development, Iraq

ABSTRACT

In the current dynamic-natured business environment, it is inevitable that today's software systems may not be suitable for tomorrow's business challenges which indicate that the software in use has aged. Although we cannot prevent software aging, we can try to prolong the aging process of software so that it can be used for longer. In this paper, we outline a conceptual software rejuvenation framework based on model **driven** development approach. The framework is simple but effective and can be implemented in a recursive five step process. We have illustrated the applicability of the framework using a simple business case study which highlights the effectiveness of the framework. This work adds to the existing literature on software aging and its preventative measures. It also fills in the research gap which exists about software aging caused by changing requirements.

Index Terms: Model Driven Development, Software Aging, Software Rejuvenation Framework

1. INTRODUCTION

In this paper, we propose a rejuvenation framework that addresses software aging on abstract level. The changing world demands faster and better alignment of software systems with business requirements to cope with the rising demand for better and faster services. This simply means that a perfectly untouched functioning software ages just because it has not been touched. The aging phenomenon occurs in software products in similar ways to human; Parnas [1] draw correlations between the aging symptoms in human and software. As demands for functionality grow software complexity rises, and as a result software, underperformance and malfunctioning became apparent [2]. Software aging is a known phenomenon with recognized symptoms such as increase in failure rate [3]. Researchers have identified a number of causes of software aging, for example, accumulation of errors over time during system operation. One other cause is "weight gain" as in human, software gains weight as more codes are added to an application to accommodate new functionalities, and consequently, the system loses performance. There are numerous examples where software aging has caused electronic accidents in complex systems such as in billing and telecommunication switching systems [4]. Beside the causes researchers in the field have identified a number of aging indicators such as increased rate of resource (e.g., memory) consumption [5]. Another aging indicator is how robust a system is against security attacks if observed over time. This is because security attack techniques are becoming more sophisticated by day. However, more is needed to be done to address the aging phenomena. Grottke *et al.* [5] claim that the conceptual aspect of software aging has not been paid adequate attention by researchers to cover the fundamentals of software aging.

Currently, addressing software aging is mostly done using reengineering techniques such as:

Access this article online

DOI: 10.21928/uhdjst.v1n2y2017.pp37-45 **E-ISSN:** 2521-4217 **P-ISSN:** 2521-4209

Corresponding author's e-mail: hoger.mahmud@uhd.edu.iq

Received: 02-07-2017 Accepted: 22-08-2017 Published: 30-08-2017

1. Forward engineering concerns with moving from high-level abstraction to physical implementation of a system
2. Reverse engineering concerns with analyzing a system to identify components and connectors of that system to represent the system in a different form or higher level of abstraction
3. Redocumentation deals with creation or revision of semantically equivalent representation within the same abstract level
4. Design recovery concerns with reproducing all required information about a system so that a person can understand what the program does
5. Restructuring concerns with transforming a representation of a system to a different one, without any modification to the functionality of the system.

Reengineering can facilitate the examination of a system and learn more about it so that appropriate changes can be made. However, it is not the ideal solution for software upgrade as the process is extremely time-consuming and resource expensive. In this paper, we present a conceptual software rejuvenation framework based on model driven development (MDD) techniques capable of addressing software aging with less time and resource. The framework is most effective where the software aging is due to changing business requirements which in effect requires the addition or omission of functionalities. We have illustrated the applicability of the framework through a simple business case study which supports the effectiveness of the framework. This work contributes to the field of software aging by presenting a novel conceptual framework to software developers that can be utilized to dilute software aging.

The rest of this paper is organised as follows, in Section 2 we provide a brief background about software aging and rejuvenation and in Section 3 we present some related works. In Section 4, we outline the frame work and in Section 5, we illustrate the applicability of the framework using a simple business case study. In Section 6 and 7, we discuss, conclude, and provide some recommendations.

2. BACKGROUND

In this section, we provide a brief background to both software aging and software rejuvenation with the aim to provide better understanding of the proposed framework later in Section 4.

A. Software Aging

Software aging was first introduced by Huang *et al.* [6] and since then the interest in the topic has risen among academics and industries. Complex systems rely on an intricate architectural setup to function, if the structure is slowly destroyed by maintaining and updating the system software aging becomes inevitable [7]. It is a known fact that a system maintainer can mess up perfectly fine functioning software through changing codes or inserting incorrect codes which is known as "ignorant injection" [8]. To provide a focus view of research areas on software aging Cotroneo *et al.* [9] have analyzed more than 70 papers in which they have concluded that overall there are two major categories of research into understanding software aging the first is model-based analysis and the second is measurement-based analysis. Several measureable techniques have been proposed to detect software aging such as "aging indicators" and "time series analysis." The techniques are used to collect data about resources used in a system, and then, analyze it to see if the consumption rate has increased over time which is a sign of aging [3]. As for the causes of software aging, there are two major classes, the first is known as "ignorant surgery" and the second is known as "lack of movement." Fig. 1 shows the major contributors to the two classes of software aging causes.

B. Software Rejuvenation

To keep critical systems functioning correctly software rejuvenation is recognized as an effective technique [10]. The objective of software rejuvenation is to rollback a system continuously to maintain the normal operation of the system and prevent failures. According to Cotroneo *et al.* [3] application-specific and application-generic are two main classes of software rejuvenation techniques in which the former works on specific system features and the latter works on the whole system (e.g., system restart).

To further elaborate on the two main classes, researchers have provided a number of examples for both; flushing of kernel, file system defragmentation and resource reprioritization are examples of application specific rejuvenation and application restart, cluster failover, and operating system reboot are examples of application generic rejuvenation [3]. Fig. 2 illustrates the two classes of software rejuvenation techniques.

3. RELATED WORK

There have been a number of attempts to tackle software aging similar to what we propose here. The authors of Huang *et al.* [6] present a model-based rejuvenation approach for billing applications and Okamura and Dohi [10] proposes dynamic software rejuvenation policies by extending models presented in Pfening *et al.* [11]. The approach is case

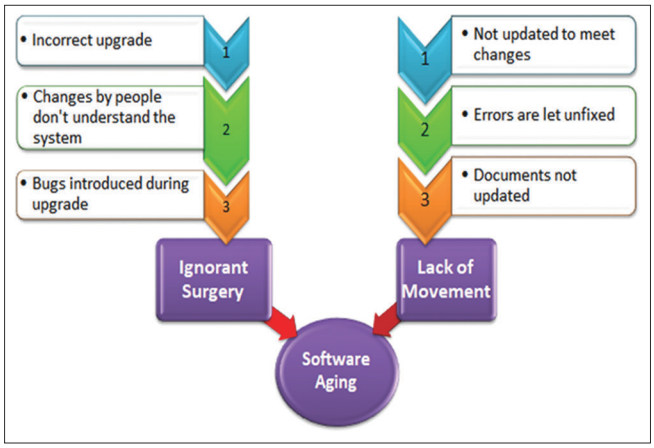

Fig. 1. Major causes of software aging

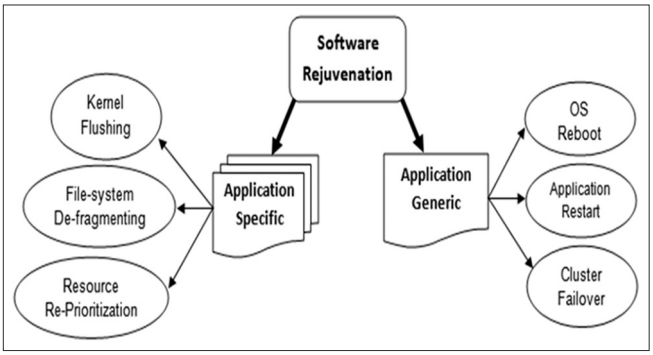

Fig. 2. Software rejuvenation techniques

specific and cannot be applied to a domain; this, however, has similarities with what we are proposing since they also use models to rejuvenate software. Saravakos *et al.* [12] proposes the use of continuous time Markov chains to model and analyze software aging and rejuvenation to better understand causes of aging which helps putting in place mitigating measures. This approach is suitable to treat symptoms of aging that happens for technical reasons rather than changes in requirements. Dohi *et al.* [13] models optimal rejuvenation schedule using semi-Markov processes to maximize availability and minimize cost. The focus here is aging caused due to processing attributes; however, unlike this work we focus on the functionality attributes of a system Garg *et al.* [14]. Adopts the periodic rejuvenation technique proposed by Huang *et al.* [6] and uses stochastic petri net to model stochastic behavior of software aging. Beside modeling techniques, others have used techniques such a time triggered rejuvenation technique used by Salfner and Wolter [15] and software life-extension technique used by Machida *et al.* [16] to counteract software aging in which they take preventative measures to ease software aging and allow more time for system rethink. Huang *et al.* [6] proposes a proactive technique to counteract software aging with the aim to prevent failure using periodic preemptive rollback of running applications. To detect symptoms of aging techniques such as machine learning is used to analyze data through adopting artificial intelligent algorithms (e.g., classifiers) [17]. Garg *et al.* [18] discuss measures for software aging symptom detection with the aim to diagnose and treat the aging taking place, others have used pattern recognition techniques to detect aging symptoms [17]. These works propose how to detect symptoms of software aging without proposing a suitable mechanism to treat the symptoms.

All the related works presented so fare address software aging from technical and performance viewpoint and none consider aging caused as a result of changing requirements. This allows us to claim that our framework contributes to the software aging and rejuvenation literature by filling in this gap and take a new direction in tackling software aging.

4. FRAMEWORK OUTLINE

The base of our conceptual rejuvenation framework is MDD technique [19], [20]. France *et al.* [21] claim that abstract design languages and high-level programming languages can provide automated support for software developers in terms of solution road map that fast-forward system developments. Following their direction we use Model Driven Development (MDD) techniques to design a rejuvenation framework to tackle requirement-based software aging. MDD simply means constructing a model of the system with fine details before transferring it into code. It provides the mapping functions between different models for integration and model reusing purposes [22]. MDD is a generic framework that can accommodate both application specific and application generic classes of software rejuvenation. Mayer *et al.* [23] states MDD is ideal for visualizing systems and not losing the semantic link between different components of the system at the same time. It is inevitable that extensive manual coding in developing a system escalates human errors in the system; this issue can be addressed through code automation which is the ultimate aim of MMD. Building and rebuilding system is an expensive process that requires time and resource; model driven aims at using, weaving and extending models to maintain, develop and redevelop systems. Experts in the field claim that MDD improves quality as models are continuously refined and reduce costs by automating the development process [22]. This process changes models from being expenses to

important assets for businesses. Researchers have identified the conceptual gap between problem domain and implementation as a major obstacle in the way of developing complex systems. Models have been utilized to bridge the gap between problem domain abstractions and software implementation through code generation tools and automated development support [2]. Models can serve many purposes such as:

1. Simplifying the concept of a complex system to aid better understanding of the problem domain and system transformation to a form that can be analyzed mechanically [24]
2. Models are platform and language independent
3. Automatic code generation using models reduce human errors
4. For new requirements only the change in model is required this reduces the issue explained previously known as weight gain.

A. Framework Steps

We propose a five step recursive software rejuvenation framework to address the issue of software aging. As mentioned the framework is based on model driven software development which is implemented in the following steps:

1. First developers gather system requirements which is one of the must do tasks in every software development
2. Developers design the entire system in great details using tools such as unified modeling language (UML)
3. The complete design is fed into code generators such as Code Cooker (http://codecooker.net) and Eclipse UML to Java Generator to generate system codes
4. Software codes are integrated, tested, and finalised, this step is necessary since a code generator tool capable of generating 100% of the code is yet to exist. This limitation is discussed in Section 6
5. In the final step where the new product is delivered and installed.

Fig. 3 illustrates the five steps explained in a recursive setting, i.e., when a new feature is required to be added to the system to address a new requirement the system is upgraded through the model rather than though code injection. The models are kept as assets and refined as new requirements come in, the next section provide more inside as to how the framework works.

5. CASE STUDY

To illustrate the applicability of the framework we present a simple none-trivial business case study specific to Kurdistan region.

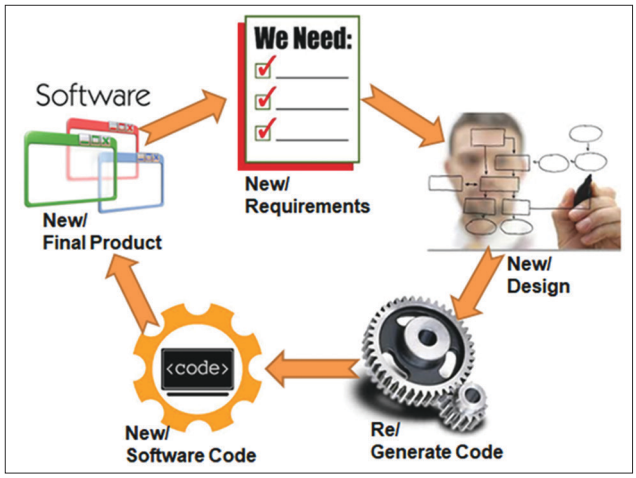

Fig. 3. Model driven development-based software rejuvenation framework

Mr. X is a supermarket owner in the city of Sulaymaniyah who sells domestic goods and he employs 10 people in his supermarket. Currently, his shop is equipped with electronic point of sale (ePOS) systems to record transactions and the form of payment by customers is cash only. Electronic payment is not feasible due to unavailability of electronic payment systems in the region's banks. His current ePOS system is capable of performing the following functionalities:

1. Store individual item details such as name, price, barcode, and expiry dates
2. Store information about employees such as name, address, date of birth, and telephone numbers
3. Retrieve and match barcodes on products to display and record item details
4. Calculate total price and print out customer receipts
5. Record all transactions and generate various reports such as daily sales report, weekly sales report, and sale by item report
6. The administration side of the system is managed through a user management subsystem which allows adding, deleting, updating, and searching on users. The system also contains a product management subsystem that allows managing products through adding, deleting, updating, and searching on item.

We make an assumption that in the next 6 months electronic payment systems (ePayment) will become available in Kurdistan for businesses to use. Now Mr. X would like to gain an edge over his competitors and add ePayment system to his current ePOS system. Fig. 4 is the UML use case diagram for the current ePOS system in Mr. X's supermarket which shows the use cases than can be performed by each actor.

Fig. 5 is the future use case diagram for the new system which shows the addition of a new actor called "customer" and a new use case called "pay electronically" coted in yellow. Now, we assume developers of the system had the framework in mind when they developed the system and have kept a design model of the system similar to the one illustrated in Fig. 6 which shows a UML class diagram design model of the ePOS system. Mr. X now goes back to them and request that the new functionality (electronic payment) to be added to the system. Using the framework the developers refine the UML class diagram model (new classes coted in yellow) to accommodate the new requirement and produce a new design similar to the one shown in Fig. 7. The new design is now ready to be fed into code generators to generate the codes for the new system. Using the framework the developers have performed a rejuvenation process on Mr. X's system without touching

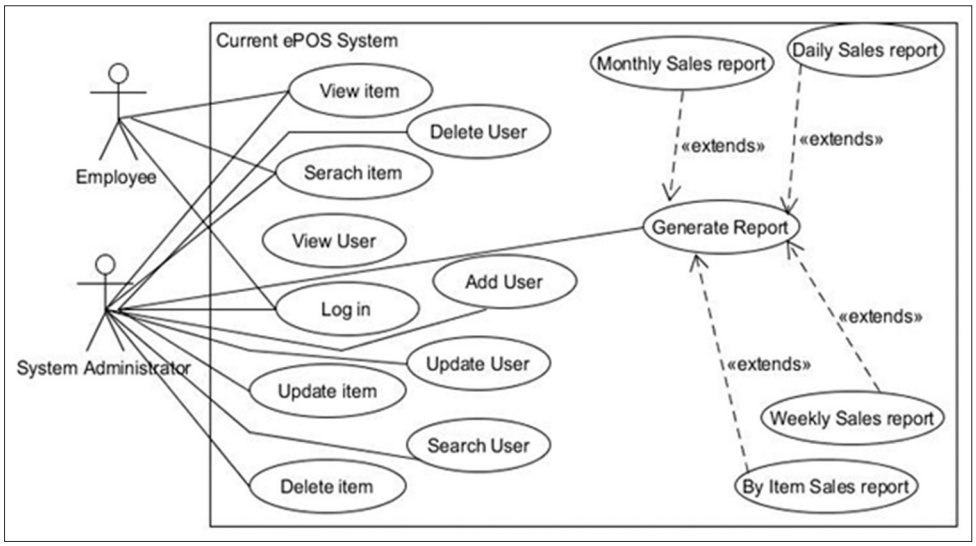

Fig. 4. Current electronic point of sale unified modeling language use case diagram

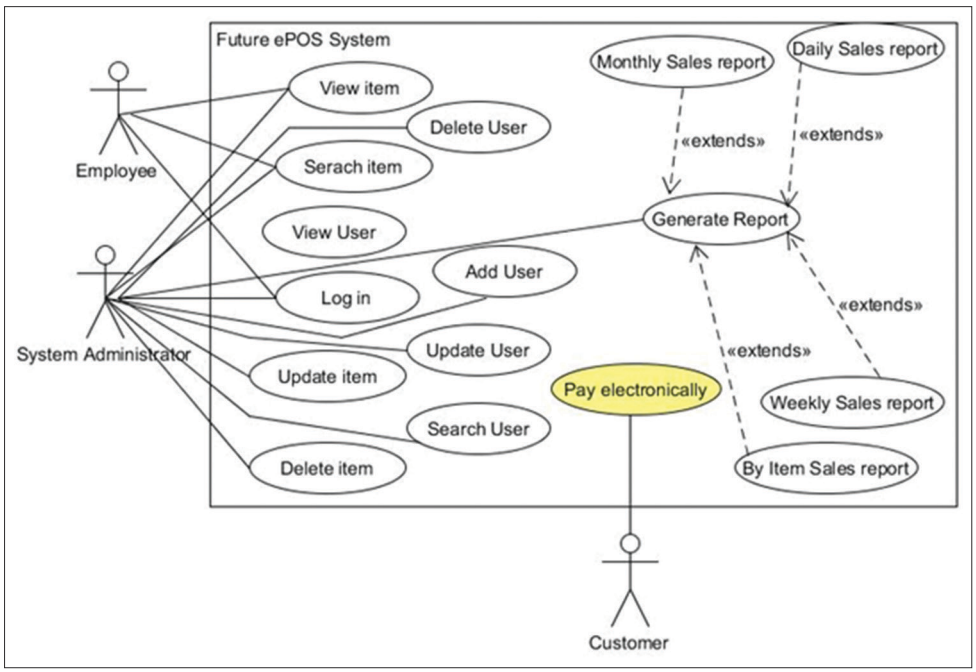

Fig. 5. Future electronic point of sale unified modeling language use case diagram

the current operating ePOS. It is important to point out that the framework tackles software rejuvenation conceptually and on abstract level which means we are bypassing all the technicalities of implementation and testing processes. During our search, we did not come across any related work that considers design for software rejuvenation rather than an actual system, which indicates that our approach is unique. However, it has to be said that although being unique is an advantage, it has made it difficult for us to compare the applicability of our framework with other existing frameworks.

6. DISCUSSION

Researchers in the field have concluded that software aging is inevitable and as software ages it loses its ability to keep up. In this paper, we have proposed a five step recursive software rejuvenation framework based on model driven software development approach. To illustrate the applicability of the framework we have outlined a simple business scenario and explained how the framework rejuvenates the current system in use by the business. The framework will provide the following advantages over existing rejuvenation techniques:

1. The model is used to redevelop the system without taking the old system out of operation which leads to reduction in down time (unavailability) which otherwise lead to lose of customers and profits
2. Using models to maintain and update software gives the development process an edge as models are language independent and can be used to develop systems in the state of the art programming languages which in turn ease software aging as the technology used in the development is current [25]
3. As codes are generated automatically human errors are reduced, which is one of the contributors of software aging
4. Redevelopment costs and times are reduced as developments are automated.

The objective of software rejuvenation is to rollback a system continuously to maintain the normal operation of the system and prevent failures. However, software rejuvenation increases system downtime as the system is taken out of operation while the rejuvenation process is performed. Knowing when to perform rejuvenation process on a system is a crucial factor recognized by researchers to minimize cost and maximize availability [3]. The framework we have proposed addresses this issue by working on the system on design level without terminating the system operation while the rejuvenation solution is finalized. It is important to stress that the framework is conceptual and requires further research as there are a number of limitations that need be addressed to make the framework fully applicable. The limitations can be summarized as follows:

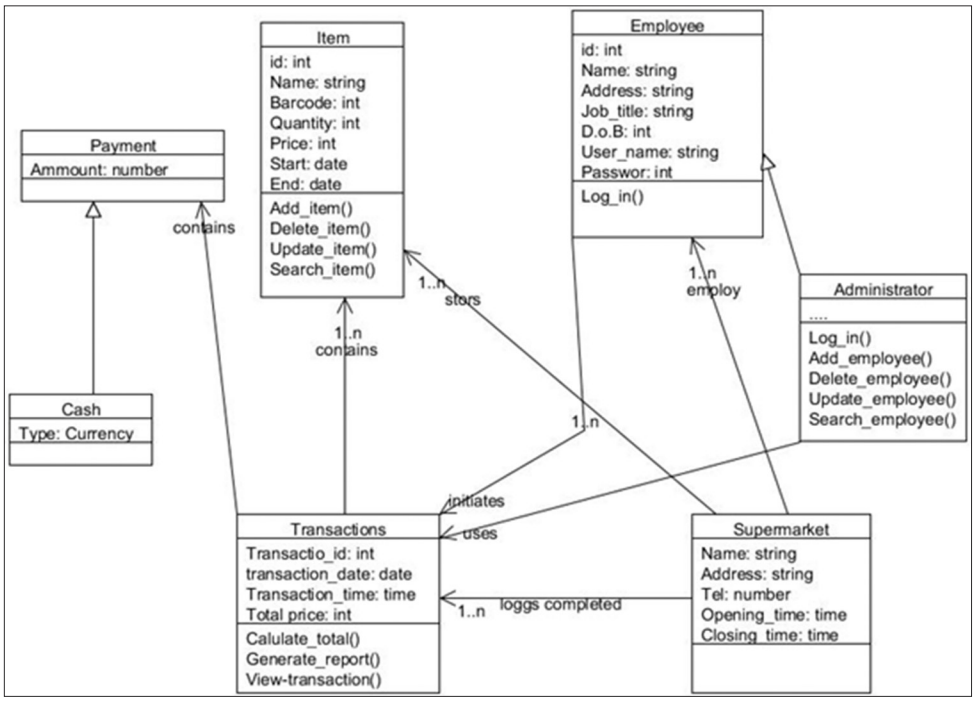

Fig. 6. Current electronic point of sale unified modeling language class diagram

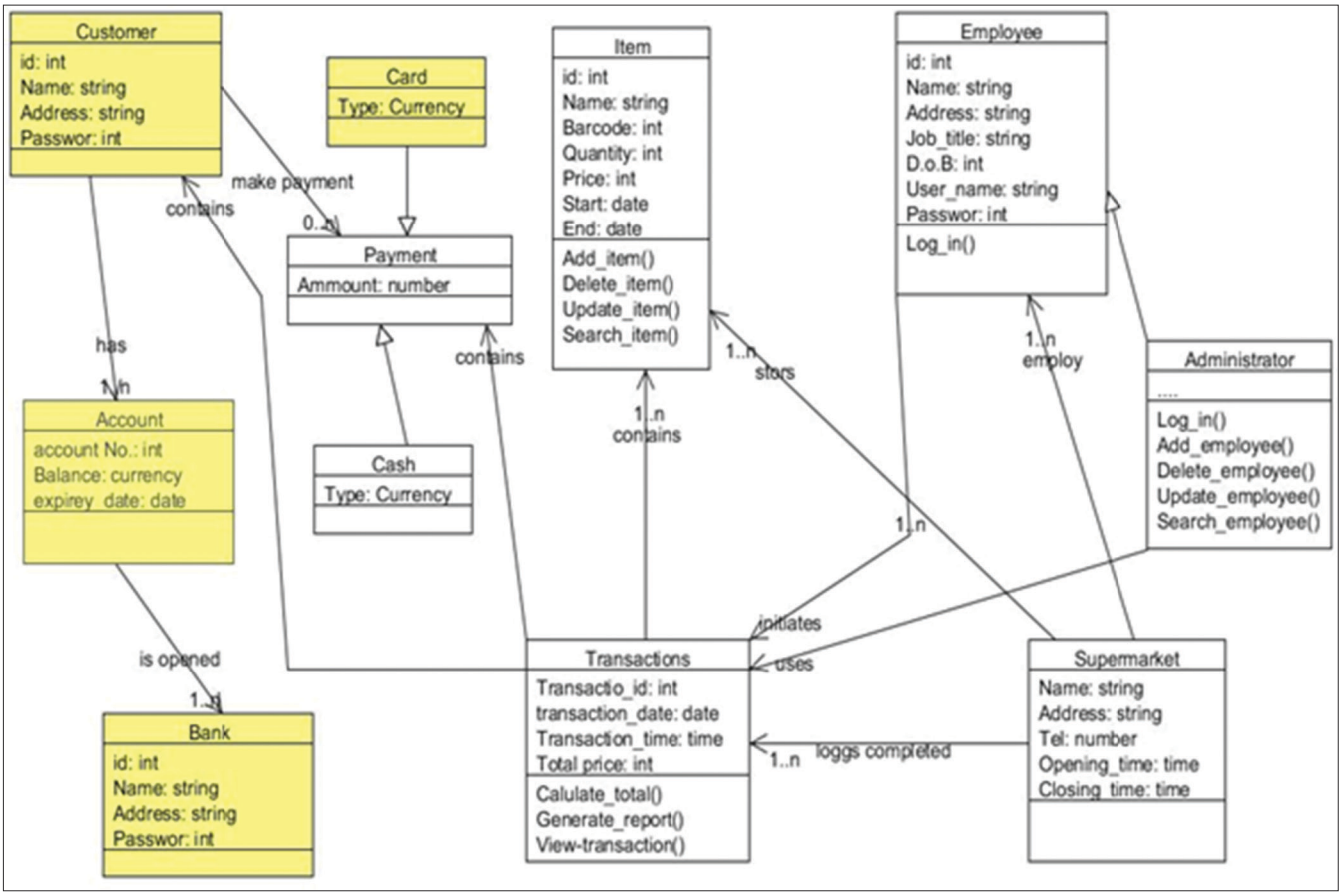

Fig. 7. Future electronic point of sale unified modeling language class diagram

1. Available software modeling tools such as UML 2.0 which is an industry standard currently does not provide the ability to model systems from user-defined viewpoint [21]
2. MDD is not widely used [22] although it has gained momentum with a potential for industry wide adaptation
3. Once the models are developed and finalized there comes the issues of translating it completely into code as a tool to generate 100% codes from a model not yet exist
4. The issue of measuring the quality of models is realized by researchers to tackle this issue France and Rumpe [2] suggests that modeling methods should come with modeling criteria that modelers can use as a guide for system modeling. However, such criterions are yet to be presented by modeling language and tool developers such as developers of UML (www.omg.org)
5. In the course of developing a system, many different models are created at varying abstract levels which creates model tracking, integration, and management issues and the current modeling tools are not sophisticated enough to deal with the issues.

Despite all the limitations, we believe the fundamental concept behind the framework has great potentials to be advanced and implemented in the future.

7. CONCLUSION AND RECOMMENDATIONS

Software aging is inevitable which occurs as a result of changing requirements, ignorant injections, and weight gain. Researchers have proposed a number of different approaches to tackle software aging; however, nearly all approaches are trying to address the aging caused by technical update or software malfunction. In this paper, we have outlined a framework for software rejuvenation that uses MDD approach as base for the rejuvenation process. The framework addresses software aging from a change in business requirement point of view which is different from what current researchers are proposing. It is simple, effective, and applicable as demonstrated by applying it to a simple business case study.

The foundation concept developed in this paper contributes to the field of software aging and paves the way for looking at software aging in a different angle.

Now to delay software aging, we recommend a number of quick mitigating actions as follows:

1. Characterize the changes that are likely to occur over the lifetime of a software product, and the way to achieve this characterization is by applying principles such as object orientation
2. Design and develop the software code in a way that changes can be carried out; to achieve this concise and clear documentation is the key
3. Reviewing and getting a second opinion on the design and documentation of a product helps in prolonging the lifetime of a software product.

When the aging has already occurred there are things we could do to treat it such as:

1. Prevent the aging process to get worse by introducing and creating structures whenever changes are made to the product
2. As changes are introduced to a product a review and update of the documentation is often a very effective step in slowing the aging process
3. Understanding and applying the principle of modularization is a good way to ease the future maintenance of a product
4. Combining different versions of similar functions into one system can increase efficiency of a software product and reduce the size of its code which is one the causes of software aging.

REFERENCES

[1] D. L. Parnas. "Software aging." in Proceedings of the 16th International Conference on Software Engineering, 1994, pp. 279-287.

[2] R. France and B. Rumpe. "Model-driven development of complex software: A research roadmap." in *2007 Future of Software Engineering*. Washington, DC, USA: IEEE Computer Society, 2007, pp. 37-54.

[3] D. Cotroneo, R. Natella, R. Pietrantuono, and S. Russo. "A survey of software aging and rejuvenation studies." *ACM Journal on Emerging Technologies in Computing Systems (JETC),* vol. 10, no. 1, pp. 8, 2014.

[4] A. Avritzer and E. J. Weyuker. "Monitoring smoothly degrading systems for increased dependability." *Empirical Software Engineering*, vol. 2, no. 1, pp. 59-77, 1997.

[5] M. Grottke, R. Matias, and K. S. Trivedi. "The fundamentals of software aging." in Software Reliability Engineering Workshops, 2008. ISSRE Wksp 2008. IEEE International Conference on, 2008, pp. 1-6.

[6] Y. Huang, C. Kintala, N. Kolettis, and N. D. Fulton. "Software rejuvenation: Analysis, module and applications." in Fault-Tolerant Computing, 1995. FTCS-25. Digest of Papers, Twenty-Fifth International Symposium on, 1995, pp. 381-390.

[7] C. Jones. "The economics of software maintenance in the twenty first century." Unpublished Manuscript, 2006. Available: http://www.compaid.com/caiinternet/ezine/capersjones-maintenance.pdf. [Last Accessed on 2017 May 15].

[8] R. L. Glass. "On the aging of software." *Information Systems Management*, vol. 28, no. 2, pp. 184-185, 2011.

[9] D. Cotroneo, R. Natella, R. Pietrantuono, and S. Russo. "Software aging and rejuvenation: Where we are and where we are going." in Software Aging and Rejuvenation (WoSAR), 2011 IEEE Third International Workshop on, 2011, pp. 1-6.

[10] H. Okamura and T. Dohi. "Dynamic software rejuvenation policies in a transaction-based system under Markovian arrival processes." *Performance Evaluation*, vol. 70, no. 3, pp. 197-211, 2013.

[11] A. Pfening, S. Garg, A. Puliafito, M. Telek, and K. S. Trivedi. "Optimal software rejuvenation for tolerating soft failures." *Performance Evaluation*, vol. 27, pp. 491-506, 1996.

[12] P. Saravakos, G. Gravvanis, V. Koutras, and A. Platis. "A comprehensive approach to software aging and rejuvenation on a single node software system." in Proceedings of the 9th Hellenic European Research on Computer Mathematics and its Applications Conference (HERCMA 2009), 2009.

[13] T. Dohi, K. Goseva-Popstojanova and K. S. Trivedi. "Statistical non-parametric algorithms to estimate the optimal software rejuvenation schedule." in Dependable Computing, 2000. Proceedings. 2000 Pacific Rim International Symposium on, 2000, pp. 77-84.

[14] S. Garg, A. Puliafito, M. Telek and K. S. Trivedi. "Analysis of software rejuvenation using Markov regenerative stochastic Petri net." in Software Reliability Engineering, 1995. Proceedings, Sixth International Symposium on, 1995, pp. 180-187.

[15] F. Salfner and K. Wolter. "Analysis of service availability for time-triggered rejuvenation policies." *Journal of Systems and Software,* vol. 83, no. 9, pp. 1579-1590, 2010.

[16] F. Machida, J. Xiang, K. Tadano and Y. Maeno. "Software life-extension: A new countermeasure to software aging." in Software Reliability Engineering (ISSRE), 2012 IEEE 23rd International Symposium on, 2012, pp. 131-140.

[17] K. J. Cassidy, K. C. Gross and A. Malekpour. "Advanced pattern recognition for detection of complex software aging phenomena in online transaction processing servers." in Dependable Systems and Networks, 2002. DSN 2002. Proceedings. International Conference on, 2002, pp. 478-482.

[18] S. Garg, A. van Moorsel, K. Vaidyanathan and K. S. Trivedi. "A methodology for detection and estimation of software aging." in Software Reliability Engineering, 1998. Proceedings. The Ninth International Symposium on, 1998, pp. 283-292.

[19] S. Beydeda, M. Book, V. Gruhn, G. Booch, A. Brown, S. Iyengar, J. Rumbaugh and B. Selic. *Model-Driven Software Development*, vol. 15. Berlin: Springer, 2005.

[20] J. P. Tolvanen and S. Kelly. "Model-driven development challenges and solutions." *Modelsward*, vol. 2016, p. 711, 2016.

[21] R. B. France, S. Ghosh, T. Dinh-Trong and A. Solberg. "Model-driven development using UML 2.0: Promises and pitfalls." *Computer*, vol. 39, no. 2, pp. 59-66, 2006.

[22] S. J. Mellor, T. Clark and T. Futagami. "Model-driven development: Guest editors' introduction." *IEEE Software,* vol. 20, no. 5, pp. 14-18, 2003.

[23] P. Mayer, A. Schroeder and N. Koch. “MDD4SOA: Model-driven service orchestration.” in Enterprise Distributed Object Computing Conference, 2008. EDOC’08. 12th International IEEE, 2008, pp. 203-212.

[24] D. Harel, B. Rumpe. “Modeling languages: Syntax, semantics and all that stuff (or, what’s the semantics of semantics?).” in Technical Report MCS00-16, Weizmann Institute, Rehovot, Israel, 2004.

[25] N. B. Ruparelia. “Software development lifecycle models.” *SIGSOFT Software Engineering Notes*, vol. 35, no. 3, pp. 8-13, 2010.

ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Building Kurdish Chatbot Using Free Open Source Platforms

Kanaan M. Kaka-Khan

Department of Computer Science, University of Human Development, Iraq

ABSTRACT

Chatbot is a program that utilizes natural language understanding and processing technology to have a human-like conversation. Nowadays chatbots are capable to interact with users in world's majority languages. Unfortunately, bots that interact with Kurdish users are rare. This paper is an attempt to bridge the gap between chatbots and Kurdish users. This paper tries to implement a free open source platform (pandorabots) to build a Kurdish chatbot. I present a number of challenges for Kurdish chatbot at the last section of this work.

Index Terms: Artificial Intelligence, Artificial Intelligence Markup Language, Chatbot, Pandorabots

1. INTRODUCTION

A. Chatbot

A chatbot is a service, powered by rules and sometimes artificial intelligence that you interact with via a chat interface [1,2]. They range from simple systems that extract a response from databases when they match certain keywords to more sophisticated ones that use natural language processing techniques [3].

B. Needs for Chatbot

And an extraordinary focus was devoted to chatbots within the tech community in recent years [4]. There is no doubt that majority of business are going to be online; if we want to make a business online we have to locate where the people are? That place now is the zone of messenger applications as mentioned by Peter Rojas "People are now spending more time in messaging apps than in social media and that is a huge turning point. Messaging apps are the platforms of the future and bots will be how their users access all sorts of services" [5]. Any user's interaction with an app or web page can utilize a Chatbot to increase the user's experience [6].

Fig. 1 shows the size of the top 4 messaging apps and social networks; big 4 messaging apps are Whatsapp, Messenger, WeChat, Viber, big 4 social networks are Facebook, Instagram, Twitter, and LinkedIn [7].

C. Applications of Chatbot

The very basic use at the early days of chatbot was almost restricted to conversations. The first chatbot in history was Eliza, a program which represents a psychologist [8]. By the time the bot provides a wide range to many important applications, some of the most important applications of chatbots are listed below:

1. Customer service
2. Mobile personal assistants
3. Advertisements
4. Games and entertainment applications
5. Talking toys
6. Call centers.

Access this article online

DOI: 10.21928/uhdjst.v1n2y2017.pp46-50

E-ISSN: 2521-4217
P-ISSN: 2521-4209

Corresponding author's e-mail: kanaan.mikael@uhd.edu.iq

Received: 09-08-201 Accepted: 24-08-2017 Published: 30-08-2017

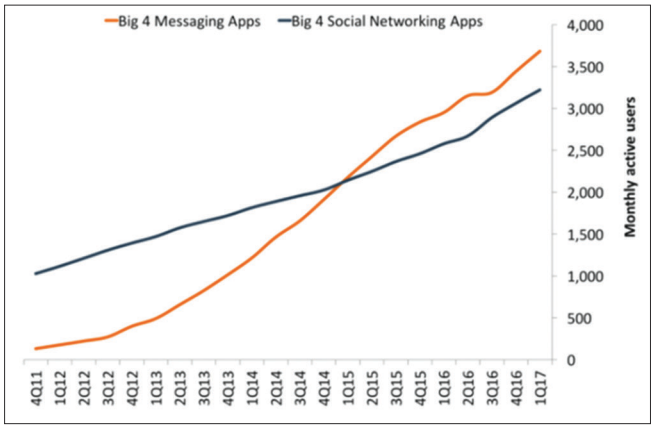

Fig. 1. Users for top 4 messaging apps and social networks in million [7]

The crucial aim of this work is to build a bot that is capable of working as a guide who is sitting on the UHD website and giving information about the University of Human Development to any user whenever asked.

2. CHATBOT HISTORY

The concept of natural language processing generally and chatbots specifically can be originated to Alan Turing question "Can machines think?" who asked in 1950 [9]. Alan's question (which is called Turing Test now) is nothing just asking questions to human and machine subjects, to identify the human. We say the machine can think if the human and machine responses are indistinguishable. In 1966, Eliza (the first chatbot) was created by Joseph Weizenbaum at MIT. For generating proper responses, Eliza uses a set of pre-programmed rules to identify keywords and pattern match those keywords from an input sentence [8]. In 1995, a new more complex bot (A.L.I.C.E) created by Richard Wallace. ALICE makes use of artificial intelligence markup language (AIML) to represents conversations as sets of patterns (inputs) and templates (outputs). ALICE got Loebner prize (yearly chatbot competition) thrice and award the most intelligent chatbot [10]. Advances in natural language processing and machine learning played important roles in improving chatbot technology; modern chatbots include Microsoft's Cortana, Amazon's Echo and Alexa, and Apple's [11].

3. RELATED WORKS AND METHODOLOGY

As in many natural language processing applications, there are many approaches to developing chatbot: Using a set of predefined rules [12], semi automatically learning conversational pattern from data [13], and full automatic chatbot (under researching). Each approach has its own merits and demerits, through manual approach more control over the language and the chatbot can be achieved, but it needs more effort to maintain a huge set of rules. The second approach which also is called corpus-based is challenged by the need to construct coherent personas using data created by different people [Botta]. Due to lack of Kurdish corpus (at least it is not available for me even if it exists), I chose manually written rules by making use of AIML, a popular programming language to represents conversations as a set of patterns (inputs) and templates (outputs).

As in other NLP applications, in the area of Kurdish chatbot, unfortunately, we find related works rarely. With the best of my knowledge this is the first Kurdish chatbot which is created academically, so sometimes I obliged to relate my work with Arabic or Persian languages. Most notably, in 2016, Dana and Habash developed Botta, the first Arabic dialect Chatbot, Botta explore the challenges of creating a conversational agent that aims to stimulate friendly conversations using the Egyptian Arabic dialect [3].

Playground and programming language are the two basic requirements for creating chatbots. Playground can be defined as a sandbox or an integrated development environment for the programming language [1]. In this work, I chose pandorabots as a playground (creating, deploying, talking with the bot) and AIML (for Making conversation) as a programming language for creating Kurdish chatbot, ALICE, an award-winning free chatbot was created using AIML [12].

After login into pandorabots playground with Facebook account, the work will be shown in the following steps:

- Step 1: I gave "kuri zanko" as the bot name.
- Step 2: In the bot editor space, I created a file named "UHD" which is AIML file to involve all the patterns (inputs) and templates (outputs).
- Step 3: I started writing an expected user input in <pattern></pattern> tag and the bot answer in <template></template> tag, both pattern and template are enclosed in a <category></category>, a category is the basic unit of knowledge in AIML [1].
- Step 4: After writing each category, I train (test) the bot to know whether it gives the correct answer.
- Step 5: After writing all the categories, the bot will be published in the pandorabots clubhouse (a public place where users can talk to the bots).

4. RESULT AND DISCUSSION

For the simple and direct user input the bot can give the answer easily, for example:

User: سڵاو

Bot: سڵاو لە بەڕێزتان،خۆتان بناسێنن

A. Pattern Matching

To form a user input matching, the bot searches through its AIML file (categories). It may happen, a user input does not match any of the pattern defined in our bot, so a default answer should be provided which is called ultimate default category:

<pattern>*</pattern>

<template>

ببورە بەڕێزم، وەڵامی پرسیارەکەتم لانیە

</template>

The star (*) determines that a user input does not match any of the bot patterns, relying on one default answer is extremely tedious for the clients. This obliges us to think about random responses to provide different responses for the same user input.

<random>

ببورە بەڕێزم، وەڵامی پرسیارەکەتم لانیە

بەڕێزم پرسیارەکەت بەجۆرێکی تر بکەرەوە

بەڕێزم پرسیارەکەت ڕون نیە

ببورە لە پرسیارەکەت نەگەشتم

</random>

These random responses make sense that the user is chatting with a human, not a bot.

B. Wildcards

Wildcards are used to capture many inputs using only a single category [1]. Through wildcards bots can be more intelligence. There are many wildcards but (* and ^) are the most two ones which are used in this work:

<pattern>ناوم *</pattern

In this example, the star(*) stands for any name that is given by the user.

<pattern>زانکۆی گەشەپێدان *</pattern>

In the second example, the star stands for any words or sentences which appear after the name "زانکۆی گەشەپێدان".

<pattern>^ کۆمپیوتەر ^</pattern>

The (^) wildcard lets the bot to capture any input containing the word "کۆمپیوتەر" and gives a proper answer.

Wildcards should be used carefully because their priority is different, Fig. 2 shows wildcard and exact matching priorities.

A category with # wild card will be matched first and * wildcard will be matched last, for example: When a user even types "سڵاو لە ئێوە"the response will be taken from "سڵاو#" pattern not "سڵاو لە ئێوە" pattern.

C. Variables

Bot intelligence can also be achieved through variables. Variables can be used to store information about your bot and the users; this gives the user a sense that he/she is chatting with a human being. Fig. 3 shows a short conversation between my bot and a user.

D. Recursion

Recursion means writing a template that is calling another category, and this leads to minimizing the number of categories in our bot AIML file.

<pattern>های</pattern>

<template><srai>سڵاو</srai></template>

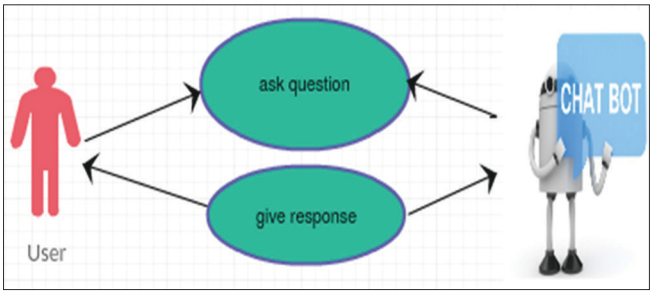

Fig. 2. Chatbot simple flow diagram

سڵاو * < سڵاو ^ < سڵاو لە ئێوە < سڵاو _ < سڵاو # <

Fig. 3. Wildcards priority

Through using recursion, no need to rewrite a new category to input "های", we just refer to the template "سڵاو" using <srai> tag, and the bot answers the user exactly as he/she said "سڵاو" to the bot.

E. Context

To make our bot capable of doing human-like conversation, it should remember the things that have been previously said. My bot is capable of remembering the last sentence it said. (Fig. 4-6) shows different conversations regarding context.

F. Challenges

- Challenge 1: The first and greatest challenge for Kurdish Chatbot is the lack of platform designed specifically to Kurdish Language, Kurdish structure extremely differs from English or any other languages, Kurdish word order is SOV [subject+ object+ verb] [14]. The reason behind the slow progress in Arabic NLP is the complexity of the Arabic language [3], same to Kurdish. Hence, it is very tough to have a very intelligent Kurdish bot using free open source platforms.
- Challenge 2: Dialectal Variation, Kurdish language has many different dialects; the gap among dialects sometimes reaches a level that speakers of a dialect do not understand another dialect, and it means that it is quite tough to build a bot capable of chatting with all different Kurdish dialects.
- Challenge 3: Normalization is one of the important processes in developing bots, normalization includes sentence splitting, correcting spelling errors, person, and gender substitution.

Fig. 4. A sample conversation between a user and the bot

Fig. 5. A sample conversation regarding context

wanna -> want to

isn't -> is not

How R U -> How Are You

With you -> with me

The user may be bad in spelling, he/she may type "how r u" instead of "how are you". These changes (normalization

Fig. 6. Detailed conversation between a user and the bot

and substitution) can be done easily in English and make the bot to interact with the user as a human not a bot, while it's a bit difficult to perform the same for Kurdish because the bot components (AIML files, Set files, and Map Files) are already exist for English language while not for Kurdish, it requires vast effort from both computer science and linguistic people to maintain such files.

- Challenge 4: In spite of majority of platforms claiming for language agnosticism, practically we face issues for Kurdish due to its own structure. For example, when a name is given, as "Alan" to the bot and later on he asks the bot about his name it says "your name is Alan." While the same name is given in Kurdish language"ئالان" to the bot and I ask the bot for his name, it should tell "تۆ ناوت ئالانە" a suffix will be seen "ە" with the name "ئالان", this seems to be an easy task but really needs a hard work to do.

5. CONCLUSION AND FUTURE WORK

Chatbots are online human-computer dialog system[s] with natural language [15]. I have presented the first Kurdish chatbot and described some of the challenges for Kurdish chatbot. Building chatbot from scratch is extremely tough, time consuming, costly. This reason led me to go for free open source platform (pandorabots). This work aims to be a basic structure for Kurdish dialect, providing future Kurdish bot masters with a base chatbot which contains basic files, general knowledge.

6. BIOGRAPHY

Kanaan M. Kaka-Khan is an associate professor in the Computer Science Department at Human Development University, Sulaimaniya, Iraq. Born in Iraq 1982. Kanaan M. Khan had his bachelor degree in Computer Science from Sulaimaniya University, and Master Degree in IT from BAM university, India. His research interest area includes Natural Language Processing, Machine Translation, Chatbot, and Information Security.

REFERENCES

[1] "How to Build a Bot using the Playground UI". Available: https://www.playground.pandorabots.com/en/tutorial. [Last Accessed on 2017 Aug 25].

[2] "The Complete Beginner's Guide to Chatbots." Matt Schlicht, Founder of Chatbots Magazine, Apr. 20, 2016. Available: https://www.chatbotsmagazine.com/the-complete-beginner-s-guide-to-chatbots-8280b7b906ca. [Last Accessed on 2017 Aug 25].

[3] "Botta: An Arabic Dialect Chatbot." Dana Abu Ali and Nizar Habash, Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, Osaka, Japan, pp. 208-212, Dec. 11, 17, 2016.

[4] "Best uses of Chatbots in the UK." Charlotte Jee. Available: http://www.techworld.com/picture-gallery/apps-wearables/9-best-uses-of-chatbots-in-business-in-uk-3641500. Jun. 08, 2017.

[5] "Chatbot Survey 2017." Ayush Jain, Co-founder and CEO at Mindbowser. Available: https://www.slideshare.net/Mobileappszen/chatbots-survey-2017-chatbot-market-research-report. [Feb. 08, 2017.

[6] "Chatbot Applications and Considerations." Josef Ondrejcka. Available: http://ramseysolutions.com/chatbot-applications-and-considerations. [Sep. 19, 2016].

[7] "Messaging Apps are Now Bigger than Social Networks." BI Intelligence. Available: http://www.businessinsider.com/the-messaging-app-report-2015-11. [Sep. 20, 2016].

[8] J. Weizenbaum. "ELIZA-a computer program for the study of natural language communication between man and machine." *Communications of the ACM,* vol. 9, no. 1, pp. 36-45, 1966.

[9] A. M. Turing. "Computing machinery and intelligence." *Mind*, vol. 59, no. 236, pp. 433-460, 1950.

[10] R. S. Wallace. "The Anatomy of A.L.I.C.E." Available: http://www.alicebot.org/anatomy.html. [Last Accessed on 2017 Aug 25].

[11] M. Weinberger. Why Amazon's Echo is Totally dominating-and what Google, Microsoft, and Apple have to do to Catch Up. Available: http://www.businessinsider.com/amazon-echo-google-home-microsoft-cortana-apple-siri-2017-1. [Jan. 14, 2017].

[12] R. Wallace. *The Elements of AIML Style,* San Francisco: Alice AI Foundation, 2003.

[13] B. A. Shawar and E. Atwell. "Using dialogue corpora to train a chatbot." In Proceedings of the Corpus Linguistics 2003 Conference, pp. 681-690, 2003.

[14] "Evaluation of in Kurdish machine translation system." Kanaan and Fatima, Proceedings of UHD 2017, the 4th International Scientific Conference, Sulaimanya, Iraq, pp. 862-868, Jun. 2017.

[15] J. Cahn. "CHATBOT: Architecture, design, and development." University of Pennsylvania School of Engineering and Applied Science Department of Computer and Information Science, Apr. 26, 2017.

p-ISSN 2521-4209
e-ISSN 2521-4217

جامعة التنمية البشرية
UNIVERSITY OF HUMAN DEVELOPMENT

UHD Journal of Science and Technology

A Scientific periodical issued by University of Human Developement

Vol.1 No.(2) December 2017

2017 2717

e.mail:jst@uhd.edu.iq