Natarajan Meghanathan
David C. Wyld (Eds)

Computer Science & Information Technology

7th International Conference on Signal, Image Processing and Pattern Recognition (SPPR-2018), December 22-23, 2018, Sydney, Australia

AIRCC Publishing Corporation

Volume Editors

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

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

ISSN: 2231 - 5403
ISBN: 978-1-921987-95-3
DOI : 10.5121/csit.2018.81701 - 10.5121/csit.2018.81716

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

Preface

The 7th International Conference on Signal, Image Processing and Pattern Recognition (SPPR-2018), was held in Sydney, Australia during December 22-23, 2018. The 7th International Conference on Soft Computing, Artificial Intelligence and Applications (SCAI-2018), The 9th International Conference on Communications Security & Information Assurance (CSIA-2018), The 10th International Conference on Wireless, Mobile Network & Applications (WiMoA-2018), The 8th International Conference on Computer Science, Engineering and Applications (ICCSEA-2018), The 9th International Conference on Internet Engineering & Web Services (InWeS-2018), The 7th International Conference of Networks and Communications (NECO-2018) and The 10th International Conference on Grid Computing (GridCom-2018) was collocated with The 7th International Conference on Signal, Image Processing and Pattern Recognition (SPPR-2018). The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

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

The SPPR-2018, SCAI-2018, CSIA-2018, WiMoA-2018, ICCSEA-2018, InWeS-2018, NECO-2018, GridCom-2018 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in computer network and communications research.

In closing, SPPR-2018, SCAI-2018, CSIA-2018, WiMoA-2018, ICCSEA-2018, InWeS-2018, NECO-2018, GridCom-2018 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the SPPR-2018, SCAI-2018, CSIA-2018, WiMoA-2018, ICCSEA-2018, InWeS-2018, NECO-2018, GridCom-2018.

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

Natarajan Meghanathan
David C. Wyld

Organization

General Chair

Natarajan Meghanathan	Jackson State University, USA
David C. Wyld	Southeastern Louisisna University, USA

Program Committee Members

Abdelmajid Hajami	FST Settat, Morocco
Abdulhamit Subasi	Effat University, Saudi Arabia
Ahmed Korichi	University of Ouargla, Algeria
Alessio Ishizaka	University of Portsmouth,United Kingdom
Annamalai	Prairie View A&M University, USA
Ashish Tanwer	Cisco Systems, United states
Azeddine Chikh	University of Tlemcen, Algeria
Barhoumi Walid	SIIVA-LIMTIC Laboratory, ENICarthage, Tunisia
Benyamin Ahmadnia	Universitat Autònoma de Barcelona,Spain
Bin Cao,Hebei	University of Technology,P.R. China
Carlo Sau	Universit degli Studi di Cagliari, Italy
Chaker LARABI	Universit de Poitiers , France
Chin-Chen Chang	Feng Chia University, Taiwan
Chuanzong Zhang	Aalborg University, Denmark
Dac-Nhuong Le	Haiphong University, Vietnam
Dac-Nhuong Le	Haiphong University, Vietnam
Dalel BOUSLIMI	Institut Mines- Telecom, France
Daniel Ekpenyong Asuquo	University of Uyo, Nigeria
Duan Keqing	Wuhan Early Warning Academy, China
Emad Awada	Applied Science University, Jordan
Fabio Gasparetti	Roma Tre University, Italy
Felix Yang LOU	City University of Hong Kong,Hong Kong, China
Hacer Yalim Keles	Ankara University, Turkey
Hamid Ali Abed AL-Asadi	Basra University, Iraq
Hanming Fang	Logistical Engineering University, China
Hassan Ugail	University of Bradford, UK
Issac Niwas Swamidoss	Nanyang Technological University, Singapore
Jose-Luis Verdegay	University of Granada,Spain
Kala Marapareddy	The University of Southern Mississippi, USA
Kimia Rezaei	Islamic Azad University, Iran
Klimis Ntalianis	Athens University of Applied Sciences, Greece
Lark Kwon Choi	The University of Texas at Austin, USA
Linkai Luo	Hang Seng Management College,Hong Kong, China
Longzhi Yang	Northumbria University,UK
Maciej Kusy	Rzeszow University of Technology,Poland
Mohamed Anis Bach Tobji	University of Manouba, Tunisia
Mohammad Javad Mahmoodabadi	Sirjan University of Technology,Sirjan, Iran

Naresh Doni Jayavelu	University of Washington,Seattle, WA
Oscar Mortagua Pereira	University of Aveiro,Portugal
Pavel Loskot	Swansea University, Canada
Pawel Karczmarek	The John Paul II Catholic University ,Poland
Pietro Ducange	SMARTEST Research Centre eCampus University,Italy
Poonam Tanwar	Manav Rachna International University,India
Qiong Zhou	California Data Science, Inc. Palo Alto, CA
Saman Babaie-Kafaki	Semnan University,Semnan, Iran
Samer Al Martini	Abu Dhabi University,UAE
Shoeib Faraj	Institute of Higher Education of Miaad, Iran
Smain FEMMAM	UHA University,France
Tarik YERLIKAY	Trakya University,Turkey
Wai Lok Woo	Newcastle University,United Kingdom
Weili Zhang	eBay Inc,San Jose, CA, US
Wonjun Lee	The University of Texas at San Antonio, USA
Yunliang JIANG	Huzhou University,P.R.China
Zhao Peng	Huazhong University of Science and Technology, China
Zhongsheng Hou	Beijing Jiaotong University,China
Aftab Alam	King Khalid University, Kingdom of Saudi Arabia
I-Cheng Chang	National Dong Hwa University, Taiwan
Juan A. Fraire	Universidad Nacional de Crdoba, Argentina
Mohamed Ashik M	Salalah College of Technology, Oman
Nadia Qadri	University of Essex, United Kingdom
Peiman Mohammadi	Islamic Azad University, Iran
Rekha P	Amrita University, India
Rim Haddad	Science Faculty of Bizerte, Tunisia
Samadhiya	National Chiao Tung University, Taiwan
Sergey Muravyov	Tomsk Polytechnic University, Russia
Sergio Pastrana	University Carlos III of Madrid, Spain
Seyyed AmirReza Abedini	Technical and vocational University, Iran
Seyyed Reza Khaze	Islamic Azad University, Iran
Solomiia Fedushko	Lviv Polytechnic National University, Ukraine
Thuc-Nguyen	University of Science, Vietnam
Ying Feng	University of Alabama, USA
Yungfahuang	Chaoyang University of Technology, Taiwan
Yuriy Syerov	Lviv National Polytechnic University, Ukraine
Yusmadi jusoh	Universiti Putra Malaysia, Malaysia
Zebbiche Toufik	University of Blida, Algeria
Hamid Ali Abed AL-Asadi	Basra University, Iraq
Mohamed Fahad AlAjmi	King Saud University, Saudi Arabia
Selwyn Piramuthu	University of Florida, Florida
Sergio Pastrana	University Carlos III of Madrid, Spain
Seyyed AmirReza Abedini	Islamic Azad University, Iran
Thuc-Nguyen	University of Science, Vietnam
Amol D Mali	University of Wisconsin, USA
Ankit Chaudhary	Truman State University, USA

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)

Networks & Communications Community (NCC)

Soft Computing Community (SCC)

Organized By

Academy & Industry Research Collaboration Center (AIRCC)

TABLE OF CONTENTS

7th International Conference on Signal, Image Processing and Pattern Recognition (SPPR-2018)

7th International Conference on Soft Computing, Artificial Intelligence and Applications (SCAI-2018)

9th International Conference on Communications Security & Information Assurance (CSIA-2018)

10th International Conference on Wireless, Mobile Network & Applications (WiMoA-2018)

8th International Conference on Computer Science, Engineering and Applications (ICCSEA-2018)

9th International Conference on Internet Engineering & Web Services (InWeS-2018)

7th International Conference of Networks and Communications (NECO-2018)

10th International Conference on Grid Computing (GridCom-2018)

TOMOGRAPHIC SAR INVERSION FOR URBAN RECONSTRUCTION

Karima Hadj-Rabah[1], Faiza Hocine[2], Assia Kourgli[3] and Aichouche Belhadj-Aissa[4]

[1,2,3,4]Department of Telecommunications, University of Sciences and Technology Houari Boumediene (USTHB), Algiers, Algeria

ABSTRACT

Given its efficiency and its robustness in separating the different scatterers present in the same resolution cell, SAR tomography (TomoSAR) has become an important tool for the reflectivity reconstruction of the observed complex structures scenes by exploiting multi-dimensional data. By its principle, TomoSAR reduces geometric distortions especially the layover phenomenon in radar scenes, and thus reconstruct the 3D profile of each azimuth-range pixel. In this paper, we present the results and the comparative study of six tomographic reconstruction methods that we have implemented. The analysis is performed with respect to the separability and location of scatterers by each method, supplemented by the proposal of a quantitative analysis using metrics (accuracy and completeness) to evaluate the robustness of each method. The tests were applied on simulated data with TerraSAR-X sensor parameters.

KEYWORDS

SAR Tomography (TomoSAR), Reconstruction Algorithms, Accuracy& Completeness

1. INTRODUCTION

In the last two decades, SAR tomography (TomoSAR) had a growing interest in remote sensing, particularly after the acquisition of very high resolution (VHR) data acquired by the latest generation of SAR radar sensors such as: TerraSAR-X and CosmoSky-Med. TomoSAR is a new data acquisition method, it exploits a series of SAR images taken with slightly different view angles to reconstruct the 3D profile of the reflectivity function for each azimute-range pixel [1].

The choice of a reconstruction method leading to conclusive results depends, on the one hand, on the nature of the area intended to be analyzed: forested, urban, etc and on the other hand, on the type of data and their characteristics in terms of spatial and temporal resolution. In the case of VHR images, it is common to adopt a deterministic scattering model to model a scene due to the high density of point scatterers with a high signal-to-noise ratio (SNR) present in the latter [2]. However, by its principle, SAR tomography exploits pairs of SAR images acquired simultaneously and / or at different time intervals, so that the temporal and spatial decorrelation problems, the atmospheric delay and the noise of various sources are the main tomographic process limitations and the inaccuracy of reconstructed scenes in terms of amplitude and altimetry.

Several parametric and non-parametric, single- and multi-looking SAR tomographic reconstruction methods have been developed and implemented in literature. Classical Fourier-

Natarajan Meghanathan et al. (Eds) : SPPR, SCAI, CSIA, WiMoA, ICCSEA, InWeS, NECO, GridCom - 2018
pp. 01–08, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.81701

Based focusing and SVD algorithm were the first approaches applied to simulated and real data acquired in the C-band by the ERS sensor [4] [5]. However due to the non-regularity of the acquisitions distribution on the Baseline axis, these two approaches introduce a degradation of the PSF (Point Spread Function) reconstruction according to the elevation in terms of resolution and side lobes. As an alternative to these two classical approaches, CAPON and MUSIC have been proposed in [6] [7] to ensure super-resolution (SR) and good side lobes suppression. Nevertheless, the use of the estimated covariance matrix represents the major disadvantage of CAPON and MUSIC. In [8] [9], Non-Linear Least-Square was proposed despite its high cost, because it provides good accuracy in altitude (Height accuracy).

In addition, the evaluation of the reconstruction methods mentioned above was carried out qualitatively. In [7], the authors used root mean square error (RMSE) to estimate the performance of tomographic reconstruction. Two new metrics have been adapted to TomoSAR in [2] and [10] to define root mean square accuracy $RMSE_A$and root mean square completeness $RMSE_C$, the two latter are based on distortion measure by the Euclidean distance in order to measure the error between the estimated targets (estimated scatterers) and the real targets (ground truth targets).

The objective of this work is to present a comparative study between the most widely used reconstruction approaches in the literature, without taking into consideration the sparse property of the radar signals. This study is based on a quantitative evaluation of the treatment results by introducing other criterias to calculate the accuracy and the completeness of the target location in elevation and the restitution of their reflectivities (their amplitudes).

The remainder of the paper is organized in the following way: the next section briefly describes the TomoSAR geometry, in section 3 we will look at the characteristics description of the reconstruction approaches studied, this section will also include a detailed analysis of the evaluation methods suggested by the authors, then the analysis results as well as a comparison between the different approaches will be presented in section 4, and finally, the paper ends with a conclusion.

2. TOMOSAR GEOMETRY AND IMAGING MODEL

Due to side-looking geometry, the projection of 3D objects on the plane (azimuth *x*, range *r*) introduces geometric problems in urban and uneven terrain areas causing an ambiguity when interpreting SAR images. To remedy these distortions, TomoSAR makes it possible to separate several scatterers located at different altitudes present in the same resolution cell, by projecting the scatterers responses on an axis perpendicular to the 'azimuth-range' plane. These responses are reconstructed from N SAR images taken at different view angles (see Fig.1). The complex valueg_n of a coordinate pixel (x, r) for the nth acquisition with n =1,..,N is a sample of the Fourier transform of the reflectivity function $\gamma(s)$ with respect to the elevation s, its expression is [1]:

$$g_n = \int_{\frac{-\Delta_s}{2}}^{\frac{\Delta_s}{2}} \gamma(s) \exp(-j2\pi\xi_n s)\, ds + w_n \quad (1)$$

With: Δ_s is the elevation interval, $\xi_n = -2b_n/\lambda r$represents the spatial frequency which depends on the sensor position $b_n$on the Baseline axis, λ is the wavelength and w_n is the noise, the latter follows a Gaussian distribution [11].

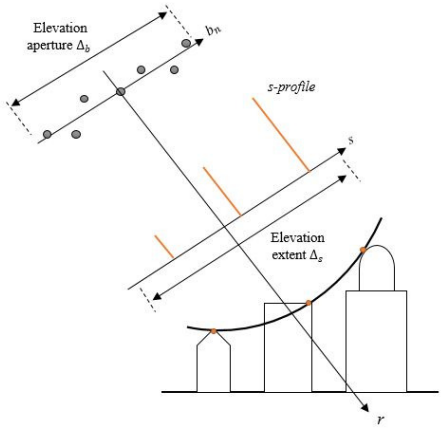


Figure 1. TomoSAR geometry

Equation (1) can be modeled by the following linear system:

$$g = AY + W \tag{2}$$

With: g is the observations vector of length N, A is the steering vectors matrix of size NxM, and Y is the reflectivity profile uniformly sampled at s_m with m= 1,.., M. To ensure SR, M must be very large ($M \gg N$). Therefore, the s-profile recovery *i.e.* the Y profile reconstruction is a spectral estimation problem.

3. TOMOGRAPHIC RECONSTRUCTION APPROACHES AND EVALUATION METHODS

3.1. Reconstruction Approaches

Tomographic reconstruction approaches can be classified into: non-parametric approaches and parametric approaches, or single-looking and multi-looking approaches. Non-parametric approaches allow the estimation of power spectral density whose statistical properties are unknown in prior, by passing the observed data through a set of bandpass filters in order to estimate the output power, while parametric approaches model the observed data by a few sinusoids in order to estimate their parameters. A better estimation can be offered by the latter only if the suggested model is close to the processed data [12].

Single-looking approaches do not exploit the correlation between the pixel set to be processed and its neighborhood, whereas multi-looking approaches require the covariance matrix estimation from the observations vector ($\hat{R}$), this ensures a SR in elevation but with a considerable loss of the reconstructed scenes spatial resolution.

The most implemented reconstruction methods in literature are Conventional Beamforming (CBF), Beamforming (BF), Adaptive Beamforming (CAPON), TSVD, MUSIC and Non-linear Least Square (NL-LS), their characteristics description is summarized in Table 1: their equations, their ranking (parametric or non-parametric, single- or multi-looking), their resolution as well as their advantages and disadvantages.

Table 1. Characteristics of Tomographic reconstruction methods

Method	Parametric	Multi-looking	Equation	Elevation resolution	Advantages and Disadvantages
Conventional Beamforming (CBF)	No	No	$\hat{Y} = A^H g$	Poor	- Simple inverse Fourier transform. - No SR. - Presence of Side lobes.
Beamforming (BF)	No	Yes	$\hat{Y} = \frac{A^H \hat{R} A}{N^2}$	Poor / Average	- Biased estimator when Number of scatterers is greater than 1.
Adaptive Beamforming (Capon)	No	Yes	$\hat{Y} = \frac{\hat{R}^{-1} A}{A^H \hat{R}^{-1} A} . g$	Average	- Requires a big number of looks. - Better scatterers separation but the amplitude estimation is biased.
TSVD	No	No	$\hat{Y} = \left(V . \frac{1}{S} . U^H\right) . (G_{m,n} . g)$ With: $G_{m,n} = \Delta_s . sinc(\Delta_s(\xi_m - \xi_n))$ And $[U, S, V] = svd(A)$	Poor / Average	- Stable for well-defined problems (well-defined truncation rank T).
MUSIC	Yes	Yes	$\hat{Y} = \left\lvert \frac{1}{A^H G^H G A} \right\rvert^2$ With G the noise subspace.	High	- Better elevation resolution and weak side lobes.
Non-Linear Least-Square (NL-LS)	Yes	No	$\hat{Y} = \min_k \{\lVert g - AY \rVert_2^2\}$ $= (A^H A)^{-1} A^H g$	High / Very High	- Better amplitude estimation (if (ξ) is precised).

3.2. Evaluation Methods

Our comparative study is based on a quantitative evaluation that uses two metrics: precision and completeness, recently proposed in [2]. Precision provides a measure that describes how the estimated targets $\hat{P}$are close to the real targets$P$, its expression is as follows:

$$A_{dist} = \frac{1}{N_p} \sum_{j=1}^{N_p} \min_k dist(\hat{P}_j, P_k) \tag{3}$$

While completeness is a measure of how real targets are modeled by estimated targets, its expression is:

$$C_{dist} = \frac{1}{N'_p} \sum_{j=1}^{N'_p} \min_k dist(\hat{P}_k, P_j) \tag{4}$$

WithN_p and $N'_p$are the number of estimated and real targets respectively, $dist$ represents the distortion measure.

We adapted these two measurements to our analysis by estimating the elevation position of each target scatterer as well as its reflectivity (amplitude) separately, therefore, we defined Elevation

Accuracy, Amplitude Accuracy, Elevation Completeness, and Amplitude Completeness. In addition, in order to obtain more consistent results, we used the Manhattan distance (D1) and the Euclidean distance (D2) to measure the distortion.

4. RESULTS AND DISCUSSION

Our study includes three phases, the first one consists of simulating tomographic SAR data, the second step involves the implementation of reconstruction methods, and the last phase is dedicated to evaluating the obtained results.

The targets reconstruction tests, by each algorithm, were performed on a simulated profile along the elevation axis from the TerraSAR-X satellite parameters (see Table 2), assuming that the scatterers number is equal to 3 in one resolution cell (see Fig. 1) with a SNR of 10 dB.

Table 2. TerraSAR-X parameters.

Distance from the scene center [m]	740000
Δ_s [m]	269,5
λ [m]	0,031
N	25

A synthesis aperture in elevation was effectuated, by the following with random samples distribution scheme on the Baseline axis, this samples distribution is represented in the graph of Figure 2. We performed an implementation of the tomographic reconstruction methods described in the previous section, the resulting profiles are shown in Figure 3.

The evaluation results of the different reconstructions accuracy are shown in Tables 3, 4 and 5. The elevation and amplitudes accuracy and completeness were calculated after applying a 3-peaks detector on the curves of Figure 3 to compare between them.

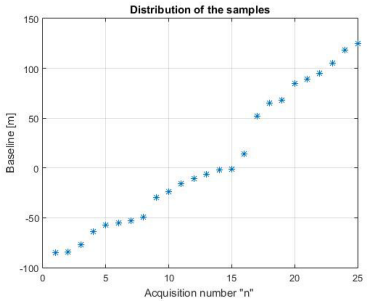
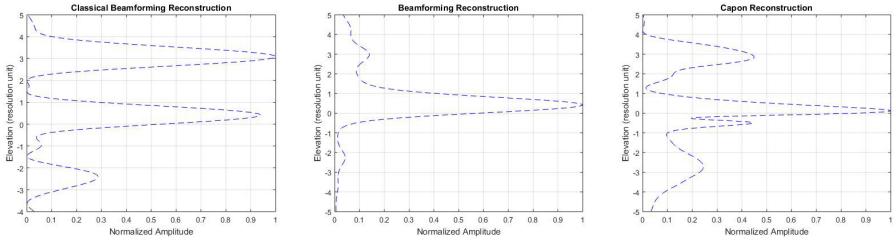
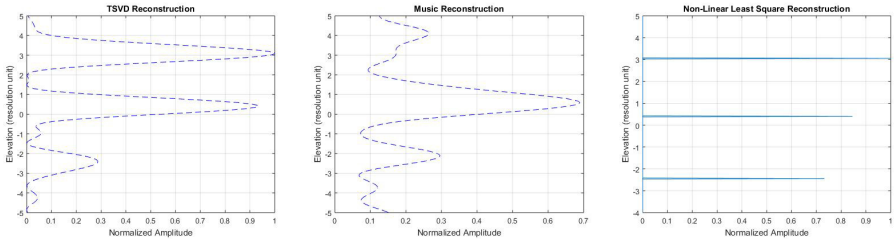


Figure 2. Samples distribution scheme



(a) (b) (c)



(d) (e) (f)

Figure 3. Reflectivity profile reconstruction by :(a) CBF, (b) BF,(c) CAPON, (d) TSVD, (e) MUSIC and (f) NL-LSmethod

Table 3. RMSE values of the different reconstruction approaches.

Approach	**CBF**	**BF**	**CAPON**	**TSVD**	**MUSIC**	**NL-LS**
RMSE	0.0824	0.0714	0.0759	0.0823	0.0657	0.0856

Table 4. Elevation Accuracy and completeness of the different reconstruction approaches with a normalization factor of (10^3).

Approach		**CBF**	**BF**	**CAPON**	**TSVD**	**MUSIC**	**NL-LS**
Elevation Accuracy	**D1**	0.0027	0.0143	0.0217	0.0023	0.0057	0.0027
	D2	0.0073	0.4857	0.6897	0.0057	0.0417	0.0073
Elevation Completeness	**D1**	0.0027	0.0410	0.0340	0.0023	0.0057	0.0027
	D2	0.0073	4.6457	2.1820	0.0057	0.0417	0.0073

Table 5. Amplitude Accuracy and completeness of the different reconstruction approaches.

Approach		**CBF**	**BF**	**CAPON**	**TSVD**	**MUSIC**	**NL-LS**
Amplitude Accuracy	**D1**	0.0929	0.2581	0.0388	0.0946	0.1925	0.0377
	D2	0.0166	0.1001	0.0020	0.0167	0.0504	0.0042
Amplitude Completeness	**D1**	0.1407	0.2091	0.1058	0.1383	0.1817	0.0778
	D2	0.0293	0.0668	0.0233	0.0283	0.0449	0.0180

The RMSE values of the estimated scatterers by each of the methods described above are presented in Table 3. The latters are close to each other, which leaves us to conclude that the different reconstructions have practically the same precision.

Although the three scatterers were well separated by all the approaches according to the graphs of Figure 3, we note that compared to the 'BF', 'CAPON' and 'MUSIC' approaches, the scatterers localization on the elevation axis is erroneous contrary to the 'CBF', 'TSVD' and 'NL-LS' approaches which effectively allow the reflectivity profile reconstruction with a good estimation of the scatterers location on the elevation axis as well as their amplitudes. We emphasize that the erroneous localization given by 'BF', 'CAPON' and 'MUSIC' is mainly due to the poor covariance matrix estimation of the observations vector caused by the strong noise presence and consequently a non-precise reconstruction. Hence, the evaluation by the RMSE calculation is not the most judicious choice.

To corroborate these graphical results, we calculated the accuracy and completeness using the expressions described in the previous section. The results of these metrics are given in Table 4. The analysis of these measurements shows that the best elevation accuracy and completeness values are those of 'CBF', 'TSVD' and 'NL-LS' methods. Also, from the results of Table 5, we notice that the 'NL-LS' approach has the best amplitude accuracy and completeness values. These results are in conformity with those of Figure 3. We can conclude that the metrics 'Elevation

Accuracy', 'Amplitude Accuracy', 'Elevation Completeness' and 'Amplitude Completeness' are a tool for performance evaluation of SAR tomographic reconstruction. This observation makes it easier to choose the most appropriate reconstruction approach depending on the surface and / or the sub-surface to be reconstructed. In this sense, if only one scatterer is present in a resolution cell, all the previously described methods reconstruct the target with very good accuracy. However, in the case of multiple scatterers, in an urban area for example it is very important to preserve the spatial resolution in order to be able to observe the urban infrastructures, in this case, the single-looking approaches are privileged. In addition, for medium-resolution applications, it is recommended to apply the TSVD method, as for high resolution applications, the NL-LS approach can provide good performance with a quite important calculation cost.

5. CONCLUSION

Tomo-SAR has shown its effectiveness in exploiting very highresolution SAR data for mapping and monitoring urban environments, and in recognizing the ambiguity caused by geometric problems, in particular layover. In this work, we have implemented the most exploited tomographic reconstruction approaches and have analyzed the results of these methods in terms of localization and amplitude of the reconstructed profile interactively, then we have quantified the quality of this reconstruction by metrics expressing accuracy and completeness. We have found a good correlation between the metrics and the reconstructed profiles representation by each method. Therefore, we concluded that accuracy and completeness represent an objective judging way in choosing the most appropriate reconstruction approach.

REFERENCES

[1] X. X. Zhu&R.Bamler, (2014) “Superresolving SAR Tomography for Multidimensional Imaging of Urban Areas”, IEEE Signal Processing Magazine.

[2] O. D’Hondt, C. Lopèz-Martinez, S. Guillaso&O. Hellwich, (2018), “Nonlocal Filtering Applied to 3-D Reconstruction of Tomographic SAR Data”, IEEE Transactionson Geoscienceand Remote Sensing, Vol. 56, No. 1, pp 272-285.

[3] L. Linze, P. Lei, Z. XueDong&L. Hui, (2017), “The Research on Accuracy Evaluation Method of the Tomographic SAR Three-Dimensional Reconstruction of Urban Building Based on Terrestrial Lidar Point Cloud”, ISPRS Geospatial week 2017, Wuhan, China.

[4] G. Fornaro, F. Serafino, et al., (2003), “Three-Dimensional Focusing with Multipass SAR Data”, IEEE Transactions on Geoscience and Remote Sensing, Vol. 41, No. 3, pp. 507-517.

[5] G.Fornaro, F.Lombardini, et al., (2005), “Three-Dimensional Multipass SAR Focusing: Experiments with Long-Term Spaceborne Data”, IEEE Transactions on Geoscience and Remote Sensing, Vol. 43, No. 4, pp702-714.

[6] S. Duque, P. Lopez-Dekker, et al., (2010), “Bistatic SAR Tomographic: Processing and Experimental Results”, IGARSS 2010.

[7] F. Lombardini& F. Cai, (2014), “Temporal decorrelation Robust SAR Tomography”,IEEE Transactions on Geoscience and Remote Sensing, Vol. 52, No. 9, pp 5412-5421.

[8] G.Fornaro&F. Serafino, (2006), “Imaging of Single and Double Scatterers in Urban Areas via SAR Tomography”, IEEE Transactions on Geoscience and Remote Sensing, Vol. 44, No. 12, pp3497-3505.

[9] X. X. Zhu&R. Bamler, (2010), “Very High Resolution Spaceborne SAR Tomography in Urban Environment”, IEEE Transactions on Geoscience and Remote Sensing, Vol. 48, No. 12, pp 4296-4308.

[10] O. D'Hondt, C. Lopèz-Martinez, S. Guillaso&O. Hellwich, (2017), "Impact of Non-local Filtering on 3D Reconstruction from Tomographic SAR Data", IGARSS 2017.

[11] D. Reale, W. Franzè, et al., (2015), "Detection of Single Scatterers in Multilook SAR Tomography",Joint Urban Remote Sensing Event (JURSE).

[12] P. Stoica& R. Moses, (2005), Spectral analysis of signals, Pearson Prentice Hall.

AUTHORS

Karima Hadj-Rabah received the M.Sc. degree in Telecommunication, networks and multimedia from University of Science and technology HouariBoumedienne (USTHB), Algeria, in 2016. She is curently working toward the Ph.D. degree in synthetic aperture radar (SAR) tomography reconstruction using very highresolution data, in laboratory of image processing and radiation, USTHB, under the supervision of Professor AichoucheBelhadj-Aissa.

Faiza Hocine Graduated from the University of Sciences and Technology HouariBoumedienne (USTHB), Algeria. PhD in image processing and remote sensing of the same university, in 2015. Currently, she is researcher in Electronics, image processing, remote sensing at USTHB. Her research interests include satellite image processing, SAR interferometry radar.

AssiaKourgli received her PhD from USTHB University, Algeria, in 2007. She is now a professor at Faculted'Electronique et d'Informatique of USTHB University. Her research interests include remote sensing and image processing. Currently, she is interested in texture analysis and synthesis, remote sensing image retrieval, and three-dimensional (3-D) terrain modelling.

AichoucheBelhadj-Aissa obtained her engineering degree in electronics from National Polytechnic School, Algiers, the magister degree and the Doctorate in image processing and remote sensing from the University of Sciences and Technology Houari Boumediene, Algiers. Currently, she is university Professor and head of the research team "GIS and integration of geo-referenced data". The main research themes focus on modeling and analysis of textures and forms, fusion and classification of objects, SAR interferometry-polarimetry and GIS.

A PSEUDO-SPLICING ALGORITHM FOR PARTIAL FINGERPRINT RECOGNITION BASED ON SIFT

Zheng Zhu, Aiping Li*, Rong Jiang, Yulu Qi, Dongyang Zhao, Yan Jia

School of Computer Science,
National University of Defense Technology, Changsha 410073, China

ABSTRACT

At present, many fingerprint recognition techniques are applied to public infrastructures. Their targets are mainly for normal-sized fingerprints. However, with the rise of small-sized fingerprint sensors, the acquired partial fingerprints containing only part of information of the finger, which causes that many researchers change their research directions to partial fingerprint recognition. This paper proposes a SIFT-based pseudo-splicing partial fingerprint recognition algorithm. This algorithm uses the SIFT algorithm to pseudo-splice the input fingerprints during the fingerprint enrollment to increase the robustness of the fingerprint feature database. The comparisons of the accuracy of the recognition among this algorithm, the minutia-based fingerprint recognition algorithm and the fingerprint recognition algorithm based on image similarity, that shows the first performs well. Moreover, this paper proposes an algorithm to evaluate the quality of partial finger print by calculating the invalid blocks of fingerprint image. The result shows that the evaluation algorithm can effectively filter out low-quality fingerprints.

KEYWORDS

Fingerprint Recognition, SIFT, Partial Fingerprint, Pseudo-splicing

1. INTRODUCTION

In recent years, fingerprint recognition techniques have been widely applied to various areas[1], such as management, access control, finance, public security and cyber security etc. With the popularity of fingerprint recognition techniques, it has been used on a large scale in the security verification of [2]mobile terminals (mobile phones, personal computers, tablet computers, etc.). In the common fingerprint recognition techniques, the size of the fingerprint scanners is generally $1" \times 1"$or even larger[3]. Many minutia-based fingerprint recognition algorithms are effective in these size[4]. With the commercialization of fingerprint recognition techniques, the requirement that have better and smaller scanners is increasing. Miniaturization of fingerprint sensors has led to small sensing areas usually varying from $1" \times 1"$ to $0.42" \times 0.42"$[5]. However, fingerprint scanners with a sensing area smaller than $0.5" \times 0.7"$, which is considered to be the average fingerprint size[6], can only capture partial fingerprints. These partial fingerprints contain much fewer features than full fingerprints and it may be rotated. Common fingerprint recognition algorithms are no longer suitable for partial fingerprints. Therefore, this paper proposed a novel fingerprint recognition algorithm suitable for partial fingerprint.

This paper proposes a SIFT-based pseudo-splicing partial fingerprint recognition algorithm to improve the accuracy of the recognition. The comparisons of the accuracy of the recognition among this algorithm, the minutia-based fingerprint recognition algorithm and the fingerprint

Natarajan Meghanathan et al. (Eds) : SPPR, SCAI, CSIA, WiMoA, ICCSEA, InWeS, NECO, GridCom - 2018
pp. 09–19, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.81702

recognition algorithm based on image similarity, that shows the deficiency of the common fingerprint recognition techniques and this algorithm is better suitable for partial fingerprint. The remainder of this paper is organized as follows. The second part introduces the related work on fingerprint recognition techniques. The third part introduces a SIFT-based pseudo-splicing partial finger print recognition algorithm and fingerprint quality assessment algorithm. The fourth part is the result of the fingerprint comparison. The fifth part is a conclusion and future works.

2. RELATED WORK

With the development of fingerprint technology, many researchers have made many contributions in the fingerprint recognition and many algorithms have been proposed[7, 8]. Jia et al.[9] proposed a double matching method that combines the local minutia matching algorithm and the global matching algorithm. Their experimental results showed that the accuracy of the recognition on normal-size fingerprints is well, but they ignored the test for partial fingerprints. Cappelli et al. [10] introduced a novel minutiae-only local representation aimed at combining the advantages of both neighbor-based and fixed-radius structures, without suffering from their respective draw backs. Gudavalli et al. [11] proposed a multi-biometric fingerprint recognition system based on the fusion of minutiae and ridges as these systems render more efficiency, convenience and security than any other means of identification. In order to reduce the number of templates compared with the input fingerprint in the verification stage and accelerate the speed of matching, Zhu et al. [12] proposed a method that splicing the features of multiple template fingerprints into one. They still ignored the test for partial fingerprints, but they mentioned the idea of splicing fingerprint template. For partial fingerprint, single image contained fewer features. It was useful for recognition by splicing multiple partial fingerprints.

Due to the unique features of the partial fingerprints, many researchers tried to recognize fingerprints based on image[13]. Zanganeh et al.[14] proposed a region-based fingerprint recognition algorithm. This method was used to image matching by calculating the local similarity in the image region and the overall correlation coefficient. Ito et al. [15] proposed an efficient fingerprint recognition algorithm combining phase-based image matching and feature-based matching. Liu et al. [16] proposed an efficient fingerprint search algorithm based on database clustering, which narrows down the search space of fine matching. Feng et al. [17] established a fingerprint matching algorithm combining ridge correlation and minutia feature points. This algorithm performed better than original minutia-based algorithm and was suitable for some nonlinear distorted images. However, the ridge direction was difficult to distinguish for partial fingerprints.

Nowadays, some researchers start studying the algorithm of fingerprint recognition based on local features[18, 19, 20, 21]. Aguilar-Torres et al. [22] presented a fingerprint recognition method using a combination of the Fast Fourier Transform (FFT) with Gabor filters for image enhancement, and fingerprint recognition was carried out using a novel recognition stage based on Local Features and Hu invariant moments for verification. Ceguerra et al. [23] presented a new approach for combining local and global recognition schemes for automatic fingerprint verification (AFV), by using matched local features as the reference axis for generating global features. Madhuri et al. [24] realized that most of the minutia-based algorithms are not suitable for partial fingerprints, then proposed a fingerprint recognition technique using local robust features. Matching algorithms based on local features include SIFT (Scale Invariant Feature Transform)[25] and SURF (Speeded Up Robust Features)[26]. Their experimental results showed that the algorithm presented better effect on rotated fingerprints and partial fingerprints, but the test data used in their experiments was inconsistent with the size of the fingerprints discussed in our paper.

3. SIFT-BASED PSEUDO-SPLICING PARTIAL FINGERPRINT RECOGNITION ALGORITHM

3.1. SIFT Algorithm

SIFT (Scale Invariant Feature Transform), proposed by David G. Lowe [25], is a description of the image processing field. The SIFT features not only keep invariance to rotation, brightness changing, but also keep a certain degree of stability for viewing angle changing, affine transformations, and noise, and it is a very stable local feature. It is not only suitable for fast and accurate matching of massive feature libraries, but also can be easily combined with other feature vectors.

The SIFT feature matching algorithm has two stages.

The first stage is the SIFT feature generation stage. This stage mainly extracts feature vectors that are rotation-independent, and luminance-independent from the images. Firstly,it searches for image locations on all identifies potential points of interest for rotation invariant using Gaussian differentiation functions; then it determines position through a fine-fitting model at each candidate location. The selection of key-points depends on their degree of stability; Next according to the gradient direction of the local image, one or more directions are assigned to each key-point position, all subsequent operations on the image data are relative to the directionand position of the key-points; Finally, the gradients of the local image are measured at selected scales within the neighborhood around each key-point. These gradients are transformed into a representation that allows for relatively large Local shape deformation and illumination change.

The second stage is the SIFT feature vectors matching stage. After the SIFT feature vectors of the two images are generated, the next step is to use the Euclidean distances of the key-point feature vectors as the similarity determination metrics of the key-points in the two images. Take a key-point of one of the images and find the closest two key-points in the other image by traversing. In the two key-points, if the value of the closest distance divided by the value of the next closest distance is smaller than a certain threshold, it is determined as a pair of matching points.

3.2. Pseudo-Splicing Algorithm for Partial Fingerprints

Due to fewer features of the partial fingerprint, one or two partial fingerprints are insufficient to cover the fingerprint information of an entire finger. A complete fingerprint database during the fingerprint enrollment stage is required. This paper proposes a pseudo-splicing algorithm to optimize the accuracy of fingerprint matching and simplify the complexity of fingerprint matching algorithm.

The SIFT algorithm has characteristics of invariant to rotationand brightness changing, so it can be used to realize pseudo-splicing of the partial fingerprint by using SIFT algorithm repeatedly. Fingerprint recognition is mainly divided into two stages, namely the fingerprint enrollment stage and the fingerprint verification stage. The fingerprint feature database needs to be constructed in the enrollment stage, and the matching result to be output by comparing raw fingerprint with the fingerprint database in the verification stage. Fingerprint splicing is a method that constructs a fingerprint feature database during the enrollment stage.

Algorithm 1 Pseudo-splicing algorithm for partial fingerprints
Input: Fingerprint feature set I_A ; Fingerprint similarity threshold θ ;
The maximum number of finger feature set μ ; Raw fingerprint fp
Output: Fingerprint feature set I_A

1. $I_A = \emptyset$;
2. When the number of I_A<μ :
3. for I_{A_i} in I_A:
4. $score_i$ = get score using SIFT matches fp and I_{A_i};
5. if maximum of $score_i$<θ:
6. $I_A = I_A \cup$ fp;

Assuminga feature set$I_A$is one of the user A's fingers. At the beginning, the I_A is an empty set, namely$I_A = \emptyset$. When a fingerprint is obtained in the enrollment stage, the SIFT matching is performed between the raw fingerprint and all the fingerprints in the set I_A, and the matching score is calculated. If the maximum value among these scores is smaller than thepredefined threshold θ (θ is the fingerprint similarity threshold), the raw fingerprint is accepted and put it into the set I_A.If the number of the set I_A is greater than the predefined threshold μ (μ is the maximum number of finger features of one person),the process of enrollment stage is terminated, and the set is now $I_A = \{ I_{A_1}, I_{A_2}, \cdots, I_{A_\mu} \}$. The pseudo-code of the algorithm is shown in Algorithm 1.

The fingerprint feature databasecreatedby pseudo-splicing can improve the accuracy of fingerprint matching in the verification stage. However, if a low-quality fingerprint is encountered, it can pollute the fingerprint feature database. Therefore, this paper proposes an algorithm to evaluate the quality of fingerprints during the fingerprint enrollmentstage and fingerprint verification stage.

3.3. Fingerprint Quality Assessment Algorithm

The fingerprint studied in this paper is the image of96 pixels × 96 pixels, as shown in Figure 1.

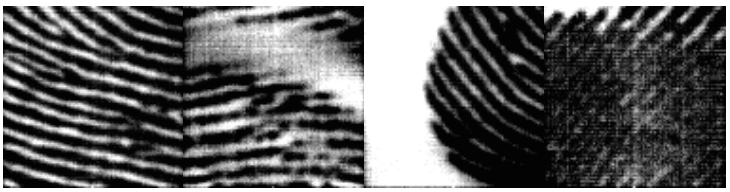


Figure 1. (a) high-quality fingerprint (b) low-quality fingerprints

Figure 1 (a) shows a high-quality fingerprint and Figure 1 (b) shows three low-quality fingerprints. As can be seen from the figure 1, the distribution of the ridges of the high-quality fingerprint is relatively homogeneous, and there are no large area of white blocks or black blocks. If we put low-quality fingerprints into the fingerprint feature database, it will pollute the fingerprint database, which will cause a great influence on the verification of the fingerprint. This is very important to guarantee the performance of a biometric system, especially during the enrollment [27]. Therefore, this paper proposes a fingerprint quality assessment algorithm to improve the accuracy of verification of the fingerprint.

When a fingerprint is acquired, the width and height of the image are calculated, then the image is traversed in step s, anda sub-image is croppedwith the side lengthof b to calculate the ratio of the black block.As shown in Formula 1.

$$\mathrm{r} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}}{\mathrm{m} \times \mathrm{n}}, where\ p_{ij} = \begin{cases} 0, & c_{ij} \geq \alpha \\ 1, & c_{ij} < \alpha \end{cases} \quad \text{（Formula 1）}$$

In formula 1, r is the calculated ratio of the black block of the sub-image, m and n are the width and height of the sub-image respectively, p_{ij} indicates whether the pixel at (i, j) is black, and c_{ij} is the value of the pixel at (i, j), where α is the threshold for judging color. If the value of one pixel is larger than the threshold, it is regarded as white, otherwise it is black.

After traversing the above rules, a set of black blocks are calculated, namely $\mathrm{R} = \{r_1, r_2, \cdots, r_L\}$, where r_k $(1 \leq k \leq L)$ is the ratio of black block calculated from formula 1, and L is the total number of traversed blocks. Then counting the number of invalid blocks using formula 2, as shown in Formula 2.

$$\mathrm{N} = \sum_{k=1}^{L} q_k, where\ q_k = \begin{cases} 1, & r_k > \beta\ or\ r_k < 1-\beta \\ 0, & Others \end{cases} \quad \text{（Formula 2）}$$

In formula 2, N denotes the number of invalid blocks, $q_k$denotes whether the k-th black block is an invalid block, and β is the ratio threshold of the invalid block. If the ratio of the black block is greater than β or less than $1 - \beta$, itis regarded as an invalid block.

Finally, determining whether the fingerprint quality is qualified by comparing the number of invalid blocks N with the fingerprint quality threshold ε. Here is given the fingerprint quality assessment algorithm as shown in Algorithm 2.

Algorithm 2 Fingerprint quality assessment algorithm
Input： Raw fingerprint fp; step s; Side length of blockb;
Color threshold α; Invalid block ratio thresholdβ;
Fingerprint quality thresholdε
Output：The result of fingerprint quality assessment

Get the widthfp_width and height fp_height of the fingerprint image fp;

The number of invalid blocks invalid_num = 0;
for i from 0 to fp_width step s:
4. for j from 0 to fp_height step s:
5. Cropping a sub-image sub_fp with the starting point(i, j) and side length b;
6. Binarizing sub-image sub_fpwith the color threshold α;
7. Calculating the ratio of the black blockblack_ratio and the ratio of the white block white_ratioof the binarized sub-image;
8. if black_ratio>β or White_ratio >β:
9. invalid_num = invalid_num + 1;
10. return invalid_num<ε;

---

4. EXPERIMENT

4.1.Experimental Data

The fingerprint data used in this paper is 96 pixels × 96 pixels images, as shown in Figure 1 above. In the fingerprint data set, not all images are available. As shown in (b) on Figure 1, these images should be discarded. Therefore, the fingerprint quality assessment algorithm is used to filter out the low-quality fingerprints during enrollment stage. Then the fingerprint recognition algorithm based on the minutia points、the similarity of the image and the pseudo-splicing partial fingerprint with SIFT are respectively tested. The experimental data of this paper uses 120 persons' fingers. Each finger has 40 fingerprints. The first 20 fingerprints are used to create fingerprint feature database, and the rest are used to test. The feature database has 120 personal features and each person has 5 fingerprint features. In the minutia-based feature database, 5 minutia features are extracted randomly; in the image similarity-based feature database, 5 fingerprints are sampled randomly; in SIFT-based pseudo-splicing feature database, 5 fingerprints are trained by the pseudo-splicing algorithm.

4.2. Experimental Strategy

This experiment matches each fingerprint in the test set with each feature in the feature database of 120 individuals. Each fingerprint in the test set will get a score, and then according to the predefined similarity threshold θ, the fingerprint is attached with a label, that is, if the fingerprint is successful matched, the label of the fingerprint is assigned to 1, otherwise it is assigned 0 . Thus we get a set of binary relations $[(x_{111}, y_{111}), (x_{112}, y_{112}), \cdots, (x_{ijq}, y_{ijq})]$, where i represents the fingerprints of the i-th person in the test set, and j represents the features of the j-th person in the feature database, andq represents the q-th fingerprint of the i-th person in the test set.So x_{ijq} represents the result of matching the q-th fingerprint of the i-th person in the test set with the features of the j-th person in the feature database, and y_{ijq} represents the value of the label corresponding to x_{ijq}. y_{ijq} is equal to 1 if i equals j, otherwise it is 0. Then a value V is calculated by Formula 3, and if V is less than 0, it is regarded as false rejection, otherwise it is regarded as false acceptance. Finally, calculating the False Rejection Rate (FRR) and the False Acceptance Rate (FAR) corresponding to the current similarity threshold θ.

$$V = \frac{x_{ijq} - \theta}{|x_{ijq} - \theta|} - y_{ijq}, \text{When } x_{ijq} = \theta, \frac{x_{ijq} - \theta}{|x_{ijq} - \theta|} \; regarded\ as\ 0 \qquad \text{(Formula 3)}$$

The false acceptance rate [28], or FAR, is the measure of the likelihood that the biometric security system will incorrectly accept an access attempt by an unauthorized user. The false recognition rate [29], or FRR, is the measure of the likelihood that the biometric security system will incorrectly reject an access attempt by an authorized user. This paper shows the accuracy of the recognition of different algorithms by calculating FRR and FAR.

4.3. Experimental Result

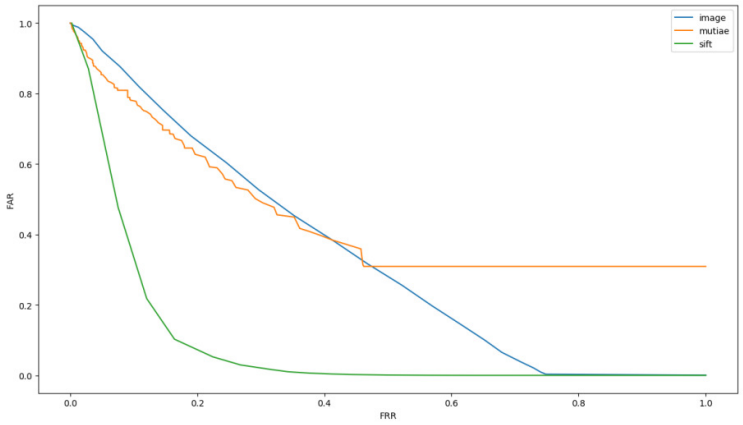


Figure 2.The result of comparison of three fingerprint recognition algorithms

The minutia-based fingerprint recognition algorithm in this experiment is tested using the edit distance algorithm. The similarity-based algorithm is tested using the perceptual hash algorithm. The result of the comparison are shown in Figure 2.

The ROC curve is used to describe the effect of the fingerprint algorithm. The ordinate represents FAR, the abscissa represents FRR, and the Equal Error Rate (EER) is a balance point between FAR and FRR. The lower the equal error rate value, the higher the accuracy of the biometric system [30].

As shown in figure 2, the yellow curve represents the result of the minutia-based fingerprint recognition algorithm; the blue curve represents the result of the similarity-based fingerprint recognition algorithm; the green curve represents the SIFT-based pseudo-splicing fingerprint recognition algorithm. The curve of the minutia-based fingerprint recognition algorithm and the curve of the similarity-based fingerprint recognition algorithm have higher ERR than the curve of the SIFT-based pseudo-splicing fingerprint recognition algorithm. Obviously, the SIFT-based pseudo-splicing fingerprint recognition algorithm works best.

There are two important parameters that affect the accuracy of recognition in the SIFT-based pseudo-splicing fingerprint recognition algorithm. One is the number of features of the fingerprint feature database, and the other is the similarity threshold in the pseudo-splicing algorithm. The rest of the paper presents the algorithm is optimized by adjusting these parameters.

This experiment tests the number of features of the fingerprint feature database differently, and keeps the other parameters unchanged. The results are shown in Figure 3.

In the legend of figure 3, the number behind N represents the number of features of the fingerprint feature database. As can be seen from the figure 3, with the increasing of the number of features of the fingerprint feature database, the lower the EER, the better accuracy of the recognition.

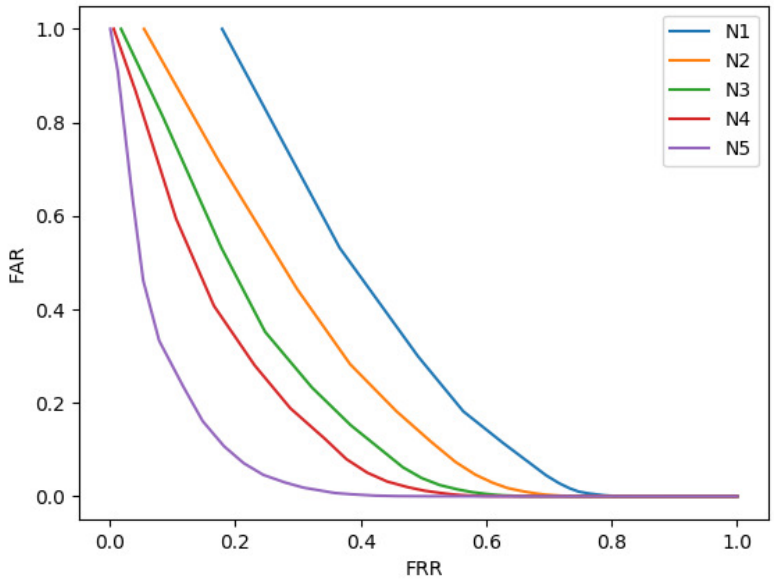
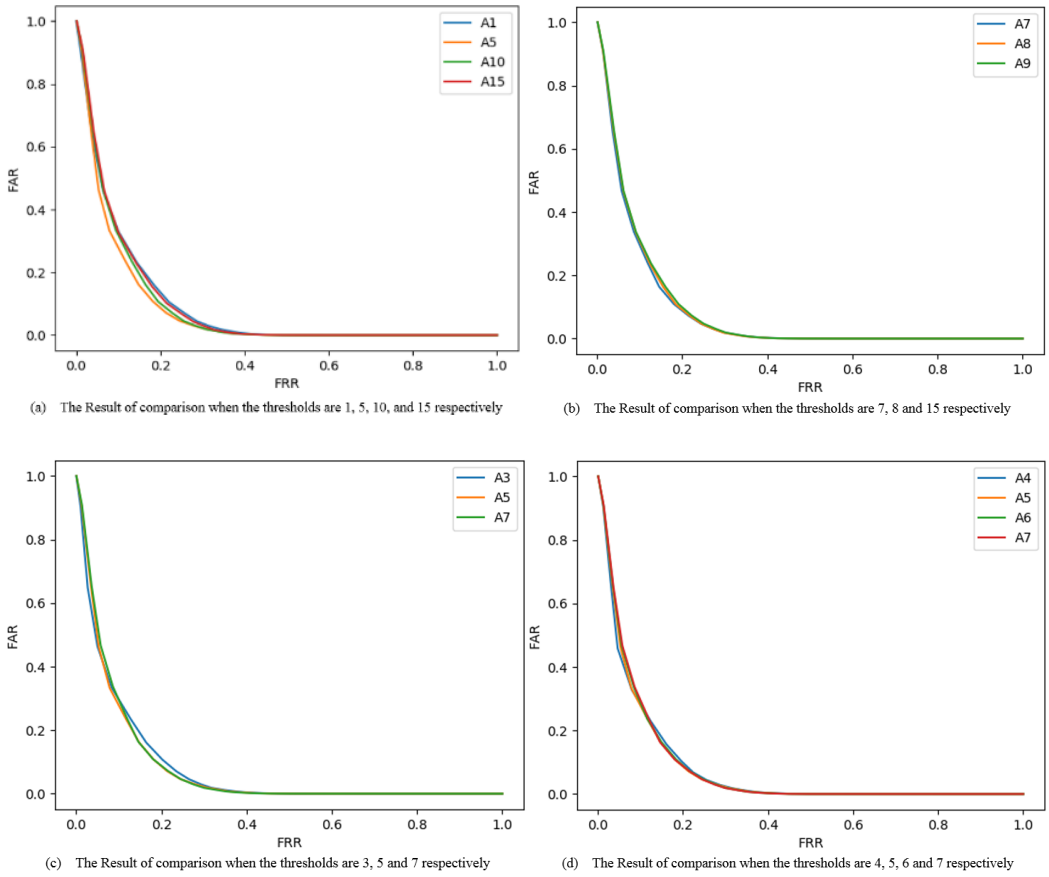


Figure 3. The result of comparison under different number of features of the fingerprint feature database

Now, keeping the number of features and other parameters of the fingerprint feature database unchanged, and testing the different fingerprint similarity threshold in the pseudo-splicing algorithm. The results are shown in Figure 4.

In the legend of figure 4, the number behind A represents the fingerprint similarity threshold in the pseudo-splicing algorithm. The curves between neighboring threshold are dense. This paper shows the result of comparison under different fingerprint similarity thresholds by dividing into four parts. It can be seen from Figure 4 (a), the lowest EER occurs when the threshold is 5. It can be seen from Figure 4(b), when the threshold is 7, the EER is the lowest. It can be seen from (c) and (d) in figure 4 that when the threshold is between 5 and 7, the EER is lower and the algorithm works well.

According to the comparison above, it shows that the more number of features of the fingerprint feature database, the better accuracy of the recognition, but with the number of features of the fingerprint feature database increasing, the consumption of the time of recognition becomes higher. So the number of features of the fingerprint feature database should be adjusted according to the specific application scenarios. In addition, the algorithm performs well when the fingerprint similarity threshold is between 5 and 7 in the pseudo-splicing algorithm.



(a) The Result of comparison when the thresholds are 1, 5, 10, and 15 respectively

(b) The Result of comparison when the thresholds are 7, 8 and 15 respectively

(c) The Result of comparison when the thresholds are 3, 5 and 7 respectively

(d) The Result of comparison when the thresholds are 4, 5, 6 and 7 respectively

Figure 4. The result of comparison under different similarity thresholds

5. CONCLUSION

This paper proposes a SIFT-based pseudo-splicing partial fingerprint recognition algorithm, aiming at improving the accuracy of the fingerprint recognition algorithm for partial fingerprints. Through the comparison, the proposed algorithm has better accuracy of recognition for partial fingerprints, and the algorithm is optimized to increase the accuracy of recognition for partial fingerprints by adjusting different parameters. In the future, we will try to combine other algorithms to improve the accuracy of the recognition for the partial fingerprint.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (No. 61732004, 61472433, 61732022, 61672020, 61502517); National Key Research and Development Program of China (2017YFB0802204, 2016YFB0800802, 2016YFB0800803, 2016YFB0800804, 2016QY03D0601, 2016QY03D0603, 2016YFB0800303).

REFERENCES

[1] Jain A K, Prabhakar S, Ross A. Fingerprint matching: Data acquisition and performance evaluation[J]. Dept. of Computer Science, Michigan State Univ., East Lansing, Tech. Rep. MSU-CPS-99–14, 1999.

[2] Lee W, Cho S, Choi H, et al. Partial fingerprint matching using minutiae and ridge shape features for small fingerprint scanners[J]. Expert Systems with Applications, 2017, 87: 183-198.

[3] Maltoni D. A tutorial on fingerprint recognition[M]//Advanced Studies in Biometrics. Springer Berlin Heidelberg, 2005: 43-68.

[4] Maltoni D, Maio D, Jain A K, et al. Handbook of fingerprint recognition[M]. Springer Science & Business Media, 2009.

[5] Jea T Y, Govindaraju V. A minutia-based partial fingerprint recognition system[J]. Pattern Recognition, 2005, 38(10): 1672-1684.

[6] Biometrikainc.,A technical evaluation of fingerprint scanners, http://www.biometrika.it/eng/wp_scfing.html, Monte Santo 21, 47100 Forli, Italy.

[7] Hrechak A K, McHugh J A. Automated fingerprint recognition using structural matching[J]. Pattern Recognition, 1990, 23(8): 893-904.

[8] Yadav S, Mathuria M. Fingerprint Recognition based on Minutiae Information[J]. International Journal of Computer Applications, 2015, 120(10).

[9] JIA Jia, CAI Lianhong. A Fingerprint Verification Approach based on Minutiae Re-matching Method [J]. Journal of Tsinghua University (Science and Technology), 2006, 46(10):1776-1779.

[10] Cappelli R, Ferrara M, Maltoni D. Minutia cylinder-code: A new representation and matching technique for fingerprint recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(12): 2128-2141.

[11] Gudavalli M, Kumar D S, Raju S V. A Multibiometric Fingerprint Recognition System Based on the Fusion of Minutiae and Ridges[C]//Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India (CSI) Volume 1. Springer, Cham, 2015: 231-237.

[12] ZhuEn, YIN Jian-ping, ZHANG Guo-min, et al. Merging Features of Multiple Template Fingerprints [J]. Journal of National University of Defense Technology, 2005, 27(6):26-29.

[13] Nanni L, Lumini A. Descriptors for image-based fingerprint matchers[J]. Expert Systems with Applications, 2009, 36(10):12414-12422.

[14] Zanganeh O, Srinivasan B, Bhattacharjee N. Partial Fingerprint Matching through Region-Based Similarity[C]// International Conference on Digital Lmage Computing: Techniques and Applications. IEEE, 2014:1-8.

[15] Ito K, Morita A, Aoki T, et al. A fingerprint recognition algorithm combining phase-based image matching and feature-based matching[C]//International Conference on Biometrics. Springer, Berlin, Heidelberg, 2006: 316-325.

[16] Liu M, Jiang X, Kot A C. Efficient fingerprint search based on database clustering[J]. Pattern Recognition, 2007, 40(6):1793-1803.

[17] Feng J, Ouyang Z, Cai A. Fingerprint matching using ridges[J]. Pattern Recognition, 2006, 39(11):2131-2140.

[18] Zhou R, Zhong D, Han J. Fingerprint Identification Using SIFT-Based Minutia Descriptors and Improved All Descriptor-Pair Matching[J]. Sensors, 2013, 13(3):3142-56.

[19] Costas A, Boult T. Improving Partial Fingerprint Recognition[J].

[20] Malathi S, Meena C. An efficient method for partial fingerprint recognition based on local binary pattern[C]// IEEE International Conference on Communication Control and Computing Technologies. IEEE, 2010:569-572.

[21] Aravindan A, Anzar S M. Robust partial fingerprint recognition using wavelet SIFT descriptors[J]. Pattern Analysis & Applications, 2017, 20(2):1-17.

[22] Aguilar-Torres G, Sanchez-Perez G, Toscano-Medina K, et al. Fingerprint recognition using local features and Hu moments[J]. Journal of applied research and technology, 2012, 10(5): 745-754.

[23] Ceguerra A V, Koprinska I. Integrating local and global features in automatic fingerprint verification[C]//Pattern Recognition, 2002. Proceedings. 16th International Conference on. IEEE, 2002, 3: 347-350.

[24] Madhuri R, Mishra. Fingerprint Recognition using Robust Local Features[J]. 2012.

[25] Lowe D G. Distinctive Image Features from Scale-Invariant Keypoints[J]. International Journal of Computer Vision, 2004, 60(2):91-110.

[26] Bay H, Tuytelaars T, Gool L V. SURF: speeded up robust features[C]// European Conference on Computer Vision. Springer-Verlag, 2006:404-417.

[27] Yao Z, Bars J M L, Charrier C, et al. Fingerprint Quality Assessment with Multiple Segmentation[C]// International Conference on Cyberworlds. IEEE, 2015:345-350.

[28] https://www.webopedia.com/TERM/F/false_acceptance.html, 2018 (accessed 10 April 2018).

[29] https://www.webopedia.com/TERM/F/false_rejection.html, 2018 (accessed 10 April 2018).

[30] https://www.webopedia.com/TERM/E/equal_error_rate.html, 2018 (accessed 10 April 2018).

INTENTIONAL BLANK

A POST-PROCESSING METHOD BASED ON FULLY CONNECTED CRFS FOR CHRONIC WOUND IMAGES SEGMENTATION AND IDENTIFICATION

Junnan Zhang and Hanyi Nie

Computer College, NUDT, Changsha, China

ABSTRACT

Chronic wound have a long recovery time, occur extensively, and are difficult to treat. They cause not only great suffering to many patients but also bring enormous work burden to hospitals and doctors. Therefore, an automated chronic wound detection method can efficiently assist doctors in diagnosis, or help patients with initial diagnosis, reduce the workload of doctors and the treatment costs of patients. In recent years, due to the rise of big data, machine learning methods have been applied to Image Identification, and the accuracy of the result has surpassed that of traditional methods. With the fully convolutional neural network proposed, image segmentation and target detection have also achieved excellent results. However, the accuracy of chronic wound image segmentation and identification is low due to the limitation of the deep convolution neural network. To solve the above problem, we propose a post-processing method based on fully connected CRFs with multi-layer score maps. The experiment results show that our method can be used to improve the accuracy of chronic wound image segmentation and identification.

KEYWORDS

Fully Connected CRFs, Chronic Wound Segmentation, Post-processing Method

1. INTRODUCTION

In recent years, with the rise of big data, the field of artificial intelligence has been aroused a broad concern. AlexNet [1], the champion of ImageNet competition in 2012, uses convolutional neural networks for image classification and recognition. Its accuracy exceeds the traditional method significantly, which makes people pay attention to the application of convolutional neural networks in the field of image classification again. Afterward, people continued to innovate (VGG[2], GoogleNet[3], Residual Net[4], DenseNet[5], CapsuleNet[6] and other deep neural networks), and further improve the accuracy of image classification. However, in the field of medical image classification and segmentation, the accuracy of using deep convolution neural networks is relatively low. Therefore, the application of convolutional neural networks in this area is not effective.

Chronic wound including Diabetic foot ulcers, venous leg ulcers, and acne, has long recovery time and need different methods of treatment at various period. The current treatment of chronic wounds usually takes up a significant amount of medical resources and is not easy to treat [7-9]. Long-term hospitalization is a burden for both hospitals and patients. On the one hand, the resources of the hospital are occupied for an extended period, and it is impossible to provide medical services to other patients in urgent need. On the other hand, the long-term hospitalization costs are too high for most of the patients to afford. It is also very common in some remote areas,

Natarajan Meghanathan et al. (Eds) : SPPR, SCAI, CSIA, WiMoA, ICCSEA, InWeS, NECO, GridCom - 2018
pp. 21–29, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.81703

patients are far away and inconvenient to see a doctor. These conditions have brought great suffering to the patients.

Hence, it is of great importance to do a post processing method based on fully-connected CRFs for chronic wound images segmentation and identification. This method helps improve the accuracy of segmentation and classification and the results obtained can assist doctors in diagnosis and treatment.

In this paper, we first introduce fully connected CRFs(conditional random fields). Then we propose a post processing method based on fully-connected CRFs for chronic wound images segmentation and identification. Finally we make compare experiment and the results prove the validity of this method.

2. RELATED WORKS

The basic CRF model contains a unary energy function at an independent pixel or image block and a paired energy function in a neighboring pixel block or image block[11-14]. The adjacency CRF structure generated based on that model limits its association between distant pixel points or image blocks within an image, and thus causes an excessively smooth object boundary. To improve the accuracy of segmentation and identification, the researchers extended the basic CRF framework by adding hierarchical connections and higher-order energy functions in the image region[15-18]. However, this method is bound to be limited by the accuracy of unsupervised image segmentation. Although some progress has been made, this limitation affects the ability of region-based CRF method to accurately assign labels to complex image boundary pixels[19]. Although some progress has been made, this limitation affects the ability of region-based methods to accurately assign labels to complex image boundary pixels.

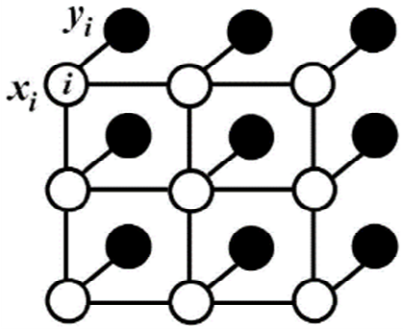


Figure 1. Fully connected CRF model on pixel image[20]

Figure 1. shows a simple fully connected CRF model on pixel image, y_i represents the label of its corresponding pixel point x_i, and all pixels x_i are connected together.

In order to solve the above problem, a fully connected CRF model based on the energy function of all pixel pairs on the image is proposed[21]. Although the fully-connected CRF model has been used for image semantic segmentation[22-25].But the computation complexity of the fully connected model limits its usage on hundreds of image regions that are segmented on the image. Therefore, the accuracy of segmentation is still limited by the accuracy of the unsupervised image segmentation algorithm that generates these regions. In order to solve this problem, the researchers connected all the pixel pairs in the image to achieve more fine segmentation and recognition, at the cost of an explosive increase in the amount of parameters. To overcome this problem, the pairwise energy function is defined as a linear combination of Gaussian kernels in

arbitrary feature space[21] and this CRF distribution is approximated by the mean field. The approximation is iteratively optimized through a series of message passing steps, each of which updates a single variable by fusing the information of all other variables. The update of the mean field can be achieved by Gaussian filtering in the feature space. The computational complexity of message passing can be reduced from quadratic to linear by employing efficient high-dimensional filtering approximations.

The fully-connected CRF model is defined as follows:

Suppose random field X, whose elements are $\{X_1, X_2, ..., X_n\}$, where the value range of X_i is the set of labels L:$\{L_1, L_2, ..., L_k\}$.Suppose another random domain I, whose elements are $\{I_1,...,I_n\}$. I is the information of the pixel on the image with the input size N, and X is the label of each pixel. I_j is the color vector of pixel j, and X_j is the label assigned to pixel j. Conditional random field (I, X) is determined by a Gibbs distribution:

$$P(X|I) = \frac{1}{Z(I)}\exp(-\sum_{c\in C_G} \emptyset_c(X_c|I))$$

G=(v, e) in the above equation is an image to be labeled, and each pixel point pair c is an element in the set of pixel pair C_G on this image G, which contains an energy function φ_CThe Gibbs energy function of the label $x\in L^N$ is:

$$E(x|I) = \sum_{c\in C_G} \emptyset_c(x_c|I)$$

The maximum posterior estimate of the conditional random field label is:

$$x^* = argmax_{x\in L^N} P(x|I)$$

For the convenience of representation, the latter part will use $\Psi_c(x_c)$ to represent $\varphi_c(x_c|I)$.

In the fully connected pixel-pair CRF model, G is a complete graph of the label X, C_G is the set of all unary and binary pixel groups, so the corresponding Gibbs energy function is

$$E(X) = \sum_i \Psi_u(x_i) + \sum_{i<j} \Psi_p(x_i, x_j)$$

The values of i and j in the above formula range from 1 to N. This unary energy function $\Psi_u(x_i)$ is calculated separately for each pixel by a classifier that generates a label assignment distribution function based on image features. The binary energy function is calculated as follows:

$$\psi_p(x_i, x_j) = \mu(x_i, x_j)\sum_{m=1}^{K} \omega^{(m)} k^{(m)}(f_i, f_j)$$

The $k^{(m)}$ in the formula is a Gaussian kernel function. Then we get:

$$k^{(m)}(f_i, f_j) = \exp(-\frac{1}{2}(f_i - f_j)^T \Lambda^{(m)}(f_i - f_j))$$

The f_i and f_j in the formula are the feature vectors of the pixels i and j in the arbitrary feature space, $\omega^{(m)}$ is the linear combination weight, and μ is a label-compatible function. Label compatible function $\mu(x_i, x_j)=1$ if $xi \neq xj$, otherwise its value is 0. Each Gaussian kernel $k^{(m)}$ is characterized by a symmetric positive definite matrix $\Lambda^{(m)}$. For image segmentation and identification problems, the energy functions of two kernels can be used. I_i and I_j in the following formula represent color vectors, and p_i and p_j represent the position of pixel points.

$$k(f_i, f_j) = \omega^{(1)} \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}\right) + \omega^{(2)} \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}\right)$$

The first half of the formula is the appearance kernel function. The neighboring pixels with the same color are more likely to be the same class of pixel. The degree of distance and color is controlled by the parameters θ_α and θ_β. The second part is the smooth kernel function, which is used to remove isolated small areas[11].

3. METHOD

The image segmentation and identification network based on the deep neural network outputs a score map and then get the pixel point semantic segmentation label. The score map is generally smooth, so applying short-range CRF for post-processing will have the opposite effect. In order to overcome the above problems, the researchers used the CRF model of [10] as a post-processing method. The score map has a score for each pixel point label classification, that is, the probability that each pixel point may be assigned a label value. This probability is calculated separately for each pixel, so it can be input as a unary energy function into the energy function of the fully-connected CRF:

$$\Psi_u(x_i) = -logP(x_i)$$

In the above formula, $P(x_i)$ is the pixel point assignment label probability output by the deep neural network based image segmentation and identification network. The binary energy function is still related to the color and position of the pixel, and the weight of the pixel color and position is modified by changing the relevant parameters. The fully-connected CRF uses the deep neural network based image segmentation and identification score map of the network output as input, and outputs the fine segmentation result after several fully-connected CRF iterations. Compared with the output of image segmentation and identification method based on the deep neural network, the details of the output of the post processing method based on fully-connected CRFs are clearer.

However, the single score map contains insufficient information on chronic wound images, and the results by the post processing method are poor. To solve the problem, we convert multiple scale feature maps into score maps, and superimpose multiple score maps as a multi-scale score map for the input of fully-connected CRF post-processing methods. The formula is as follows:

$$P(x_i) = \sum_{j=1}^{M} \varphi^{(j)} P_j(x_i)$$

In the above formula, $P_j(x_i)$ is the assigned label probability of pixel i on the jth score graph, and $\lambda^{(j)}$ is the weight of this probability value. The binary energy function remains unchanged, and the parameter weights are modified according to the characteristics of the chronic wound image.

In the optimization of the parameters, we use the method of from coarse to fine. The score map is superimposed using the last three layers. The sum of the three weights is 1. The initial value ratio is (2:2:6), and then the parameters are optimized. The update step is 0.05 until the optimal value is selected. According to the paper[10], the weight parameters ω_2 and θ_γ of the binary energy function are fixed to the default value of 3. The parameter ω_1 has a value range of [5, 10], and the update step is 1 each time. The parameter θ_α has a value range of [50, 100] and the update step size is 10. The parameter θ_β has a value range of [3, 10] and the update step size is 1.

4. EXPERIMENT

To verify the effectiveness of post-processing methods for chronic wound image segmentation and recognition based on fully connected CRF, we designed two comparison experiments: (1) Compare the post-processing results obtained by the post-processing method based on fully connected CRF with the single-layer score map and the original segmentation results of the deep neural network. (2) Compare the post-processing results obtained by the post-processing method based on fully connected CRF with the single-layer score graph and the post-processing results obtained by the post-processing method based on fully connected CRF with the multi-layer score graph. We use the original results of paper[26], and choose the best chronic wound image segmentation and identification network:MobileNet-0.75-skip-fcn16 as the chronic wound image segmentation and identification network to verify the effectiveness of our method.

EXPERIMENT ENVIRONMENT

We use an NVIDIA Geforce 1080Ti GPU to speed up parameter learning and evaluate the learned model on a computer with Intel Core i7-8700K CPU @ 3.70GHz and 32GB RAM. The program runs on a 64-bit windows10 home operating system with CUDA 9.0 and Tensorflow 1.7.0-GPU installed.

DATA SET

We use the data set that is built by the article[26]. The dataset is collected partly from cooperated medical institutions and partly from Medetec Wound Database. We made the size of the images a uniform resolution (512 by 512 pixels).

GROUND TRUTH.

The chronic wound image is manually selected for the chronic wound area and the background area. The red area on the ground truth is the chronic wound area, and the black area is the background.

The evaluation standard is as follows:
TP: the ground-truth is positive and the prediction is positive.
FN: the ground-truth is positive but the prediction is negative.
FP: the ground-truth is negative but the prediction is positive.
TN: the ground-truth is negative and the prediction is negative.
We use accuracy, mean intersection-over-union (mIoU), and dice similarity coefficient (DSC) to compare the result. They are computed as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{mIoU} = \frac{TP}{TP + FP + FN}$$

$$\mathrm{DSC} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

5. RESULTS

The figure below shows a comparison of post-processing results obtained by the post-processing method based on fully connected CRF with the single-layer score map and the original segmentation results of the deep neural network. The first column(A) on the left are the original images, the second column(B) are the ground truth, the third column(C) are the Mobilenet origin results, and the fourth column(D) images are images obtained by merging the first and the third columns of images, which is convenient for viewing the segmentation effect. The fifth column(E) are the results obtained by using the post-processing method with single-layer score map and the sixth column(F) are the images obtained by merging the first and the fourth column images. Tt can be found that the image processed by the post-processing method with single-layer score map has a certain improvement in the edge detail compared with the original segmentation result. However, it could cause anti-effect such as first and third lines of the images.

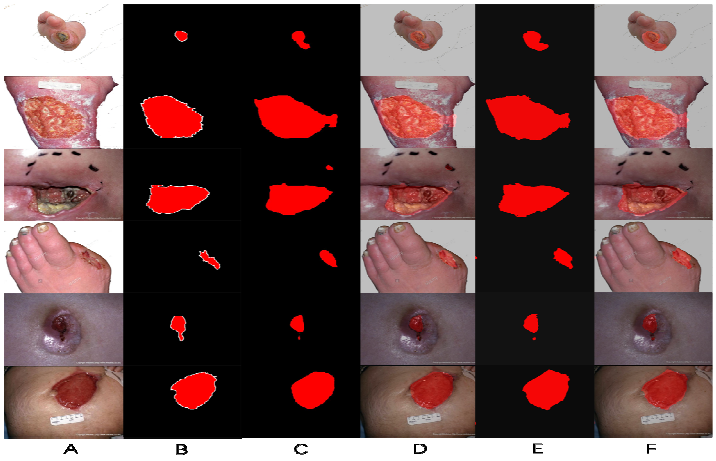


Figure 2. Comparison of the results of neural networks and post-processing method with single-layer score map

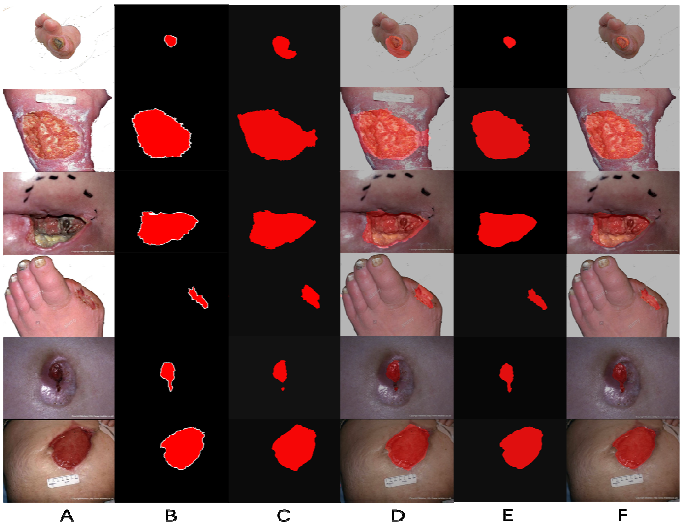


Figure 3. Comparison of the results of post-processing method with single-layer score map and post-processing method with multi-layer score map

The first column(A) on the left are the original images, the second column(B) are the ground truth, the third column(C)(also E in figure 2) are the results obtained by using the post-processing method with single-layer score map, and the fourth column(D)(also F in figure 2) images are images obtained by merging the first and the third columns of images, which is convenient for viewing the segmentation effect. The fifth column(E) are the results obtained by using the post-processing method with multi-layer score map and the sixth column(F) are the images obtained by merging the first and the fourth column images.

Compared with the two results in the above figure, we can find that it is better to use the post-processing method with multi-score map to process the original segmentation image. It can also correct some misclassifications and make the contours of chronic wound areas more detailed. The table below compares the accuracy of the two methods.

Table 1. Accuracy comparison

Method	Accuracy (%)	mIoU (%)	DSC (%)
Origin Results	98.26	85.76	92.33
Single-layer score map	98.23	85.80	92.35
Multi-layer score map	**98.36**	**86.08**	**92.52**

The comparison results of the show that the post-processing method with multi-layer score map is better than the post-processing method with single-layer score map and achieves the highest of the three indexes. That proves the validity of our method.

6. CONCLUSION

In this paper, we first introduce the CRF algorithm and its improved version of the fully connected CRF algorithm. Then we innovatively combined multi-scale score map with the fully connected CRF algorithm and propose a chronic wound image segmentation and identification post-processing method based on fully connected CRF. This post-processing method overcomes the insufficient information using the single score map as the input of the fully connected CRF, and thus can get better results. The post-processing method is divided into two steps: Firstly, we calculate the score maps of the corresponding last three layers of the feature maps, then combine the three-layer score maps as the input of the fully connected CRF, and refine the parameters to obtain the best results. The results of the experiments have proved that the post-processing method based on fully connected CRF with multi-layer score map can optimize the origin segmentation results obtained by deep convolution neural networks.

ACKNOWLEDGEMENTS

Thank to everyone who helped me during writing this paper.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks.” in NIPS, 2012.

[2] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv: 1409.1556 (2014).

[3] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[4] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[5] Huang, Gao, et al. "Densely connected convolutional networks." arXiv preprint arXiv:1608.06993 (2016).

[6] Sabour, Sara, Nicholas Frosst, and Geoffrey E. Hinton. "Dynamic routing between capsules." Advances in Neural Information Processing Systems. 2017.

[7] V. Shukla, M. A. Ansari, and S. Gupta, “Wound healing research: a perspective from india,” International Journal of Lower Extremity Wounds, vol. 4, no. 1, pp. 7–9, 2005.

[8] C. K. Sen, G. M. Gordillo, S. Roy, R. Kirsner, L. Lambert, T. K. Hunt, F. Gottrup, G. C. Gurtner, and M. T. Longaker, “Human skin wounds: a major and snowballing threat to public health and the economy,” Wound Repair and Regeneration, vol. 17, no. 6, pp. 763–771, 2009.

[9] J. Posnett and P. Franks, “The burden of chronic wounds in the uk,” Diabetic Medicine, vol. 14, no. 5, pp. S7–S85, 2008.

[10] Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 40(4): 834-848.

[11] Shotton J, Winn J, Rother C, et al. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context[J]. International Journal of Computer Vision, 2009, 81(1): 2-23.

[12] Triggs B, Verbeek J J. Scene segmentation with crfs learned from partially labeled images[C]//Advances in neural information processing systems. 2008: 1553-1560.

[13] Gould S, Rodgers J, Cohen D, et al. Multi-class segmentation with relative location prior[J]. International Journal of Computer Vision, 2008, 80(3): 300-316.

[14] Fulkerson B, Vedaldi A, Soatto S. Class segmentation and object localization with superpixel neighborhoods[C]//Computer Vision, 2009 IEEE 12th International Conference on. IEEE, 2009: 670-677.

[15] He X, Zemel R S, Carreira-Perpiñán M Á. Multiscale conditional random fields for image labeling[C]//Computer vision and pattern recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE computer society conference on. IEEE, 2004, 2: II-II.

[16] Kumar S, Hebert M. A hierarchical field framework for unified context-based classification[C]//null. IEEE, 2005: 1284-1291.

[17] Kohli P, Torr P H S. Robust higher order potentials for enforcing label consistency[J]. International Journal of Computer Vision, 2009, 82(3): 302-324.

[18] Russell C, Kohli P, Torr P H S. Associative hierarchical crfs for object class image segmentation[C]//Computer Vision, 2009 IEEE 12th International Conference on. IEEE, 2009: 739-746.

[19] Ladicky L, Russell C, Kohli P, et al. Graph cut based inference with co-occurrence statistics[C]//European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2010: 239-253.

[20] Zheng S, Jayasumana S, Romera-Paredes B, et al. Conditional Random Fields as Recurrent Neural Networks[J]. 2015:1529-1537.

[21] Krähenbühl P, Koltun V. Efficient inference in fully connected crfs with gaussian edge potentials[C]//Advances in neural information processing systems. 2011: 109-117.

[22] Rabinovich A, Vedaldi A, Galleguillos C, et al. Objects in context[C]//Computer vision, 2007. ICCV 2007. IEEE 11th international conference on. IEEE, 2007: 1-8.

[23] Toyoda T, Hasegawa O. Random field model for integration of local information and global information[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(8): 1483-1489.

[24] Galleguillos C, Rabinovich A, Belongie S. Object categorization using co-occurrence, location and appearance[C]//Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008: 1-8.

[25] Payet N, Todorovic S. ^ 2--Random Forest Random Field[C]//Advances in Neural Information Processing Systems. 2010: 1885-1893.

[26] Liu X, Wang C, Li F, et al. A framework of wound segmentation based on deep convolutional networks[C]//Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2017 10th International Congress on. IEEE, 2017: 1-7.

INTENTIONAL BLANK

POSSIBILITIES OF PYTHON BASED EMOTION RECOGNITION

Primož Podržaj and Boris Kuster

Faculty of Mechanical Engineering, University of Ljubljana,
Askerceva 4, 1000 Ljubljana, Slovenia

ABSTRACT

Vision is probably the most important sense for human beings. As a consequence, our way of behaviour and thinking is also often based on visual information. When trying to perform complex information especially in situations where humans are involved, it is of great benefit if some information can be obtained from images. This is the field of image processing and computer vision. There are various libraries available for these tasks. Probably the best known one is OpenCV. It can also be used in Python programming language. Simple and more complex image processing algorithms are already available in the library. One of the more complex ones is face detection. In this paper it shown how face detection can be executed within Python with OpenCV library. This is the first step needed in emotion recognition. When face is detected, we can determine the emotional state of the subject using a special purpose library.

KEYWORDS

Image processing, Python, OpenCV, face detection, emotion recognition.

1. INTRODUCTION

Of the five senses (vision, hearing, smell, taste, and touch) vision is undoubtedly the one that man has come to depend upon above all others, and indeed the one that provides most of the data he receives [1]. Actually almost all animal species use eyes in fact evolution has invented the eye many times over [2]. The main reason for this is that eyes are very effective sensors for recognition, navigation, obstacle avoidance and manipulation. Artificial sensors that mimic the function of an eye are cameras. Computer vision is inspired by the capabilities of the human vision system and could be defined as the automatic analysis of images and videos by computers in order to gain some understanding of the world [3]. An important part of computer vision is image processing, which means transforming an image in some way.

In this paper some Python libraries that make it possible to perform basic image processing tasks will be introduced. OpenCV, which is the most important one will be presented in some detail. It enables some not so basic image processing (we could say it already enters computer vision). One example is face detection. This is the first step in reaching the main topic of the paper, which is emotion recognition.

Natarajan Meghanathan et al. (Eds) : SPPR, SCAI, CSIA, WiMoA, ICCSEA, InWeS, NECO, GridCom - 2018
pp. 31–40, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.81704

2. IMAGE ACQUISITION AND MANIPULATION

In order to analyze any visual information, we must first get it into an appropriate form. A typical setup for obtaining in image is shown in Fig. 1. The image acquisition process starts with an illumination source, where the light rays are coming from. The scene element is the object under observation. The rays that come from the illuminations source are either reflected or absorbed by the object. Then the imaging system (typically optical lens) collects the incoming light an focuses it on the imaging plane. This is the plane where the sensor should be located. The sensor actually measures the amount of energy received at a specific location. If we want to get a color image, we would have to use filters for each of the three primary colors ((R)ed, (G)reen and (B)lue). As this is too cumbersome, the sensor elements can be arranged in the so called Bayer pattern as shown in Fig. 2. So, for each element we know only one of the colors (note that due to the sensitivity of the human eye there are twice as many green pixels than red or blue).

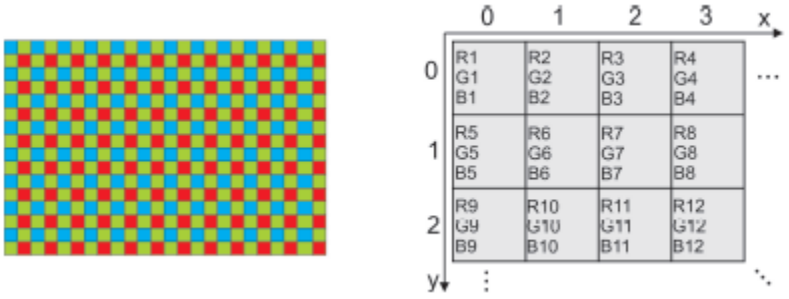


Fig. 2. Bayer pattern (left) and mathematical representation of digital image (right) [5]

The process of obtaining the values for other pixels is called demosaicing. There are various algorithms for this task, which of course differ in execution time. The most simple one just infers the missing color components from the nearest pixel with that color. After demosaicing we are left with a three dimensional matrix. It is however better to speak of a two dimensional matrix, where each element has three components as shown in the left image of Fig. 2. This is also a mathematical representation of an image and also represents the starting point for any image manipulation. In general image manipulation can be defined as any mathematical operation that transforms the original two dimensional image into another one. There are many possible divisions of these opeartions. In this short introduction we will divide them into two groups:

1. Point processing operations
2. Neighborhood processing operations

Point processing is an operation where the intensity value (in monochromatic image) or intensity values (color image) in the transformed image i_t depend only on the intensity value (values) of the corresponding pixel in the original image i_o.

$$i_t(x, y) = f(i_o(x, y)) \tag{1}$$

Although, being a very simple operation, it has some applications. A typical one is full-scale histogram stretch which is easily the most common linear point operation [6]. Another important operation is for example the conversion of a color image to a grayscale (monochromatic) one using the following equation [7]

$$i_t(x,y) = 0.299 \cdot i_{oR}(x,y) + 0.587 \cdot i_{oG}(x,y) + 0.114 \cdot i_{oB}(x,y) \quad (2)$$

Neighborhood processing is an operation where the intensity of a specific pixel in the original image depends on the intensity of more than pixel in the original image. Typically the corresponding pixel and some pixels in the neighborhood (that's where the operation gets its name from). A typical operation of this kind is filtering used to remove or at least decrease the noise in an image. One possible (very simple) formula for such an operation (using only four neighboring pixels) can be written as follows:

$$i_t(x,y) = \frac{i_o(x,y-1) + i_o(x-1,y) + i_o(x,y) + i_o(x+1,y) + i_o(x+1,y+1)}{5} \quad (3)$$

It probably doesn't need to be stressed out that neighborhood processing is much more powerful. As a consequence almost all the applications need that kind of operations.

3. PYTHON PROGRAMMING LANGUAGE

3.1 Some Basic Information

Python is a high-level general-purpose programming language created by Guido van Rossum in 1991. It has a design philosophy that puts emphasis on code readability. It supports multiple programming paradigms including object-oriented, imperative, functional and procedural and has a large standard and comprehensive library. The first release was followed by Python 2.0 in 2000 and Python 3.0 in 2008. At the time of writing this paper the latest version is Python 3.7. Python is a good choice for all the researchers in the scientific community because it is [8]:

- free and open source
- a scripting language, meaning that it is interpreted
- a modern language (object oriented, exception handling, dynamic typing etc.)
- concise, easy to read and quick to learn
- full of freely available libraries, in particular scientific ones (linear algebra, visualization tools, plotting, image analysis, differential equations solving, symbolic computations, statistics etc.)
- useful in a wider setting: scientific computing, scripting, web sites, text parsing, etc.
- widely used in industrial applications

In comparison with other programming languages such as C/C++, Java, and Fortran, Python is a higher-level language. The computation time is therefore typically a little longer, but it is much easier to program in. In the case of C and Fortran, wrappers are also available. PHP and Ruby on the other side are high-level languages as well. Ruby can be compared to Python but lacks scientific libraries. PHP on the other hand is a more web-oriented language.

Python can also be compared to Matlab, which has a really extensive scientific library. It is however not open source and free. Scilab and Octave are open source environments similar to Matlab. Their language features are however inferior to the ones available in Python. People in general tend to think that complex problems demand complex processes in order to produce complex solutions. Python was developed with exactly the opposite philosophy. It has an extremely at learning curve and development process for software engineers [9]. It is used for

system administration tasks, by NASA both for development and as a scripting language in several of its systems, Industrial Light & Magic uses Python in its production of special effects for large-budget feature films, Yahoo! uses it (among other things) to manage its discussion groups and Google has used it to implement many components of its web crawler and search engine [10]. As Python is also a language that is easy to learn and both powerful and convenient from the start [11], we might soon be asking, who is not using it.

As already mentioned Python has an extensive set of libraries which can be imported into a project in order to perform specific tasks. The ones that really should be mentioned in any scientific paper dealing with mathematics are NumPy and SciPy. NumPy is a library which provides support for large, multi-dimensional arrays. As images are in fact large two (greyscale) or three (color) dimensional matrices, this library is essential in all image processing tasks. It should also be emphasized that many other libraries (not limited to image processing) use NumPy array representation. SciPy is a library build on the NumPy array object and contains modules for signal and image processing, linear algebra, fast Fourier transform, etc. The last library mentioned in this introductory section is Matplotlib. As the name suggests this library is a plotting library. Although it is used a lot in all areas of science, Image processing relies heavily on it.

3.2 Basic Image Processing Libraries

There are several Python libraries related to Image processing and Computer vision.
The most important ones are however:

- PIL/Pillow
  This library is mainly appropriate for simple image manipulations (rotation, resizing, etc.) and very basic image analysis (histogram for example)

- SimpleCV
  It's a library intended (as the name suggests) to be a simplified version of OpenCV. It doesn't offer all the possibilities of OpenCV, but it is easier to learn and use.

- OpenCV
  It is by far the most capable and most commonly used computer vision library. It is written in C/C++, but Python bindings are added during the installation. It also gives emphasis on real time image processing.

Among the ones, which will not be presented it might be worth mentioning Ilastik. It is a simple, user-friendly tool for interactive image classification, segmentation and analysis. By far the most important library is however the OpenCV library. Some of its capabilities will be demonstrated.

3.3 OpenCV library

OpenCV is an open source computer vision library available written in C and C++ which runs under Linux, Windows, Mac OS X, iOS, and Android. Interfaces are available for Python, Java, Ruby, Matlab, and other languages. A very simple program used just to show an image (see Fig. 3) can be written as follows:

```
import numpy as np
import cv2
img = cv2.imread ('lena-color.jpg')
cv2.imshow('image', img)
cv2.waitKey(0)
cv2.destroyAllWindows()
```

It is also very easy to form a binary image with a certain predefined threshold. The following commands give a binary image (threshold is set to 127) for the image shown in Fig. 3.

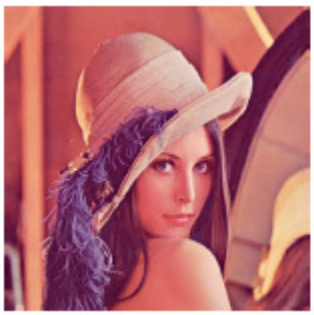


Fig. 3. Lena color image

Fig. 4. Binary image

```
gray_image = cv2.cvtColor (img, cv2.COLOR_BGR2GRAY)
ret, binary_image = \
       cv2.threshold (gray_image, 127,255,cv2 .THRESH_BINARY)
cv2.imshow( 'image', binary_image)
```

One of the important image processing tasks is edge detection. In OpenCV this task can be performed using a simple command:

```
edges = cv2.Canny(gray_image, 100, 200)
```

If the source image is the one shown in Fig. 3, we get the resulting image as shown in Fig. 5. The parameters 100 and 200 define the limiting values of the intensity gradient. Pixels below the lower value are non-edge pixels. Pixels above the upper value are edge pixels. Pixels in between are edge pixels if they are connected to the pixels with the intensity gradient above the upper limiting value. Of course they can be set arbitrarily.

4. COMPUTER VISION TASKS RELATED TO EMOTION RECOGNITION

Emotion recognition is quite a complex task to achieve. Basically it is divided into two steps (face detection and emotion recognition).

4.1 Face Detection

The task of face detection is (as the words suggest) to find the face(s) (sometimes also eyes) in an image. The following sequence of commands does just that.

```
face_cascade = cv2 . CascadeClassifier ( 'C: \\ Users \\ . . .
\\ haarcascade_frontalface_default . xml ' )
eye_cascade = cv2. CascadeClassifier ( 'C: \\ Users \\ . . .
\\ haarcascade_eye. xml ' )
img = cv2 . imread ( ' lena - color.jpg' )
gray = cv2 . cvtColor ( img, cv2 .COLOR_BGR2GRAY)
faces = face_cascade. detectMultiScale ( gray, 1.3, 5)
for (x,y,w,h) in faces :
    cv2 . rectangle ( img, ( x,y), ( x+w, y+h ), ( 255 , 0 , 0 ) , 2 )
    roi_ gray = gray [ y : y+h , x : x+w]
    roi_color = img [ y : y+h , x : x+w]
    eyes = eye_cascade.detectMultiScale ( roi_ gray )
    for ( ex, ey, ew, eh ) in eyes :
        cv2 . rectangle ( roi_color, ( ex, ey ) , ( ex+ew, ey+eh ),
                          ( 0, 255, 0 ), 2 )

cv2 . imshow( ' img ' , img )
cv2 . imwrite ( ' face_lena. jpg ', img )
cv2 . waitKey ( 0 )
cv2 . destroyAl lWindows ( )
```

The result is shown in Fig. 6. For the image in Fig. 6 algorithm works perfectly.

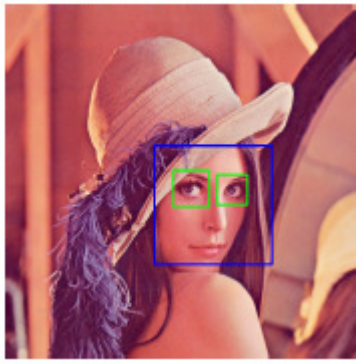


Fig. 6. An example of face detection

If however a more complex image is used, the result (especially for eyes) is not so good. See for example Fig. 7. The algorithm itself applies the so called Haar feature-based cascade classifiers. It was proposed as an effective object detection method by Paul Viola and Michael Jones [12]. The number of these features can be enormous. But most of them are irrelevant. A good feature is for example the fact that the region of the eyes is usually darker than the region of the nose and cheeks. A second good feature could for example be based upon the fact that the eyes are usually darker than the bridge of the nose. With the increasing number of such features, we can increase the reliability of the algorithm. Misclassifications are of course always a possibility. It should also be noted, that the reliability decreases with the decreasing amount of pixels in the face area.

Fig. 7. An example of face detection for many people being on the image

5. EMOTION RECOGNITION

Emotion recognition is very difficult computer vision task. In essence, what we expect from this task is to group faces in an image into one of the following seven emotions: angry, disgust, fear, happy, sad, surprise, neutral. The principle behind the algorithm is the so called deep learning shown schematically in Fig. 8. The input signal are the faces detected. After they are detected (as described above), they are cropped out of the image and fed into a convolutional neural network. Convolutional networks are a neural network architecture particularly well suited to processing images. A "kernel" is slid over the image and multiplied with its pixel values. A kernel's weights are learned using backpropagation. In effect, kernels find patterns in the image regardless of their position in an image, which allows the convolutional neural network to be spatially invariant.

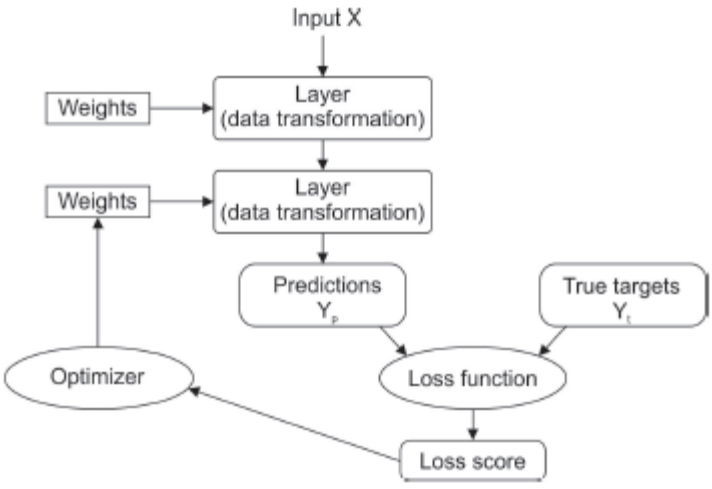


Fig. 8. Schematic representation of deep learning concept [13]

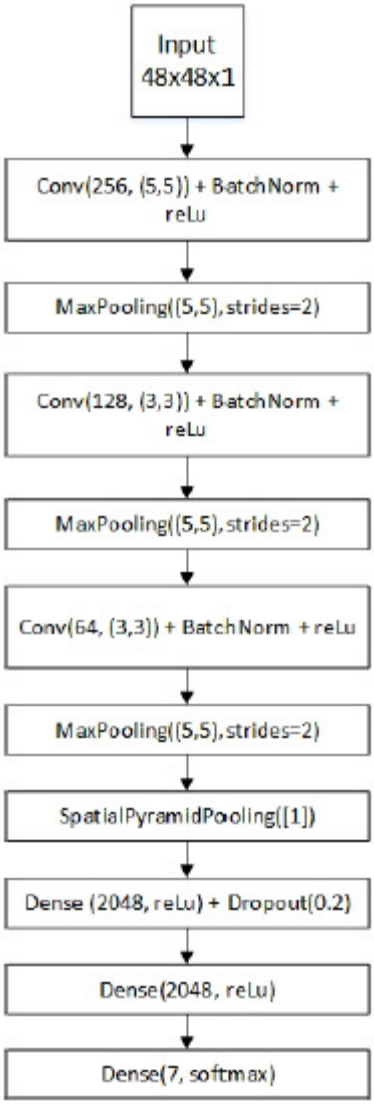


Fig. 9. Convolutional Neural Network used in the paper

Our particular implementation of the CNN is shown schematically in Fig. 9. It consists of three convolutional layers, each followed by a Batch Normalization layer, Rectified Linear Unit (relu) activation layer and max pooling, to reduce the representation size. After this, a Spatial Pyramid Pooling layer enables us to input pictures of arbitrary input shape. Some regularization is applied to the convolutional kernels to reduce the possibility of over-fitting. The CNN layers extract features from the images, which are then fed into a multi layer perceptron neural network, with the layers containing 2048, 2048 and 7 neurons, respectively. Some Dropout (0.2) is applied after the first dense layer, to combat over-fitting. The output layer contains 7 neurons and a softmax activation function, since we classify faces into one of the above mentioned seven emotions. We use the Adam optimizer with the default learning rate of 0.01. In order to train the network, the Facial Emotion Recognition dataset from Kaggle has been used. It contains around 30000 grayscale pictures of 48x48 dimension, each containing a centered face. The faces emotions are labeled.

Fig. 10. Results of the program

Some examples of images being fed to our program are shown in Fig. 10. Not all of the images are classified correctly. An example of such a case is shown in Fig. 11. The accuracy of the network on test set is around 50%, while the best result on Kaggle leaderboard achieved around 70% accuracy. We suspect a dataset of higher resolution face images would increase the recognition ability of the network.

Fig. 11. An example of a wrong emotion recognition

6. CONCLUSION

The concept of deep learning has probably very bright future ahead. Some people even say that it might be the concept which will finally rival human intelligence in many areas. In this paper the deep learning concept has been used to asses the possibilities of emotion recognition in images.

At first some very basic information about image acquisition and its mathematical representation is given. This is followed by the introduction of python and its comparison with some other popular programming languages. As the topic of the paper is image processing, related libraries are given. OpenCV, being most popular is described in some detail. Face detection and emotion recognition are outlined as well. The results show that further research is needed in order to increase the accuracy of the approach.

REFERENCES

[1] E. R. Davies, Computer and machine vision: theory, algorithms, practicalities. Academic Press, 2012.

[2] P. Corke, Robotics, Vision and Control: Fundamental Algorithms In MATLAB, 2nd Ed..Springer, 2017.

[3] K. Dawson-Howe, A practical introduction to computer vision with opencv. John Wiley & Sons, 2014.

[4] R. C. Gonzalez and R. E. Woods, Digital Image Processing, 3rd Ed.. Pearson Education, 2008.

[5] T. B. Moeslund, Introduction to Video and Image Processing - Building Real Systems and Applications. Springer, 2008.

[6] A. C. Bovik, The essential guide to image processing. Academic Press, 2009.

[7] G. Blanchet and M. Charbit, Digital signal and image processing using MATLAB, Vol. 1 - Fundamentals. ISTE, 2014.

[8] C. Fuhrer, J. E. Solem, and O. Verdier, Scientific Computing with Python 3. Packt Publishing Ltd, 2016.

[9] S. Nagar, Introduction to Python: For Scientists and Engineers. Bookmuft, 2016.

[10] M. L. Hetland, Beginning Python: from novice to professional, 3rd Ed.. Apress, 2017.

[11] R. V. Hattem, Mastering Python: master the art of writing beautiful and powerful Python by using all of the features that Python 3.5 offers. Packt Publishing, 2016.

[12] P. Viola, and M. Jones, Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. I-511-I-518, 2001.

[13] F. Chollet, Deep learning with Python. Manning Publications Co., 2017.

AUTHORS

Primož Podržaj received his PhD from the University of Ljubljana where he is an Associate Professor in the field of Control Theory. He is the head of the Laboratory for Process Automation. His research interest includes machine vision, image processing and artificial intelligence.

Boris Kuster received his bachelor's degree in mechanical engineering from the University of Ljubljana, where he is now studying for a master's degree in mechatronics. His primary interests are machine learning, artificial intelligence and robotics.

PHISHING DETECTION FROM URLS BY USING NEURAL NETWORKS

Ozgur Koray Sahingoz, Saide Işılay Baykal and Deniz Bulut

Department of Computer Engineering, Istanbul Kultur University,
Istanbul, Turkey

ABSTRACT

In recent years, Internet technologies are grown pervasively not only in information-based web pages but also in online social networking and online banking, which made people's lives easier. As a result of this growth, computer networks encounter with lots of different security threats from all over the world. One of these serious threats is "phishing", which aims to deceive their victims for getting their private information such as username, passwords, social security numbers, financial information, and credit card number by using fake e-mails, webpage's or both. Detection of phishing attack is a challenging problem, because it is considered as a semantics-based attack, which focuses on users' vulnerabilities, not networks' vulnerabilities. Most of the anti-phishing tools mainly use the blacklist/white list methods; however, they fail to catch new phishing attacks and results a high false-positive rate. To overcome this deficiency, we aimed to use a machine learning based algorithms, Artificial Neural Networks(ANNs) and Deep Neural Networks(DNNs), for training the system and catch abnormal request by analysing the URL of web pages. We used a dataset which contains 37,175 phishing and 36,400 legitimate web pages to train the system. According to the experimental results, the proposed approaches has the accuracy in detection of phishing websites with the rate of 92 % and 96 % by the use of ANN and DNN approaches respectively.

KEYWORDS

Phishing Detection System, Artificial Neural Networks, Deep Neural Networks, Big Data, Machine Learning, Tensor flow, Feature Extraction

1. INTRODUCTION

Due to the extensive growth in the number of internet users, lots of our daily life operations are transferred from the real world to the cyber world such as communication, coordination, commerce, banking, registrations, applications, etc. Because of this, the malicious peoples and attackers also transferred to this world and make their threats and crimes easily anonymously. To ensure the security and privacy of cyber data, technology must be used and organized carefully by using "Cyber Security" concept [1].

According to ITU-T, cyber security is the accumulation of tools such as policies, security safeguards, training, risk management approaches guarantee and technologies that can be used to protect the cyber organization and environment. [2] Another source explains this concept as follows: Cyber security is the body of technologies about processes, networks, computers programs and data. Its aim is designed for protect these components of technologies from attack, damage and unauthorized access [3]. According to Craigen et.al. cyber security is the organization and collection of resources, processes and structures used to defend cyberspace and cyberspace enabled systems from events that misrelate by default ownership rights [4]. Cyber

Natarajan Meghanathan et al. (Eds) : SPPR, SCAI, CSIA, WiMoA, ICCSEA, InWeS, NECO, GridCom - 2018
pp. 41–54, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.81705

security applies precaution methods used to protect data from being stolen, concurred or attacked [5]. All the definitions of cyber security say about prevent and protect: Cyber security prevents from fraud or thief who wants to seize person/public/national information or connection.

"Identity theft" or specifically "phishing" is one of the most threatening security deficits of the users in the Internet. In this type of crimes, attackers use some malicious web pages which impersonate as legitimate web sites, to collect the victims' critical information such as username, passwords, financial data, etc. Typically, a phishing attack starts with an electronic mail which seems to come from a reputable company as depicted in Figure 1. The content of the mail encourages the victim to click on the address, which can also be hidden as a hypertext. This address directs the victim to a fake web site, which is designed exactly similar with a valid website, such as an e-mail site social engineering site of generally financial institutions web sites.

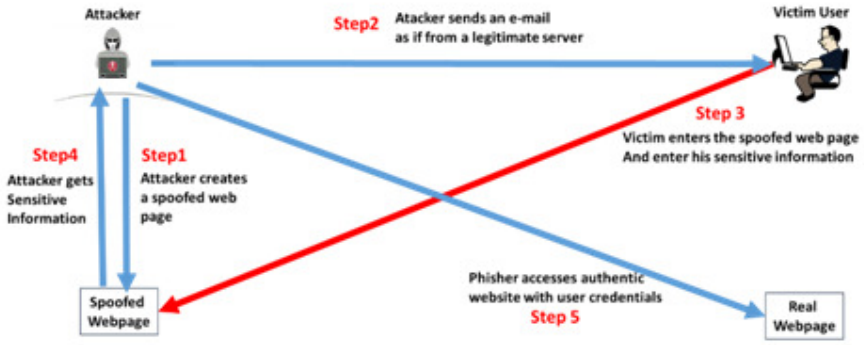


Figure 1: Life Cycle of a Phishing Attack

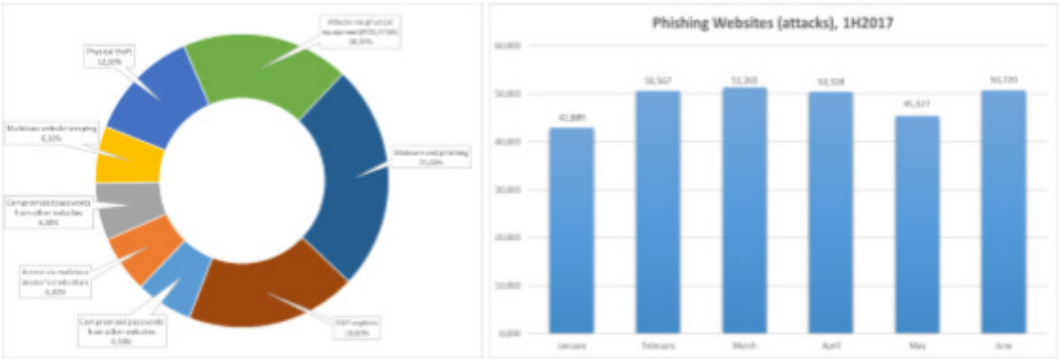
As can be seen from this life cycle, even experienced computer user can fall into the phishing attack and be a victim. Therefore, for detection of phishing attacks, a dynamic support and security mechanism is needed. As a phishing detection algorithm, generally blacklists/whitelists are used. This is an effective prevention mechanism and it quickly classify an URL as a phishing or legitimate. However, as emphasized in [18] between 47% and 83% of phishing web pages are blacklisted in 12 hours, which is enough duration for deceiving most of the people. Additionally, within the first 2 hours, about 63% of phishing campaigns are finished. Therefore, blacklists/whitelists are not effective especially for zero-day attacks.

To overcome this type of attack there is need to construct a dynamic and efficient algorithm which can learn the structure of the legitimate web pages and classifies the abnormal ones. Therefore, in this project, we aimed to set up a classification system, which can identify whether an URL is either phishing or legitimate. To train the system we have used a dataset which contains about 74,000 items in both these types. To compare the efficiency of the different algorithms and select the best one, we used both Artificial Neural Network (ANN) and Deep Neural Network(DNN) approaches for training and testing the system with the help of Tensorflow framework. And experimental results showed that the proposed approaches produce very good accuracy rates for detecting phishing URLs. Within the proposed approaches, DNN gives better accuracy rate than ANN with the related values as depicted in the results section.

The rest of the paper is organized as follows: In the next section the background knowledge is given. Section 3 depicts the design details of the proposed system. Experimental Result are shown with comparative graphic in Section 4. Finally, Conclusion and future works are listed

2. BACKGROUND

In oxford dictionary, phishing means "an effort by hackers to destroy or damage a computer system or network". It means broking the "confidentiality, integrity, and availability"-CIA triad rules. In the real world there are many attack types for broking this CIA such as Sniffing, Denial of Service (DoS), Sql Injection, Spyware, Viruses, Trojans, Social Engineering, Worm, Botnet and Phishing [8]. However, as can be seen from Figure 2.a. Phishing attacks are located at the first position.



a) Phishing vs other attacks b) # of Phishing web sites in 1st half of 2017

Figure 2. Phishing Statistics

Also, as can be seen from Figure 2.b. Phishing attack is a continuing process, in every part of the year this attack takes its place in the cyber world. Phishing is an attack type using both social engineering and technical hints to have users' personal identity information and bank account details [9]. There are many phishing attack types in the literature. The most preferred one the use of emails. Attacker prepares an email which urges the user for entering his valuable information on a malicious webpage as depicted in Figure 3.a. In this e-mail there are some hyperlink which directed the user to this malicious webpage, which is exactly similar with the original one. After the user enters the information on the webpage, attacker can access the victim's sensitive information

a)Phishinge-Mail b)SpoofedWebpage

Figure3.DeceivinguserwithE-mailandspoofedwebpage

2.2. TYPES OF PHISHING ATTACKS

Phishing attacks can be divided in 2 layers: social engineering and technical subterfuge. Social engineering layer includes attackers, victim, sending fake email, which contains spoofed webpages. This process starts by sending this email, which comes from a legal and famous organizations for gathering some sensitive information such as user name, id, password, credit card information etc. Second layer is about spoofed webpage. Fake e-mail directs the victim to the spoofed webpage which appears visually very similar to the original webpage. This layer also uses cross-site scripting, session hijacking, malware phishing, DNS poisoning and key/screen loggers' techniques. These layers send the obtained information and get remote access by attackers to victim's computer or original webpage [12, 13]. According to [14], mostly attacked websites are shown in Figure 4.

.

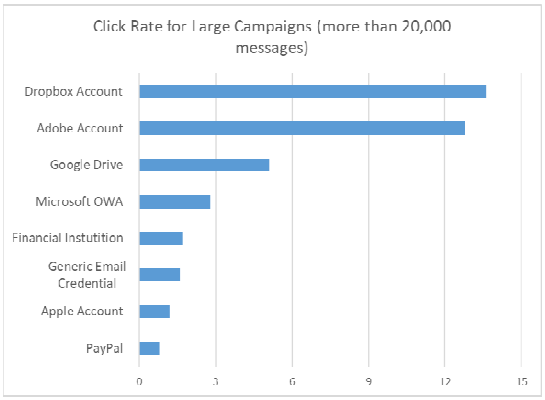


Figure 4: According to statistics of company phishing

2.2. DETECTION OF PHISHING ATTACKS

Phishing attacks can be applied by using a lot of methods. Detection ways have been found on the basis of this attack types. This section will explain the detection methods of phishing attacks. The main vulnerability of phishing is about the Human Factor. Therefore, the main prevention is about the education of the workers, how to avoid from this type of attacks. However, due to the type of new attacks, even experienced used can fall into this type of attack. Therefore, a cyber support will be helpful for the users.

The mostly preferred methods to prevent phishing is the use of Blacklists, which are periodically updated list which includes some keywords lists, URLs and IP addresses. The famous blacklist using methods are: Google Safe Browsing API, DNS-Based Blacklist, Phish Net: Predictive Blacklisting, Automated Individual White-List. However, due to its deficiency for detecting zero day attack, some security managers prefer the use of Heuristics approach, which analyses and investigates the feature of the web page and detect whether the page use this information or not [19]. The reputable heuristics anti-phishes are Spoof Guard, Collaborative Intrusion Detection, Phish Guard, Phish wish, CANTINA, and etc.

Visual Similarity method uses the visual similarity of the webpage like its source code, contained pictures, text and additionally some formatting, logo, CSS and HTML tags, etc. These features are compared with the previous form of the web page or its stored copy in the local server. However, this technique has an important deficiency that it cannot detect the phishing attacks of

the newly generated web pages. Besides, its image-based operation, comparison gets too much time for detection.

The dynamic approach can be seen as the use of data mining and/or machine learning techniques. If there are sufficient number of legitimate and non-legitimate web pages and their related features, it can be easy to train the system with this dataset by the use of some machine learning techniques. Support Vector Machines, Bayesian Classifier, KNN techniques, Ad boost, Random Forest, decision tree, neural networks, etc.

2.3. NEURAL NETWORKS

Machine learning is one of the very important field of computer science, which allows software to learn and adapt to inputs and improve performance on a specific task. Machine learning is highly used to follow human behaviours and to make some predictions by using either supervised or unsupervised algorithms. Neural networks are designed influenced from biological neural networks. In real neurons, the input data are processed and transmitted by use of electrical signals. In artificial neural networks, system works with input nodes –it is called as neuron-, edges as functions, layers, and output neurons. All these components related with nodes and edges. Input neurons connecting other neurons via functions. A simple diagram of a neuron is shown in Figure 5.

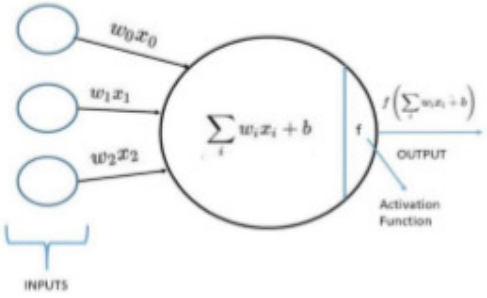


Figure 5: A simple neuron structure.

Even though given inputs are the same, weight and bias criteria can be changed to calculation. Almost all neurons calculate for the next neurons by that formula. And they are collecting activation functions such as RELU, TANH, etc. According to activation functions all these multiplication and addition process collecting fully connected layer. Than predicting output decreased by some loss functions. This output is gathering and comparing real value. At the end of the output, this result optimizing and so on. Figure 6 shows structure of neural networks. [15]

Neural networks are divided into two sub networks, which artificial neural networks and deep neural networks, which use multiple layer in its framework as depicted in Figure 6. According to the parameters and size of the problem the number of hidden layers and also the number of neurons in each layer can be changed. If you only use a single hidden layer, this is mainly called as ANN structure.

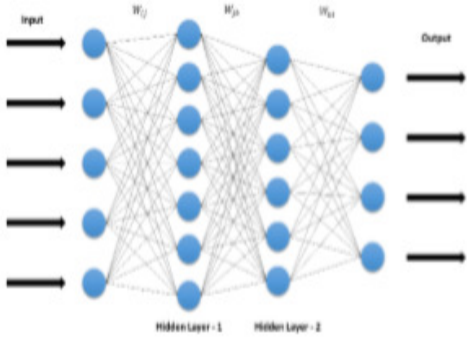


Figure 6: Two Hidden Layer Deep Neural Network

2.4. RELATED WORK

There are many works in the literature, which are focussed on phishing detection. According to study of James et. Al. attackers are swindled via e-mails to get individual information and bank account details like usernames, userid and passwords [16]. This paper contains information about machine learning methods, which is used for detecting phishing websites. In this paper two success rate are analysed which is WEKA and MATLAB. The J48 Decision Tree gave best result in WEKA. When dataset is splitted 60% for testing, detection accuracy was 93.2% in lexical features. Regression Tree was given best result with 91.08% accuracy in MATLAB when using 40% dataset for training however accuracy was decreased when using 10% of dataset for testing.

Buber et.al. suggest that, cyber-attacks affect to many people and foundation and this attack can cause financial damages in this work [17, 21, 22]. There are a lot of cyber-attack types. Purpose of phishing attack, which is one of them, is getting confidential information of users by using people's weaknesses. In this paper, a machine learning based system was developed for detecting phishing attacks. Some features were generated by using taking advantages of Natural Language Processing (NLP) in this system. For detecting URL which is used in phishing attacks, a system was developed by using these features. According to tests Random Forest Algorithm showed the highest result success rate.

3. METHODOLOGY

In the implementation phase we developed two different classifiers with: Artificial Neural Networks and Deep Neural Networks. Due to their structure we need to use some numeric values as the input of our system. Therefore, we need to select some features from the URL and then train and execute our system based on these parameter values.

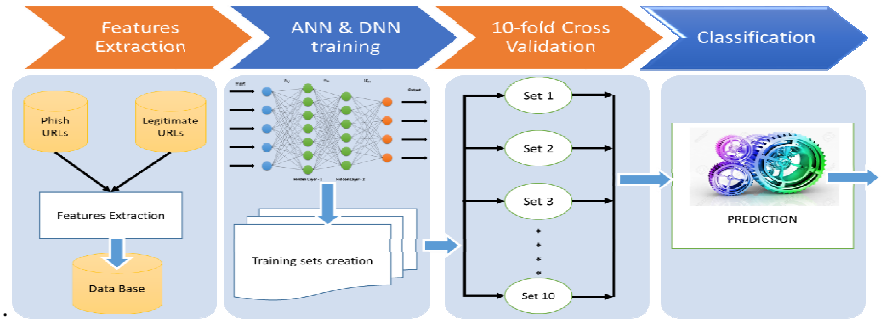


Figure 7. Execution Diagram of the Proposed System

To understand the meaning of each feature, firstly we need to identify the parts of URLs. In the next subsection, this concept is explained. After that, the selected features are detailed.

3.1. URLS

To understand the approach of phishers, firstly, the components of URLs and their aim should be understood. The basic components of a URL is depicted in Figure 8.

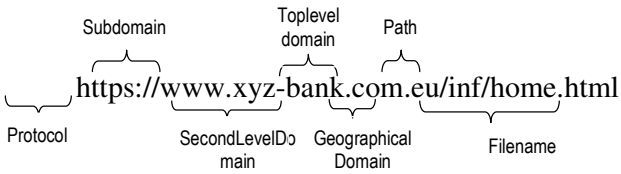


Figure 8. Components of a URL

In the standard form, a URL starts with its protocol name, such as hypertext transfer protocols, file transfer protocols, etc., which are used to access the web page. Consequently, the subdomain and the Second Level Domain (SLD) names identify the server hosting the web site. SLD name is very important for us, because this part mainly contains the name of the firm, therefore, phishers focussed on this part and try to produce different forms of name which are like original ones. The Top-Level Domain (TLD) name shows the domains in the Domain Name System root zone of the Internet such as educational, commercial government, etc. Finally, Geographical Domain name shows the geographical location of the web site such as, Germany, Turkey, France, etc.The previous four parts compose the domain name (host name) of the web page; however, the inner address is represented by the path of the page in the server and with the name of the page in the html form. The ongoing part is like a folder a file name which shows the location of the file in the server.

3.2. SELECTED FEATURE

In this subsection, we detailed the selected features that are used in the implementation of the proposed system. There are total 27 features, and they are detailed as follows.

1. Length of the URL: Phishers generally hide the address of their spoofed web page by increasing the length of the address. In this long text they also add the name of the attacked web page, but this is not the domain name part of the URL. Additionally, if this length is increased too much, then it will not fit the address bar, and the victim cannot see the domain part. Some researchers focussed on this size and they grouped the URL according the following rule [20]:
If the length of the URL<54, then it is classified as “legitimate”, If the length of the URL is between 54 and 75, then it is classified as “suspicious”, If the length of the URL>75, then it is classified as “phishing”,

However, in our study, we don’t make this type of classification. Classification is executed by the classifier, and this value is only a parameter for our classifier. Shorter URLs have the greater possibility for being “legitimate”.

2. Punctuation character count: Phishers use some meaningless characters for confusing the victim. Therefore, they can also use some punctuation characters, especially “.”, ”;”, ”!”, ”&”, ”%”, etc. Increased value has more tendency to be a phishing webpage as depicted in Figure 9.

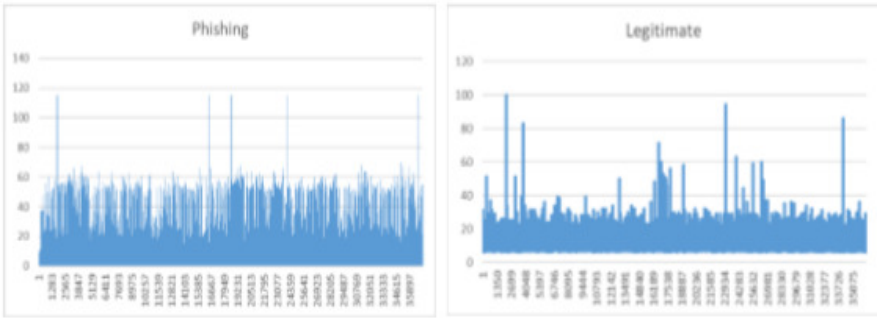


Figure 9. Number of Punctuation Characters in the URLs

3. Is it an IP address: This is a binary feature, if it is an IP address then its value is 1, else 0. For deceiving the users, phisher generally try to use some part of the original URL in the spoofed URL addresses. Therefore, they do not prefer the use of IP address in their attack.

4. Suspicious words count: Phisher prefers some specifics words such as 'confirm', 'account', 'secure', 'admin', 'login', 'submit', 'update', 'setup', 'secure', etc. These words helps for the victims to think the related web page is legitimate. Therefore, we get ?? suspicious words which are selected in the study of Buber et.al. The number of these words are used as a feature in our system.

5. Alexa ranking: There are more than 1.7 billion websites all around the world. Alexa holds popular websites and ranking them. Generally, the popular websites are not preferred for phishing attacks. Most of the phishing campaigns execute their attack in the first 2 hours and after 12 hours it can enter blacklists. Therefore, these sites cannot get upper location in this list. If a website has a higher location in the list, this increase the probability of being legitimate.

Apart from the others, this is a domain-based feature. This feature is not directly derived from the URL. We need to use a third-party service to calculate the Alexa ranking. Therefore, use of this feature slow down the execution of system.

6. Number of brands: Use of brand names is generally preferred by the phishes. We collected our brand name list from the first 500 firms in the Fortune, some brands from the Alexa ranking system, some banks (international), some social networking and micro blogging sites.

7. (3 Features) Average/Longest/Shortest Word lengths: For confusing the victims mind, Phishers use different length of works in their address. The length of the words in the URL is also an important feature for us. We get three different features as average, shortest and longest words in the URL.

8. Number of keywords: Use of some special keywords can also deceive the computer users. Therefore, we identify some keywords such as "login, secure, account, server" which are mostly preferred in the malicious URLs and then construct a keyword list. This list contains about 176 words and is constructed especially from the URLs in the Phish tank and this list only contains English words.

9. (8 features) Number of special characters and words ('.', '=', '_', '-', '\\', '@', 'com', 'cmd'). While investigating the phishing URLs which are get from the Phishtank, it is seen that some special characters and words are mostly preferred. Therefore, we get the number of all of these as different features in the proposed system. For example, if we look at "paypal.com-login.com", we

can see that “paypal” is only a subdomain and original host name is “com-login.com”. However, use of “paypal”, “.” and “com” together results the user to see the host name as “paypal.com”. A standard computer user is hardly seeing this fact, therefore a software based support is important for us. For example, the comparison of the “number of @ characters” between the legitimate and phishing URLs is depicted in Figure 10.

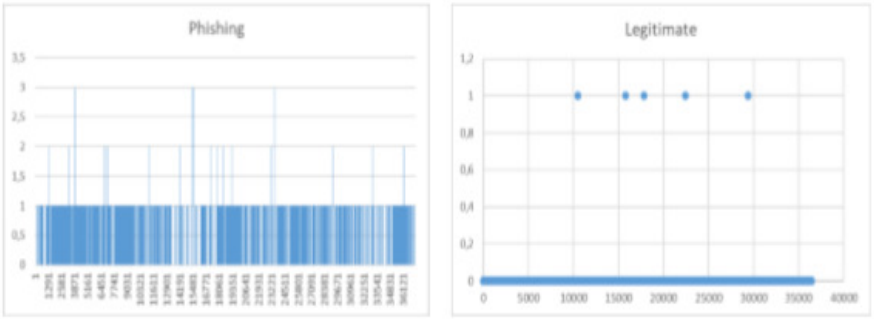


Figure 10. Number of @ Characters in the URLs

Additionally, the use of special characters can also be so deceptive. For example, “mail.google.com” is a legitimate webpage, however, phishers can change it as “mailgoogle.com” with different host name, which is hard to distinguish from the original one.

10. Subdomain number: Legitimate URLs generally have a smaller number of subdomains, however, as explained in the previous example phishers can use the subdomain names as if the domain names. Additionally, they can use several subdomains name the address similar to the original ones. Therefore, a smaller number of subdomains increase the probability of being legitimate web page.

11. Number of Digits: To pass some spamming filters, phishers use some numeric characters in their URLs. Generally, there is no occurrence of numeric characters in the domain name of the legitimate web site.

12. Standard deviation of the words’ length: In the URL (especially in long phishing URLs) there are a number of words. The standard deviation of them is get as a feature in the system.

13. Number of words: The number of words is also an important feature This feature also contains the compound words, which are two or more words that are combined to form a new word with different/similar meaning. To deceive the users, phishers also use compound words in the URL. Therefore, there is need to find each word (and compounds words) in this address. The comparison of phishing dataset and legitimate dataset is shown in Figure 11.

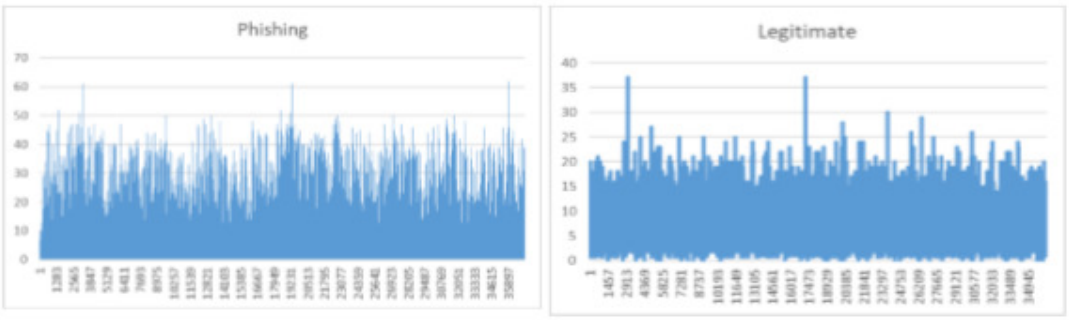


Figure 11. Number of words in the URLs

14. Average length of the compound words: In the previous feature we get the number of compound words. However, the size, especially the average size, of these words is also important to detect the phishing attacks.

15. Character Repetition: To cheat the use phisher can repeat some characters in the domain name. For example, “apple.com” can be repeated by “applle.com” or “applee.com”. This type of names can also be distinguished by the use of similarity index. Usage of some distance measures can ease the calculation of this value.

16. (2 Features) Use of “username” and “userid”: While analysing the phishing URLs from gathered from the Phishtank, it is seen that many of the URLs contains these specific words, “as “username” and “userid”, which are used for deceiving the user. Therefore, these features are defined as binary features and if these words exist then their values are 1, else 0.

3.3. TRAINING THE SYSTEM

The success of the system depends on the learning/training mechanisms used. In the proposed system we used two different learning mechanism: Artificial Neural Network and Deep Neural Network. In Artificial Neural Network(ANN) approach we used a one hidden layer framework, which contains 20 neurons in it. Due to its structure, we trained the system with only 100 epochs and we preferred the use of “adam” optimizer. As an activation function different functions can be selected: RELU, SIGMOID and TANH. Therefore, we tested all of them and found that SIGMOID function gives the best performance among them.

In the Deep Neural Networks design we increased the number hidden layers to two and at every layer, ‘RELU’ activation function is used. Each hidden layer contains 20/40 neurons is used. In the output layer, the activation function is preferred as sigmoid while the optimizer function is preferred as ‘adam’. Training is executed for 100 epochs and we can reach about 91% accuracy rate. To train and test the proposed system we used the Tensorflow, which is an open-source library for data science. It contains some learning algorithms that can be used in different application areas. As an important advantage, system can be run not only on multiple CPUs but also on Graphics Processing Units (GPUs).

3.4. CROSS VALIDATION

Cross-validation is a statistical method to evaluate a stability of the training models by splitting the original dataset into two parts: a training set and a test set. Due to its simplicity and understandability, it is a popular method, which results in a less biased or less optimistic experimental results. To reach a randomness free experimental result we used these set as 10-fold cross validation and divide original data to the ten parts and get one of them as test set while using the other nine as train set.

3.5. CLASSIFICATION

After training the system, we can easily classify any URL in the system. Before executing the classification, firstly related features must be extracted from the URL. After that according to used third party depended features, such as Alexa Ranking, there is a need to connect with this part. After collecting each features classification algorithm is executed. 4

4. EXPERIMENTAL RESULTS

In this study, we compared deep neural network approach with the artificial neural network approach by using the defined features. To train the system we need to use a dataset. Therefore, we prefer the up to date dataset of Buber et. al., which contains 36400 legitimate and 37175 phishing URLs in it.

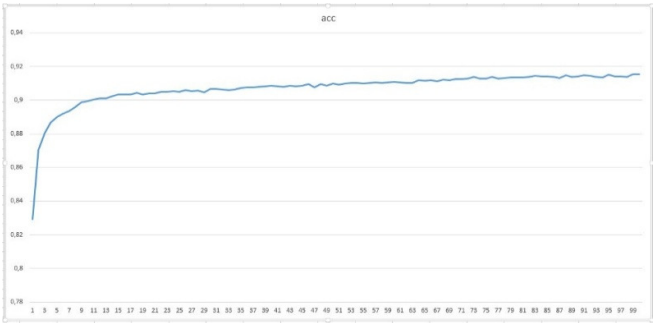


Figure 12. Accuracy Rate of ANN Approach with Sigmoid Activation Function

After, training and testing the data set, best result is reached in Deep Neural Network approach up to 96% accuracy rate with 100 epochs as depicted in Figure 13 with different number of neurons in the hidden layers. If we increase the epoch number, this rate is increasing a little bit more.

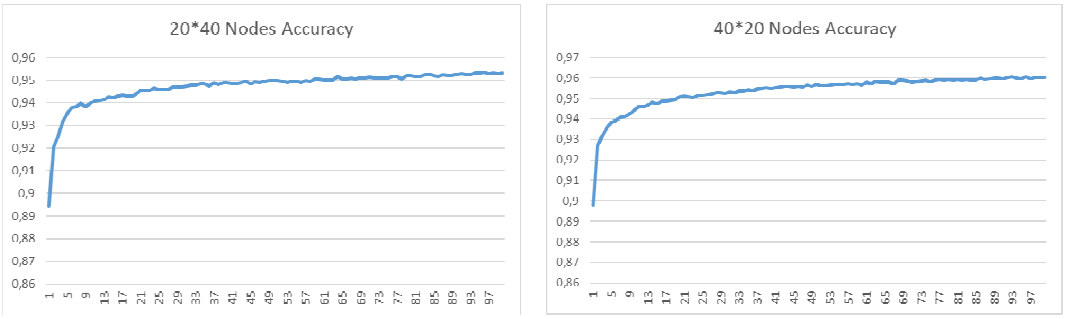


Figure 13. Accuracy Rate of Deep Neural Network with two hidden layers

The execution time of the proposed system is also an important parameter for selection of the phishing detection system. This execution time can be divided into two parts: the feature extraction time and classification time. To measure the feature extraction time, we tried to classify 100 different URLs and measure the all required time needed for calculating the related features in the system, and also total time for all features. The average time in calculated as about 0.6 sec for feature extraction of a URL. We also investigate the Feature based time need and result is depicted in Table 1.

Table 1. Some important features' calculation time

Feature	IP Address	Total Word	Standard Deviation	Number of Brands	Longest Word	Shortest Word	Alexa Rank
Average	**0.1263**	**0.0107**	**0.0036**	**0.0002**	**0.0105**	**0.0105**	**0.4080**
Max	0.7416	0.1371	0.0159	0.0010	0.0338	0.0339	2.1824
Min	0	0	0	0	0	0.00099	0.2813

As can be seen from this table the dominant factor of the feature is the Alexa Ranking part. Due to its need for connecting the third-party services it needs almost 2/3 of all calculation time. Therefore, if it is wanted to decrease the decision time this feature can be disabled. In the table some other time-consuming features are also shown. The other features are calculated less than 10-4 sec, therefore, they not listed.

5. CONCLUSIONS AND FUTURE WORKS

Due to the growing use of Internet in our daily life, cyber attackers aim their victim over this platform. One of the mostly encountered attack is named as "phishing" which creates a spoofed web page to obtain the users sensitive information such as userid and password in financial websites by using social networking facilities. The malicious web page is created as if a legitimate web page, especially copying the original web page one to one. Therefore, detection of these pages is a very trivial problem to overcome due to its semantic structure which takes the advantage of the humans' vulnerabilities.

Software tools can only be used as a support mechanism for detection and prevention this type attacks, and these tools especially use whitelist/blacklist approach to overcome this type of attacks. However, they are static algorithms and cannot identify the new type of attacks in the system. Therefore, as an efficient solution, we propose the use of Artificial Neural Network and Deep Neural Network based system for classifying the incoming URLs. The experimental results show that both these approaches result satisfactory accuracy rate and DNN with 40*20 hidden layer structure produce best solution with about 96% of accuracy.

The latency of the execution time of the algorithm is also an important metric for selection of the detection algorithms. As seen from the results use of Alexa Ranking results a great increase in the execution time, although it has a great importance for detection of phishing. Therefore, according to aim of the system this feature can be disabled for decreasing the execution time.

As the Future works, to decrease the execution time and increase the efficiency of the system, the power of the Graphics Programming Units can be used. Additionally, the other approaches of Deep Learning, such as recurrent neural networks, convolutional neural networks and LSTM can be tested for increasing the performance of the system.

REFERENCES

[1] "USOM," 2018. [Online]. Available: https://www.usom.gov.tr/dosya/1418807122-USOM-SGFF001-Siber%20Guvenlige%20Giris%20ve%20Temel%20Kavramlar.pdf. [Accessed May 2018].

[2] "ITU-," 2008. [Online]. Available: https://www.itu.int/rec/T-REC-X.1205-200804-I. [Accessed May 2018].

[3] Rouse, Margaret, "whatis," November 2016. [Online]. Available: http://whatis.techtarget.com/definition/cybersecurity. [Accessed May 2018].

[4] Diakun, Nadia – Thibault, Purse, Randy & Craigen, Dan, "Defining Cybersecurity," October 2014. [Online]. Available: http://www.timreview.ca/sites/default/files/article_PDF/Craigen_et_al_TIMReview_October2014.pdf. [Accessed May 2018].

[5] "Technopedia," [Online]. Available: https://www.techopedia.com/definition/24747/ cybersecurity. [Accessed May 2018].

[6] Stallings, William, "Introduction," in Network Security Essentials: Applications and Standards, New York, Pearson, 2011, pp. 4-5.

[7] Chia, Terry. "IT Security Community Blog," Stack Exchange, 20 August 2012. [Online]. Available: http://security.blogoverflow.com/2012/08/confidentiality-integrity-availability-thethree-components-of-the-cia-triad/. [Accessed May 2018].

[8] Arslan, Mehmet Emin, "Cyber Security and Cyber Attack Types," Gazi University, Ankara, 2016.

[9] Arachchilage, Nalin Asanka Gamagedara, Psannis, Konstantinos E. & Gupta B. B., "Defending against phishing attacks: taxonomy of methods, current issues and future directions," Springer Science Business Media, New York, 2017.

[10] Podjarny, Guy., "SNYK," SNYK, 10 May 2017. [Online]. Available: https://snyk.io/blog/owasptop-10-breaches/. [Accessed 19 May 2018].

[11] Anti Phishing Working Group, "Phishing Activity Trends Report 1st Half," Anti Phishing Working Group, San Francisco, 2017.

[12] Khonji, Mahmoud, Iraqi, Youssef, Senior Member, IEEE, & Jones, Andrew, "Phishing Detection: A Literature Survey," IEEE COMMUNICATIONS SURVEYS & TUTORIALS, vol. 15, no. 4, pp. 2091-2092, 2013.

[13] Jain, Ankit Kumar & Gupta B. B., "Phishing Detection: Analysis of Visual Similarity," Security and Communication Networks, p. 4, 10 January 2017.

[14] Crowe, Jonathan., "Blog of Barkly," Barkly Protects, July 2017. [Online]. Available: https://blog.barkly.com/phishing-statistics-2017. [Accessed 20 May 2018].

[15] Ivan Galkin, "Crash Introduction to Artificial Neural Networks," Ulcar, [Online]. Available: http://ulcar.uml.edu/~iag/CS/Intro-to-ANN.html. [Accessed 27 May 2018].

[16] Sandhya, L., Thomas Ciza & James, Joby, "Detection of phishing URLs using machine learning techniques," in Control Communication and Computing (ICCC), India, 2013.

[17] Buber, Ebubekir, Diri, Banu & Sahingoz, Ozgur Koray, “NLP based Phishing Attack Detection from URLs”, 17th International Conference on Intelligent Systems Design and Applications (ISDA), Delhi, India,

[18] Khonji, Mahmoud, Iraqi, Youssef., & Jones, Andrew, (2013). Phishing detection: a literature survey. IEEE Communications Surveys & Tutorials, 15(4), 2091-2121.

[19] Nguyen, Luong Anh Tuan, et al. "A novel approach for phishing detection using URL-based heuristic." Computing, Management and Telecommunications (ComManTel), 2014 International Conference on. IEEE, 2014.

[20] Rami M. Mohammad, Fadi Thabtah, Lee McCluskey, (2012) An assessment of features related to phishing websites using an automated technique. In: The 7th international conference for internet technology and secured transactions (ICITST-2012), London

[21] Ozgur Koray Sahingoz, Ebubekir Buber, Onder Demir, Banu Diri, (2018) "Machine learning based phishing detection from URLs", Expert Systems with Application, 2018, https://doi.org/10.1016/j.eswa.2018.09.029.

[22] Ebubekir Buber, Banu Diri and Ozgur Koray Sahingoz, (2017) "Detecting phishing attacks from URL by using NLP techniques," 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, 2017, pp. 337-342.

AUTHORS

Saide Isilay Baykal is a computer engineer. She graduated from Computer Engineering Department of Istanbul Kultur University in 2018. Her research areas are Deep Learning, Machine Learning, Artificial Intelligence and Cyber Security.

Deniz Bulut is a computer engineer. She graduated from Computer Engineering Department of Istanbul Kultur University in 2018. She graduated from Kadriye Moroglu High School. Her research areas are Deep Learning, Machine Learning, Artificial Intelligence and Cyber Security

Ozgur Koray Sahingoz is currently an associate professor in the Department of Computer Engineering at Istanbul Kultur University. He graduated from the Computer Engineering Department of Bogazici University in 1993. He received his M.Sc. and Ph.D. degree from Computer Engineering Department of Istanbul Technical University, in 1998 and 2006, respectively. His research interests lie in the areas of Artificial Intelligence, Deep Learning, Parallel and Distributed Computing, Soft Computing, Information Systems, Wireless Sensor Networks, Intelligent Agents, Multi Agent Systems.

ADAPTABASE - ADAPTIVE MACHINE LEARNING BASED DATABASE CROSS-TECHNOLOGY SELECTION

Shay Horovitz, Alon Ben-Lavi, Refael Auerbach, Bar Brownshtein, Chen Hamdani and Ortal Yona

School of Computer Science, College of Management Academic Studies

ABSTRACT

As modern applications and systems are growing fast and continuously changing, back-end services in general and database services in particular are being challenged with dynamic loads and differential query behaviour. The traditional best practice of designing database – creating fixed relational schemas prior to deployment - becomes irrelevant. While newer database technologies such as document based and columnar are more flexible, they perform better only under certain conditions that are hard to predict. Frequent manual modifications of database structures and technologies under production require expert skills, increase management costs and often ends up with sub-optimal performance. In this paper we propose AdaptaBase - a solution for performance optimization of database technologies in accordance with application query demands by using machine learning to model application query behavioural patterns and learning the optimal database technology per each behavioural pattern. Experiments present a reduction in query execution time of over 25% for the relational-columnar model selection, and over 30% for the relation-document based model selection.

KEYWORDS

Database, Cross-Technology, Machine Learning, Adaptive

1. INTRODUCTION

Throughout the digital age, efficient mechanisms to store and organize data were always vital [1]. In 1970, Edgar Codd described a new method [19] for storing data, suggesting that records would be stored in tables with fixed length records and based the Relational database model. This initiated the development of new Relational model database management systems (RDBMS). RDBMSs were very efficient in storing and processing structured data and as a result became very popular. Along with the development of the internet, accompanied with demand for greater flexibility, a new type of data started to gain volume rapidly - unstructured data. This type of data is both non-relational and schema-less, which the traditional table-based RDBMS can't manage efficiently. Consequently, alternatives - named as No-SQL databases began to emerge.

With the presentation of new types of databases [22,37], came the industry recognition that different database types are applicable for different conditions; Relational databases fit well for applications that involve many complex queries, transactions and data analysis [8], yet - they suffer from lack of ORM orientation, as they were not originally designed to support OOP principles. Moreover, with a dramatic increase in the size of data, query performance degrades accordingly, which may cause query failures and service crashes due to timeout. Yet, the

Natarajan Meghanathan et al. (Eds) : SPPR, SCAI, CSIA, WiMoA, ICCSEA, InWeS, NECO, GridCom - 2018
pp. 55–72, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.81706

alternative of No-SQL databases also fails to serve as a one stop shop for database applications, as they come with major concerns [31] such as absence of complete ACID, limited query language, deficient support, and lack of standards. As such, modern application design accommodates multiple database model types [26,14].

Business requirements change frequently [28], hence - bringing changes in organization's data models and database schema respectively; Thus, database performance reduction is expected along time, since the original database models were designed in mind of different assumptions and data is not stored in its optimal structure any longer. For the time being, manual changes are required to overcome this problem, such as changing tables' schema or optimizing indexes - this must be done by database experts and it's a fragile, expensive [7] and complex task, thus commonly avoided. The operational cost of such database changes can be expressed with the following formula:

$$(\sum_{i}^{N} a_i \times b_i) \times x + C + D$$

Expression 1 – operational cost

Where: N is the number of available DBAs to work on the current problem; x is the problem complexity; a_i is the experience of the DBA and b_i is the estimated work time. C is the estimated extra space to store duplicate data and D is the data transfer factor.

Due to the above, it would be beneficial to have a system and methods capable of learning the application query behavior, and adaptively fit the optimal database type in accordance with query behavior evolutionary changes, while saving on operational costs. AdaptaBase - an adaptive database model optimizer is a solution for meeting the above challenge. In this paper we focus on typical query behavioral patterns that are dominated by read operations such as SELECT and SELECT JOIN queries as this is the most popular setting [5], and we examine the performance potential and feasibility of an adaptive selection of database model between relational, document-based and columnar models. Adaptabase employs machine learning classification and clustering algorithms in order to map between the characteristic query behavioral patterns or query distributions to the optimal database technology or model type. First, queries are being extracted from the MySQL relational database, then clustered into query types, grouped into query distribution patterns, tested per each pattern and database model, and lastly fits given patterns to the optimal database model and technology taking assuming seasonality of query behavioral patterns.

We tested our proposed solution on a NodeCellar [4] application - built with modern technologies such as Backbone.js, Twitter Bootstrap, Node.js, Express, and MongoDB, and adapted another version of it with MySQL for comparisons. Experiments are twofold: first, we evaluate the performance of our solution on dynamic model selection of Relational and Document based models with MySQL and MongoDB accordingly; Next, we test our solution on dynamic selection of Relational and Columnar models; Our columnar model is represented by lean tables in MySQL rather than Cassandra - which is based on BigTable and Dynamo, enclosing additional technologies side by side with the columnar structure and effect on performance. Cost wise, referring Expression (1), in Adaptabse, a_i, b_i and x are 0 since the solution is automatic - saving working hours and training leading to reduced OPEX.

The remainder of this paper examines these issues both analytically and empirically. In Section 2 we discuss related work in this field. Section 3 elaborates on the problem and present different scenarios where query behavior has an influence on performance. Section 4 presents AdaptaBase

design and algorithms, and discusses its implementation internals. Section 5 presents our experiments on a real application and last, Section 6 summarizes this work.

2. RELATED WORK

Integration of relational and NoSQL databases has been studied deeply. In [23] a load balancer is used to monitor the performance of a hybrid db detecting hot spots for data migration. [2] tested the ways of integration of relational and NoSQL databases. [29] presented a solution to query MongoDB by SQL language. [33] converted structural to non-structural db. [20] allows migration relational to Document-oriented database. [20] presented approaches to data integration between relational and NoSQL.

The challenge of converting data between SQL and NoSQL databases has been addressed in [24,34,30]. In [30] an autonomous SQL-to-NoSQL schema migration is proposed. [12] seek most suitable NoSQL structure to migrate from relational Database. [35] presented a SQL-to-HBASE data-schema migration. [27] presented RDBMS-to-NoSQL schema and query migration. Hybrid SQL and NoSQL databases are described in [17,32,39]. Performance comparisons for relational and NoSQL can be found in [36,38].

In contrast to the above, our approach adapts alternative columnar and document-based models to a given relational model and dynamically routes the queries to the model that provides the best performance for current query distribution behavioural patterns.

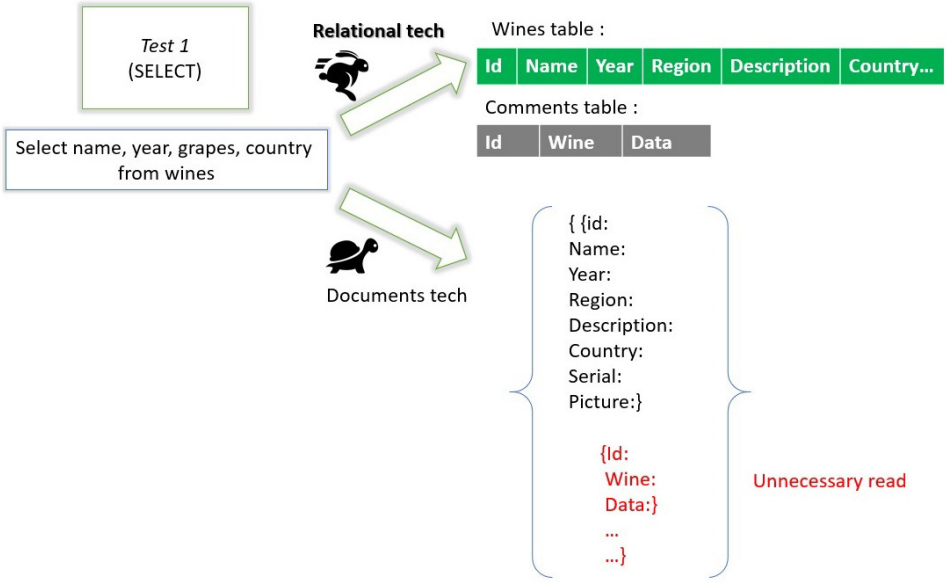
3. PROBLEM

Throughout the process of exploring the benefits and flaws of different database model types, we focused on three types of database models: **Relational** model driven databases are based on storing data in tables - sets of records, each having different attributes. Tables are durable, fast and well suited for transactional operations [25], and the popular SQL language allows a rich and diversified queries, supporting ACID. Yet, relational databases expect fixed predefined schema definition, not tolerant to model changes and are not suitable for dynamic environments with changing query distribution behavioral patterns. In addition, since each row attributes are stored in disk with a continuous form, querying specific attributes is inefficient. **Columnar** databases utilize column oriented model - data is stored and indexed in columns as oppose of rows in the relational model. This allows processing selected columns fast by skipping non relevant attributes that were not requested by the query. While the DBA can partition the relational data in lean tables having small amount of columns - supporting queries that require many columns will end up with subordinate performance due to the need to perform JOIN operations between the lean structured tables. The columnar model is ideal for data analysis applications - suitable for data mining and analytic applications. Columnar databases are not a good fit for transactional workload applications[16]. **Document-based** databases utilize a document model - data is stored in the format of XML or JSON that allows hierarchy and is best suited for schema less, non-structured and non-relational data. While this allows great flexibility, it may be unreliable and index management can be very expensive [31].

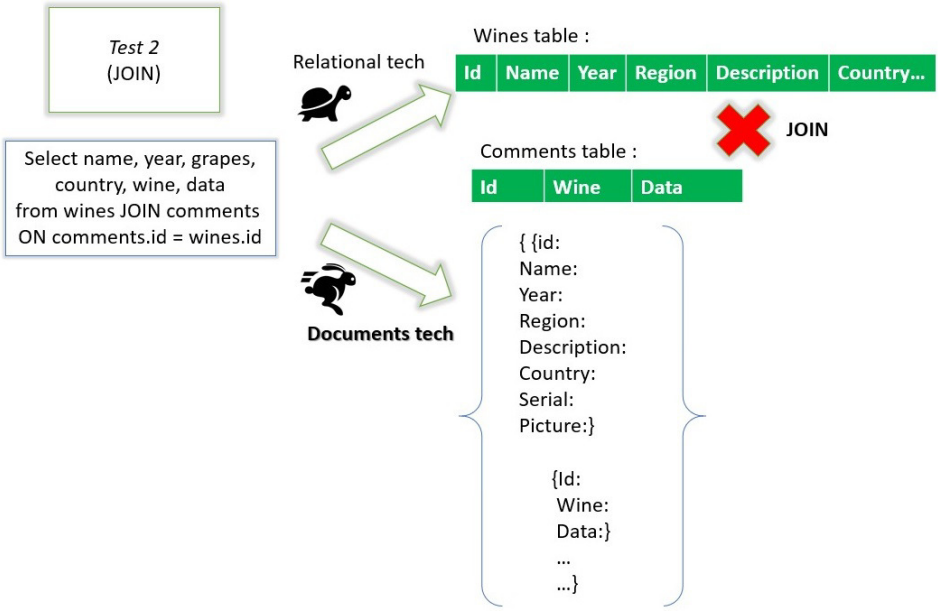
In order to get deeper insights into each database model performance, we executed a set of experiments with different query distributions for Relations vs Document-based and Relational vs Columnar models:

3.1. Relational vs Document-based models comparison

For the relational model we used MySQL and for the Document based model we used MongoDB. The experiments measure execution time of each application instance as a function of distribution of different queries by firing application events using Apache JMeter. All runs are separated by a pause of 10 seconds.



(a) Relational model faster than Document based model scenario



(b) Document-based model faster than Relational model scenario

Figure 1: Relational vs Document-based models performance per query type

In the case of a query asking for data of a single relational table, the relational model in MySQL will end up with faster execution time, whereas the hierarchical representation in MongoDB will be slower due to reading unnecessary data as in Figure 1(a). In contrast, querying data from multiple tables, the relational model requires a JOIN operation ending up with slower execution time compared to the document based model that reads the data that was asked in a single document, as in Figure 1(b).

The experiment depicted in Figure 2 consists of 200 splitted queries, ranging between 0 and 200 SELECT JOIN queries, complemented by INSERT queries. While MySQL performance is heavily dependent on the portion of SELECT JOIN queries, MongoDB in far less affected, presenting 10 times faster execution than MySQL.

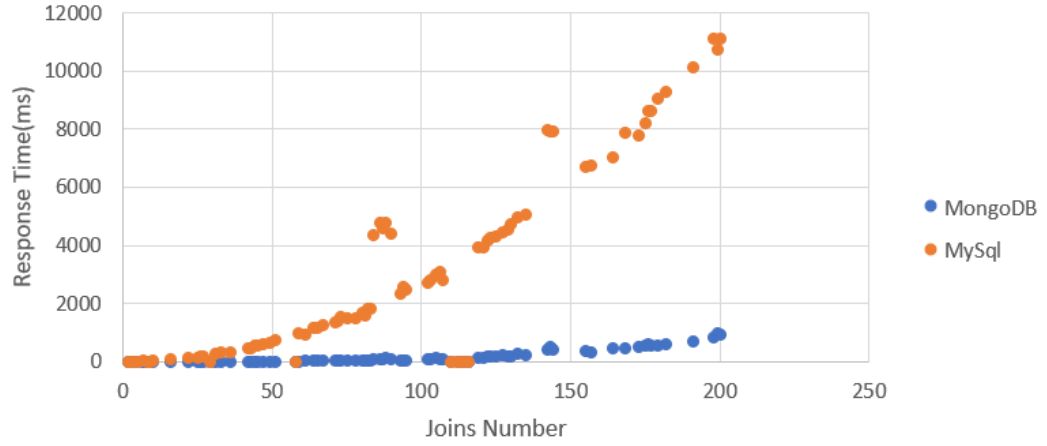


Figure 2: Execution time for distribution of SELECT-JOIN,INSERT

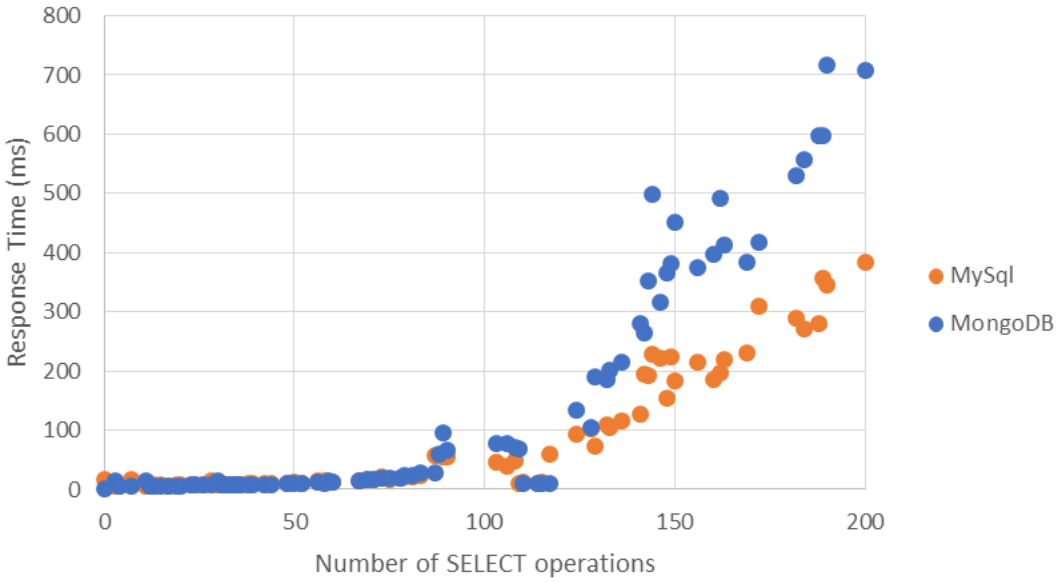


Figure 3: Execution time for distribution of SELECT,INSERT

In the next experiment, depicted in Figure 3, we execute again splitted queries, ranging between 0 and 200 SELECT (no embedded Joins) queries, complemented by INSERTs this time.

While for small ratio of INSERT queries the difference between MongoDB and MySQL is insignificant, in the case of dominant INSERTs, MongoDB performance worsens substantially, with execution time more than double compared to MySQL.

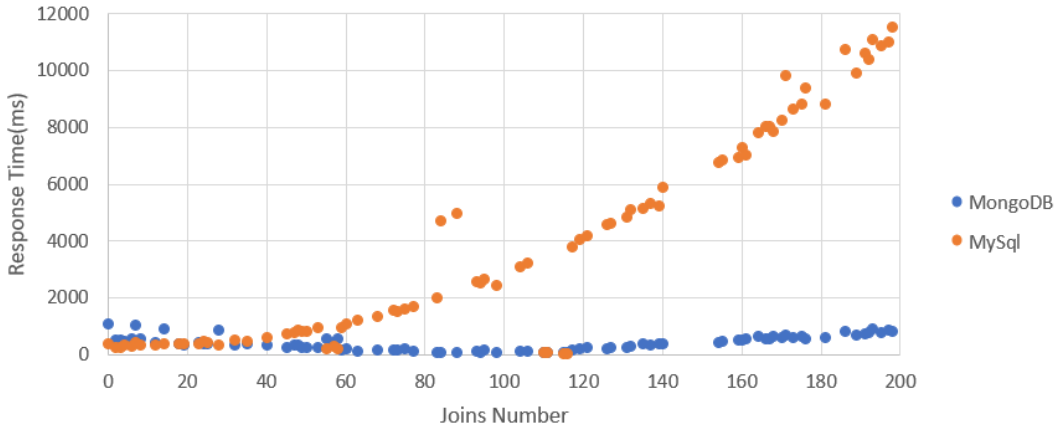


Figure 4: Execution time for distribution of SELECT-JOIN,SELECT

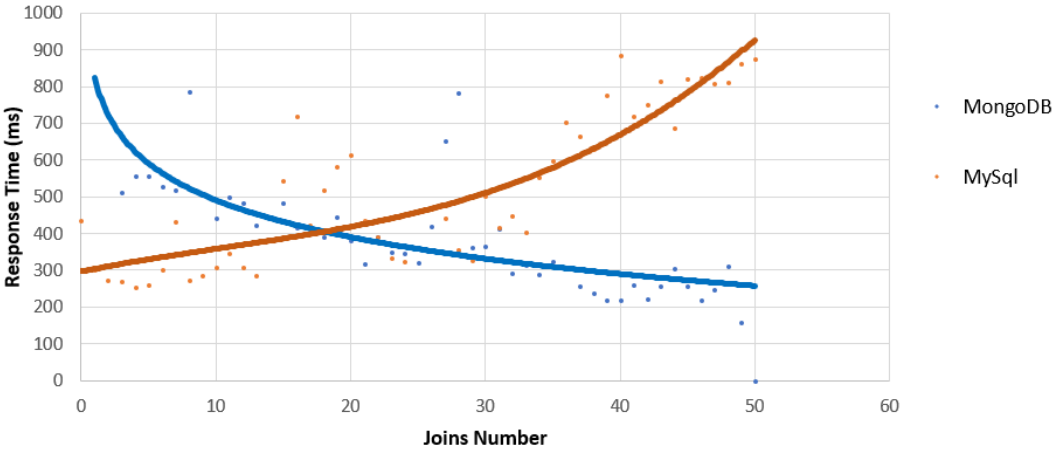
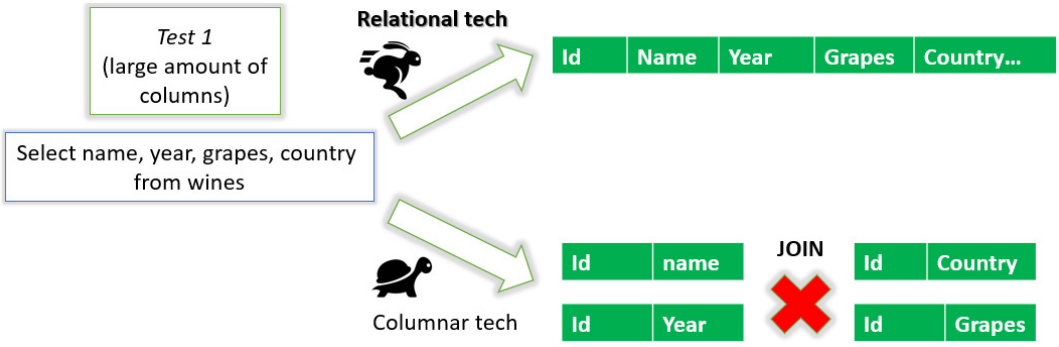


Figure 5: Execution time for distribution of SELECT-JOIN,SELECT (Zoomed on [0-50] Joins)

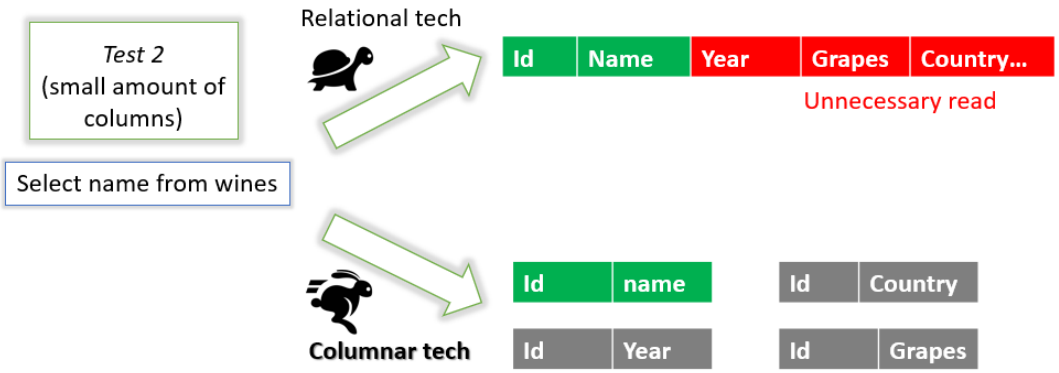
As we focus in read-only queries, we compared simple SELECT with no JOIN and SELECT with JOIN on both databases, as can be seen in Figure 4. While for small proportion of JOINs, MySQL presents better performance than MongoDB, as the proportion increases MongoDB gains extremely better performance, due to JOIN queries being slower than SELECT. A zoom in for infrequent JOINs is in Figure 5, where the cross between the performance of models is visible.

3.2. Relational vs Columnar based models comparison

For the Relational vs Columnar based model comparison, first - we compared SELECT with JOIN and no JOIN queries on both model types in order to estimate the effect of breaking a relational table to lean columnar tables on the performance.



(a) Relational model faster than Columnar model scenario

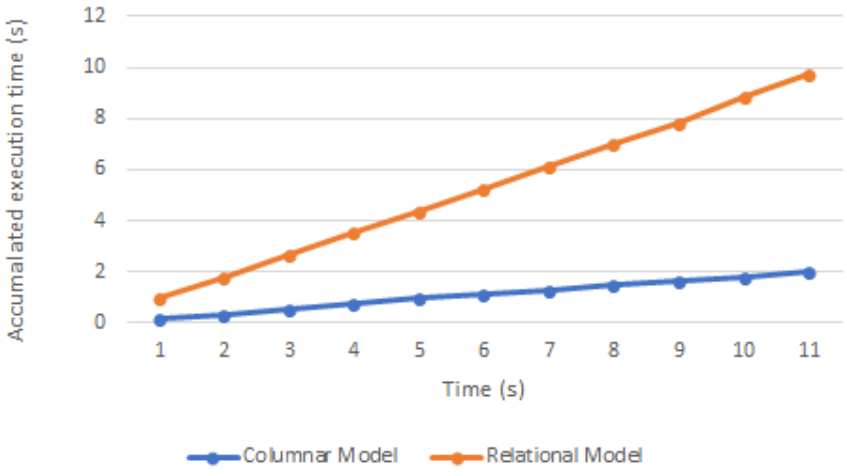


(b) Columnar model faster than Relational model scenario

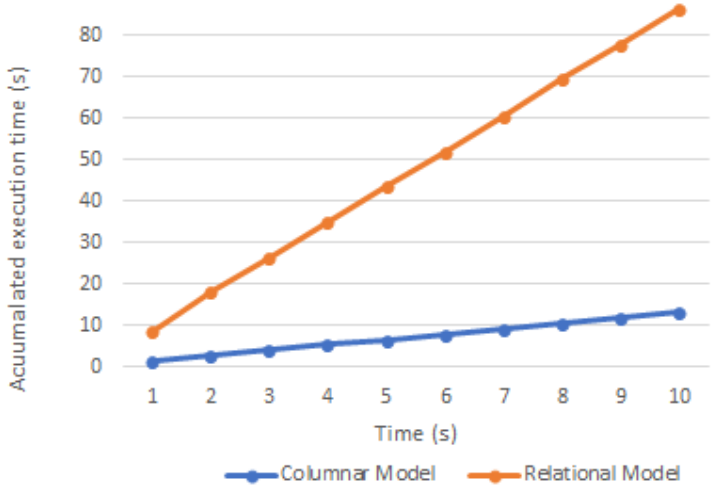
Figure 6: Relational vs Columnar models performance per query type

We used two types of tables - Fat table - that consists all the columns in a single table and Thin tables - breaking the fat table into sub-tables such that JOIN can reconstruct the original table.

As in Figure 6(a), Fat tables are common in relational databases due to representing an entity by a single table. In contrast, in Figure 6(b), thin tables are the best practice of the Columnar approach - where each column is stored separately in the disk. In our experiments, we executed identical queries into the two table types and compared the execution time. Queries that referred to a larger number of columns performed better running times in the fat table than the thin, because no JOIN action were required to join the split columns. In cases where the queries referred to a small number of columns, running times were better in the thin table, because in this case there was no unnecessary reading of information from the disk.



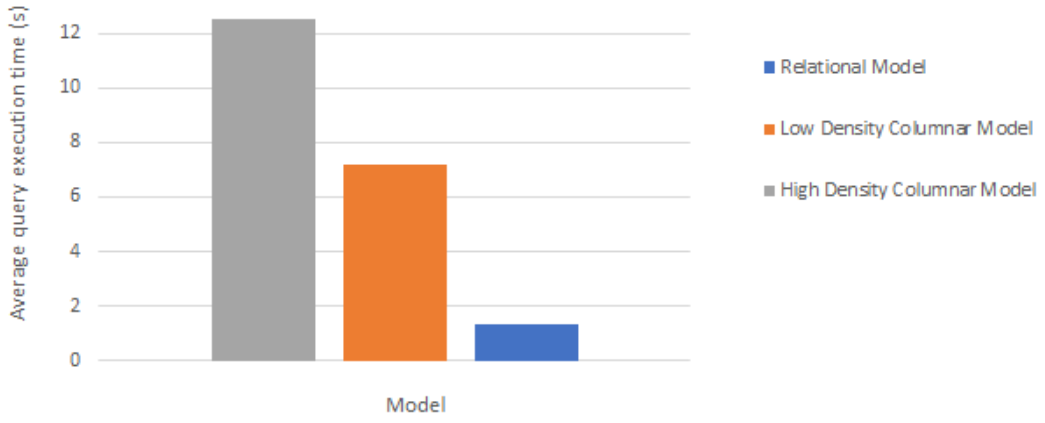
(a) Medium number of rows



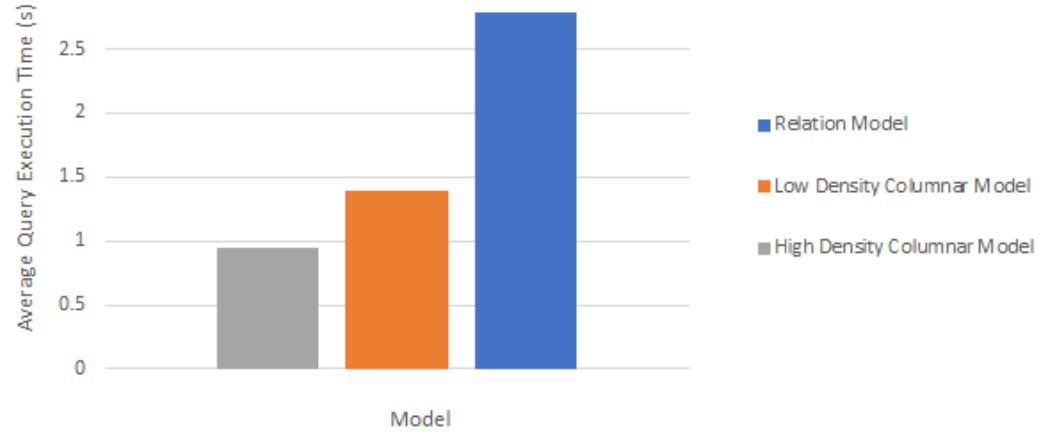
(b) Large number of rows

Figure 7: Relational vs Columnar models accumulated time comparison

In Figure 7 we execute SELECT JOIN queries for several minutes. The columnar model represents a table of 20 columns and the relational model consists of two tables, each of 10 columns - when combined yield the original table. Both tables contain the same amount of rows - 131,072 in case (a) and 1,024,576 in case (b). The columnar model's performance is much better than the relational model as it contains no JOIN. Relational model performance worsens from (a) to (b) up to being 9 times slower compared to the columnar model. The more rows the table consists, the slower query execution times we see.



(a) Large number of columns in query



(b) Small number of columns in query

Figure 8: Average query execution time per model type

In Figure 8 we executed JOIN queries against several tables - having 1,048,576 rows in (a) and 4,194,304 rows in (b). The relational model is represented by a single table; the low density columnar model is represented by 2 tables - that require JOIN in order to return the original table; The high density columnar model consists of 3 tables that require 2 JOIN operations in order to return the original full table. The SELECT queries we run in this case returns the full set of columns as in the original - relational table. We can clearly see in (a) that the more JOINS are apparent in the query, the slower execution time we observe - when querying for large no' of columns using different tables - this is caused due to the JOIN action. Yet, in (b) due to having small no' of columns in the query increases the in efficiency of the relational model.

4. SOLUTION

AdaptaBase provides machine learning based prediction of the optimal database model for given query behavioural patterns - the distribution between query types.

In AdaptaBase, the data & analysis analysis follows the process depicted in Figure 9, and is composed of three phases: first, we learn the query behavioral patterns - the dominant

distributions of SQL queries of the application along time; then we test the performance of each query behavioral pattern with each database model type, and last, with the learned mapping of query distribution to database model, match the current query distribution in time, and switch to the optimal database model.

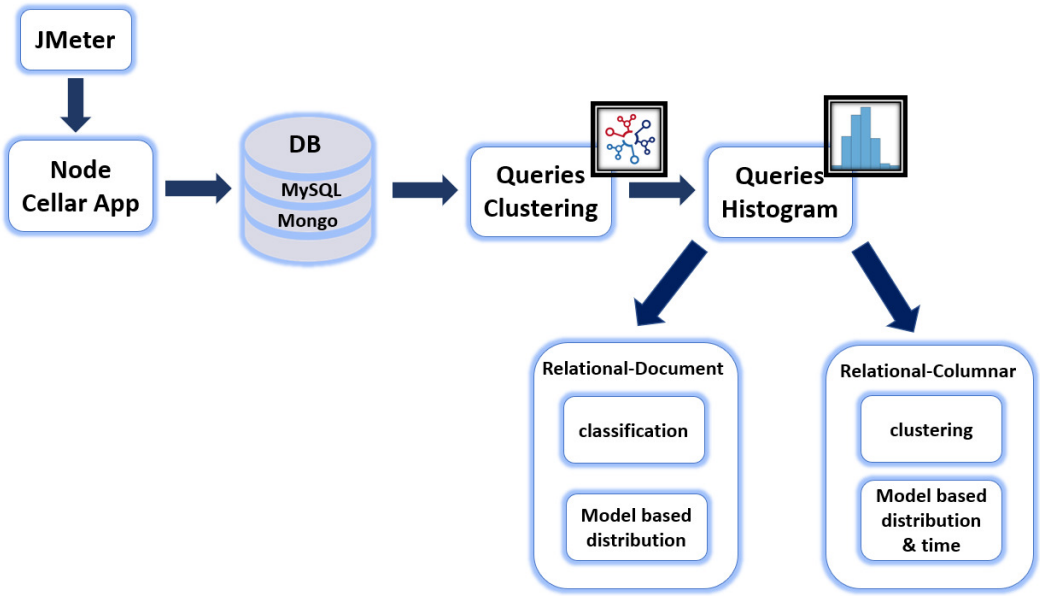


Figure 9: Adaptabase high level data flow

JMeter is used for sending scheduled HTTP requests to Node.JS based NodeCellar application server. The requests follow predefined seasonality patterns. Accordingly, the application server executes different queries against the database; query events are logged and AdaptaBase collects those logs automatically, and stores them for later use by the learning process.

Upon fixed time intervals, the query clustering module is executed, in order to learn about the different query types, and allow us to distinguish between different query distributions or query behavioural patterns in the next phase. SQL query clustering is a well studied issue, and has several solutions, ranging query clustering based on a comparison of query structures, the associated table schemes and statistics such as the sizes of tables that appear in the queries [3], performing query rewrites to standardize query structure [10], using sets of features for query clustering [18], clustering based on attributes for materialized views [13] and clustering based on similarity of the same work plan [21]. Our set of queries in the NodeCellar application was fairly simple and didn't require a heavy query clustering mechanism; as such we performed the query clustering with as the following: First, each query is converted to a vector. Each word in that query gets a certain index in that vector, and the value in that index is the number of occurrences of the word in that query. Afterwards, the vectors are being clustered using DB-SCAN algorithm. After query clustering is done, we compute the different query distributions (behavioural patterns) over time periods, forming a set of histograms of query types counts and write those distributions to a table. We experimented two separate techniques in order to create a model for predicting the optimal database model type for a given query behavioural pattern:

In the **Relational-Columnar** case, we performed clustering on the table of query distribution counts (histograms), by running random K-Means [9] algorithm to identify the bold behaviors given query distributions and time of day. The algorithm selects the number of clusters with

silhouette analysis in order to choose the optimal *k* parameter value with the highest silhouette score. This provides us *k* dominant query distributions. Each distribution is tested against the relational and the columnar models ending up with a mapping of each distribution and its optimal database model.

In the **Relational-Document based** case, first we choose a sample of rows from the table of query distribution counts; then per each sampled distribution, we test the performance of this query distribution on each database technology - MySQL and MongoDB and set the label of the technology that had the minimal execution time as the target for each row in that sample; then we run a classification algorithm (experimented different algorithms) with *k=10* cross validation on the sampled rows of the table - allowing it to map the relation between each distribution and the optimal database model.

While the learning model has been achieved, upon each time window - a distribution of the actual current queries is computed - and then served for inference by the learning mode - yielding a decision of the predicted optimal database model for that current query distribution.

As for the transformation of queries between the different database models and technologies - In the Relational-Document based cases, for queries migrations there are industrial tools [6]. For data migration it would be possible to use [11,15] .In the Relational-Columnar based experiment, queries transformation isn't needed since we are simply dividing the table into multiple lean tables within MySQL in order to gain a columnar structure.

Since in this work we focus on read-only queries, the price of data transfer between the db technology/model types is not taken into consideration. In order to support cost-efficient transfers between the database types, one may either maintain dual copies of the data - which may be adequate in case where the query behaviour is mainly selection/reads and insertions (which price is insignificant for the DB technologies reviewed in our solution) or when required to cover update/delete operations - the synchronization of data between models may become expensive - here, in addition to the current separation of relational schema into multiple tables, one may learn which of the separated (projected) tables may be efficiently managed with document-db only model.

5. EXPERIMENTS

5.1. Environment settings

For the Relational-Document based model we conducted out tests using VM (2X4, CentOS 3.10) for the server and a separate machine (HP 14bf1xx 16GB RAM, DDR4, 512 GB SSD, Intel Family 6, 2000Mhz) that served as a client. The VM contained a layer of docker (v17.03)-running the containerized application server of NodeCellar application and the databases - MongoDB and MySQL. The client machine contains Apache JMeter testing tool to send HTTP requests that emulate user activity on the NodeCellar application. For the Relational-Columnar based model we used a VM (1X8, Ubuntu 4.4.0) on a server with 4 cores, 23 Ghz, 6GB RAM, 128GB HDD. The document based environment setting is depicted in Figure 10.

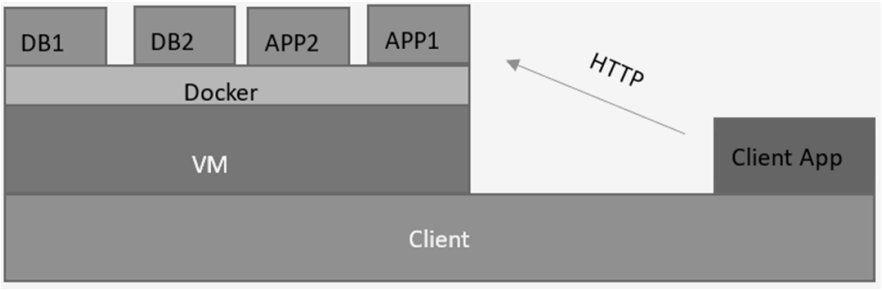


Figure 10: Adaptabase Relational-Document based test environment

5.2. NodeCellar application and queries

For testing we used NodeCellar - a wine collection managing application. For the document-rational experiment three major changes were added to the original application: Trnaslating the original MongoDB based data access layer to MySQL for the columnar case; Adding additional entity of comments. The schemas appear in Figure 11.

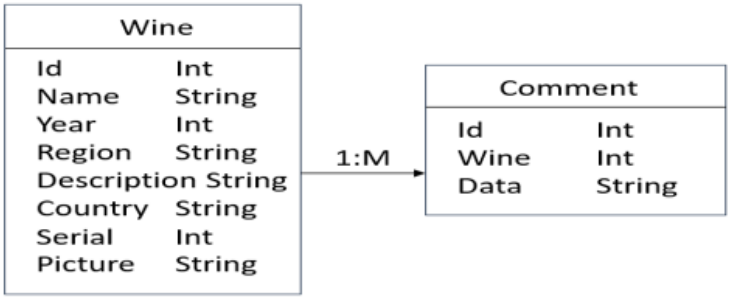


Figure 11: NodeCellar relations

Each entity is created as a single table in MySql. In MongoDB - both were combined into a single collection. In MongoDB this model was implemented using a single collection named Wines. The collection consists a complex document scheme which represents the wine and its comments. The app initializes the database with 1000 wine records and 2983 comment records. In figures 12 the queries used for the tests are described.

App Route	Category	MongoDB query	MySQL query
GET /wines	Select no join	db.collection('wines').find({},{ _id:0, serial:0 ,comments:0 })	SELECT name, year, grapes, region, description, picture, country FROM wine
GET /wines/comments	Select with join	db.collection('wines').find({}, { _id:0, name:1, comments:1 },{ })	SELECT wine.name, comment.data FROM wine INNER JOIN comment ON comment.WineID=wine.Serial
POST /wines	Insert	collection.insert(wine, {safe:true})	INSERT INTO wine set ?

(a) Relational to Document based test queries

App Route	Category	MySQL query
GET /wines	Select no join	Select name0, year0, grapes0, country0, region0, picture0, name1, year1, grapes1, country1, region1, picture1, name2, year2, grapes2, country2, region2, picture2, name3, year3, grapes3, country3, region3, picture3,count(*) from wine_test_1 where year0 = 2009 and country0 = 'Italy' and id < 300000 Group by name0, year0, grapes0,country0, region0, picture0, name1, year1, grapes1, country1, region1, picture1, name2, year2, grapes2, country2, region2, picture2, name3, year3, grapes3, country3, region3, picture3
GET /wines	Select with join	Select t1.name0, t1.year0, t2.grapes0, t2.country0, t2.region0, t2.picture0, t2.name1, t2.year1, t2.grapes1, t2.country1, t2.region1, t2.picture1, t2.name2, t2.year2, t2.grapes2, t2.country2, t2.region2, t2.picture2, t2.name3, t2.year3, grapes3, t2.country3, t2.region3, t2.picture3 from wine_test_111 as t1 inner join wine_test_112 as t2 on t1.id = t2.id where t1.year0 = 2009 and t2.country0 = 'Italy' and t1.id < 300000 Group by t1.name0, t1.year0, t2.grapes0, t2.country0, t2.region0, t2.picture0, t2.name1, t2.year1, t2.grapes1, t2.country1, t2.region1, t2.picture1, t2.name2, t2.year2, t2.grapes2, t2.country2, t2.region2, t2.picture2, t2.name3, t2.year3, grapes3, t2.country3, t2.region3, t2.picture3

(b) Relational to Columnar test queries

Figure 12: NodeCellar application queries used in tests

5.3. Relational-Document based Experiments

In Figure 13, the accuracy of 5 different machine learning algorithms is depicted for the Relational-Document based case. All the algorithms performed well (accuracy of 0.8-0.95), and for each distribution predict which database would be best suited.

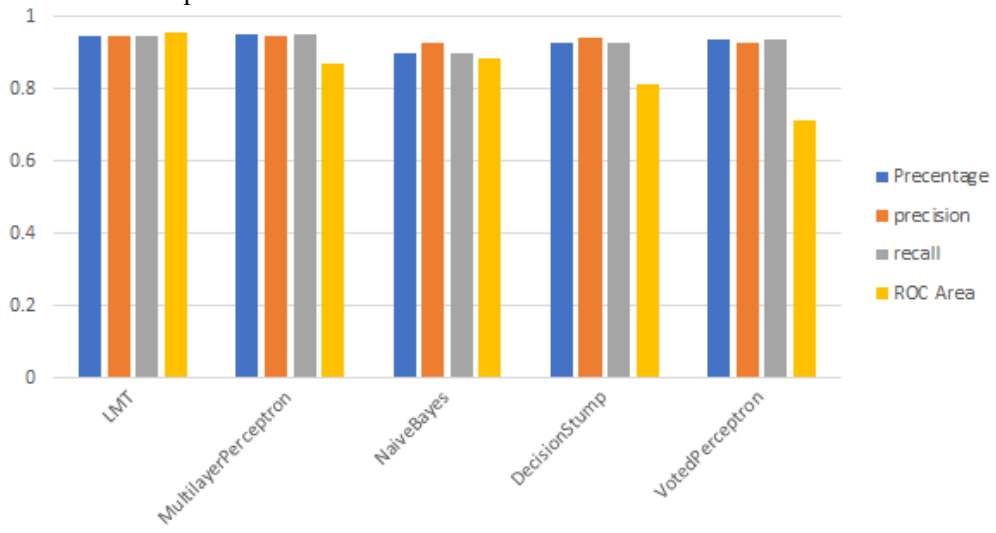
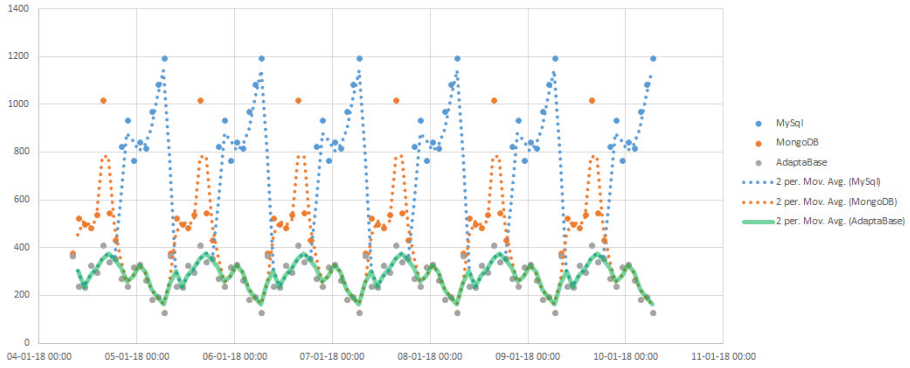


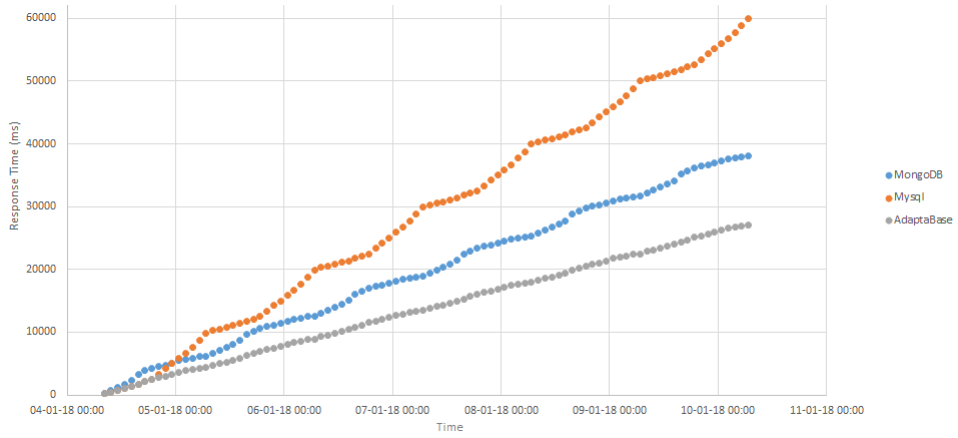
Figure 13: Machine Learning algorithms for Relational-Document prediction

In Figure 14 in (a) the query performance (execution time) was measured during a week of user work. The execution time was measured for each instance and in addition for each distribution the model predicted which database to use. The machine learning algorithm nicely adapts to the optimal database model that provides the minimal query execution times. (b) is the cumulative version - the algorithm performance gains an improvement factor of 1.2-2 compared to alternative predefined database model. Notice that the aggregated execution time of AdaptaBase is significantly shorter than the best aggregated one – which in this case is MongoDB. While

MongoDB performs better than MySQL in this specific scenario over time in general, occasionally MySQL performs better than MongoDB and for those occasions as well, AdaptaBase selects the optimal database.



(a) Regular

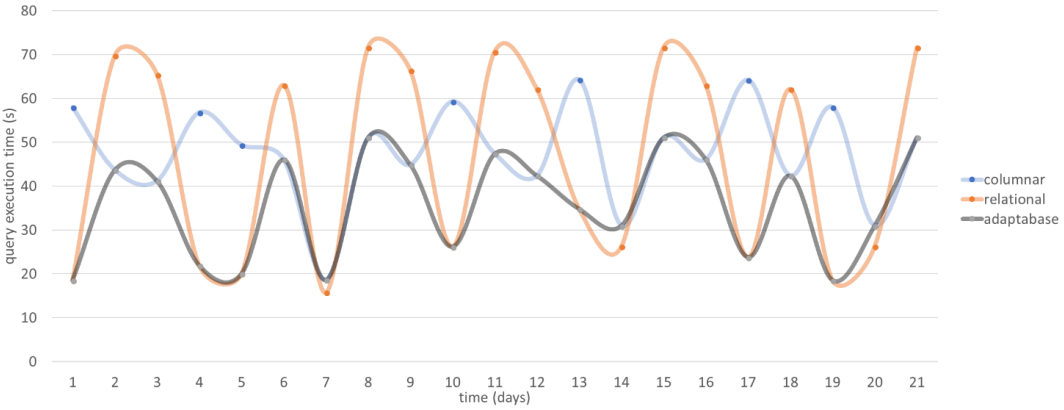


(b) Cumulative

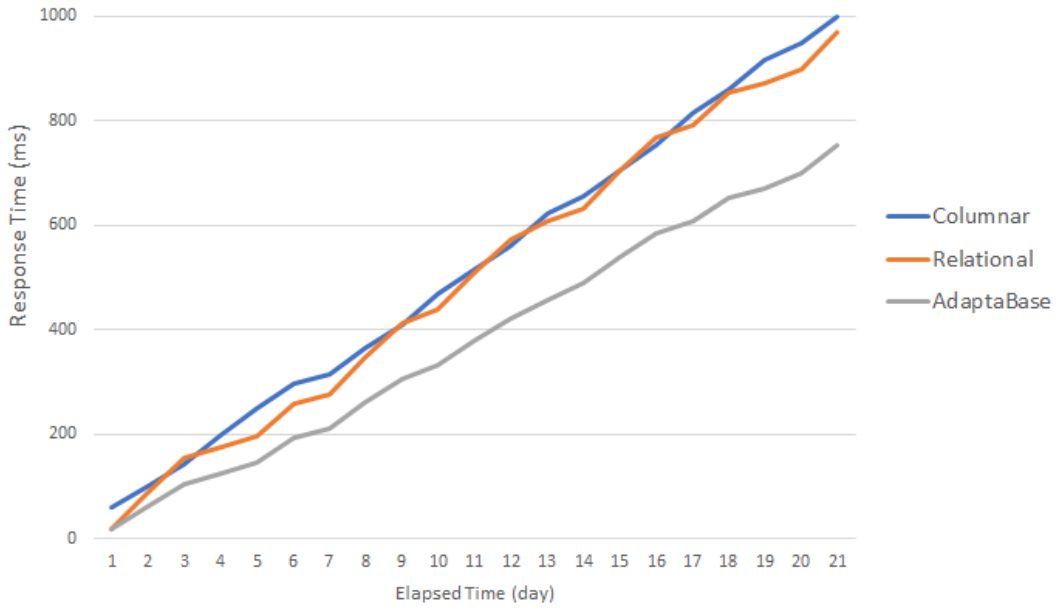
Figure 14: AdaptaBase performance vs alternatives (fixed Relational or Document-based

5.4. Relational-Columnar based Experiments

In Figure 15(a), we compare the performance of the technologies we examined - Columnar and Relational over time - to our machine learning based algorithm. In these experiments, all of the three technologies are tested in parallel. Figure 15(b) presents this experiment in accumulated execution time. In total, our solution achieves improved performance of over 25% in the period of 21 days.



(a) Regular



(b) Cumulative

Figure 15: AdaptaBase performance vs alternatives (fixed Relational or Columnar)

6. SUMMARY

In this paper we have presented AdaptaBase - a solution that can reduce query execution times and eventually save on OPEX. Our solution is based on machine learning based prediction of the optimal db model for a given query behavioural patterns.

Our experiments based on actual query execution on real DB systems- i.e. MySql and MongoDB - presented a reduction in query execution time of 25% for the relational-columnar model selection, and up to 30% for the relation-document based model selection.

Next, we intend to evaluate modifying commands such as INSERT, UPDATE, DELETE and extend our experiments to other database types such as graph and key-value databases.

REFERENCES

[1] A Brief History of Database Management. http://www.dataversity.net/brief-historydatabase-management.

[2] Bridging Relational and NoSQL Worlds: Case Study. https://www.igiglobal.com/chapter/bridging-relational-and-nosql-worlds/191986.

[3] Efficient Query Recommendation. http://www.cs.technion.ac.il/users/wwwb/cgi-bin/trget.cgi/- 2015/MSC/MSC-2015-14.pdf.

[4] Node Cellar. http://nodecellar.coenraets.org.

[5] On workload characterization of relational database environments. http://ieeexplore.ieee.org/- abstract/document/129222/.

[6] Query Translator. http://www.querymongo.com.

[7] Relational Databases Are Not Designed To Handle Change https://www.marklogic.com/-blog/relational-databases-change.

[8] Relational vs. non-relational databases: Which one is right for you? https://www.pluralsight -.com/blog/ software-development/relational-non-relationaldatabases.

[9] Selecting the number of clusters with silhouette analysis on KMeans clustering. http ://scikit – learn.org/stable/autoexamples/cluster/plotkmeanssilhouetteanalysis.html.

[10] Similarity Metrics for SQL Query Clustering . https://odin.cse.buffalo.edu/papers/2018- /TKDEQuerySimilarity.pdf.

[11] Warehouse. https://github.com/dundalek/warehouse.

[12] Al Shekh Yassin, F.J.: Migrating from sql to nosql database: Practices and analysis (2017).

[13] Aouiche, K., Jouve, P.E., Darmont, J.: Clustering-based materialized view selection in data warehouses. In: East European Conference on Advances in Databases and Information Systems. pp. 81–95. Springer (2006).

[14] Arnold, J., Glavic, B., Raicu, I.: Hrdbms: Combining the best of modern and traditional relational databases. Illinois Institute of Technology, Department of Computer Science, PhD Oral Qualifier (2015).

[15] Arora, R., Aggarwal, R.R.: An algorithm for transformation of data from mysql to nosql (mongodb). International Journal of Advanced Studies in Computer Science and Engineering 2(1) (2013).

[16] Bhatia, A., Patil, S.: Column oriented dbms an approach. International Journal of Computer -Science & Communication Networks 1(2), 111–116 (2011).

[17] Bjeladinovic, S.: A fresh approach for hybrid sql/nosql database design based on data structuredness. Enterprise Information Systems pp. 1–19 (2018).

[18] Chu, W.W., Zhang, G.: Associative query answering via query feature similarity. In: Intelligent Information Systems, 1997. IIS'97. Proceedings. pp. 405–409. IEEE (1997).

[19] Codd, E.F.: A relational model of data for large shared data banks. Communications of the ACM 13(6), 377–387 (1970).

[20] El Alami, A., Bahaj, M.: Migration of a relational databases to nosql: The way forward. In: Multimedia Computing and Systems (ICMCS), 2016 5th International Conference on.pp. 18–23. IEEE (2016).

[21] Ghosh, A., Parikh, J., Sengar, V.S., Haritsa, J.R.: Plan selection based on query clustering. In: VLDB'02: Proceedings of the 28th International Conference on Very Large Databases.pp. 179–190. Elsevier (2002).

[22] Han, J., Haihong, E., Le, G., Du, J.: Survey on nosql database. In: Pervasive computing and applications (ICPCA), 2011 6th international conference on. pp. 363–366. IEEE (2011).

[23] Huang, H.S., Hung, S.H., Yeh, C.W.: Load balancing for hybrid nosql database management systems. In: Proceedings of the 2015 Conference on research in adaptive and convergent systems. pp. 80–85. ACM (2015).

[24] ISLAM, M.S.: Techniques for converting big data from sql to nosql databases.

[25] Kemper, A., Neumann, T.: Hyper: A hybrid oltp&olap main memory database system based on virtual memory snapshots. In: Data Engineering (ICDE), 2011 IEEE 27th International Conference on. pp. 195–206. IEEE (2011).

[26] Ko, C.Y.: Three approaches to a multidatabase system. In: Proceedings of the Philippine Computer Science Congress (PCSC), www. citeseer. ist. psu. edu/ko00three. html (2000).

[27] Kuderu, N., Kumari, V.: Relational database to nosql conversion by schema migration and mapping. International Journal 3(9), 506–513 (2016).

[28] Law, J., Rothermel, G.: Whole program path-based dynamic impact analysis. In: Proceedings of the 25th International Conference on Software Engineering. pp. 308–318. IEEE Computer Society (2003).

[29] Lawrence, R.: Integration and virtualization of relational sql and nosql systems including mysql and mongodb. In: Computational Science and Computational Intelligence (CSCI), 2014 International Conference on. vol. 1, pp. 285–290. IEEE (2014).

[30] Lee, C.H., Zheng, Y.L.: Sql-to-nosql schema denormalization and migration: a study on content management systems. In: Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on. pp. 2022–2026. IEEE (2015).

[31] Nayak, A., Poriya, A., Poojary, D.: Type of nosql databases and its comparison with relational databases. International Journal of Applied Information Systems 5(4), 16–19 (2013).

[32] Okeke, K.K., Ejiofor, V.E.: Implementation of cross-platform language between sql and nosql database systems. In: OcRI. pp. 239–240 (2016).

[33] Potey, M., Digrase, M., Deshmukh, G., Nerkar, M.: Database migration from structured database to non-structured database. In: International Conference on Recent Trends & Advancements in Engineering Technology (ICRTAET 2015). pp. 1–3. Citeseer (2015).

[34] Schreiner, G.A., Duarte, D., dos Santos Mello, R.: Sqltokeynosql: a layer for relational to key-based nosql database mapping. In: Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services. p. 74. ACM (2015).

[35] Serrano, D., Stroulia, E.: From relations to multi-dimensional maps: a sql-to-hbase transformation methodology. In: Proceedings of the 26th Annual International Conference on Computer Science and Software Engineering. pp. 156–165. IBM Corp. (2016).

[36] Srividyaa, S., Varalakshmi, R.: A study on output performance of nosql data processing over rdbms in big data.

[37] Strauch, C., Sites, U.L.S., Kriha, W.: Nosql databases. Lecture Notes, Stuttgart Media University 20 (2011).

[38] Wu, C.M., Huang, Y.F., Lee, J.: Comparisons between mongodb and ms-sql databases on the twc website. American Journal of Software Engineering and Applications 4(2), 35–41 (2015).

[39] Wu, H., Ambavane, A., Mukherjee, S., Mao, S.: A coherent healthcare system with rdbms,nosql and gis databases (2017).

AUTHORS

Dr. Shay Horovitz, PhD is head of Data Science specialization at the Computer Science School at the College of Management and a senior researcher & expert at Huawei. His research area is Machine Learning algorithms for the cloud, large scale networks and big data.

Alon Ben-Lavi, graduated B.Sc. at the College of Management - Academic Studies. His research focus on the effects of database models on applications performance.

Refael Auerbach, web solutionist and Big Data expert. A software engineer with experience over a decade in web development, distributed computing and machine learning. B.Sc in Computer Science.

Bar Brownshtein, owns Bs.c in computer science at Israel college of management - academic studies. Specializes in data science. Works at Indusify as software developer.

Chen Hamdani, holds a bachelor's degree in computer science with a specialization in data science at the College of Management. Works as a developer in the Prime Minister's Office.

Ortal Yona, Graduated Bsc in computer science with specialization in data science from the college of management academic studies in Israel.

RESIDENTIAL LOAD PROFILE ANALYSIS USING CLUSTERING STABILITY

Fang-Yi Chang[1], Shu-Wei Lin[1], Chia-Wei Tsai[2] and Po-Chun Kuo[3]

[1]Digital Transformation Institute,
Institute for Information Industry, Taipei, Taiwan
[2]Department of Computer Science and Information Engineering
Southern Taiwan University of Science and Technology, Tainan, Taiwan
[3]Department of Information Management
Southern Taiwan University of Science and Technology, Tainan, Taiwan

ABSTRACT

Clustering is an useful tool in the data analysis to discover the natural structure in the data. The technique separates given smart meter data set into several representative clusters for the convenience of energy management. Each cluster may has its own attributes, such as energy usage time and magnitude. These attributes can help the electrical operators to manage their electrical grids with goals of energy and cost reduction. In this paper, we use principle component analysis and K-means as dimensional reduction and the reference clustering algorithm, respectively, and several choices must be considered: the number of cluster, the number of the leading principle components, and whether use normalized principle analysis schema or not. To answer these issues simultaneously, we use the stability scores as measured by dot similarity and confusion matrix as our evaluation decision. The advantage is that it is useful for comparing the performance under different decisions, and thus provides us to make these choices simultaneously.

KEYWORDS

Smart meter; Unsupervised; Nonparameter; Clustering; PCA; Stability; Smart Grid; Value-Add Electricity Services; Energy Saving; Energy management

1. INTRODCUTION

The research of smart meter data has been stimulated by the need for electrical grid operators for energy management as the era of smart grid coming. Smart meter can send the fine grained energy consumption data back near real-time to the electrical operators or the electrical retailers. The amount of smart meter is massive and is accumulated at very faster speed, thus how to utilize and manage the smart meter efficiently has become an important topic worldwide. In Taiwan, Taiwan Power Company, the largest electrical utility in Taiwan, has been aware of the need of clustering of smart meter to better understand the energy usage patterns of low-voltage customers. The company have found that the patterns are so complex, diverse and dynamic that artificial-based methods are inefficient to deal with them. Taiwan now has been undergoing the green

Natarajan Meghanathan et al. (Eds) : SPPR, SCAI, CSIA, WiMoA, ICCSEA, InWeS, NECO, GridCom - 2018
pp. 73–81, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.81707

energy transition since few years ago, and hence Taiwan Power Company or the related bureaus need to know the end user's usage behavior, especially during peak hours, to complete this transition aimed with energy reduction.

Clustering is an useful tool to separate the smart meter data set into several representative groups to reflect the attributes of energy usage time and magnitude for the convenience of energy management. Here we briefly introduce the recent research works of clustering of smart meter. These works can be mainly divided into two approaches. One is to identify the relationship between energy usage patterns and socio-demographics [7][8]. Another is to identify suitable and representative groups using smart meter data only [5][6][9]. Both two approaches are to find energy reduction solutions and hence optimize the electrical network and reduce the energy. Since the amount of smart meter is massive, it is necessary to reduce the dimensionality of the raw data to provide a more robust and efficient clustering. The methods of dimensional reduction among the recent works are principle component analysis (PCA)[9] and artificial-based variable selection [5][7][8]. PCA is an classical technique of dimensional reduction and has achieved a significant success in many field, including bio-statistics, signal process and image process. It reduces dimensionality or selects variables by using the leading high variance principle components to perform data analysis and filtering the rest components which act as noisy signal. Artificial-based variable selection in these research works is to artificially find the representative attributes. For example, Stephen et al. [5] chosen the four time period and other seven attributes by specific mathematical formulas. On the other hand, there have been a variety of clustering techniques, including K-means, finite mixture model (FMM) and Hidden Markov Model (HMM), and evaluation decisions, which have been applied in smart meter in accordance with the experimental data and main purpose. For example, Stephen et al.[5] used FMM and Bootstrap to estimate the validation scores, which is the relative entropy. Adrian Albert and Ram Rajagopal [8] used HMM and BIC score to perform spectral clustering. Charalampos et al.[9] used K-means, Hierarchical Clustering and Hausdorff-based K-medoids, and evaluated these performances by Dunn Index, Calinski Harabasz Index and Energy Variance Index.

In this paper, we use PCA and K-means as our dimensional reduction technique and reference clustering algorithm, respectively. For the optimal decision, we use 'stability' to select to number of clustering and the corresponding clustering and claim that the it is advantageous in reality. The advantages are 'stability' that is naturally led by the evaluation decision, which is convenience for energy management in reality. It provides the 'trade-off' between the number of clusters and the corresponding stability scores, avoiding leading to providing the simplest clustering result only. For example, if the optimal number of cluster is 2, the corresponding clustering is too simple for electrical operators or retailers.

2. DIMENSIONAL REEDUCATION

A. Principle Component Analysis

Principle component analysis is a classical technique of dimensional reduction by orthogonal transformation into a new set of coordinates which are linearly uncorrelated [1]. These new coordinates are called principle components and are arranged such that the kth principle component has the kth largest variance among all principle components. The larger variance implies that the corresponding variable has more information and have more relevant to clustering. In the paper, we denote the PCA relative to the covariance matrix and the correlation

matrix as center PCA and normalized PCA, respectively, and compare the performances between two schemes.

3. CLUSTERING

A. Clustering Algorithm

K-means is attempt to minimize the objective function:

$$Q_K^{(n)}(c_1, ..., C_K) = \frac{1}{n} \sum_{i=1}^{n} \min_{k=1,...,K} ||X_i - c_k||^2 \quad (1)$$

where K is the fixed number of clusters, X_1, … $X_n \in R$ are the data points and c_1, …. c_k denote the centers of the K clusters.

Several research papers have tried to model energy usage pattern in parameter probabilistic model, including Gaussian, beta, gamma and log-normal distribution [10][11][12]. However, since the distribution of our experimental data are complex and diverse, and no other information is available about our data, we do not assume any parameter probabilistic model to our experimental data in the paper, which motivates us use clustering stability as the evaluation decision. In fact we do not know whether the given data set can be represented as any mathematical models, and the K-means algorithm is used anyway. However, what we concern is that the whole data set can be represented using K-means to split each true cluster in several smaller and representative groups, rather than select the 'correct' number of clusters. It is acceptable even with the fact that the true clusters are split in smaller groups, or to afterwards join these groups to form a bigger group.

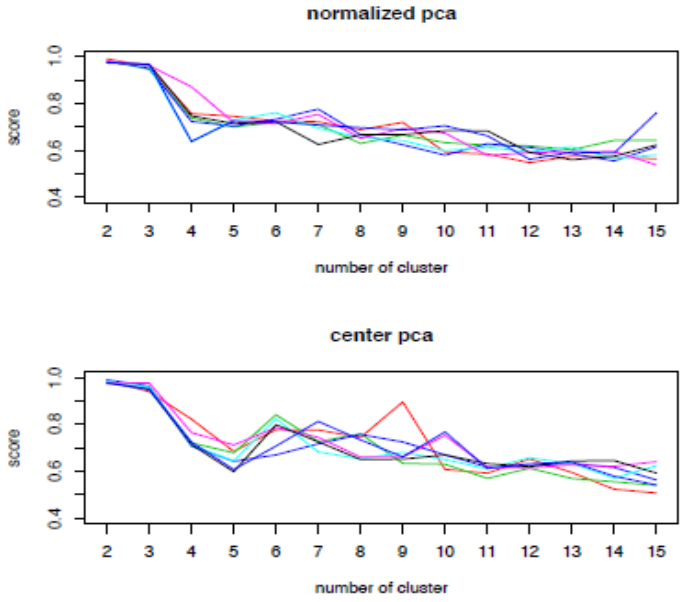


Figure 1. The minimum of dot similarities are estimated for a varying number of principal components p for each number of cluster with center and normalized principle components

B. Clustering Stability

In this work, several issues need to be considered:

- Whether the PCA should be normalized?
- How many leading principle components should be selected?
- How many clusters? or if the optimal decision is two clusters, is there any secondary decision at certain acceptable level?

Since there are more than one decisions that should be simultaneously determined and no mathematical model can be used to describe the complex relationship, using clustering stability may help us to make these decisions. Here is the algorithm to calculate stability score for a fixed number of clusters k [3]:

1) Perform K-means on the original data set (the referencing clustering).
2) Randomly sample a fraction f, larger than 0.5, of the original dataset for t times
3) Perform K-means on the each sub-sample subset (the resampling clusterings)
4) Calculate similarities between the referencing and the resampling clustering

The essential concept in the theory of clustering stability is resampling. It is reasonable to assume that if the decisions respect to the above questions are suitable, the results under the same framework are similar. More specifically, when the structure in the given data is represented well by k clusters, the reference clustering will be similar to the result obtained from the sub-sample data. The similarity measures considered in the papers are dot similarity and confusion matrix.

Table 1: the values for dot similarity and confusion matrix

dot similarity								
		2	3	4	5	6	7	8
2 pc	n	0.99	0.96	0.76	0.75	0.73	0.72	0.69
	c	0.99	0.94	0.82	0.68	0.78	0.78	0.74
3 pc	n	0.98	0.94	0.74	0.70	0.72	0.71	0.63
	c	0.98	0.95	0.72	0.68	0.84	0.73	0.76
4 pc	n	0.98	0.95	0.72	0.70	0.73	0.78	0.67
	c	0.99	0.96	0.71	0.64	0.67	0.72	0.76
5 pc	n	0.98	0.96	0.64	0.73	0.76	0.69	0.66
	c	0.98	0.96	0.71	0.64	0.82	0.68	0.65
6 pc	n	0.98	0.96	0.87	0.72	0.71	0.75	0.65
	c	0.97	0.98	0.76	0.71	0.79	0.74	0.66
9 pc	n	0.97	0.97	0.75	0.71	0.72	0.63	0.67
	c	0.98	0.95	0.71	0.60	0.80	0.73	0.65
12 pc	n	0.97	0.97	0.64	0.73	0.72	0.71	0.70
	c	0.98	0.95	0.72	0.61	0.71	0.81	0.73

confusion matrix								
		2	3	4	5	6	7	8
2 pc	n	0.99	0.96	0.78	0.72	0.73	0.72	0.68
	c	0.99	0.95	0.83	0.65	0.82	0.76	0.72
3 pc	n	0.98	0.95	0.69	0.68	0.71	0.68	0.63
	c	0.98	0.96	0.60	0.65	0.84	0.71	0.78
4 pc	n	0.98	0.96	0.73	0.67	0.71	0.75	0.67
	c	0.99	0.97	0.62	0.70	0.59	0.70	0.67
5 pc	n	0.98	0.97	0.58	0.70	0.73	0.69	0.64
	c	0.98	0.97	0.64	0.67	0.86	0.70	0.68
6 pc	n	0.98	0.97	0.87	0.70	0.71	0.73	0.67
	c	0.97	0.98	0.71	0.67	0.81	0.70	0.64
9 pc	n	0.97	0.97	0.71	0.69	0.72	0.63	0.67
	c	0.98	0.96	0.65	0.58	0.79	0.68	0.67
12 pc	n	0.97	0.97	0.60	0.70	0.72	0.66	0.65
	c	0.98	0.96	0.67	0.64	0.72	0.84	0.73

4. EVALUATION

A. III Dataset

The dataset was collected by Institute For Information Industry which consists of 109 different households in Northern Taiwan between Aug 2017 and September 2018. We treat each record, one day, as an individual points in the process of clustering. Thus, the records in the same customers over periods of time could be belong to different clusters. Although our sample rate is 1 minute, we perform our experiment at sample rate 15 minute to simulate the real condition in Taiwan. The sample rate of smart meter in Taiwan is 15 minute.

B. Experimental result

The stability of the clustering as measured by the minimum among similarity measures for these two measures are plotted for a varying number of principal components in Figure 1-2. These two have similar tendency. The values are briefly presented in table 1. Partitions into 2 and 3 clusters were stable regardless of the number of principal components and normalization, as evidenced by similarity scores being close to 1. Partitions into 4, 6 clusters were most stable for 2, 3 PCs, respectively, with similarity scores above 0.8 with center pca.

Fig 3-6 respectively show the average of each group obtained from the partitions into 2, 3, 4, 6 by K-means with 3 leading principle components. Using center principle components provides more clear separation based on energy usage time and magnitude. On the other hand, although the six clustering is not the optimal decision, it reveals more structures inside the whole data. Intuitively, the six-clustering may be interpreted as the result by splitting the groups obtained from the partitions into 2, 3 or 4, but it need more research work to verify the point.

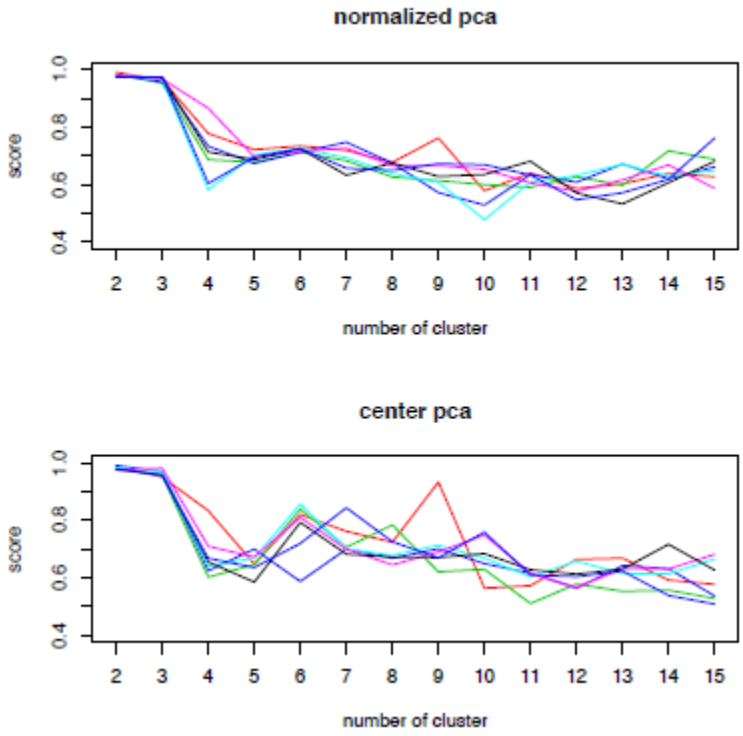


Figure 2. The minimum of confusion matrices are estimated for a varying number of principal components p for each number of cluster with center and normalized principle components

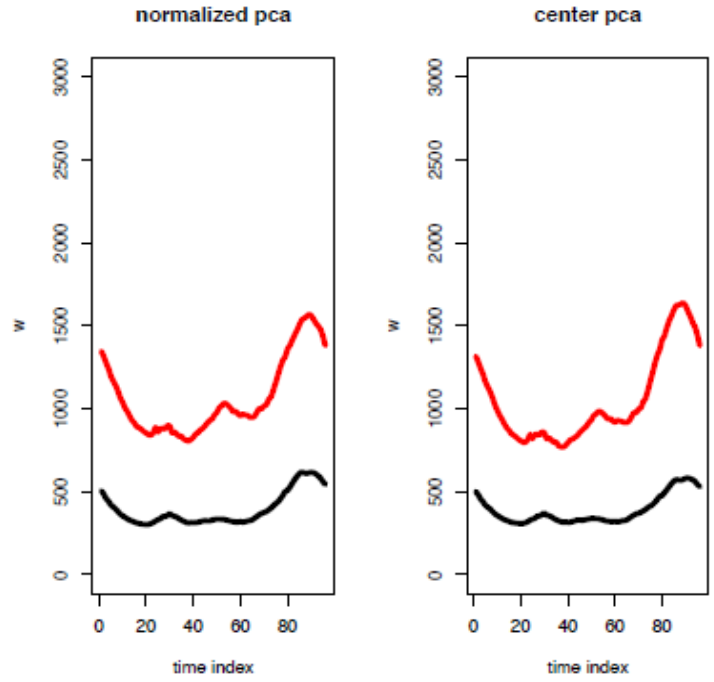


Figure 3. The average energy patterns as the number of cluster is two

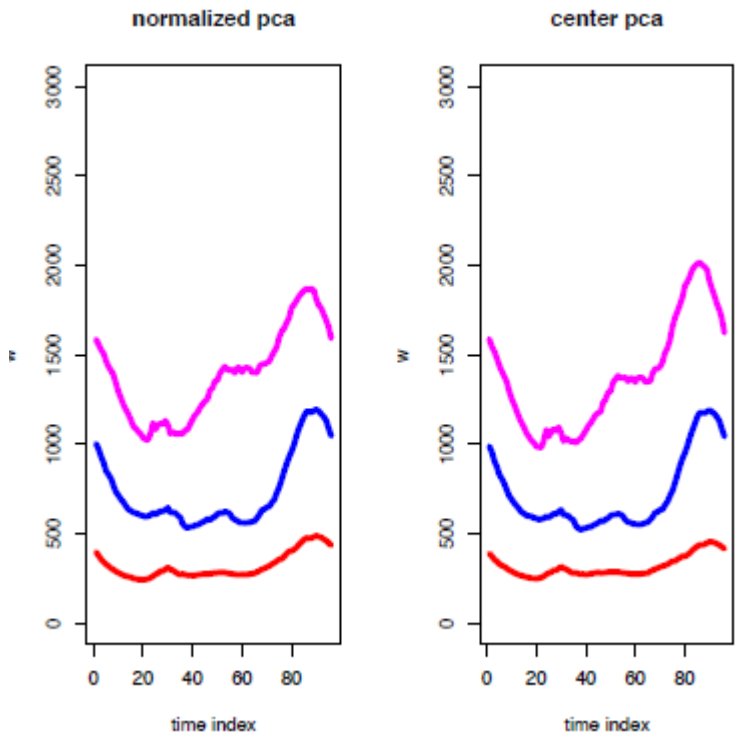


Figure 4. The average energy patterns as the number of cluster is three

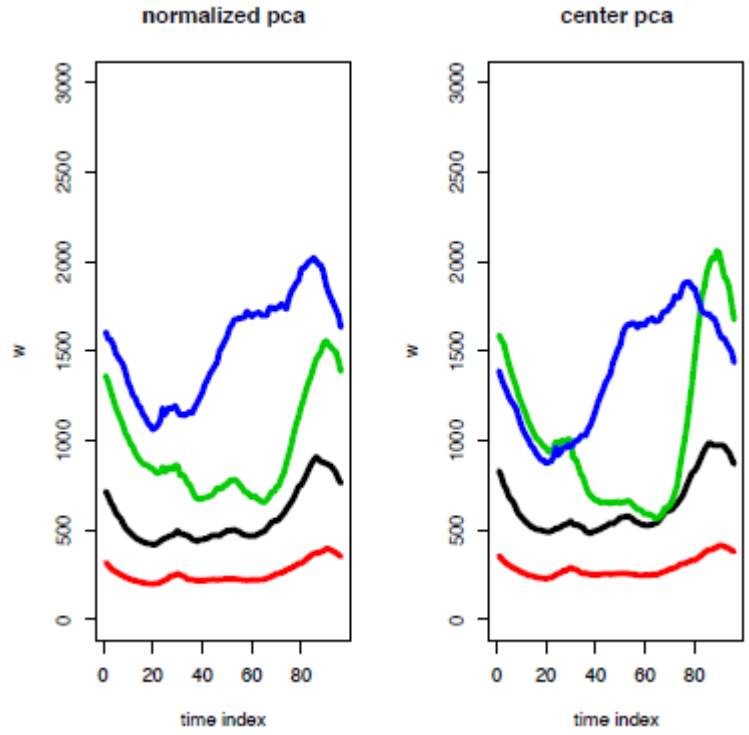


Figure 5. The average energy patterns as the number of cluster is four

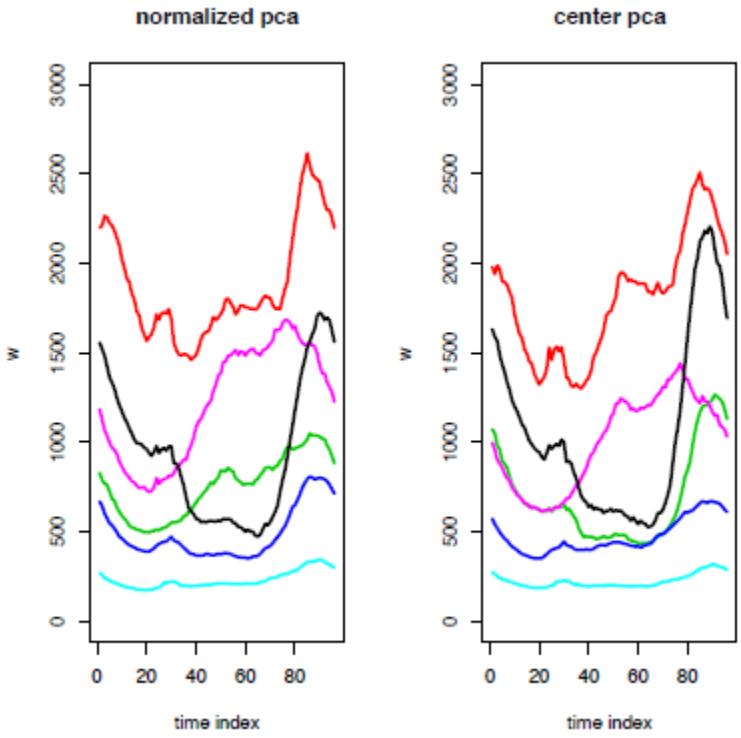


Figure 6. The average energy patterns as the number of cluster is six

5. CONCLUSION

It is the first try to cluster smart meter data with the evaluation by stability, and this paper shows the empirical results. We think that there has space either to develop theory of clustering stability or to use the evaluation in more smart meter data set.

Back to our problems as mentioned in section II, clustering stability indeed help us to make these decisions at the same time. There are no big different performance between center and normalized pca in term of stability, but the center one can provide more clear separation clustering based on energy usage time and magnitude. According to our empirical results, 2 or 3 cluster are the optimal decisions. The transition from average 0.9 level to average 0.7 level occurs between k=3 and k=4. However, the secondary decisions may consider 4, 6 with 2, 3 pcs, respectively, with the stability scores above 0.8. As mentioned in the section I, partitions into 2 or 3 are too simple for electrical operators or retailers to make any management decision, and thus we must to find the secondary solutions with certain reliability.

The clustering of smart meter is a nonparameter or unsupervised learning problem. There are no suitable mathematical model describing the complex relationship mentioned in section II up to now. Hence, we think that clustering stability is a nice try for clustering of smart meter.

ACKNOWLEDGMENT

We thank the Bureau of Energy, Ministry of Economic Affairs of Taiwan, ROC for the financial support under Contract No.107-E0215.

REFERENCES

[1] I. Jolliffe. Principle component analysis. Wiley Online. 1967.

[2] Ulrike von Luxburg. Clustering stability: an overview. Found. Trends Mach. Learn., vol. 2, 235-274. 2010.

[3] Asa Ben-Hur, and Isabelle Guyon. Detecting Stable Clusters Using Principal Component Analysis. In Functional Genomics: Methods and Protocols. M.J. Brownstein and A. Kohodursky (eds.) Humana press, pp.159-182, 2003.

[4] Yi Wang, Qixin Chen, Tao Hong, Chongqing Kang. Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges. IEEE Transactions on Smart Grid, 2018.

[5] Stephen Haben, Colin Singleton, Peter Grindrod. Analysis and clustering of residential customers energy behavioral demand using smart meter data. IEEE Transactions on Smart Grid, 235-274. 2015.

[6] Simmhan, Yogesh, and Noor, Muhammad Usman. Scalable Prediction of Energy Consumption using Incremental Time Series Clustering. In IEEE International Conference on Big Data, 2013.

[7] Christian Beckel, Leyna Sadamori, Thorsten Staake, Silvia Santinic. Revealing household characteristics from smart meter data. Energy, 78:397-410, 2014.

[8] Adrian Albert and Ram Rajagopal. Smart Meter Driven Segmentation: What Your Consumption Says About You. IEEE Transactions on Power System, 2013.

[9] Charalampos Chelmis, Jahanvi Kolte, and Viktor K. Prasanna. Big data analytics for demand response: Clustering over space and time. In IEEE International Conference on Big Data, 2015.

[10] Viktoria Neimane . Distribution network planning based on statistical load modeling applying genetic algorithms and Monte-Carlo simulations. In 2001 IEEE Porto Power Tech Proceedings, 2001.

[11] Schalk W. Heunis and Ron Herman A probabilistic model for residential consumer loads. In IEEE Trans. Power Syst., vol. 17, no. 3, pp. 621625, Aug. 2002.

[12] E. Carpaneto and G. Chicco Probabilistic characterisation of the aggregated residential load patterns. In IET Gen., Transm. Distrib., 2007.

INTENTIONAL BLANK

CYBER-ATTACKS ON THE DATA COMMUNICATION OF DRONES MONITORING CRITICAL INFRASTRUCTURE

Hadjer Benkraouda, Ezedin Barka and Khaled Shuaib

College of Information Technology,
United Arab Emirates University, P.O. Box 15551, Al Ain, UAE

ABSTRACT

With the exponential growth in the digitalization of critical infrastructures such as nuclear plants and transmission and distribution grids, these systems have become more prone to coordinated cyber-physical attacks. One of the ways used to harden the security of these infrastructures is by utilizing UAVs for monitoring, surveillance and data collection. UAVs use data communication links to send the data collected to ground control stations (GCSs). The literature [1] suggests that there is a lack of research in the area of the cybersecurity of data communication from drones to GCSs. Therefore, this paper addresses this research gap and analyzes the vulnerabilities and attacks on the collected sensor data, mainly on: data availability, data integrity and data confidentiality, and will propose solutions for securing the drone's data communication systems.

KEYWORDS

Information security, UAV Security, Critical Infrastructure Security.

1. INTRODUCTION

Unmanned aerial vehicles (UAVs), also known as drones, have seen an increase in use in the last few years [2]. UAV functions range from entertainment for hobbyists to critical mission for the military. In recent years, UAVs have seen technical development that made them eligible to be used in many fields to help in reducing risk and cost, accomplishing dangerous and expensive missions by replacing human operators.

For example, UAVs are used as first responders after disasters like earthquakes, floods or fires for survivor location and rescue missions [3]. [4] reports that UAV aided sensing was also used to log telemetry data of the levels of toxic gases to determine gas leakages. More recently and in line with the big data movement, the data collected from UAV sensors has been used to perform analyses for predicative development and preventative maintenance [5]. Another important field that UAVs are entering is the field of surveillance and monitoring. This only became possible with the advancements in battery life (trip length), autonomous charging methods and fast communication mediums. These advanced UAVs are suitable for monitoring critical infrastructure, such as the power grid, water management systems and transportation systems. Industrial systems are all moving towards digitalizing their processes to offer the prospect of smoother operations, improved efficiency, and better economics. However, this growth in connectivity within industrial operations has opened a door to cyber threats. Cyber-attacks can

Natarajan Meghanathan et al. (Eds) : SPPR, SCAI, CSIA, WiMoA, ICCSEA, InWeS, NECO, GridCom - 2018
pp. 83–93, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.81708

damage hardware and lead to downtime both causing economic losses, and in more serious cases can lead to human fatality.

While industrial control systems, such as the ones used to control the smart grid, are becoming more connected, they are still dispersed and located in remote areas. In recent decades, vital components of critical infrastructure such as power generation plants and substations have been heavily protected with physical barriers: gates, CCTV, two-factor authentication entry access, and guards. These solutions are less effective in ensuring the physical security dispersed and remote areas, making critical infrastructure ICSs more prone to coordinated cyber-physical attacks. The low cost and technological advances in UAVs made them strong contenders to be used to augment security in these systems by monitoring and providing real-time data to operators. However, surveillance UAVs like other connected devices are themselves prone to cyber-attacks. This paper will analyze the attacks that target the data communication link in surveillance UAVs and propose solutions.

The rest of the paper is organized as follows. Section II reviews previous research and related work in the area of UAV cybersecurity. Section III gives an overview of the Unmanned aerial system (UAS) architecture. Section IV presents the types of UAV reconnaissance. Section V analyzes the security threats on data communication between UAVs and the GCS. In section VI, we propose solutions that address the identified vulnerabilities and provide insights on how to secure UAV data communications. Finally, section VII concludes this paper and suggests future work.

2. RELATED WORK

Since the area we are looking into is novel, we looked into adjacent security issues in UAVs and papers that review the communication mediums from UAVs to GCSs. Rudinskas et al performed a security analysis of UAV radio communication systems. The research paper studies the security of transferred information between the SAMONIT (Polish UAV project “Aircraft for monitoring”) and other entities. One of the key conclusions that the researchers highlight, is the importance of ciphering transferred information to ensure security [6]. [7], [8] both model the threats in UAVs, and AVs respectively and propose solutions. They both aim at giving a better understanding of the security vulnerabilities, attack types and the counter solutions that mitigate them. The aim of both papers is to help technologists make informed design and deployment decisions.

The most researched areas in UAV cybersecurity are GPS jamming and spoofing [1]. [9] demonstrates that UAVs that rely primarily on commercial GPS systems for positioning are vulnerable to jamming attacks. [10] exhibits the viability of spoofing commercial GPS due to the lack of encryption. Both these attack can lead to the crash or capture of critical UAVs by malicious users.

A lot of researchers have explored security vulnerabilities related to UAVs, [1] surveys all research that has been done in the area of UAV security and concludes that there is a scarcity of research about security threats related to UAV data communication. To this end, this paper analyzes these security threats.

3. UAS ARCHITECTURE

In this paper, the term UAS is used to refer to the system that is comprised of: an unmanned aerial vehicle (UAV), a ground control station (GCS) and communication links for the UAV-GCS interactions. UAVs have infiltrated many fields and this has made UAVs very diverse in their components each to suit its functionality. But most UAVs share some main components.

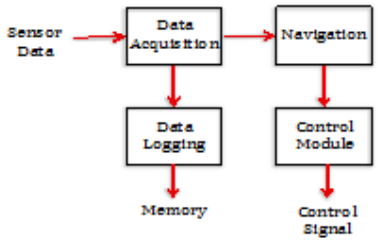


Figure 1: UAV entities

3.1.Unmanned aerial vehicles entities:

The basic UAV model can be defined as a combination of four separate, but dependent systems, figure 1 gives an overview of the interactions between UAV entities [7]:

Data acquisition: The system responsible for gathering data from the environment of the UAV. This is usually from sensors that the UAV is equipped with.

Navigation system: The system used to help in piloting the UAV. This is typically done by providing the UAV's orientation (roll, the pitch and the yaw positions). The accuracy of the navigation system is further improved by using a GPS system.

Control module: This module uses the data from the navigation system to pilot the UAV either manually through an operator or autonomously through a program.

Data logging module: This module is used for temporarily saving data before sending it to the GCS.

3.2.Ground control station entities:

Operator: This can be either a person or a program that is used to control and/or monitor UAVs during their operations.

Data storage module: This module stores data that can be used for inspection and monitoring or analysis

Data analysis module: This module is compromised of workstations that use the data received from the UAV and the data storage module for analytics.

3.3.UAV communication networks

There are two different directions of communication between UAVs and GCS (either GCS or HQ GCS) as can be seen from figure 2. The first type is control signal communication; this is usually sent from the GCS to the UAV and is used to control the UAV's motion. The other type, the category that this paper focuses on, which is data communication. This type usually refers to data sent from the UAV to the GCS. The data that is sent is mostly composed from sensor data that either aids in UAV control or telemetry data that is collected for monitoring or mission aiding purposes.

The data is communicated in phases. The sensor first collects the data. Next, it is delivered to the UAV, which either communicates it to GSC or HQ GCS. At each stage the communication

methods will vary based on the sensor's functionality, size of the data packets to be sent and the distance that the data packet has to travel. Below are the most common data communication methods.

Wired communication: This method is a physical connection and is more effective for short and immobile connections. This method is used to connect sensors to the UAV.

Wireless communication: This form of communication is commonly used to communicate data between UAVs and GCSs. Different wireless communication technologies are used. For example Bluetooth, Zigbee and WiFi are used for short ranges. WiMAX and Cellular, on the other hand, are used for longer distances and can accommodate higher data rates. Satellite communications are mostly used to communicate GPS coordinates to the UAVs and in areas where WiMAX and cellular networks are not available. Furthermore UAV manufacturers have developed proprietary transmitters and receivers used to accommodate for UAV environments such as Ocusync and Lightbridge by DJI.[11]

4. TYPES OF RECONNAISSANCE

To have a better understating of the security requirements and possible cyber-attacks the use cases of data communication links are described in this section. The UAVs that send data signals to the GCS can be categorized into three types monitoring, surveillance and data acquisition UAVs.

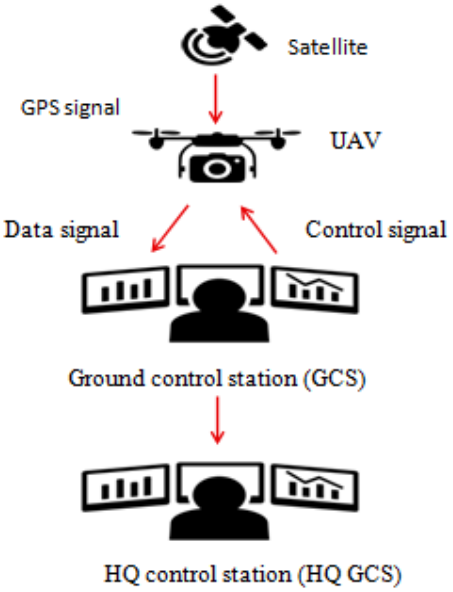


Figure 2: An overview of UAS

4.1 Monitoring

These UAVs are used to monitor the state of a target area. Monitoring UAVs are equipped with sensors; these sensors differ based on the mission of the UAV. For example, a UAV that is equipped with a temperature sensor (thermometer) can measure temperature and send it to the GCS. The operator will receive an alarm in case the readings exceed the normal range. The success of any monitoring mission depends on the correctness and availability of the sensor data. Any attacks that compromise these aspects will cause a failure of the mission.

4.2 Surveillance

Surveillance drones are used to keep a target area under observation using live stream video data that is sent to the GCS. Drones used for surveillance have become more common and they are used in many fields. For example, they are used by law enforcement agencies as part of their investigations, stakeouts and criminal pursuits. These UAVs are also used in securing critical infrastructure in remote areas like wind farms or PV plants. These critical infrastructures need continuous observation because they have become more susceptible to coordinated cyber-physical attacks [12]. In many cases surveillance data is time sensitive and needs to be sent to the operators instantaneously, this makes any attacks that result in delays can cause mission failure. Additionally, the video data is confidential and contains sensitive data, attacks that compromise the confidentiality of the video data can result in sharing the data with threat actors.

4.3 Data acquisition:

UAVs can be used for data acquisition and logging where the data is saved in the memory. In the age of big data, data is collected for many reasons such as aligning with compliance and analysis. UAVs can be particularly suitable for data collection in dangerous areas such as disaster sites, war zones or hazardous power plants. The success of any monitoring mission depends on the correctness and availability of the sensor data. Any attacks that compromise these aspects will cause a failure of the mission.

5. THREAT ANALYSIS

In this section, the vulnerabilities of UAV data communication are explored. These vulnerabilities can lead to cyber-attacks that can be classified into three categories attacks that compromise data 1) confidentiality, 2) integrity or 3) availability. Figure 3 depicts the taxonomy of the attack types in order to effectively map our proposed mitigations scheme.

5.1. Availability attacks

Attacks that compromise the availability of sensor data in UAVs can be achieved in two ways; namely through controlling the UAV or communication interruption.

In the first method of attack, the attacker compromises the UAV or the GCS. The attacker is able to gain control of the UAV and modify the functionality of its components. In the case of a camera sensor, after gaining control of the system, the attacker is able to turn-off the camera. This attack can be part of a robbery attack where the building being robbed is monitored by surveillance drones. When the attackers gain control of the UAV-GCS system they can turn off the camera during their robbery and the video becomes unavailable to the security team.

In another scenario the attacker, after taking control of the UAV, can relocate the UAV to a different geographical location. In this scenario, the data collected by a temperature sensor, for example, is not representative of the actual intended parameter. The data collected can give the operator a false sense of a safe operation environment or cause the operator to send a repair team all in vain.

In the second method of attack that compromises the availability of the system, the attackers interrupt the communication link between the UAV and the GCS. This can be done in different ways, most prominently through jamming and GPS spoofing.

Jamming aims at disrupting communication through interference or collision before reception. During a jamming attack the adversary can, for example, block or delay critical fault detection

from propagating towards the ground station. Jamming attacks can be launched without extensive knowledge or information about the attack target, this makes this attack easy to perform successfully.

DoS/DDoS are types of jamming attacks, their realization happens by flooding the network with bogus requests to make the system appear unavailable and disallow other legitimate packets from being sent. There are three ways that a DoS attack can be launched: flooding, spoofing and buffer overflow.

GPS is among the most ubiquitous technologies used for path finding in transportation. Most devices, including commercial UAVs, use civilian GPS that is unencrypted and this makes them prone to attacks.

Based on [1], *GPS spoofing* is among the most researched attacks. But in this case, GPS spoofing is used to alter the geo-location of the UAV to contaminate the data collected by sensors.

5.2.Integrity attacks:

This type of attack is achieved by either modifying the data being sent or by fabricating malicious data to replace the legitimate data. There are 2 prominent ways to compromise the integrity of the data communication of a UAV; Sensor replay attack and replacing authentic sensory data with bogus data (spoofing).

In one scenario, attackers target a camera that is used for the surveillance of critical infrastructure; the camera sends a live stream video feed to a security team that ensures that no intruders can launch a coordinated cyber-physical attack. One way around this security safeguard is by recording the monitored area aforetime and then executing a video replay attack

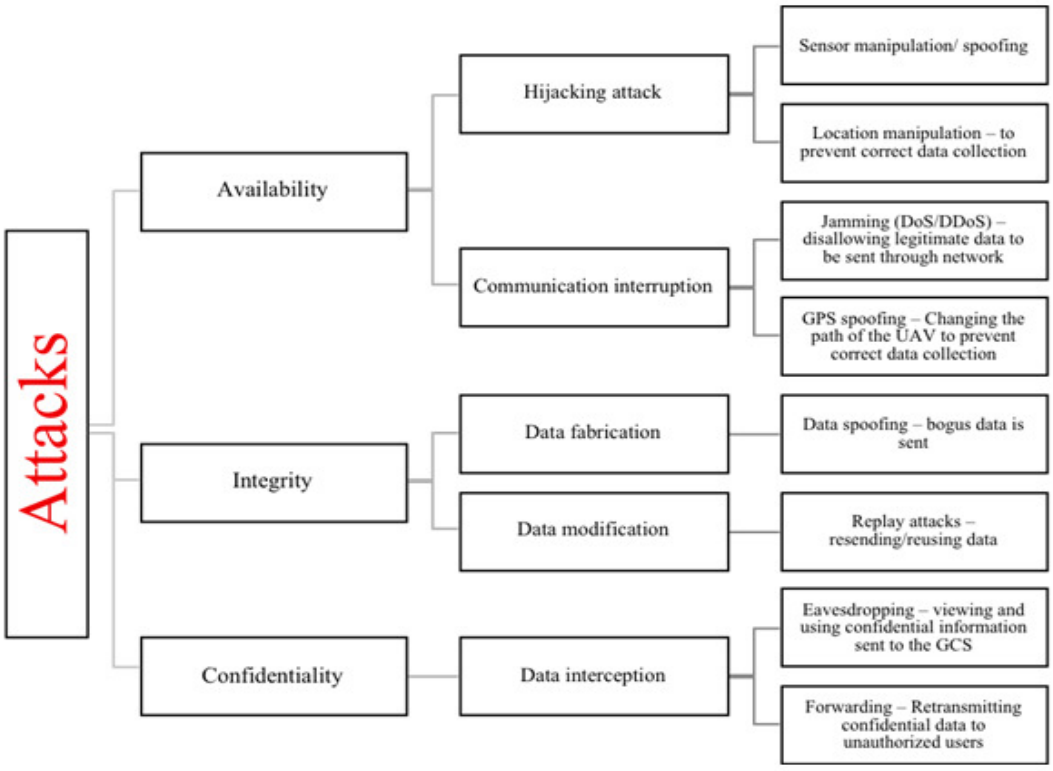


Figure 3: Taxonomy of attacks on the UAV-to-GCS communication links and data.

where the pre-recorded video is played instead. Without additional cyber-security safeguards, the security/inspecting team will not be able to detect that the video is replayed and the attackers are successful.

In another scenario, the attackers fabricate sensor data and send it to the inspector/security team. In this attack the adversary performs an active man-in-the middle attack where they intercept the information sent from the sensor, block or redirect that data and send fabricated data instead. This can cause the inspectors/security team to take misinformed decisions that may lead to expensive or harmful actions.

5.3.Confidentiality attacks:

In this attack, the adversary gets unauthorized access to confidential information by intercepting the sensor data. Attacks under this category can be either passive or active. In a passive attack, the malicious actor can eavesdrop on communication links between the UAV and GCS. In the case of a camera sensor the attacker will have live video feed of critical infrastructure. An attacker can use this as part of intelligence gathering in the reconnaissance phase of an attack targeting a critical infrastructure such as a nuclear power plant. In an active attack the adversary intercepts the signal and forwards the data to another unauthorized entity. This can be done for monetary gain; the data can be sold on the black market for example.

6. PROPOSED SOLUTIONS FOR SECURE DATA COMMUNICATION SYSTEM

This section describes safeguards that ensure data communication security. The proposed solution is depicted in figure 4. The attacks that the solution addresses are those mentioned in the previous section and they fall into three categories: availability, integrity and confidentiality.

6.1. Safeguards ensuring availability:

In real-time systems such as UAVs, the availability of data becomes of critical importance. Therefore, protecting the UAV and ensuring its resiliency against availability attacks is vital for the success of UAV missions. There are different attacks that target the availability of sensor data; likewise, there are different safeguards that help in preventing these attacks.

Hijacking attacks, either to manipulate sensors or to execute relocation attacks can be prevented by ensuring that only authorized users can modify UAV operation. A step that has to precede authorization is user authentication, to confirm that the users are valid users.

Jamming attacks, including DoS and DDoS attacks, can be prevented in several ways. One solution is by placing a firewall in the network. A firewall is a network security safeguards that filters and controls ingress and egress network traffic based on a set of rules. For example, if many packets are sent from the same address (DoS), the firewall can block incoming traffic from that source. Another way to detect that a jamming attack is happening is by setting a window and checking the rate of collision; if the rate becomes higher than normal, this would indicate that the system is under a jamming attack. The operator can then block the source of jamming attack [13]. Or adjust the transmission rate in order to contain jamming interference (Li et al. 2007).

6.2. Safeguards ensuring integrity:

UAV data communication links are being used to deliver important data that drives the decisions for missions in the army and in the industry. The integrity of data that is sent by the UAV is vital to mission success. As section V discusses, there are 2 types of attacks on data integrity: modification and fabrication attacks. Modification attacks, like replaying the same data packet or altering the contents of the data packet, can be prevented by adding a nonce/timestamp and a hash

(irreversible mathematical function/ one-way function) to each packet. The timestamp makes sure that the data packet can only be used once while the hash ensures the integrity of the message.

Another way to reinforce security is by increasing fault tolerance by introducing hardware redundancy [14]. This solution is successful in some cases, while in other cases when the size, weight, and battery constraints of UAVs are exceeded they become infeasible. Alternatively, [15] proposes that an analytical redundancy is introduced instead, that would compensate for the failures by reconfiguring the control scheme of the UAV. A hybrid of these two solutions can be used.

To prevent fabrication attacks, the sender has to be authenticated. Authentication can be achieved by using public key infrastructure, certificates and certificate authorities. Or through hard coding pre-approved MAC addresses in both the UAVs and the GCS since the numbers of UAVs and GCSs are limited.

6.3. Safeguards ensuring confidentiality:

UAV data is not of itself confidential therefore in many cases; efficiency and speed of communication are favored over confidentiality in commercial UAVs. Nowadays, commercial UAVs are part of critical missions like critical infrastructure and police surveillance. To ensure confidentiality the UAV data has to be encrypted (encoding data so that only authorized users can access it) before being sent and then authenticated and decrypted at the receiver's end. It is important to note that many encryption algorithms are computation and communication intensive and can throttle the bandwidth of a network and cause delays. It is therefore advisable for symmetric encryption to be primarily used while reserving the use of asymmetric encryption only for sensitive operations like digital signatures. In applications where live steam video is being transferred encryption can cause delays that make data to be unusable. Therefore, the encryption algorithm has to be chosen accordingly. Selective encryption is one way to reduce delays caused by encryption. This is achieved by minimizing the data that needs to be encrypted while still achieving sufficient security. This encryption algorithm works by applying encryption to a subset of the live data [16].

ACKNOWLEDGEMENTS

This research is funded by the UAEU research grant number 31T065.

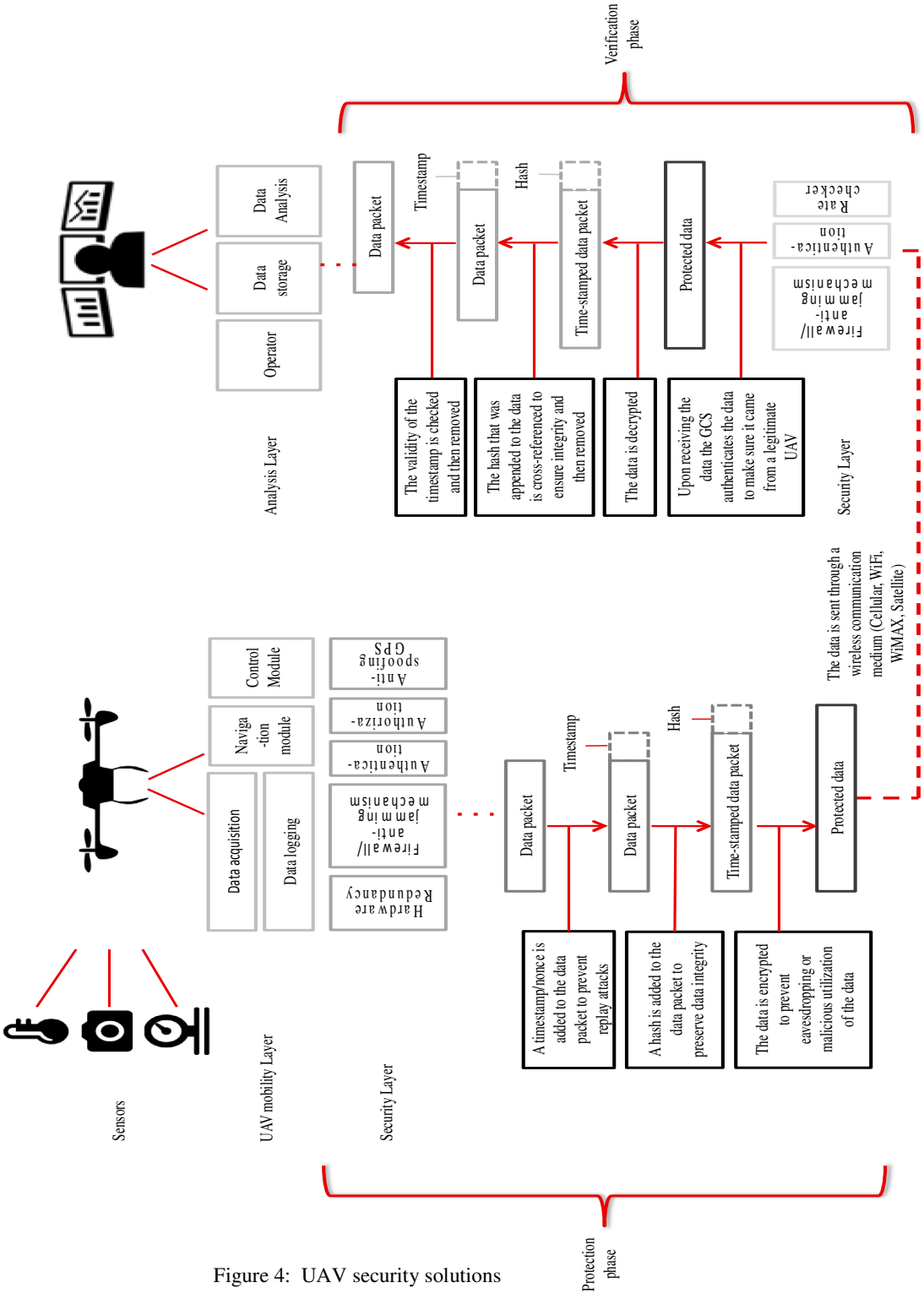


Figure 4: UAV security solutions

REFERENCES

[1] Krishna, C. G., & Murphy, R. R. (2017). A review on cybersecurity vulnerabilities for unmanned aerial vehicles. 2017 IEEE International Symposium on Safety, Security and Rescue Robotics (SSRR). doi:10.1109/ssrr.2017.8088163

[2] Rise of the Drones - Managing the Unique Risks Associated with Unmanned Aircraft Systems.(n.d.). Retrieved April 15, 2018, from http://www.agcs.allianz.com/insights/white-papers-andcase-studies/rise-of-the-drones/

[3] 5 Ways Drones Could Come to Your Rescue." Popular Mechanics. November 14, 2017. Accessed April 24, 2018. https://www.popularmechanics.com/military/g1437/5-ways-dronescould-come-to-your-rescue/.

[4] Gas-Drone: Portable gas sensing system on UAVs for gas leakage localization - IEEE Conference Publication. (n.d.). Retrieved April 19, 2018, from http://ieeexplore.ieee.org/document/6985282/

[5] Upstream Oil, Gas Companies Keep Exploring Benefits of UAVs. (n.d.). Retrieved April 16, 2018, from https://www.rigzone.com/news/oil_gas/a/146416/upstream_oil_gas_companies_keep_exploring_benefits_of_uavs/?all=hg2

[6] Rudinskas, D., Goraj, Z., & Stankūnas, J. (2009). Security analysis of uav radio communication system. Aviation,13(4), 116-121. doi:10.3846/1648-7788.2009.13.116-121

[7] Javaid, A. Y., Sun, W., Devabhaktuni, V. K., & Alam, M. (2012). Cyber security threat analysis and modeling of an unmanned aerial vehicle system. 2012 IEEE Conference on Technologies for Homeland Security (HST). doi:10.1109/ths.2012.6459914

[8] Thing, V. L., & Wu, J. (2016). Autonomous Vehicle Security: A Taxonomy of Attacks and Defences. 2016 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData). doi:10.1109/ithings-greencom-cpscomsmartdata. 2016.52

[9] K. Wesson and T. Humphreys, “Hacking drones,” Scientific American, vol. 309, no. 5, pp. 54–59, 2013.

[10] Y. Javaid, F. Jahan, and W. Sun, “Analysis of global positioning system-based attacks and a novel global positioning system spoofing detection/mitigation algorithm for unmanned aerial vehicle simulation,” Simulation, vol. 93, no. 5, pp. 427–441, 2017.

[11] Jawhar, I., Mohamed, N., Al-Jaroodi, J., Agrawal, D. P., & Zhang, S. (2017). Communication and networking of UAV-based systems: Classification and associated architectures. Journal of Network and Computer Applications,84, 93-108. doi:10.1016/j.jnca.2017.02.008

[12] Researchers Found They Could Hack Entire Wind Farms. (n.d.). Retrieved April 25, 2018, from https://www.wired.com/story/wind-turbine-hack/

[13] Petnga, L., & Xu, H. (2016). Security of unmanned aerial vehicles: Dynamic state estimation under cyber-physical attacks. 2016 International Conference on Unmanned Aircraft Systems (ICUAS). doi:10.1109/icuas.2016.7502663

[14] SADEGHI, M., SOLTAN, H., & KHAYYAMBASHI, M. (n.d.). The study of hardware redundancy techniques to provide a fault tolerant system. Retrieved from http://dergi.cumhuriyet.edu.tr/cumuscij/article/view/5000121174

[15] Evans, J., Inalhan, G., Jang, J. S., Teo, R., & Tomlin, C. (n.d.). DragonFly: A versatile UAV platform for the advancement of aircraft navigation and control. 20th DASC. 20th Digital Avionics Systems Conference (Cat. No.01CH37219). doi:10.1109/dasc.2001.963312

[16] Massoudi, A., Lefebvre, F., Vleeschouwer, C. D., Macq, B., & Quisquater, J. (2008). Overview on Selective Encryption of Image and Video: Challenges and Perspectives. EURASIP Journal on Information Security,2008, 1-18. doi:10.1155/2008/179290

AUTHORS

Hadjer Benkraouda. Msc

Hadjer Benkraouda received her B.Sc. in Electrical Engineering from the United Arab Emirates University (2015), and an M.Sc. in Cybersecurity from New York University (2017). She held cooperate (Bloomberg) and research positions (UAEU and NYU-AD). Her current research interests include Industrial Control Systems security and Network Security.

Ezedin Barka, PhD

Dr. Barka is currently an Associate Professor at the United Arab Emirate University. He received his Ph.D. in Information Technology from George Mason University, Fairfax, VA in 2002, where he was a member, and still associated, with the Laboratory for Information Security Technology (LIST). His current research interests include Access Control, where he published some papers addressing delegation of rights using RBAC. Other research areas include Digital Rights Management (DRM), Large-scale security architectures and models, Trust engineering, B2C e-commerce, and Network "Wired & Wireless" and distributed systems security. Dr. Barka is an IEEE member, member of the IEEE Communications Society and member of the IEEE Communications & Information Security Technical Committee (CISTC). He serves on the technical program committees of many international IEEE conferences such as ACSAC, GLOBECOM, ICC, WIMOB, and WCNC. In addition, he has been a reviewer for several international journals and conferences.

Khaled Shuaib, PhD

Khaled Shuaib Received his Ph.D. in Electrical Engineering from the Graduate Center of the City University of New York, 1999, his ME and BE in Electrical Engineering from the City College of New York, 1993 and 1991 respectively. Since September 2002, Khaled has been with the College of Information Technology (CIT), at the UAEU where he is currently a Professor and a Department Chair. His research interests are in the area of network design and performance, wireless communication networks, Blockchains, IoT, network security and Smart Grid. Khaled is a Senior member of IEEE. Prior to joining the UAEU, Khaled had several years of industrial experience in the US working as a Senior Member of Technical staff at GTE Labs, Waltham, MA (1997-1999), and as a Principle Performance Engineer for Lucent Technologies, Westford, MA (1999-2002).

INTENTIONAL BLANK

ANDROID MALWARE DETECTION USING MACHINE LEARNING AND REVERSE ENGINEERING

Michal Kedziora, Paulina Gawin, Michal Szczepanik and Ireneusz Jozwiak

Faculty of Computer Science and Management
Wroclaw University of Science and Technology
Wroclaw, Poland

ABSTRACT

This paper is focused on the issue of malware detection for Android mobile system by Reverse Engineering of java code. The characteristics of malicious software were identified based on a collected set of applications. Total number of 1958 applications where tested (including 996 malware apps). A unique set of features was chosen. Five classification algorithms (Random Forest, SVM, K-NN, Nave Bayes, Logistic Regression) and three attribute selection algorithms were examined in order to choose those that would provide the most effective malware detection.

KEYWORDS

Malware Detection, Android, Random Forest, SVM, K-NN, Naive Bayes, Logistic Regression

1. INTRODUCTION

The basis for the effective detection of malware is the analysis of malware on a given platform. The main malware detection techniques consist of static and dynamic analysis [16]. Dynamic analysis techniques rely on monitoring the application in real time, working in an isolated environment [1]. Static analysis works on decompiled source code, without launching applications [2] analyzing the case of reporting rights, components, API calls. In this paper we are focused on static approach case based on the automatic analysis of decompiled mobile application code. Based on reference items [3], [4] and [5], a unique feature vector derived from the application Java code was constructed. The total number of features is 696. We divided them into three categories: First one is model implementation of onReceive() methods for BroadcastReceiver components. As demonstrated in [3], in malware applications, calls to certain methods more often occur in the overridden onReceive() method than in secure applications. The full list of wanted calls in the onReceive() method of components extending the BrodacastReceiver class can be found in first part of Table 1. Second one is Linux system commands - as the Android system uses the Linux kernel, there is an API available to execute Linux-specific commands on the Android mobile device. Some of the commands under examination relate to operations on the file system. There is also a group of commands that are used to obtain administrative access to the device (rooting), then to increase the possibilities of attack, and to hide the operation of malware on the device. The full list of searched Commands is available in second part of table 1. Selected on the basis of [6]. Third one is API Calls - the largest group of features (616). Includes methods from classes, some of which have been

Natarajan Meghanathan et al. (Eds) : SPPR, SCAI, CSIA, WiMoA, ICCSEA, InWeS, NECO, GridCom - 2018
pp. 95–107, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.81709

indicated in [5]. Third part of Table 1 contains a list of classes, whose selected methods were included in the extraction of features. These were classes characteristic of the context of the mobile application, for objects of type Intention, for HTTP protocol operations, for telephone operations (SMS, connection, MMS), for device network settings, for data encryption or for dynamic code loading [19].

2. PREPARING TESTING ENVIRONMENT

Based on the literature on the subject, especially on [4], [7], [8], [9], [10], [11], [5], [12], [18], five classification algorithms were selected for testing under this work. These algorithms were among the most popular in the field of malware detection on the Android system. A collection of tested data consist of secure and malicious applications. Safe applications have been downloaded from two sources. First one is APKPure - an alternative Android application store. Second one is F-Droid - the directory of the FOSS Android application (Free and Open Source Software), which includes free and open software [17]. To increase the likelihood that applications are considered safe indeed they are not malicious, each of the downloaded files has been uploaded to the web application VirusTotal. Its task consists of analyzing the file using over 70 antivirus scanners, indicating whether the file is malicious, additionally revealing a label, indicating the species of malware to which the given scanner has classified a dangerous file. Secondly, VirusTotal provides additional information about the file, such as: the date of the first and last file upload operation to VirusTotal, the number of these operations, the results of static analysis (eg internal structure of the file), the results of dynamic analysis (behavioral characteristics of the application). The condition for joining the application to the test set of this work was not to detect malicious activity by any of the more than 70 scanners. Malicious applications also came from two sources: VirusShare - malware repository, currently containing over 24 million and Contagio Mobile - a blog that is part of the Contagio Dump project, which is a collection of malware samples. Contagio Mobile focuses on mobile malware, especially on Android and iOS. Applications downloaded from the above two sources have also been analyzed in VirusTotal. They were added to the test database if at least one of the scanners classified them as malware.

2.1. Java Code Features extraction

To be able to extract the features, the files should be prepared properly. To this end, BASH shell scripts have been developed and auxiliary scripts that organize files. Scripts gets the .apk file to the input, returning the corresponding .jar file. For this purpose, the dex2jar tool is used. It's used for work with .dex and .class files. It enables: reading and writing to a .dex file (Dalvik Executable format, executable format for Dalvik), conversion from a .dex file to .class files (compressed as a .jar file), disassembling the .dex file to a smali format, as well as the decryption of code strings present in the code, which have been obfuscated through encryption and whose decryption was to take place only after execution.

Table 1. Methods, commands and classes from which methods where extracted belonging to the feature vector from java code.

BroadcastReceiver	psneuter	StringBuilder
startService	wpthis	Process
bindService	exploid	Context
schedule	rageagainstthecage	Intent
startForegroundService	motofail	ActivityManager
registerReceiver	GingerBreak	PackageManager
goAsync	**Classes**	SmsManager
startActivity	ContentResolver	TelephonyManager
startActivieties	Cipher	DexClassLoader

Commands	Class	BaseDexClassLoader
su	File	ClassLoader
mount	FileOutputStream	Runtime
insmod	DataOutputStream	System
rebot	psneuter	ConnectivityManager
chown	wpthis	NetworkInfo
pm install	exploid	WifiManager
zergRush	rageagainstthecage	HttpURLConnection
m7	motofail	Socket
fre3vo	GingerBreak	handler

A Java program was prepared for the extraction of the tests. The external library used for the extraction of features is JD-core-java - a package decompiling Java decompiler called Java Decompiler (Java Decompiler authors did not provide a tool in the form of a library that can be used inside the code, only a decompiler as a plug-in for selected programming environments or graphics tool). Necessary to obtain the application code from the .jar file, on which the analysis of malware in the second research case is based. To focus on the analysis of features selected by all three selection methods, in Table 2 there are features present in the top characteristics for each selection method and their participation in malware applications and secure applications. All features from Table 2 are much more common in malware than in secure applications.

Table 2. Percentage of occurrences of features derived from java code, which are in the top features chosen by 3 selection algorithms in the malware and safe application sets.

Feature	**Percentage in malware**	**Percentage in safe apps**
startService	0.34809	0.00118
getString	0.35412	0.04471
setPackage	0.25553	0
putExtra	0.27767	0.01059
startActivity	0.19215	0.01882
getSystemService	0.17907	0.02000
append	0.20121	0.04588
indexOf	0.11268	0.00941
getInputStream	0.09558	0.00235

Methods from the String class, such as append or indexOf, which are much more prevalent in malicious applications than in safe ones, show a legal manipulation on the string of characters to obfuscate the code. Operations on strings are used to avoid detection by dynamically creating URLs, providing parameters to the reflection mechanism API, or to hide Linux commands. The proof of probable manipulations with Linux commands is that the method for calling them appeared in 1% of malicious applications, while the extractor detected ten times less true Linux commands (at the level of 0.1%) - thus the vast majority of applications that calls the method to Linux commands do not contain these explicit commands (or they are very unpopular and unusual commands - but this alternative is less likely). In addition, the Context.getString method allows to extract a string from application resources that are outside the Java code. Therefore, this is a great opportunity to save a dangerous string of characters, e.g. the URL of a malicious server, in the application resources, so that it will not be detected during Java code analysis, and this method allows you to download this string to the code.

The Context.startService method, the frequency of which in malware applications was mentioned in [5], is used to start the service. In the dataset of this work, it occurred about 350 times more often in malware. Knowing that the service is a component running in the background, possibly

without the user's knowledge, it seems clear that malicious applications will want to reach for the described method in order not to alert the user about malicious activity by using a component that will work in hiding.

The getSystem method. The Context class service, appearing in malicious applications 9 times more often than in safe applications, allows access to given system services. Without examining the parameters of this method, it is difficult to conclude what specifically access was requested for. However, among the system services that this method gives access to, there are those that can be potentially dangerous: window visibility management, network connections, Wi-Fi connectivity, HTTP download process, location data.

The getInputStream method from the HttpURLConnection class, occurring in almost 10% of malicious applications, and only in 0.2% of secure applications, is important in the process of data transfer (both sending and receiving) via the HTTP protocol. Thanks to this method, on the one hand, the application can receive harmful packages (payload), on the other hand, send sensitive data about the user to the external server.

Intent: setPackage and putExtra methods, used mostly in malware in comparison with secure applications, may have their justification in intentional intentions. Intentional intentions are those that do not indicate a specific component that the intention can pick up. Therefore, it is possible that the intention will be received by another application. The threat occurs when a secure application uses implicit intent, does not specify which component can perform the action, and then such an intention intercepts the malware. Then he will be able to send the application in response to inaccurate data or send information about the success of the operation at the moment when the operation did not take place. The result of intention can be saved using the putExtra function. The setPackage function is used to determine which components can receive the intention. Perhaps such a large presence in the malware serves to specify exactly which component should be responsible for the actions in order to have full control over the course of malicious activity, so that no other application could accidentally intercept the expected event.

It is noted that only two patterns listed at all appear in the code of the tested applications - and these are the startService and schedule methods. Appear on the level of 3% and 1% respectively in the malware code. They are used to activate the service accordingly and scheduling the service. Statistical tests were performed on the characteristics of Table 2 using the Mann-Whitney U-test. The confidence level at 0.05 was assumed. For each of the features in Table 2, the null hypothesis of median equality was rejected and an alternative hypothesis with a larger median in the malware population was adopted than in the population of safe applications.

2. PRACTICAL MALWARE DETECTION

The research will cover features obtained from the application code. Three methods of feature selection where be tested. Then, for each classifier, its selected parameters will be tested, and with the adopted determined parameters, the classification will be determined depending on the number of features taken into account. Next, the most common features in malware will be listed. The best results in terms of the number of correctly classified instances will be compiled for the 5 tested classifiers, along with the time of the algorithm's operation and the time of the pre-processing process and the extraction of features. For testing each classifier, 10-fold cross validation will be used, repeated 10 times. The stages are as follows: First is selection of features, second is characteristics of malware, third is classifiers and their parameters (see Table 3), fourth is summary of the best results and time data for classifiers and last each of the 5 classifiers will be tested for selected parameters (see second column of Table 3).

Table 3. Parameters tested for classifiers.

Classifier	**Parameter**
Random Forest	Iterations / max depth
Naive Bayes	-
logistic regression	Iterations
k-NN	number of neighbors
SVM	Kernel function

2.1. Random Forest

First test case included changing of the maximum depth of the tree with assumptions: number of iterations: 100, number of features: 696. Table 4 shows that after reaching the maximum depth of 50, the percentage of correctly classified instances was stable - and in the range of maximum depth from 80 to 200 and equal to infinity even identical. Some of the other indicators also remained at the same level from a depth greater than or equal to 80: TP, FN and F-measure. Interestingly, among these research cases, the lowest time was recorded for a maximum depth of 100. Measures TN and FP had the best result for a depth of 20.

Table 4. Results of the examination of the influence of the maximum depth for a random forest on the percentage of correctly classified instances, the root of mean square error and the time of learning and testing.

Max Depth	Correctly classified	mean square error	Learning and testing time [ms]	TP	TN	FP	FN	F-measure
10	78,1598	0,4268	15227,6102	0,6305	0,9582	0,0418	0,3695	0,7555
20	80,1345	0,3917	31206,3691	0,6564	0,9708	0,0292	0,3436	0,7797
30	80,2212	0,3731	46360,0157	0,667	0,9604	0,0396	0,333	0,7833
40	80,5359	0,3656	62293,7623	0,677	0,9554	0,0446	0,323	0,7885
50	80,6118	0,3636	93321,374	0,68	0,9536	0,0464	0,32	0,7899
100	80,6444	0,3637	104305,4938	0,6808	0,9534	0,0466	0,3192	0,7904
200	80,6444	0,3637	122231,6517	0,6808	0,9534	0,0466	0,3192	0,7904

However, it is worth noticing a very large difference between the matrices of the confusion matrix. The average value of TP was 68%, while TN - 96%. Similarly, the average FP value was 4%, and FN - 32%. The high FN value, i.e. the second type error, seems to be more dangerous in the field of malware detection - not detecting software malware and using it may result in the unconscious infection of the system or even the network of systems, while the error of the first type, i.e. the recognition of a safe application as dangerous is a false alarm, which can be further examined and reversed the diagnosis, and then use the application securely.

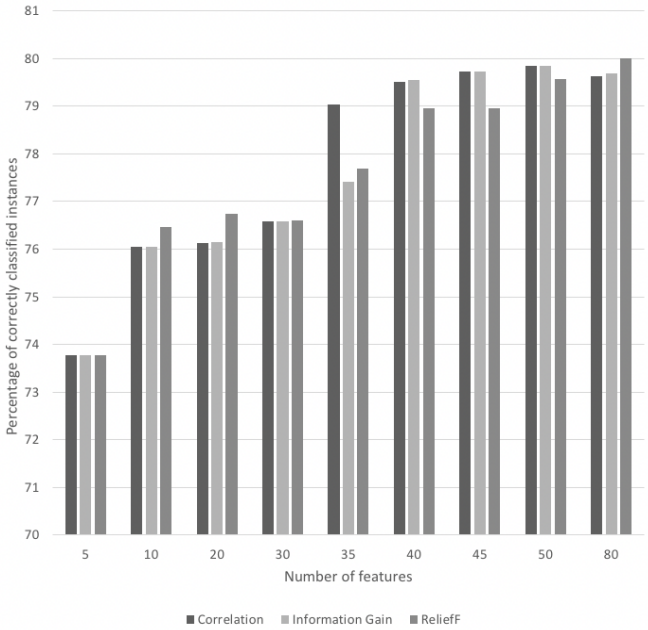


Figure 1. Percentage of correctly classified instances with changing number of attributes and attribute selection method (Random Forest)

Second test case included changing of the number of iterations with assumptions: maximum depth: 80, number of features: 696. The percentage of correctly classified instances, varying with the number of iterations. The highest percentage of correctly classified instances was recorded for 55 iterations and it was 80.6662%. 55 iterations are also the case of the highest TN index value (0.9542) and the lowest FP index value (0.0458). For the 80 iterations, the best results were obtained for FN (0.319) and F (0.7906). Again, the second type of error is much greater than the error of the first type, and the percentage of correctly classified instances never exceeded 81%, which proves the poor quality of the classification.

Figure 2. Percentage of correctly classified instances with varying number of attributes and attribute selection method (SVM)

Third test case included changing of the number of features with assumptions: maximum depth: 80, number of iterations: 55. Fig. 2 presents the summary results from the research on the change in the number of features (and the algorithm of feature selection) and its impact on the percentage of correctly classified instances. 5 to 80 best traits with step 5 were examined. Only at 80 guilds

selected using the ReliefF algorithm, this index was achieved at over 80%. It is not possible to distinguish a strong favorite among the algorithms of feature selection - they came out on the lead at different times all the algorithms studied.

2.2 Naive Bayian Classifer

This test case included changing of the number of features. Figure 3 shows how the percentage of correctly classified instances has changed depending on the number of features selected by each selection method. The highest score achieved over 80 attributes is 76.12% for 10 features selected by the ReliefF algorithm. In two cases - 10 and 35 attributes - it dominated competitors. Despite that, from 45 to 80 features, there was no significant improvement in the quality of the classification.

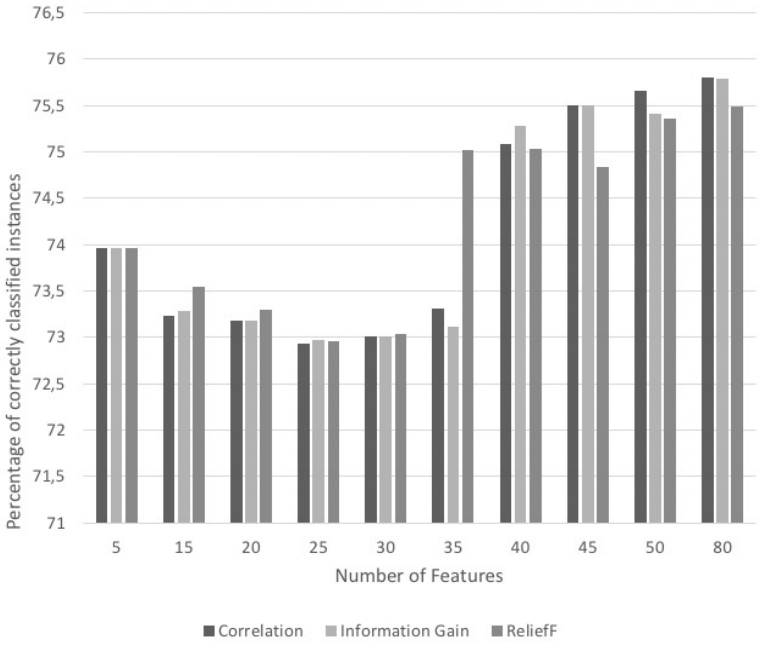


Figure 3. Percentage of correctly classified instances with varying number of attributes and attribute selection method (Naive Bayesian classifier, features from Java code)

2.2 Logistic Regression

First test case included changing of the number of iterations with assumptions: Number of features: 696. The quality of classification for logistic regression with the increase in the number of iterations decreased slightly. This is best seen in Fig. 4, where for the initial 10 and 20 iterations, more than 79.1% of correctly classified instances were achieved, and for each of the subsequent cases it was about 78.9%. With such small differences it can be said that changes in the records were negligible.

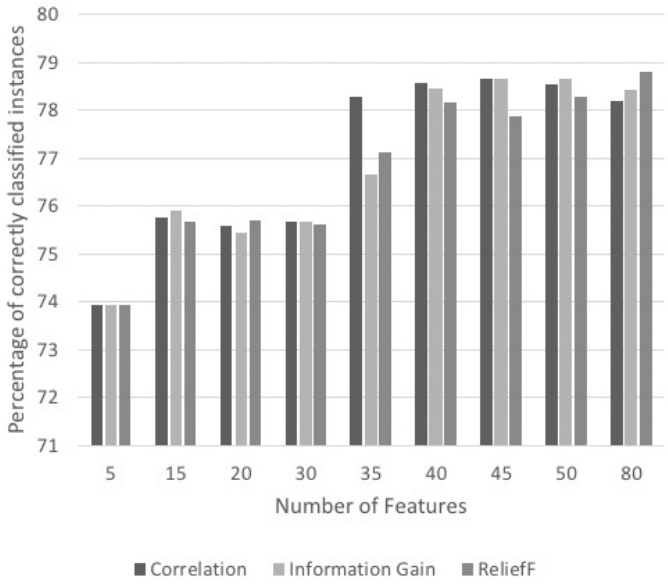


Figure 4. Percentage of correctly classified instances with varying number of attributes and attribute selection method (Logistic regression)

Second test case included changing of the number of features with assumptions: Number of iterations: 20. Figure 5 shows a deterioration in the quality of the classification with a number of features less than 35. For 35 and more features, the metric value was reached between 77% and 79%.

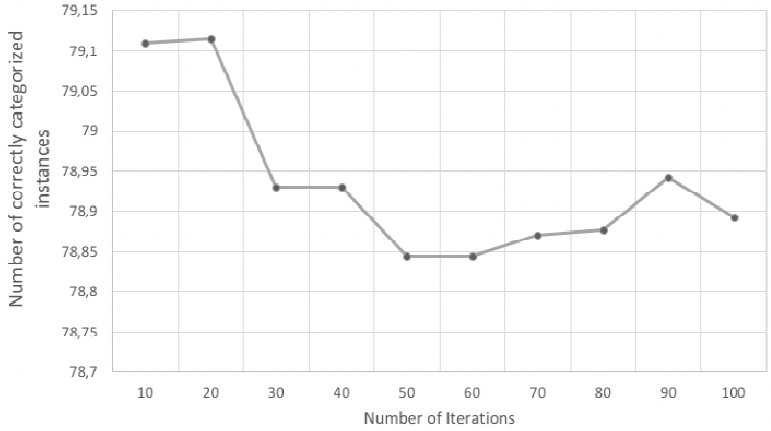


Figure 5. Percentage of correctly classified instances from the number of iterations (logistic regression)

2.5. K Nearest Neighbors

First test case included changing of the number of neighbors with assumptions: Number of features: 696. Euclidean distance function Fig. 6 show that the best results in percent of correctly classified instances, TP, FN and F, were achieved for k = 1, and these statistics deteriorated with increasing parameter k. When the number of neighbors is equal unity, there is a risk of overfitting, which is too much a fit of the model to the learning data.

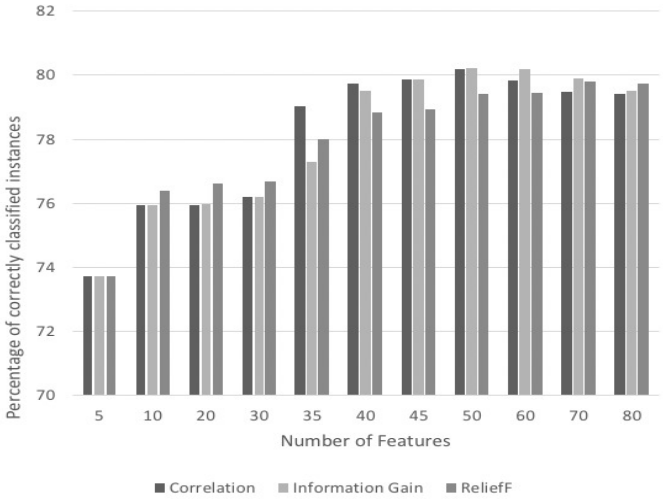


Figure 6. Percentage of correctly classified instances with changing number of attributes and attribute selection method (K-NN)

Second test case included changing of the number of features with assumptions: Number of neighbors: 1. Research on the number of features selected by 3 selection methods for K = 1 in the k nearest neighbors algorithm. With 50 features, the highest result was obtained in terms of properly classified instances - 80.1995% for information profit. Then, up to 80 traits, the value of this metric ranged from 78.3% to just over 80%. None of the selection methods was not the absolute best in this case.

2.6. SVM

First test case included Kernel Function with assumptions: Number of features: 696. Figure 7. shows that the best result by percentage of correctly classified instances was achieved for the Puk function (79.8579%), however, it is only slightly better than the Polykernel function (79.7879%), but the difference in time is large - for the Puck function this is over 60 seconds, and for Polykernel, 23 seconds. The Puk function also proved to be the best according to TP, FN, F-measure and error indicators. According to TN, at the high level of 98%, the function Normalized Polykernel won, but this should be combined with an extremely low TP rate - 55%.

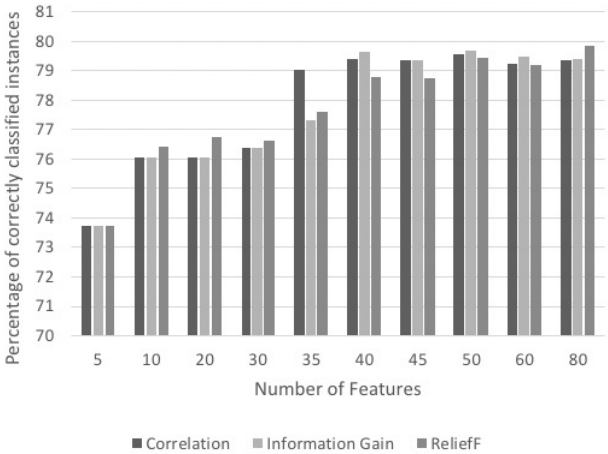


Figure 7. Percentage of correctly classified instances with varying number of attributes and attribute selection method (SVM)

Second test case included changing of the number of features. No clear favorite among the algorithms of feature selection is noticed. By reducing the number of features, a better result was not obtained according to the percentage of correctly classified instances than in sub-point a, where all 696 features were used. The highest result in this case is 79.8091% for 75 features selected by the ReliefF algorithm - it is worth noting that it is only 0.05 percentage point worse than the best result using 696 features, and the time decreased from 60 seconds to 23 seconds.

2.7. Summary of the Best Results

As part of this paper, selected aspects of the test outputs were verified by statistical tests. The tests carried out belong to the two groups: comparison of the medians of the occurrence of features in safe applications and malware and comparison of the accuracy of classifiers. Statistical methods (Shapiro-Wilk test and Lilliefors test) were checked for distribution normality for features whose size was to be compared in populations. After receiving a negative answer to the question about the normality of the distribution, the Mann-Whitney U-test was used to compare the population. It is used to answer the question whether observations in one population are greater than in the second population, which is interpreted as a comparison of medians in populations [13]. Based on [14] and [15], the McNemar test was selected to compare the accuracy of classifiers. Most of the features were dichotomous and did not have a normal distribution, which spoke for the use of the test.

Table 5. Accuracy of the classifiers

Classifier	Options	Correctness	Learning and testing time
Random Forest	max depth = 80 Iterations = 55 Features = 696	80.6662	74552.3246
Naive Bayes	Features = 10 (RelieF)	76.1217	9.5848
Logistic Regression	Iterations = 20 Features = 696	79.1152	2441.0845
k-NN	neighbors = 1 features = 696	80.3301	20979.0845
SVM	Kernel Function = PUK features = 696	79.8579	60714.3058

All statistical tests were carried out in the MATLAB environment. The best result in terms of percentage of correctly classified instances, equal to 80.6662% was obtained by random forest, with an iteration number equal to 55, maximum depth equal to 80 and 696 features. At the same time, the learning and testing time was the highest, at 75 seconds. It is worth comparing this result with the algorithm k nearest neighbors, which acted three times shorter, and correctly classified instances are lower by only 0.4 percentage points. Unfortunately, none of the algorithms exceeded 81 according to the discussed indicator. A comparison of the classification results times is shown in Table 5.

Table 6. Results of statistical surveys

	RF	NBC	KNN	LR	SVM
RF		←	=	←	←
NBC			↑	↑	↑
KNN				←	←
LR					↑
SVM					

Then the classifiers were statistically compared to the accuracy of the classifiers with the McNemar test. Table 6 shows the results of statistical surveys. The equality sign says that there were no grounds for rejecting the null hypothesis about the equality of the classifiers accuracy. The arrow indicates the classifier for which an alternative hypothesis has been adopted with greater accuracy than for the second classifier. According to statistical surveys, there are the following relationships between the accuracy of classifiers: Random forest algorithm and k nearest neighbors have the same accuracy, and both are more accurate than logistic regression, SVM and the naive Bayesian classifier. The naive Bayesian classifier has an accuracy lower than all other algorithms.

3. CONCLUSION AND FUTURE WORK

Paper is focused on the issue of malware detection for currently the most popular mobile system Android, using static analysis. In this thesis, an overview of Android malware analysis was presented, and a unique set of features was chosen that was later used in the study of malware classification. Five classification algorithms (Random Forest, SVM, K-NN, Nave Bayes, Logistic Regression) and three attribute selection algorithms were examined in order to choose those that would provide the most effective malware detection. The characteristics of malicious software were identified based on a collected set of applications. This analysis was conducted for features extracted from Java class code. It was determined which source of features provides higher quality of classification.

Research has been carried out to select the best classification algorithms for application, which is detection of malware on the Android platform, indication of the applications features of the highest usefulness in the classification of malware. Among the classification algorithms, the best proved to be: random forest and k nearest neighbors. They obtained the highest scores on the percentage of correctly classified instances (at the level of 80.3% - 80.7% for Java code). The accuracy of these classifiers was examined statistically and turned out to be the same. With the use of the naive Bayesian classifier and logistic regression, the classification accuracy was lower. It was noticed, to a small extent, the advantage of the existence of patterns of implementation of the onReceive method in malware, namely calling the function of starting or scheduling a new service.

The research on Java code has shown a strong presence of methods for manipulation on strings, as well as for downloading them outside of Java code. Such actions are manifestations of attempts to hide the real purpose of the application, i.e. obfuscation of the code. In addition, there has been a high use of methods that give access to and launch services (including system services). There is an increased presence of the method for data transfer over the HTTP protocol compared to secure applications, as well as methods for handling intentions, especially secret ones [20-23]. However, the quality of malware detection based on Java code proved to be low. None of the algorithms did exceed 81% of correctly classified instances. There are many reasons for this: the transformation and obfuscation of the code, the mechanism of reflection, manipulation on the chains of characters make the extraction of features a difficult task. Calling the API method can be implemented in several ways, and code transformation additionally increases the difficulty.

REFERENCES

[1] Kabakus, Abdullah Talha, & Ibrahim Alper Dogru, (2018), “An in-depth analysis of Android malware using hybrid techniques”, Digital Investigation.

[2] Verma, Prashant & Akshay Dixit, (2016), “Mobile Device Exploitation Cookbook”, Packt Publishing Ltd.

[3] Mohsen, Fadi, (2017), "Detecting Android Malwares by Mining Statically Registered Broadcast Receivers", Collaboration and Internet Computing (CIC), IEEE 3rd International Conference.

[4] Yerima, Suleiman Y, (2013), "A new android malware detection approach using bayesian classification", Advanced Information Networking and Applications (AINA), 2013 IEEE 27th International Conference on. IEEE.

[5] Aafer, Yousra, Wenliang Du, & Heng Yin, (2013), "Droid apiminer: Mining api level features for robust malware detection in android", International conference on security and privacy in communication systems, Springer.

[6] Seo, Seung-Hyun, (2014), "Detecting mobile malware threats to homeland security through static analysis", Journal of Network and Computer Applications 38, pp: 43-53.

[7] Aung, Zarni, & Win, Zaw, (2013), "Permission-based android malware detection", International Journal of Scientific & Technology Research 2.3, pp: 228-234.

[8] Feizollah, Ali, (2017), "Androdialysis: Analysis of android intent effectiveness in malware detection", Computers & Security 65, pp: 121-134

[9] Mas' ud, Mohd Zaki, (2014), "Analysis of features selection and machine learning classifier in android malware detection", Information Science and Applications (ICISA), 2014 International Conference on. IEEE.

[10] Al Ali, Mariam, (2017), "Malware detection in android mobile platform using machine learning algorithms", Infocom Technologies and Unmanned Systems (Trends and Future Directions)", (ICTUS), 2017 International Conference on. IEEE.

[11] Li, Yiran, & Zhengping Jin, (2015), "An Android Malware Detection Method Based on Feature Codes.", Proceedings of the 4th International Conference on Mechatronics, Materials, Chemistry and Computer Engineering.

[12] Nezhad Kamali, Maryam, Somayeh Soltani, & Seyed Amin Hosseini Seno, (2017), "Android malware detection based on overlapping of static features", 7th International Conference on Computer and Knowledge Engineering (ICCKE 2017), October 26-27, 2017, Ferdowsi University of Mashhad.

[13] B.H. Robbins, (2010), "Non Parametric Tests", B.H. Robbins Scholars Series, Dept. of Biostatistics, Vanderbilt University.

[14] Bostanci, Betul, & Erkan Bostanci, (2013), "An evaluation of classification algorithms using McNemars test", Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012). Springer, India.

[15] Dietterich, Thomas G, (1998), "Approximate statistical tests for comparing supervised classification learning algorithms." Neural computation 10.7, pp: 1895-1923.

[16] La Polla, Mariantonietta, Fabio Martinelli, & Daniele Sgandurra, (2013), "A survey on security for mobile devices", IEEE communications surveys & tutorials 15.1, pp: 446-471.

[17] Tam, Kimberly, (2017), The evolution of android malware and android analysis techniques", ACM Computing Surveys (CSUR) 49.4, pp: 76

[18] Liang, Shuang, & Xiaojiang Du, (2014), "Permission-combination-based scheme for android mobile malware detection", Communications (ICC), 2014 IEEE International Conference on. IEEE.

[19] Saracino, Andrea, (2016), "Madam: Effective and efficient behavior-based android malware detection and prevention", IEEE Transactions on Dependable and Secure Computing.

[20] Linn, Cullen, & Saumya Debray, (2003), "Obfuscation of executable code to improve resistance to static disassembly", Proceedings of the 10th ACM conference on Computer and communications security, ACM.

[21] Enck, William, Machigar Ongtang, & Patrick McDaniel, (2009), "On lightweight mobile phone application certification", Proceedings of the 16th ACM conference on Computer and communications security. ACM.

[22] Vidas, Timothy, & Nicolas Christin, (2014), "Evading android runtime analysis via sandbox detection." Proceedings of the 9th ACM symposium on Information, computer and communications security. ACM.

[23] Burguera, Iker, Urko Zurutuza, & Simin Nadjm-Tehrani, (2011), "Crowdroid: behavior-based malware detection system for android", Proceedings of the 1st ACM workshop on Security and privacy in smartphones and mobile devices. ACM.

INTENTIONAL BLANK

SAITE STORE 2.0: EXPERIENCE REPORT ON THE DEVELOPMENT OF AN IMPROVED VERSION OF A DIGITAL LIBRARY APPLICATION

Ana Emilia Figueiredo de Oliveira[1] Katherine Marjorie Mendonça de Assis[2] Camila Santos de Castro e Lima[2] Carla Galvão Spinillo[3] Elza Bernardes Monier[2] Maria de Fatima Oliveira Gatinho[2] and Marcelo Henrique Monier Alves Junior[4]

[1]Department of Dentistry I, Federal University of Maranhão, São Luís, Brazil.
[2] UNA-SUS/UFMA, Federal Universityof Maranhão, São Luís, Brazil.
[3]Department of Design, Federal University of Paraná, Curitiba, Brazil.
[4] Campus of Porto Franco, Federal Instituteof Maranhão, São Luís, Brazil.

ABSTRACT

The use of mobile technologies in the educational process has been generating positive results. In this context, institutions that work with Distance Education (DE) need to update and adapt their processes to these innovations. Considering this trend, UNA-SUS/UFMA has built, in partnership with the Saite Group, a virtual library that enables fast and free access to its contents using mobile devices. With the expansion of the tool's use, the institution invested in a new version to provide a better experience to its users. This paper aims to perform a detailing of the Saite Store new version, showing its operation and the technological implementations carried out. Finally, the importance of performing updates in the application is highlighted, considering its great potential for use by more than twenty thousand users.

KEYWORDS

Distance education, virtual library, e-learning, mobile.

1. INTRODUCTION

Currently, the ease of access to these devices has caused a growing number of people to have one or more devices and know how to handle them [1]. Therefore, mobile learning becomes a reality, allowing access to educational content and knowledge sharing with autonomy, regardless of economic, social and geographical factors, through mobile interaction platforms.

Distance Education (DE), already widespread in adult education, especially in higher education, has excelled in health continued education [2]. In Brazil this trend is confirmed and because of that the regulatory institutions such as the Ministry of Education, have become aware to the creation and imposition of rules to regularize the practices on Health Distance Education, in order to ensure quality in this mode, aligning to the presential initiatives. Regulation is essential to ensure that the economic appeal offered by DE do not encourage the multiplication of courses without enough theoretical and scientific basis to form good professionals [3].

Natarajan Meghanathan et al. (Eds) : SPPR, SCAI, CSIA, WiMoA, ICCSEA, InWeS, NECO, GridCom - 2018
pp. 109–118, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.81710

The modality has expanded in various ways, following the major advances both with regard to the evolution of technologies, and in relation to the renewed conceptions in the area of education. In this sense, one of the branches of DE that has gained more and more space is the m-Learning (Mobile Learning). The term refers to the use of mobile devices (such as cell phones, smartphones, PDAs, handheld computers, tablet PCs, laptops, and personal media players) to mediate the learning process. The m-Learning expands as mobile devices become indispensable in the routine of individuals. The need to update and sync with the virtual world experienced in the contemporary world makes people rarely distance themselves from their mobile devices. The m-Learning initiatives are based on the observance of the ubiquitous character of these devices, taking advantage of the fact that their presence is so latent in the lives of individuals, using that to bring them closer to educational practice [4].

The possibility to facilitate the access of the student to course content through mobile applications known as m-Learning presents itself as an important and powerful tool for Distance Education. M-Learning is defined as the ability to learn anywhere at anytime without the need for a physical connection to network cables [5].

Recent studies suggest that medical science students have positive perceptions regarding mobile learning [6][7].In addition, students report that m-Learning tools have been as effective as traditional teaching in clinical environments and formal learning environments [8][9].

The Saite Store is one of the tools that enable the dissemination of knowledge with quality. This virtual library makes e-books available for free. Its content is mostly aimed at health professionals, since the training of workers in this area imposes the need for permanent qualification and innovation of educational practices, in a dynamic and systematized way [10]. The objective of this work is to describe the experience of developing a new version of the Saite Store, as well as describing the virtual library.

2. MOBILELEARNING

Mobile learning or m-learning, is the learning through the use of mobile devices as a platform for studies [11], through the integration of several processing and data communication technologies. With this, it allows students and teachers a greater possibility of interaction, providing education at a more comprehensive level. It enables learning to no longer be limited by location, time, and a physical connection line, with more expensive equipment such as a desktop computer or a laptop [12] [13].

M-learning is an even more flexible form of education, which creates new spaces and ways of interaction, not just an extension of formal education. Among the advantages of using m- learning as a teaching-learning strategy, it is worth to mention: acquiring knowledge anywhere, requiring only that the users have their device at hand, even in movement; being able to access the learning system you want, when and where you need it; expanding the learning resources, each student can choose the material and method he/she finds most convenient; enabling learning in real context, since it is possible to exchange experiences with other users at different levels [14].

From these considerations, it is possible to see the relevance of m-learning implementation in the model of distance education.

3. SAITE STORE1.0

The Open University of Brazilian National Health System of the Federal University of Maranhão (UNA-SUS/UFMA), attentive to the expansion of smart phones market and following an

increasingly current tendency of mobile devices use in education, developed the Saite Store: an application that offers content mostly in health area, which can be downloaded through the Play Store and Apple Store. The platform was developed in partnership with the Saite Group - research group focused on health, innovation, technology and education. In this application, users can download free e-books and manage them according to their interest by creating their own virtual library.

A multidisciplinary team formed by professionals from the areas of Instructional Design (ID), Graphic Design (GD) and Information and Technology (IT) participated in the development of Saite Store. The application was built using HTML5, CSS3 and JavaScript technologies in view of the strong application of these technologies in the market and their compatibility with major browsers such as Chrome, Firefox, Safari and Internet Explorer.

UNA-SUS/UFMA develops applications for mobile devices addressing mostly health care subjects since 2013. One of the main purposes to develop these learning tools is to provide material elaborated by experts in health care, primary care and public health, for free, and to be accessed through any type of mobile devices online and offline. In this way, it became possible to reach even those who live or work in remote regions, where Internet access is limited.

Nevertheless, despite of being innovative learning objects, which have become popular in a short period, there were some limitations to be overcome. Before SAITE Store creation, it was noted that as each application had to be searched and downloaded individually and it became uncomfortable for users to have an application for each book in their mobiles. Another limitation was the search in Google and Apple stores, which could direct users to various applications, not necessarily those produced by UNA-SUS/UFMA.

SAITE Store has emerged as a solution to these problems. It collects all the applications developed by UNA-SUS/UFMA, making easier for users to find and download the e-books they are interested in.

SAITE Store is, therefore, an application that works as a virtual store, providing access to e-books (interactive classic books) categorized by subject areas. The books address health issues and also cover topics from other areas such as Portuguese.

The production of first grade, free of access, free and unrestricted, educational materials prepared for professionals applied in the topics addressed; secondly, a web of international teaching dynamics, transforming learning into a routine, through the insertion of educational elements in the day-to-day of music lessons, with the use of mobile devices.

Currently, Saite Store already has more than 200 e-books distributed in 12 thematic areas. It is used by more than 22,000 users, proving to be useful in the democratization of the knowledge offered by UNA-SUS/UFMA. In this platform, users have the possibility to access free, interactive and updated content at any time, as well as it allows students of the institution to continue their studies even through other platforms.

4. SAITE STORE2.0

4.1 Features

In its first release, Saite was already a virtual library that featured easy-to-use e-books - with engaging design and interactive graphical animations - that used the least scrollbars in the text. This allowed the user to access all content on a single screen. Through the methodology of exploratory and stress tests with the application, the inability to implement new functionalities and to update the application with new e-books was detected. It has led to the decision to develop a more potent version of the application that would allow compatibility with a wider range of devices.

A traceability process of version 1.0 weaknesses was then performed, listing requirements for the new version to be developed.

In this updated version, the application gives the user a more flexible and completely offline browsing experience, increasingly providing the user with autonomy to access intended content whenever and wherever they want.

In the previous version of Saite, it was only possible to navigate through the store with internet connection. When there was no connection, the user had access only to downloaded books.

With this upgrade, navigating through the application has become easier and totally off-line. This means a lot more independence: the user can perform any activities that the application makes available even without internet access; the connection is needed only for downloading e- books.

A new database architecture has been implemented for the server environment and the user's mobile device. With the operation of this data structure, the use of the offline application was allowed, since a mirror of the service is provisioned at the client.

In addition to the change in navigability, the new Saite Store also has a greater capacity of books storage than the previous version, mainly because of the improvement in the application's speed.

The new version has optimization algorithm for data provision and only once a day a mirror of the data that the application needs is generated. Two algorithms are used on the server: one to create a file with all available information and another only for new information.

Another interesting change is that the new Saite Store brings a more flexible design than the previous one and now the student has a tool more compatible and adaptable to different devices. It was sought to understand how the user uses his device and then a version that is better suited to reading was built - be it horizontal for tablets or vertical for smart phones. Specific versions were also developed for smart phones, adapted to the guidelines of each of the operating systems (iOS, Android, and Windows). According to the Google publisher (Play Store), the application's acceptance in approved devices has reached 100%. In this paper, we dealt with the details of the version for Android devices.

4.2 SaiteStore 2.0Detailing

When the users open the app, they come across the home screen, which contains a preview of the thematic areas in the app that are displayed in rotating banner format. Below are e-books divided by some thematic areas, with an initial highlight for the recent e-books category, which presents the novelties of the store. This way, the user keeps informed about new content options.

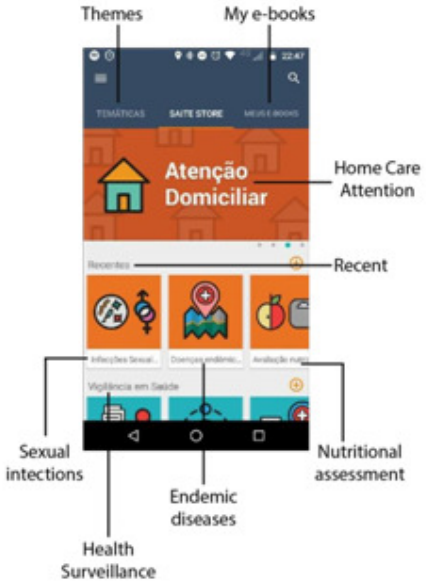


Figure 1: Saite Store Home screen.

The navigation through these lists occurs in the horizontal direction. So that the user can intuitively understand the navigation direction, the last e-book card in the line appears partially on the screen, encouraging him/her to scroll to get the rest of the information. Above these, the button with the plus symbol is included, to present the other e-books that compose the list.

The application's main functions are presented in the top menu for easy navigation. In this menu, the user will find: "Search bar", which works with data entry through typing or voice search, saving recent searches performed by the user; the "Themes" option that lists all the subject areas in which e-books are currently categorized in the store; and the "My e-books" option, which organizes all books that the user has already downloaded.

Together with the search button, there is a menu, indicated by an icon adapted to the language and layout on each device that presents a description of the application on the "About" button, a "tutorial", and the "exit" option.

The transition between these screens (main, thematic and my e-books) is evident: each one has a distinct appearance, to emphasize that they have different purposes, allowing a quick identification of the section where the user is.

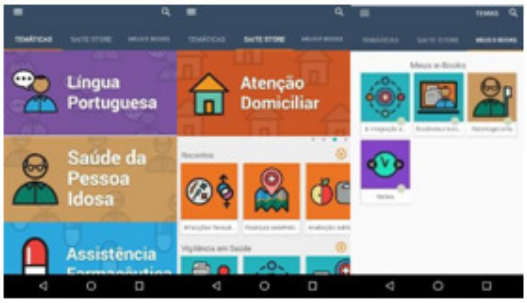


Figure 2: App’s home screens.

In the thematic areas screen, a list is presented containing the topics in which e-books are classified. Currently, the store presents 200 e-books divided into 12 thematic areas related to health, distance education, Portuguese language and scientific methodology. By clicking on the intended subject area, the user will find descriptive information about the content and all the e-books that are part of the area.

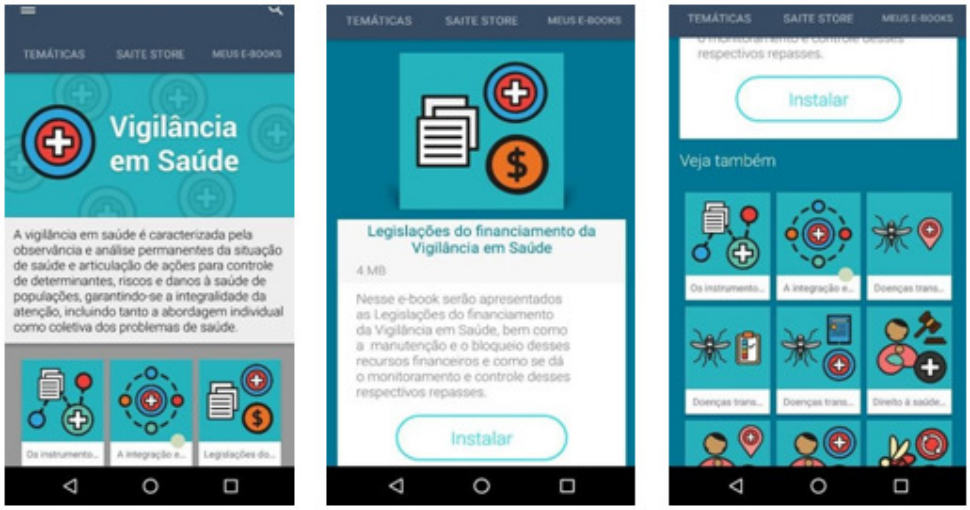


Figure 3: Screen with thematic area information, with e-book information and with related e-books.

When choosing an e-book, the user is redirected to the screen with information about its content, size, a button with installation action and a list of other e-books related to the chosen material.

By clicking on install, the user follows the installation of the material and is notified as soon as this process ends. With the e-book installed, the users can read or uninstall it. If they choose to read, they are directed to the e-book, where the content can be accessed. If they choose to uninstall the book, a warning screen pops up to confirm the action.

The e-books have a dynamic and interactive format and are updated constantly, according to the need for content renewal. Before starting to read the material, the user is informed when the last update occurred.

While reading the e-book, the application allows, by means of fixed buttons, the user to return to the homepage or return to the book description area.

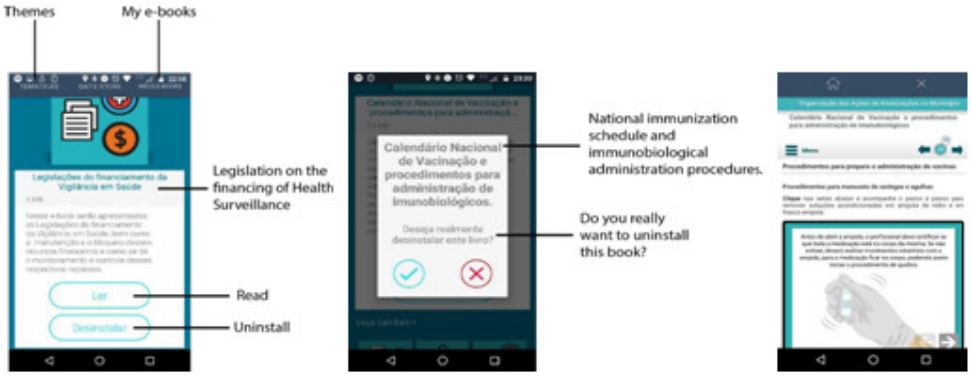


Figure 4: Read and uninstall functions, E-book uninstall alert and E-book and fixed buttons to return and.

By clicking on install, the user follows the installation of the material and is notified as soon as this process ends. With the e-book installed, the users can read or uninstall it. If they choose to read, they are directed to the e-book, where the content can be accessed. If they choose to uninstall the book, a warning screen pops up to confirm the action.

All downloaded eBooks are flagged with a green indicator and listed in the "My eBooks" area, which allows the creation of a custom library. These books can be arranged alphabetically or by theme, according to user's viewing needs.

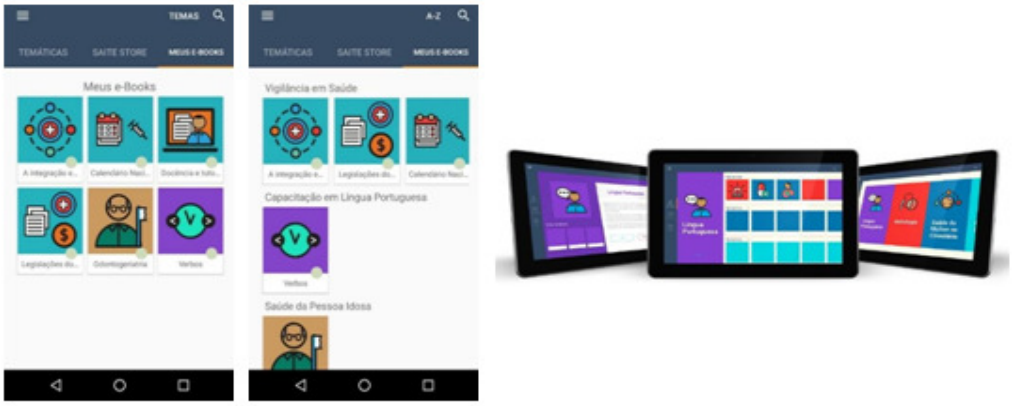


Figure 5: “My e-books” area arranged alphabetically, “My e-books" area organized by themes and Screens in tablet version.

For the tablet version, we worked on landscape orientation by re-adapting the elements to fit this guideline while retaining the application's functions and principles.

5.CONCLUSIONS

The use of applications that assist in students' studies routine has become a reality in education. In this context, Saite Store presents itself as a tool that allows access to quality free content, in an easy way. Currently, the store already has more than 21,860 downloads and presents great potential for new implementations.

For a version 3.0 of the store, some features have already been listed: the possibility of user sharing of information with other people; greater interaction with the users with dialog boxes that stimulate the permanence and facilitate even more their navigation in the platform; a module to collect data about the users' behavior in relation to the application (number of pages read, most downloaded and most uninstalled e-books, time of navigation in the application, etc.); internationalization of the store for English and Spanish versions; insertion of functionality that allows the users to reopen the book on the page where they stopped, if they close the application and open it again.

For version 2.0, it is planned to carry out surveys with users to collect information and feedback about their experiences and perceptions regarding the application and to validate the implementation of the new features listed for a new version.

SAITE Store figures as tool to promote continuing education in health, since it provides various e-books which address many topics in health field, and by no cost. The e-Books offered in the virtual store are powerful tools for health professionals learning, using any type of mobile device, with or without access to Internet, which results in more possibilities to improve knowledge, hence, resulting in better assistance to the population assisted.

REFERENCES

[1] UNESCO. Diretrizes de políticas para a aprendizagem móvel. Brasil, 2014. Available at: <http://unesdoc.unesco.org/images/0022/002277/227770por.pdf>.

[2] Moran J. A EAD no Brasil: cenário atual e caminhos viáveis de mudança. Available at: <http://files.educacao-e-tics.webnode.com/200000119-e8e55e9e03/Ead%20-%20Moran.pdf>.

[3] Mesquita KC; Silva JA, Igreja ACSM. Aplicabilidade da educação a distância na educação médica continuada. Brasília Med2012;49(2):111-117

[4] Park Y. A Pedagogical Framework for Mobile Learning: Categorizing Educational Applications of Mobile Technologies into Four Types. The International Review of Research in Open and Distance Learning, 12(2),78-102.

[5] Georgiev T, Georgieva E, Smrikarov A. M-Learning: A new stage of E-Learning. Proceedings of the 5th international conference on Computer systems and technologies.2004.

[6] Chase TJG, Julius A, Chandan JS, et al. Mobile learning in medicine: an evaluation of attitudes and behaviours of medical students. BMC Medical Education. 2018;18:152.

[7] Koohestani HR, SoltaniArabshahi SK, Fata L, Ahmadi F. The educational effects of mobile learning on students of medical sciences: A systematic review in experimental studies. Journal of Advances in Medical Education & Professionalism. 2018;6(2):58-69.

[8] Schooley B, Walczak S, Hikmet N, Patel N. Impacts of mobile tablet computing on provider productivity, communications and the process of care. Int J Med Inf. 2016;88:62–70.

[9] Lee L-A, Chao Y-P, Huang C-G, Fang J-T, Wang S-L, Chuang C-K, et al. Cognitive style and mobile E-learning in emergent otorhinolaryngology-head and neck surgery disorders for millennial undergraduate medical students: randomized controlled trial. J Med Internet Res. 2018;20(2):e56.

[10] Campos FE, Lemos, AF, Vianna RF et al. Experiências exitosas da Rede UNA-SUS: trajetórias de fortalecimento e consolidação da Educação Permanente em Saúde no Brasil / Organização de Francisco Eduardo de Campos. [et al..]. - São Luís: EDUFMA,2017.

[11] Houser C, Thornton P, Kluge D. Mobile Learning: Cell Phones and PDAs for Education. In International Conference on Computers in Education. Japão,2002.

[12] Ribeiro OS, Medina DR. Mobile Learning Engine Moodle (MLE - Moodle): das funcionalidades a validação em curso a distância utilizando dispositivos móveis. Porto Alegre: UFRGS,2009.

[13] Shippee, M.; Keengwe, J. m-Learning: anytime, anywhere learning transcending theboundaries of the educational box. EducInfTechnol (2014) 19: 103. https://doi.org/10.1007/s10639-012- 9211-2.

[14] Kwon, S.; Lee, J. Design principles of m-learning for ESL. In Procedia Social and Behavioral Sciences 2, 2010. p.1884-1889.

AUTHORS

Ana Emilia Figueiredo de Oliveira - Full Professor atthe Federal Universityof Maranhão. She holds an undergraduate degree in Dentistry from the Fluminense Federal University (UFF); a Specialization in Systems Management and Health Services (UFMA); a Masters and PhD degree in Dental Radiology from the State University of Campinas (UNICAMP) and a Post-Doctorate degree/Visiting Professor for the University of North Carolina/Chapel Hill-EUA (UNC-Chapel Hill/USA). Coordinator of the Open University of SUS/UFMA. Scientific Director of the Brazilian Telemedicine and Telehealth Association. Leader of the SAITE Research Group - Technology and Innovation in Health Education (CNPq / UFMA). As a researcher, works mainly in the following subjects: Distance Education,Technology and Innovation in Health Education, Repercussions of oral alterations on women´s health, Imaging, Primary Health Care, Mobile Applications, Management and Educational Monitoring Systems, Openinnovation.

Katherine Marjorie Mendonça de Assis- She holds a bachelor degree in Administration from the Federal University of Maranhão – UFMA; has experience in team management, communication, marketing and project management. Currently, she works at the Open University of SUS - UNA- SUS/UFMA as Coordinator of the Communication and Design Center. She previously worked at the institution as Secretary of Interinstitutional Relations and Production Supervisor.

Camila Santos de Castro e Lima - Graduated in Design from the Federal University of Maranhão (UFMA). Master's Degree in Design in the line of research Information and Communication, in the Graduate Program in Design of the Federal University of Maranhão (PPGDg / UFMA).

Carla GalvãoSpinillo- PhD, University of Reading, GB, 2000 and postdoctoral fellow at the University of Avans, Netherlands, 2010) is a research professor at UFPR-Federal University of Paraná and collaborator of PPGDesign-UFMA and UNASUS / UFMA. She is a co-founder of the Brazilian Journal of Information Design - InfoDesign, and general editor of the IDJ-Information Design Journal (John Benjamins Publishers, The Netherlands). She is a member of the Health Design Network (Canada), the Saite Group (UNASUS / UFMA) and the National Science Education Network. Participated in the elaboration of regulation of Bolts of Magisterial Medicines of Paraná (Resolution SESA No 062/2013) and the Nutritional Labeling WG of ANVISA(2015).

Elza Bernardes Monier- Graduated in Dentistry from the Federal University of Maranhão (UFMA); Specialization in Distance Education (SENAC); Master in Health Sciences (UFMA) and PhD in Medical Sciences from the State University of Rio de Janeiro (UERJ). Currently, she works at the Open University of SUS - UNA-SUS / UFMA as Coordinator of Management of Educational Offers.

Maria de Fatima Oliveira Gatinho- Graduated in Economic Sciences from the Federal University of Maranhão (UFMA); Specialization in Health Systems and Services Management (UFMA) Currently working at the Open University of SUS - UNA-SUS / UFMA as Coordinator of Financial Management.

Marcelo Henrique MonierAlves Junior - Graduated in Information System from the University Center of Maranhão (CEUMA); MBA in Project Management by FundaçãoGetúlio Vargas / FGV; Master in Computer Science, Federal University of Maranhão - UFMA. Currently, he works as a professor at the Federal Institute of Maranhão (IFMA).

REAL-TIME P2P STREAMING BASED ON PLAYBACK-RATE IN MANETS

Chia-Cheng Hu[1,2,*], Zhong-bao Liu[2,3], Hong-Bo Zhou[2]and Chong-JieZhang[2]

[1]College of Information Engineering, Yango University, Fuzhou, China
[2]School of Software, Quanzhou University of Information Engineering, Quanzhou, China
[3]School of Software, North University of China, Taiyuan, China
*Corresponding author

ABSTRACT

In a QoS-intensive multimedia application, Media-on-Demand (MoD) streaming can be delivered to asynchronous users with asynchronous requirement of MoD and VCR-like operation support. It is a critical challenge to propose a segment scheduling algorithm for real-time Peer to Peer (P2P) streaming services in mobile ad hoc networks (MANETs).However, it is a big challenge to provide MOD multimedia streaming to a large population of clients due to the asynchronous users. In this paper, we propose areal-time P2P scheduling algorithm by scheduling the segments evenly transmitted into the network according to the playback-rate of the real-time streaming service. The proposed algorithm schedules the segments from the peer with less bandwidth consumption to the network for further saving the limited bandwidth. On the other hand, it is adaptive to host mobility. Extensive simulations illustrate the effectiveness of the proposed scheme.

KEYWORDS

Mobile Ad Hoc Networks, Playback Rate, Segment Scheduling and Timely P2P Streaming

1. INTRODUCTION

In a QoS-intensive multimedia application, Media-on-Demand (MoD) streaming on Internet can be delivered to asynchronous users with asynchronous requirement of MoD and VCR-like operation support. However, it is a big challenge to provide MOD multimedia streaming to a large population of clients due to the asynchronous users and the limited provider capacity. To tackle the issue, peer-to-peer (P2P) networks has emerged as a powerful and popular paradigm for many scalable problems over Internet. The basic design philosophy of P2P networks is to encourage users to act as both clients and servers, namely as peers. In a P2P network, the peers collaborate for the purpose of providing streaming service to each other in the sense that they can all behave as clients and servers. Since a peer downloads some segments for its own playing and then caches them to serve future requests from the other peers, the segments (which constitute the multimedia file) can be spread out quickly among the peers. A mobile ad-hoc network (MANET) is a kind of networks, and it is composed of mobile devices that can arrange themselves in various ways and operate without strict top-down network administration. Nearly, P2P streaming applied in MANETs has become a new focus in the P2P research field [1-4]. Real-time streaming is a necessary requirement for viewing multimedia files, in which each segment should be received

Natarajan Meghanathan et al. (Eds) : SPPR, SCAI, CSIA, WiMoA, ICCSEA, InWeS, NECO, GridCom - 2018
pp. 119–126, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.81711

before its playback deadline. Today, several works [5-14]have been proposed for P2P applications in MANETs. However, these works do not focus mainly on the scheduling scheme, and they cannot achieve a timely delivery. In this paper, we focus on segment scheduling and aim to propose a delay-sensitive segment scheduling algorithm for real-time P2P streaming applications in MANETs. In the proposed algorithm, we adopt a distinct strategy of using the limited bandwidth in MANETs more efficiently by a rate control mechanism, which determines the number of segments transmitted into the network according to the playback-rate of the requested media file. By the aid of the rate control mechanism, the bandwidth consumption and large buffer size problems can be alleviated significantly. On the other hand in order to satisfy the real-time requirement of the MoD multimedia file, the segments with approaching play-deadlines are scheduled with high priorities. Since a segment may be cached at multiple peers in a P2P system, it will schedule the segment to be transmitted from the peer with less bandwidth consumption to the network. In the next section, the previous P2P scheduling algorithms in MANETs are first reviewed. The detailed design of proposed scheduling algorithms is given in Section 3. In Section 4, the performance evaluation is carried out. Finally, this paper concludes with some remarks in Section 5.

2. RELATED WORKS

Recently, several P2P scheduling algorithms have been proposed in MANETs [5-14]. Reference [5] achieves this goal by a layered coding method for streaming a multimedia object between two mobile peers. The method divides a video segment into a base layer and several enhancement layers. The base layer has a lower bit rate compared to the original stream. It can be decoded independently to reconstruct the original video stream with a lower resolution or frame rate. However, main challenge brought on by adopting layered coding method in P2P streaming is a need for more complex scheduling algorithms. Whereas a scheduling algorithm for non-layered streaming only is concerned with throughput maximization, scheduling layered streaming has to take other constraints into account. For example, scheduling a higher layer to be delivered in addition to a lower layer may render a vain transmission if packet loss occurs in the lower layer.

In [6], the V3 architecture for live video streaming is a cooperative streaming architecture among moving vehicles. It incorporates a signaling mechanism to continuously trigger video sources to send video data back to receivers. Broadcasting and multicasting have also been used for streaming media services to mobile users [7, 8]. They are bandwidth efficient transmission mechanisms for applications where there are multiple participants. With the services, applications send one copy of the information from the provider to the receivers. However, the receivers are required to be connected and tuned at the right channel of broadcasting or multicasting. Repeated broadcasting or multicasting is hard to solve the issue of asynchronous users due to asynchronous requirement of the same media at different times and locations.

In [9], the paper proposes a middleware that is adapted to the characteristics of the wireless medium and resource restrictions of mobile nodes. It provides secure access to the stored information for the operations of demanding applications such as multimedia and cooperative services. In [10], the paper studies the performance effects of caching derived from the of different wireless MAC layers, such as 802.11-based ad hoc networks and multi-interface-multichannel-based mesh networks. Then based on the results, it proposes an asymmetric approach to identify the best nodes to cache the data. In [11], the authors propose a protocol for wireless P2P resource sharing. The idea is to make use of locality by assigning peers, which are close in the physical network.

In [12], the authors propose a distributed algorithm for scheduling the multiple senders for multi-source transmission in mobile P2P networks. The proposed algorithm aims to maximize the data

rate and minimize the power consumption. In [13], the system takes advantage of node mobility by designating stable nodes as community coordinators for file searching. In [14], the authors propose a new concept of data file replication, which considers both node storage and meeting frequency. Then, a distributed data file replication protocol is proposed to implement the concept. However, these works [13-20] do not focus mainly on the data scheduling scheme, and they cannot achieve a timely delivery.

3. DELAY-SENSITIVE SEGMENT SCHEDULING ALGORITHM

In this section, we propose a delay-sensitive segment scheduling algorithm for real-time P2P streaming applications in MANETs. The lookup function is provided by P2P lookup protocols to collect the information of the segments cached on potential peers. Some P2P lookup protocols, such as Chord [15] and Pastry [16], should be extended for the proposed algorithm to enable the lookup result to cover multiple potential peers.

The proposed algorithm adopts a distinct transmission rate control strategy, in which the number of segment s transmitted into the network is decided by the playback-rate of the requested streaming service. The rate control strategy aims to efficiently use the MANET limited bandwidth by scheduling only an enough amount of segments for satisfying the playback-rate of the requested streaming service.

It schedules the segments with approaching play-deadlines first in order to satisfying the real-time requirement of the MoD multimedia file. Based on this consideration, the priority of a segment is computed as the inverse of the period from the current time to the playback deadline of this segment. DSSSA schedules the segments in the following sequence. First, it schedules the segments whose play-deadlines are approaching. These segments are called as urgent segments. Second, it schedules more segments by selecting those with higher priorities for satisfying the playback-rate of the requested streaming service. Third, it schedules the segments, which are only available at few peers, in order to disseminate segments quickly. These segments are defined as scare segments, if they are available only at not more than γ peers, where γ>1 is a predefined integer.

Since a segment may be cached at multiple peers in a P2P system, the proposed algorithm will schedule the segment to be transmitted from the peer with less bandwidth consumption to the network. In order to achieve the purpose, it selects the route with minimal total medium time to the network. The total medium time of a route is to sum the end-to-end delay of the route and the blocking time on all the neighbors of the forwarders along the route. In MANETs, when a host is transmitting packets, its neighbors are blocked since it shares the radio channel with its neighbors. A route with minimal total medium time can reduce the bandwidth consumption to the network.

When a new coming peer intends to request P2P streaming service, it first invokes a P2P lookup protocol for obtaining the values of the playback-rate of the requesting streaming service, the playback-rates of the segments, the playback deadlines of the segments, the information of the segments stored in the peers. Further, the delay-sensitive routing protocol of our previous work [17] for obtaining the values of the end-to-end delays and total medium time of the routes from the peers to the new coming peer.

The outlines of the proposed algorithm are as follows:

Input: (1) the playback-rate of the requesting streaming service,
(2)the playback-rates of the segments,
(3) the playback deadlines of the segments,
(4)the information of the segments stored in the peers,
(5)the end-to-end delays and total medium time of the routes from the peers to the new coming peer.

Output: the schedule of the un-scheduled segments.

Begin

Step 1. Set the priorities of the un-scheduled segments. The priority of a segment is computed as the inverse of the period from the current time to the playback deadline of the segment.

Step2. Determine an unscheduled segment *s* with the highest priority.

Step3. Determine a peer *p* with the minimal total medium time among the peers keeping segment *s*.

Step4. Schedule the segment *s* from the peer *p*.

Step 5. If there are the un-scheduled segments and the rate of transmitting the scheduled segments is smaller than the playback-rate, go to Step 2.

Step 6. If there is the un-scheduled segment *s* who is available only at not more than γ peers, determine a peer *p* with the minimal total medium time among the peers keeping segment *s* and schedule the segment *s* from the peer *p*.

Step 7. If there is the un-scheduled segment *s* who is available only at not more than γ peers, go to Step 6.

end

If a potential provider peers provides a timely segment transmission from, i.e., the time of receiving the requesting segment exceeds the playback deadline, it stops the segment requesting from the requesting peer. Once a schedule for satisfying the playback-rate of the requested streaming service cannot be made from the rest potential provider peers, peers execute the P2P lookup protocol and the delay-sensitive routing protocol again.

4. PERFORMANCE ANALYSIS

In this section, we illustrate the performance of our proposed algorithm by simulation examples. Simulations are implemented using the Network Simulator 2 package (ns-2, version 2.29) [18].In the simulations, our proposed algorithm uses Chord [15] as the P2P lookup protocol, and use the delay-sensitive routing protocol of our previous work [17] as the routing protocol. In the simulation environment, 50 hosts are randomly distributed over a 1000 m×1000 m area. The IEEE 802.11b is used as our MAC/PHY protocol, i.e., there are four available data rates 1, 2, 5.5, and 11 Mbps. Packets are sent using the un-slotted Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA). Each host has a First-In-First-Out transmission queue of no more than 64 packets at the MAC layer.

The simulations are performed by measuring the following three indices: admission rate, successful receiving rate, and buffer size. Successful receiving rate is the ratio of the number of the segments that are received before their playback deadlines to the number of the requesting segments. Admission rate is the ratio of the number of requested peers admitted to the number of peer requested. When the admission rate goes up, the network capacity increases. In the simulations of Figures 1-3, 50 peers are generated as a Possion process with an arrival rate of 5 requests per minute. Each peer is hosted on each of the 50 hosts. A special video stream is constructed and re-used for all simulation runs. Its stream content is 30 minutes of a movie with an accurate rate-control package to generate a stream with a constant bit-rate of 200 Kbps. Each peer is operated according to an On-Off model that simulates peer departing and joining. During

each 6 minutes time-slot, a peer is on or off with probabilities 0.9 and 0.1 respectively. Also, it could switch off at the beginning of a time-slot for the rest of the run at any time slot with probability 0.05. In order to simulate the end-game characteristics when peers rapidly depart the P2P overlay at the end of the video stream's transmission, each peer could switch-off with a probability of 0.5 in the last 5 minutes of the run.

Figure 1demonstratesthe admission rates of our proposed algorithm. The results show that our proposed algorithm is effective in using the limited bandwidth of MANETs. The result is derived from its rate control mechanism by only scheduling the enough number of segments to be transmitted into the network for satisfying the playback-rate of the requested media file. In the results of Figure 2, our proposed algorithm obtains small buffer size. Figure 3demonstrates that DSSSA is sensitive in providing timely segment transmission by the aid of our previous work [17] that can estimate the end-to-end delays and determine delay-sensitive routes.

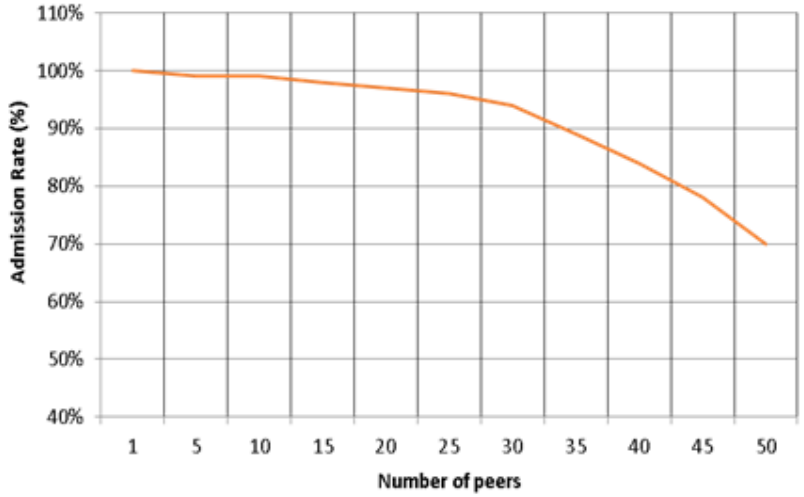


Figure 1. Admission rate.

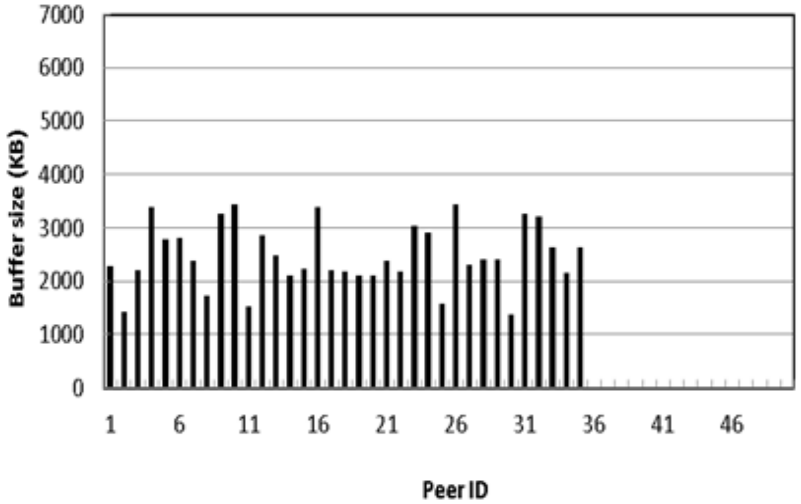


Figure 2. Buffer size.

The simulation environment in Figure 4 is the same as that adopted in Figure 1. But the difference is that the mobility of the 50 hosts is based on the random waypoint model [19], in which a host's

movement consists of a sequence of random length intervals, called mobility epochs. During each epoch, a host moves in a constant direction and at a constant speed. Their speed varies from 5 to 20 meters per second.

Figure4 compares the successful receiving rates of our proposed algorithm under the assumption of mobile hosts. Compared with the simulation results of Figure 1, our proposed algorithm is adaptive on relieving the mobility problem since it has slightly decreasing of successful receiving rate.

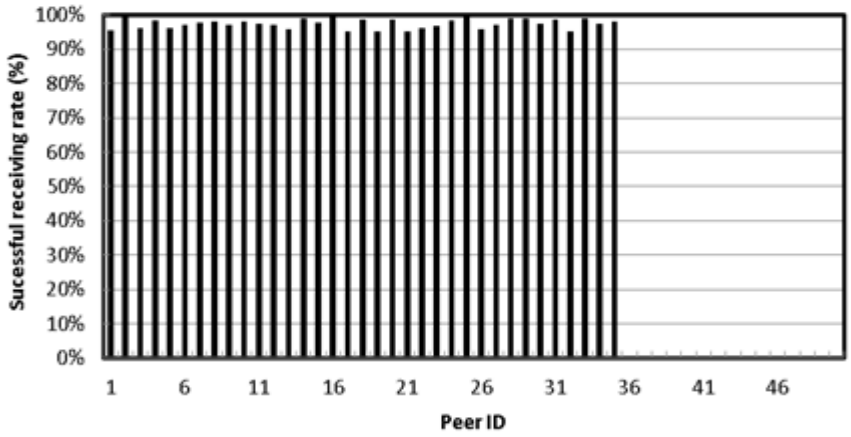


Figure 3. Successful receiving rates.

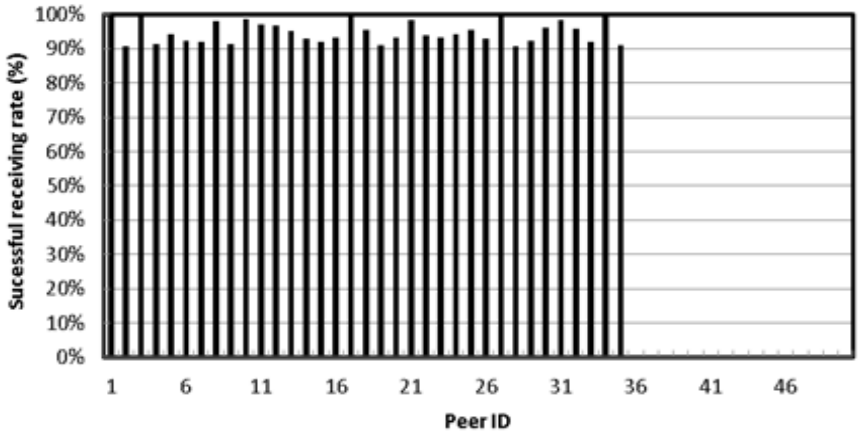


Figure 4. Successful receiving rates under mobile hosts.

5. CONCLUSIONS

In order to exploit wireless resource efficiently and provide timely P2P streaming services in MANETs, we proposed a delay-sensitive segment scheduling algorithm. Taking the playback-rate of the real-time streaming service into consideration, the proposed algorithm schedules the segments of the service evenly transmitted into a MANET. As a consequence of the consideration, the limited bandwidth in the network can be used efficiently.

ACKNOWLEDGEMENTS

This work was supported in part by the Fujian Provincial Key Laboratory of Cloud Computing and Internet-of-Thing Technology, China.

REFERENCES

[1] I. M. Moraes, M. E. M. Campista, L. H. M. K. Costa, O. C. M. B. Duarte, J. L. Duarte, D. G. Passos, C. V. N. de Albuquerque& M. G. Rubinstein, (2008) "On impact of user mobility on peer-to-peer video streaming," IEEE Wireless Communication, vol. 15, no. 6, pp. 54-62.

[2] X. F Liao, H. Jin& W. B. Jiang, (2008) "Moving P2P Live Streaming to Mobile and Ubiquitious Environment," IEEE. Computing and Informatics, vol. 27, no. 5, pp. 823-835.

[3] Z. J. Chen, C. Lin& X. G. Wei, (2009) "Enabling on-demand internet video streaming services to multi-terminal users in large scale," IEEE Transactions on Consumer Electronics, vol. 55, no. 4, pp. 1988-1996 .

[4] G. T. Xue, H. Shi, J. Y. You & W. S. Yao, (2003) "Distributed stable-group differentiated admission control algorithm in mobile peer-to-peer media streaming," Chinese Journal of Electronics, vol. 12, no. 4, pp. 517-521.

[5] M. Qin & R. Zimmermann, (2007) "Improving mobile adhoc streaming performance through adaptive layer selection with scalable video coding," Proc ACM Multimedia.

[6] M. Guo, M. H. Ammar & E. W. Zengura, (2005) "V3: A vehicle-to-vehicle live video streaming architecture," Pervasive and Mobile Computing, vol. 1, no. 4, pp. 404-424.

[7] H. Jenkac, T. Stockhammer & W. Xu, (2006) "Asynchronous and reliable on-demand media broadcast," IEEE Network, vol. 20, no. 2, pp. 14-20.

[8] M. F. Leung& S. H. G. Chan, (2007) "Broadcast-based peer-to-peer collaborative video streaming among mobiles," IEEE Transactions on Broadcasting, vol. 53, no. 1, pp. 350-361.

[9] D. F. Macedo, A. L. dos Santos& J. M. S. Nogueira, (2009) "A distributed information repository for autonomic context-aware MANETs," IEEE Transactions on Network and Service Management, vol. 6, no. 1, pp. 45-55.

[10] J. Zhao, P. Zhang, G. H. Cao& C. R. Das, (2010) "Cooperative caching in wireless p2p networks: design, implementation, and evaluation," IEEE Transactions on Parallel and Distributed Systems, vol. 21, no. 2, pp. 229-241.

[11] C. Canali, M. E. Renda, P. Santi & S. Burresi, (2010) "Enabling efficient peer-to-peer resource sharing in wireless mesh networks," IEEE Transactions on Mobile Computing, vol. 9, no. 3, pp. 333-347.

[12] P. Si, F.R. Yu, H. Ji & V.C.M. Leung, (2009) "Distributed sender scheduling for multimedia transmission in wireless mobile peer-to-peer networks," IEEE Transactions on Wireless Communications, vol. 8, no. 9, pp. 4594-4603.

[13] K. Chen, H. Shen& H. Zhang, (2014) "Leveraging social networks for P2P content-based file sharing in disconnected MANETs," IEEE Transactions on Mobile Computing, vol. 13, no. 2, pp. 235-249.

[14] K. Chen & H. Shen, (2015) "Maximizing P2P file access availability in mobile ad hoc networks though replication for efficient file sharing," IEEE Transactions on Computers, vol. 64, no. 4, 2015, pp. 1029-1042.

[15] I. Stoic, R. Morris, D. R. Karger, M. F. Kaashoek & H. Balakrishnan, (2003) “Chord: a scalable peer-to-peer lookup protocol for Internet,” IEEE Transactions on Networking, vol. 11, no. 9, pp. 17-32.

[16] A. Rowstron & P. Druschel, (2001) “Pastry: scalable, distributed object location and routing for large-scale peer-to-peer systems," Proc. 18th IFIP/ACM Conf. Dist. Sys. Platforms (Middleware 2001), Heidelberg, Germany, pp. 329-350.

[17] C. C. Hu, (2011) ”Delay-sensitive routing in multi-rate MANETs,” Journal of Parallel and Distributed Computing, vol. 71, pp. 53-61.

[18] Network Simulator (Version 2), http://www-mash.cs.berkeley.edu/ns/.

[19] C.Bettstetter, G. Resta & P. Santi, (2003) “The node distribution of the random waypoint mobility for wireless ad hoc,” IEEE Transactions on Mobile Computing, vol. 2, no. 3, pp. 257-269.

MRI AND CT IMAGE FUSION BASED STRUCTURE-PRESERVING FILTER

Qiaoqiao Li[1], Guoyue Chen[2], Xingguo Zhang[2], Kazuki Saruta[2] and Yuki Terata[2]

[1]Graduate School of Systems Science and Technology,
Akita Prefectural University, Akita, Japan
[2]Information and Computer Science,
Akita Prefectural University, Akita, Japan

ABSTRACT

Medical image fusion plays an important role in clinical application such as image-guided radiotherapy and surgery, and treatment planning. The main purpose of the medical image fusion is to fuse different multi-modal images, such as MRI and CT, into a single image. In this paper, a novel fusion method is proposed based on a fast structure-preserving filter for medical image MRI and CT of a brain. The fast structure preserving filter is a novel double weighted average image filter (SGF) which enables to smooth out high-contrast detail and textures while preserving major image structures very well. The workflow of the proposed method is as follows: first, the detail layers of two source images are obtained by using the structure-preserving filter. Second, compute the weights of each source image by calculating from the detail layer with the help of image statistics. Finally, fuse source images by weighted average using the computed weights. Experimental results show that the proposed method is superior to the existing medical image fusion method in terms of subjective evaluation and objective evaluation.

KEYWORDS

Multimodal image fusion, structure-preserving filter, weighted average.

1. INTRODUCTION

With the development of the computer science, there are many modalities of medical images to support more accurate clinical information to physicians for better medical analysis and diagnosis. Today many kinds of modalities of medical images are existing, such as computed tomography (CT), magnetic resonance angiography (MRA), magnetic resonance imaging (MRI), positron emission tomography (PET) and single photon emission tomography (SPECT) [1-3]. Different modality medical images can provide different perspectives on the human body, such as CT image can provide sense structures like bones and implants with less distortion, while the MRI image can provide normal and pathological soft tissue information [1-5]. Therefore, in order to fully diagnose the condition of patients, it is desired to fusing different modality medical images into a single image, called image fusion, such that all the information is available.

Image fusion can be divided into three levels: pixel levels, feature levels and decision levels [5-6]. Due to the advantage of pixel level method, such as containing the original measured quantities, easy implementation and computationally efficient, we focus the pixel level method in this paper.

Natarajan Meghanathan et al. (Eds) : SPPR, SCAI, CSIA, WiMoA, ICCSEA, InWeS, NECO, GridCom - 2018
pp. 127–135, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.81712

Pixel-level image fusion method can be divided into two categories: spatial domain algorithms and transform domain algorithms [8]. In the spatial domain, Calhoun et.al use the technology of independent component analysis (ICA) for their fusion method [9]. Patil et.al introduce the principal component analysis (PCA) technology in their fusion method [10]. Recently, structure-preserving smoothing filter technology is applied in the fusion method. For example, Zhan et.al [11] apply a fast filter to accelerate their fusion method. In transform domain method, Alfano et.al [12] and Vekkot [13] propose the fusion method based on wavelet and Das et.al use nonsubsampled contourlet transform (NSCT) in their fusion method [14] and so on. In addition to the fusion methods used wavelet and contourlet transform, many researches also have introduced the structure-preserving smoothing filter into their fusion methods. Such as, Li et.al [15], Bavirisetti et.al [3] and Zhan et.al [16] use the guide filter (GF) to obtain the fusion image, Kumar et.al [6] introduce the cross bilateral filter (CBF) into their fusion scheme.

Considering the edge preserving filter can extract effectively salient information from the source images, a new fast structure-preserving filter [17], a novel double weighted average image filter (SGF) based on the segment graph which is introduced into the modalities medical image fusion in this paper. In [17], Zhang et.al have proved that SGF can keep the major edges better than the GF and CBF. In this paper, a new method is proposed. First, use the SGF smooth the source image. Second, subtract the smooth image from the source image to obtain the detailed information. Third, use a weighted average method to fuse the source image. The weighted average based fusion method has been employed in [36] and the fusion results have shown good performance.

2. ALGORITHM

2.1 DOUBLE WEIGHTED AVERAGE IMAGE FILTER (SGF)

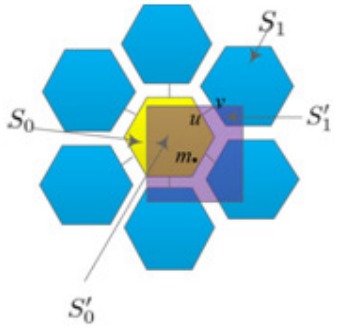


Figure 1. Filter kernel of structure-preserving filter (SGF)

Zhang et.al proposed a novel double weighted average image filter used the segment graph [17]. Because super pixel decomposition of a given image has been studied and the superpixel can run very fast in linear, they use the super pixel decomposition to construct the segment graph. A detailed introduction of the segment graph can be seen from Zhang et.al's literature.

Because the novel structure preserving structure based on the double weight, i.e. internal weight and external weight, we will introduce them in the next. Considering the tree distance which has the edge-aware property, the internal weight function w_1 can be defined by

$$w_1(m,n) = \exp(-\frac{D(m,n)}{\sigma}) \ , \tag{1.1}$$

where $D(m,n)$ represents the tree distance between pixels m and n. As σ controls the attenuation speed of $D(m,n)$, the w_1 is inversely proportional to the tree distance $D(m,n)$.

In order to describe external weight, a smoothing window w_n is introduced which is shown in Fig. 2 (purple square). As shown in Fig.2, several super pixel regions are denoted as $\{S_0, S_1, \mathrm{L}\ , S_k\}$ and the overlapped regions are represented by $\{S_0^j, S_1^j, \mathrm{L}\ , S_k^j\}$, namely $S_i^j = \mathrm{w}_n \cap S_i$, the external weight function w_2 can be defined by the area size ration S_i^j and S_i :

$$w_2(m,n) = \frac{\left|S_i^j\right|}{\left|S_i\right|} \ . \tag{1.2}$$

Once the double weights are obtained, the filter output of an input image *I* at pixel *n* can be given by

$$J_n = \frac{1}{K_n} \sum_{0 \le i \le k} w_2(n, S_i) \sum_{m \in S_i} w_1(n,m) I_m, \tag{1.3}$$

where K_n, S_i and J_n represent a normalizing term, super pixel region and filter output. w_1 and w_2 are the internal weight function and external weight function, respectively. The output J_n at the pixel n is the double weighted average of the intensity value I_m in a specific neighbour region $\Omega = \mathrm{U}_{0 \le i \le k} S_i (m \in S_0)$.

2.2. FOCUS RULE

In this paper, we adopt a weighted average method to fuse the images. The weighted average fusion rule is proposed by Shah et.al [18] and Kumar has used this fusion rule for their method [6].

Shah et.al compute the weight of wavelet coefficient [18], instead of it we compute the weight of the detail coefficient. The weight is computed in a window of size $w \times w$ around a detail coefficient $A_d(i,j)$ or $B_d(i,j)$ which is denoted as a matrix R. Let us treat each row of R as an observation and column as a variable, and then unbiased estimate $C_h^{x,y}$ of its covariance matrix [19] can be computed by

$$\mathrm{cov}\, ariance(X) = E((R - E(R))(R - E(R))^T) \tag{1.4}$$

$$C_h^{x,y} = \frac{\sum_{i=1}^{n} (r_i - \overline{r})(r_i - \overline{r})^T}{n-1} \tag{1.5}$$

where r_i is the i -th observation of the n -dimensional variables and $\overline{r_i}$ is the mean of observations. It can be observed that diagonal of $C_h^{x,y}$ is a variance vector of each column of the matrix R. Then compute eigenvalues of $C_h^{x,y}$, denoted by λ_H^j, and the number of eigenvalues depend on the size of it. Since the size of $C_h^{x,y}$ is $w \times w$, the number of Eigen values is w. The sum of these eigenvalues is directly proportional to the strength of horizontal edges and the sum can be named by $edgestrength_h$

$$edgestrength_h(\mathrm{x},\mathrm{y}) = \sum_{j=1}^{n} \lambda_H^j \ . \tag{1.6}$$

Take consideration of the vertical edges, an unbiased covariance estimate $C_v^{x,y}$ is computed under the condition of treating each column R as an observation and row as a variable, and then vertical edge strength can also obtained by summing the eigenvalues λ_V^j of it

$$edgestrength_v(x;\ y) = \sum_{j=1}^{n} \lambda_V^j \tag{1.7}$$

For a particular detail coefficient at the location (x, y), the weight is obtained by adding the horizontal edge strength and vertical edge strength

$$wd(x, y) = edgestrength_h(x, y) + edgestrength_v(x, y)\,. \tag{1.8}$$

Considering the Eq.2.6 and Eq.2.7, the weight can be rewritten by

$$wd(x,\ y) = \sum_{j=1}^{n} \lambda_H^j + \sum_{j=1}^{n} \lambda_V^j \ . \tag{1.9}$$

Then, the fused image can be obtained by

$$F(x, y) = \frac{I_1(x, y) * wd_1(x, y) + I_2(x, y) * wd_2(x, y)}{wd_1(x, y) + wd_2(x, y)}\,. \tag{1.10}$$

3. PROPOSED MULTI-FOCUS FUSION SCHEME

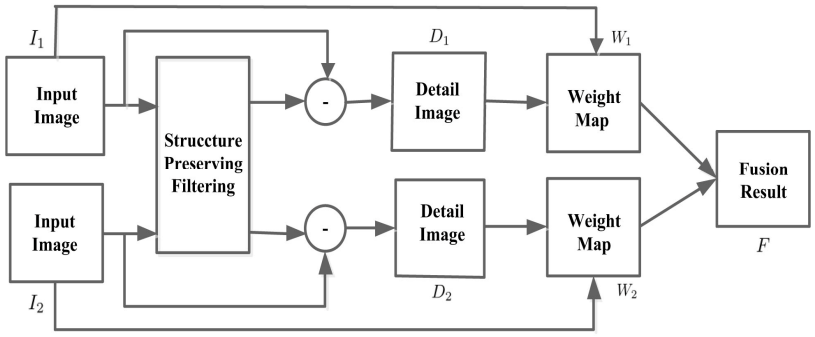


Figure 2. System diagram of the proposed fusion framework.

In this section, the proposed fusion method will be introduced in detail. Figure 2 shows the framework of the proposed fusion method. For two perfectly registered source images denoted by I_1 and I_2, the proposed algorithm consists of three main steps as shown in Figure 2.

Step 1: The two source images are first decomposed into approximation images (A_1 and A_2) and derail images (D_1 and D_2) by structure-preserving filter

$$A_k = SGF(I_k), \quad k = 1, 2. \tag{1.11}$$

And then, the detail images are computed by subtracting the approximation images from source images

$$D_k = I_k - A_k, \quad k = 1, 2. \tag{1.12}$$

Step 2: Use the detail images to computed the weight map of each image at the location (x, y)

$$W_k(x, y) = \sum_{j=1}^{n} l_H^j + \sum_{j=1}^{n} l_V^j, \quad k = 1, 2. \tag{1.13}$$

Step 3: Once the weight map is obtained, the fusion image can be computed as followings

$$F(x, y) = \frac{I_1(x, y) * W_1(x, y) + I_2(x, y) * W_2(x, y)}{W_1(x, y) + W_2(x, y)} \tag{1.14}$$

4. EXPERIMENT RESULTS

4.1. FUSION EVALUATION METRICS

In order to evaluate the fusion result performance of the proposed method, two objective image fusion performance metrics are adopted to evaluate performances of different fusion, i.e. structure-based metric $Q_w^{xy|f}$ [20] and normalized mutual information Q_{MI} [21].

4.1.1 STRUCTURAL SIMILARITY-BASED METRIC ($Q_w^{xy|f}$)

The structural similarity (SSIM) metric measures the corresponding regions in a reference source image A or B and the fusion image F with a sliding window **w** which can be defined by

$$SSIM(A, F|w) = \frac{(2\overline{w}_A\overline{w}_F + C_1)(2\delta_{w_A}\delta_{w_F} + C_2)(\delta_{w_A w_F} + C_3)}{(w_A^2 + w_F^2 + C_1)(2\delta_{w_A}\delta_{w_F} + C_2)}, \tag{1.15}$$

the detailed parameter settings of it can be seen from [22] [23]. Yang et.al [20] proposed a new metric based on SSIM which can be written by

$$Q_w^{ab|f} = \begin{cases} \lambda_w SSIM(A, F|w) + (1-\lambda_w) SSIM(B, F|w), & SSIM(A, B|w) \geq 0.75 \\ \max(SSIM(A, F|w), SSIM(B, F|w)), & SSIM(A, B|w) < 0.75 \end{cases} \tag{4.2}$$

where the weight λ_w is defined by

$$\lambda_w = \frac{s(\mathrm{A}|w)}{s(\mathrm{A}|w) + s(\mathrm{B}|w)}. \tag{4.3}$$

In implementation, $s(\mathrm{A}|w)$ and $s(\mathrm{B}|w)$ are the variance of images A and B with the window **w**.

4.1.2 NORMALIZED MUTUAL INFORMATION (Q_{MI})

Mutual information (MI) is a quantitative measure of the mutual dependence of two variables. And the mutual information for two discrete random variables U and V is defined by

$$MI(U,V) = \sum_{u \in U} \sum_{v \in V} p(u,v) \log \frac{p(u,v)}{p(u)\,p(v)} \quad , \tag{4.4}$$

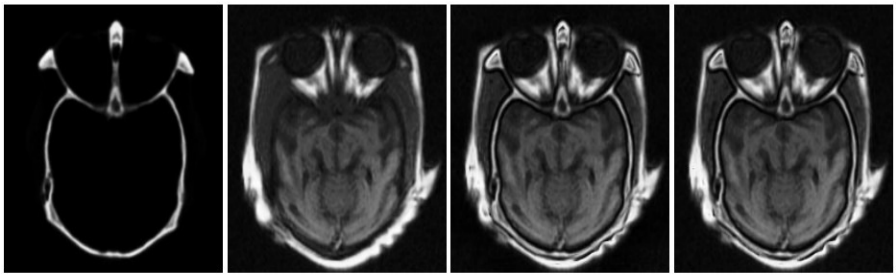
where $p(u,v)$ is the joint probability distribution function of U and V, and $p(u)$ and $p(v)$ represent the marginal probability distribution function of U and V, respectively. Based on the above definition, the normalized mutual information of the fusion image regarding to the source image A and B is computed as

$$Q_{MI} = 2\left[\frac{MI(A,F)}{H(A)+H(F)} + \frac{MI(B,F)}{H(B)+H(F)} \right],$$

where the $H(A)$, $H(B)$ and $H(F)$ are the marginal entropy of images A, B and fusion image F.

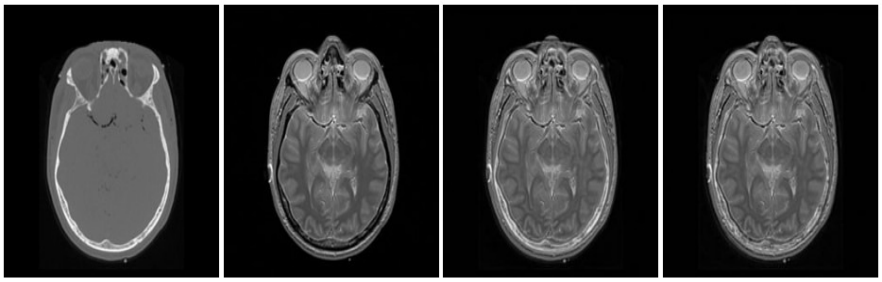
4.2 EXPERIMENT RESULT

Experiments are carried out on two pairs of CT and MRI modality medical images as shown in Figure 3 (a), (b) and Figure 4 (a), (b). The fusion image obtained by the proposed method is compared with the method proposed by Bacirisetti et.al [3]. Bacirisetti et.al use the guide filter to get the detail images and then compute the weight of them using the image statistics (GFS). The experiment results are shown in Figure 3 (c) and Figure 4 (c).



(a) (b) (c) (d)

Figure 3. Fusion result of image A



(a) (b) (c) (d)

Figure 4. Fusion result of image B

Figure 3 shows the fusion results of the test modalities images A. Figure 3 (c) shows the fusion result of GFS. Some ``halo" artifacts arise the bright region of the left edge of the fusion result.

Figure 3 (d) shows the fusion result of the proposed method. It can be seen that the proposed method can obtain satisfactory result with keeping more details, such as the fusion result is no ``halo" artifacts.

Figure 4 shows the fusion results of the test modality images B. The fusion result of method GFS and the proposed method are shown in Figure 4 (c) and (d), respectively. It can be seen from the Figure 4 (d) that the proposed method's fusion result is not well as the GFS's method in visual quality, but it is not bad.

The objective evaluation of the fused result for the test two pairs images are shown in Table 1. From the Table 1, the fusion result of the proposed method achieves the highest values almost all of the evaluation metrics. Considering the analysis of subjective evaluation and objective evaluation, the proposed algorithm obtains the better fusion result than the GFS method.

Table 1. Objective Performance

Image	Metric	GFS	Proposed
Image A	$Q_w^{xy\|f}$	0.8516	0.9170
	Q_{MI}	0.7047	0.8660
Image B	$Q_w^{xy\|f}$	0.8689	0.8868
	Q_{MI}	0.9277	0.9754

5. CONCLUSION

In this paper, we have proposed a new fusion method based on structure-preserving filter (SGF). A weighted average method is used as the fusion rule in the proposed method. The weights of the source images are computed from the detail images of them.

In order to demonstrate the effectiveness of the proposed algorithm, two pairs of MRI and CT images have been considered. As shown in the experimental results, the proposed method has obtained better performance than the method in terms of the both visual performance and objective metrics.

ACKNOWLEDGMENTS

This work has been supported by the China Scholarship Council for Ph.D. program.

REFERENCES

[1] S. Liu & T. Zhang & H. Li & J. Zhao & H. Li, (2015) "Medical image fusion based on nuclear norm minimization," International Journal of Imaging Systems and Technology, Vol.25, No.4, pp310–316.

[2] S. Liu, J. Zhao, and M. Shi, (2015) "Medical image fusion based on improved sum-modified-laplacian," International Journal of Imaging Systems and Technology, Vol.25, No.3, pp206–212.

[3] D. P. Bavirisetti, V. Kollu, X. Gang, and R. Dhuli, (2017) "Fusion of mri and ct images using guided image filter and image statistics," International Journal of Imaging Systems and Technology, Vol.27, No.3, pp227–237.

[4] A. Wong and W. Bishop, (2008) "Efficient least squares fusion of mri and ct images using a phase congruency model," Pattern Recognition Letters, Vol.29, No.3, pp173–180

[5] G. Bhatnagar, Q. J. Wu, and Z. Liu, (2015) "A new contrast based multimodal medical image fusion frame- work," Neuro computing 157, pp143–152.

[6] B. S. Kumar, (2015) "Image fusion based on pixel significance using cross bilateral filter," Signal, image and video processing, Vol.9, No.5, pp1193–1204.

[7] R. Redondo, F.Sroubek, S. Fischer, and G. Cristobal, (2009) "Multifocus image fusion using the log gabor transform and a multisize windows technique," Information Fusion, Vol.10, No.2, pp163–171.

[8] K. S. Tamilselvan and G. Murugesan, (2014) "Survey and analysis of various image fusion techniques for clinical ct and mri images," International Journal of Imaging Systems and Technology, Vol.24, No.2, pp193–202.

[9] V. D. Calhoun and T. Adali, (2009) "Feature-based fusion of medical imaging data," IEEE Transactions on Information Technology in Biomedicine, Vol.13, No.5, pp711–720.

[10] U. Patil and U. Mudengudi, (2011)"Image fusion using hierarchical pca.," in image Information Processing (ICIIP), 2011 International Conference on, pp1–6, IEEE.

[11] K. Zhan, Y. Xie, H. Wang, and Y. Min, (2017) "Fast filtering image fusion," Journal of Electronic Imaging, Vol.26, No.6, 063004.

[12] B. Alfano, M. Ciampi, and G. De Pietro, (2007) "A wavelet-based algorithm for multimodal medical image fusion," in International Conference on Semantic and Digital Media Technologies, pp117–120, Springer.

[13] S. Vekkot, (2010) "Wavelet based medical image fusion using filter masks," in FIRA RoboWorld Congress, pp298–305, Springer.

[14] S. Das and M. K. Kundu, (2012) "Nsct-based multimodal medical image fusion using pulse-coupled neural network and modified spatial frequency," Medical & biological engineering & computing, Vol.50, No.10, pp1105–1114.

[15] S. Li, X. Kang, and J. Hu, (2013) "Image fusion with guided filtering," IEEE Transactions on Image Processing, Vol.22, No.7, pp2864–2875.

[16] K. Zhan, J. Teng, Q. Li, J. Shi, et al., (2015) "A novel explicit multi-focus image fusion method," Journal of Information Hiding and Multimedia Signal Processing, Vol.6, No.3 , pp600–612.

[17] F. Zhang, L. Dai, S. Xiang, and X. Zhang, (2015) "Segment graph based image filtering: fast structure- preserving smoothing," in Proceedings of the IEEE International Conference on Computer Vision, pp361–369.

[18] P. Shah, S. N. Merchant, and U. B. Desai, (2011) "An efficient adaptive fusion scheme for multifocus images in wavelet domain using statistical properties of neighborhood," in Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on, pp1–7, IEEE.

[19] S. J. Devlin, R. Gnanadesikan, and J. R. Kettenring, (1975) "Robust estimation and outlier detection with correlation coefficients," Biometrika, Vol.62, No.3, pp531–545.

[20] C. Yang, J.-Q. Zhang, X.-R. Wang, and X. Liu, (2008) "A novel similarity based quality metric for image fusion," Information Fusion, Vol.9, No.2, pp156–160.

[21] M. Hossny, S. Nahavandi, and D. Creighton, (2008) "Comments on information measure for performance of image fusion," Electronics letters, Vol.44, No.18, pp1066–1067.

[22] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, (2004) "Image quality assessment: from error visibility to structural similarity," IEEE transactions on image processing, Vol.13, No.4, pp600–612.

[23] Z. Liu, E. Blasch, Z. Xue, J. Zhao, R. Laganiere, and W. Wu , (2012) "Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: a comparative study," IEEE transactions on pattern analysis and machine intelligence , Vol.34, No.1, pp94–109.

INTENTIONAL BLANK

ANTI-VIRUS TOOLS ANALYSIS USING DEEP WEB MALWARES

Igor Mishkovski[1], Sanja Šćepanović[2],
Miroslav Mirchev[1] and Sasho Gramatikov[1]

[1]University Ss. Cyril and Methodius, FCSE, Skopje, 1000, Macedonia
[2]Department of Computer Science, Aalto University, Finland

ABSTRACT

Knowledge about the strength of the anti-virus engines (i.e. tools) to detect malware files on the Deep web is important for people and companies to devise proper security polices and to choose the proper tool in order to be more secure. In this study, using malware file set crawled from the Deep web we detect similarities and possible groupings between plethora of anti-virus tools (AVTs) that exist on the market. Moreover, using graph theory, data science and visualization we find which of the existing AVTs has greater advantage in detecting malware over the other AVTs, in a sense that the AVT detects many unique. Finally, we propose a solution, for the given malware set, what is the best strategy for a company to defend against malwares if it uses a multi-scanning approach.

KEYWORDS

Malware, Community detection, Anti-virus engines, data science, multi-scanning approach.

1. INTRODUCTION

AntiVirus products are essential in every business deployment connected to the Internet. Nowadays, with the increase in the number and diversity of malware on the Web [1], there are more AntiVirus Tools (AVT) becoming available to protect users and/or companies from malware. However, the quarterly growth at around 12% for known unique malware samples, according the Intel Security Group's *McAfee Labs Threat Report: August 2015*, and the fact that some AntiVirus companies use the same or significantly similar AntiVirus engines leave us in some way vulnerable to the existing security threats.

Another factor that exposes even more users and companies to security threats is the Deep Web. The size of the indexed (surface) Web is currently estimated to 4.59 billion pages. At the same time, it is estimated that the non-indexed, Deep Web, is 400 or even 550 times larger [2] and rapidly expanding at a rate that cannot be quantified. The Deep Web besides offering to cyber-crime great business opportunity, hacking services, stolen credit cards and weapons, it also represents nest for malware. The hidden nature of Tor and other services means it is easy to host and hide malware controlling servers on the Deep Web. The malware from the Deep Web is not widely accessible and thus, these kind of files, coming from the Deep Web, are still not fully scanned for detection.

Thus, it is of crucial importance to everyone exposed on the Internet to know more details about the available AntiVirus Tools (AVT) on the market, their business and technical relations in terms

Natarajan Meghanathan et al. (Eds) : SPPR, SCAI, CSIA, WiMoA, ICCSEA, InWeS, NECO, GridCom - 2018
pp. 137–151, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.81713

of similarity and possible groupings. In addition to this, for a better protection companies could use a multi-scanning approach, for instance use multiple antivirus engines on email gateways in order to enable a faster reaction to the most recent security threats by drastically shortening the time required to obtain the latest virus definitions and wider detection scope. Since many of the engines have their own heuristics and detection methods, in this way companies can gain maximum protection for their email environment.

In this work, using graph analysis and visualization methods, on one hand we empirically infer detection engine similarity and the existing groupings and/or overlapping between them, while on the other hand we infer which AVTs differentiate from the other AVTs and a have greater advantage in detecting malware compared to others. Moreover, using the AVT responses to our malware file set we optimize the combination of AVTs in order to obtain maximum detection rate (i.e. coverage). We strongly believe that this approach can be used by companies who want to implement a multi-scanning approach on their email gateways. The analysis is done on a malware file set provided by F-Secure and the AVTs responses on this file set obtained using the Virus Total API.

Researchers have undertaken evaluating and/or comparing the existing AVTs, some of them using malware samples and the VirusTotal service [3, 4]. We stress that in this work we are not trying to evaluate or compare the existing AVTs. Instead, we present results that undoubtedly show how our analysis can identify if some AVTs either use the same detection engine or quite similar engines between themselves and/or grouping between them. With simple graph analysis we can easily identify which AVT have a greater advantage, i.e. are unique compared to the others. Both results, with similarity and advantage, contribute to the multi-scanning approach in choosing the appropriate AVTs for a given price. We present this problem as a Mixed Integer Linear Programming (MILP) optimization problem and give an empirical solution. The solution shows that if a multi-scanning approach is to be implemented by a company, then the grouping according to the similarity and the advantage matters, besides the detection rate.

The work is organized as follows. In Section 2 we present some related work and then in Section 3 we present what are the novelties and the main contribution of our study. The dataset used for the study is described in Section 4. The results concerning similarity and communities between different AVTs are shown in Section 5, whereas Section 6 is dedicated for the uniqueness and coverage (multi-scanning approach). Section 7 concludes this work.

2. RELATED WORK

The analysis of decision from several Anti-Virus Tools has been addressed for several purposes over the last decade and mainly since the apparition of the VirusTotal service [4]. Submitting a set of known malicious files and performing quantitative comparison to deduce the best/worst AVT was the first purpose. Malware samples collected from Honeypot were submitted to VirusTotal to infer good and bad detection performance in [5]. The authors also explore if the combination of several AVT can improve protection and showed that AVT diversity and a combination of AVTs indeed improve detection without being able to reach 100% though. Similar empirical analysis using honeypot data [6] brought the same conclusions that diversity improves protection.

Our analysis brought similar conclusions while in contrast to previous work, the scale of the data analysed was orders of magnitude larger and considered files coming from the Deep Web, showing likely more diversity than honeypot data.

Previous work showed that detection performance comparison from different AVTs using VirusTotal is irrelevant due to the fast evolution of malware and AVT decision over time [7]. When it comes to the approaches taken to evaluate and compare the AVTs based on published malware samples, it is shown that creating a representative sample is a difficult task, especially nowadays since new malware samples are created on a daily basis [7]. In addition to that, malware creators are also finding ways to obfuscate existing malware with different type of techniques (such as bytecode conversion) to avoid signature-based detection. Hence, AVTs need to adapt to this type of malware detection (research suggests, for instance, using Opcode-sequences to detect malware [8]). In [9], findings on the stress test of AVTs with respect to such slight malware modifications is discussed.

Different AVT present inconsistency in labelling a given file as malicious or not and this label evolves over time. There is even more inconsistency between vendors in the correct identification of a malware family for a file while using different naming [10]. Mohaisen et al. [3] investigated inconsistency in malware family labelling of malicious files from different AVT and questioned the relevancy of using AV labels to build malware ground truth unless several tools are combined.

Typical method to build malware ground truth is to submit several unlabelled files to a set of anti-virus tools and consider the files malicious if at least "*k out of n*" AVTs detect it as a malware [1, 9]. Other approaches [11] have proposed to use anti-virus label decision over a set of files in a generative Bayesian model to improve ground truth composition.

In recent studies [6, 7, 13, 14] it is shown that the results have also temporal scale, i.e. AV regression exists, in a way that a given AVT can declare one file as a malware in a given instance of time, but later fail to recognize the file as a malware.

3. ANTI-VIRUS TOOLS STUDY

This paper presents a comparative analysis of several Anti-Virus Tools (AVT) based on a set of files coming from the Deep Web. In this study, we use a large dataset produced by crawling Web hosts through DNS brute force, hence containing potential malware files both from the Surface and the Deep Web. The resulting dataset consists of 1.64 Million files which were subjected to the VirusTotal API in order to get the decision from the plethora of AVTs on the maliciousness of these files. This work does not present a comparative performance analysis. It has been shown that the labelling of a given file can evolve overtime and performances per AVT for a given set of files are only valid at a given time [12]. Moreover, VirusTotal implements the command line interface of AVTs which is different from the desktop version that can implement more detection capabilities such as signature matching that could be bypassed in VirusTotal [7]. This could lead to an apparent performance degradation for a given AVT that is not actually true. Hence, the comparison of the detection rate against a given set of files cannot be performed using the VirusTotal interface and is out of the scope of this paper, which focuses on inferring AVT detection engine similarities and complementarity.

Given a set of files we seek to reveal several characteristics of AVT detection engine including:

- **Similarity**: The common detection capabilities two different AVTs present. Analysing the set of files detected by different AVTs we seek to infer the similarity in their detection engines operation. This analysis can infer as well *communities* of AVT having similar detection capabilities with community leaders presenting common characteristics from many community members. This can highlight the use of several third party engines in a single product.

- **Coverage**: Given a set of pieces of malware, infer which combination of AVTs can be used to optimize the protection against the largest number of malicious program in a multi-scanning approach. This involves analysing the complementarity of different AVTs detection regarding a given set of files in order to combine AVTs presenting different capabilities.

- **Advantage**: present the capability for one AVT to detect malicious files that other AVTs are not able to detect presenting its advantage over other tools.

This analysis displays the collaboration that different Anti-Virus companies have in developing their products through showing similarities in their detection capabilities. It denotes as well the use of a given AV engine in different tools. The competitive advantage that some products may have, by developing their own techniques, differentiates them for the competitors and underlines their usefulness in a multi-scanning approach. According to a recent survey [13] three main defence mechanisms against Web malware are presented: signature-based detection, code analysis of both client and server-side Web applications, and reputation-based URL blacklists. These defence mechanisms are differently used by different AVTs and thus, big corporations sometimes use a multi-scanning approach in order to protect their assets.

4. DATASET DESCRIPTION

The file set we use for the *similarity*, *coverage* and *advantage* study of existing Anti-Virus Tools (AVT) is crawled from the Deep and Surface Web from the company F-Secure, and consists of L = 1.64M files for each of which we have the file itself, its URI, its SHA1 hash value as a unique identifier. In this set, which we will refer as F-*Secure set*, there are $L = 990$ files, that were examined by F-Secure in details and were labelled as malicious files. We call this subset a *ground-truth set*. The complete *F-Secure set* is collected from 19 June 2015 till 12 October 2015.
In order to tackle the challenges described in details in Section 3, we have used the VirusTotal API [14], which is currently the largest freely available AVT service aimed to provide the users with results from different engines. The service enables the users to upload a file (or its unique hash) for a scan with a number of engines/tools supported by the service. As a final result, the user receives classification of the file as a malware or not by each of the AVTs, together with their own malware type label if the file is marked to be malicious. Thus, we have scanned both file sets (*F-secure* and *ground-truth*) using the Virus Total API where as an input we used the file's SHA1 value. We then, processed the JSON output from the Virus Total API obtaining the following additional information for each SHA1 value (i.e. file): $[AVT_1, Descr_1]$, $[AVT_2, Descr_2]$, ..., $[AVT_k, Descr_k]$, where AVT_i is the name of the AVT that labelled the file as malware, $Descr_i$ is the description of the type of the malware as reported by AVT_i (one example is *Win32: Trojan.Badur*, though there is not standardization between big anti-virus companies), and k is the number of AVTs that reported the file as a malware (some of which are: McAfee, Sophos, GData, VIPRE, Fortinet, Avast, Comodo, Symantec, ESET-NOD32, F-Secure, etc.). From the 1.64M files in the *F-Secure set* only 24.176 files were declared as a potential malware by at least two of the AVTs ($k \geq 2$). We call this set the *similarity set*. The *similarity set* is later used in Section 5 for the similarity and community analysis of the AVTs. The labelling using the VirusTotal API is done only once in May 2016.

However, in the *similarity set* there might be lot of files, which were erroneously declared as malware. The labelling of a file by an AVT as malign or benign evolves overtime. Benign files can further be declared as malicious because they belong to an unknown emerging malware family for which AVTs do not have any signature yet [7]. In contrast, benign files can be wrongly classified as malware (false positives) due to an overly broad detection signature or algorithm used in an anti-virus product. After a short period of time, vendors can be notified of the

mislabelling to correct the error or add an exception. This is likely to happen for newly developed program for instance. On the other hand, the *ground-truth set* contains only small fraction of the existing malware and thus, might impose the problem of under-sampling, leading to higher number of false negative errors. Thus, in order to tackle better the false positives and false negative errors we have chosen the threshold for the detection rate to be $k \geq 5$ as in [1], i.e., 5 or more AVTs must label a file as a potential malware. By thresholding the *F-secure set* we obtain new *malware set*, having $L = 10.745$ potential malware files, which will be used in the later analysis for AVT coverage and advantage over other AVTs, see Section 6. The cumulative distribution function CDF for the AVT detection rate on the *malware set* is shown in Fig. 1. We see that each file is detected by an average of 15.3 AVTs, with a median of 14 AVTs and standard deviation of 9.43 AVTs.

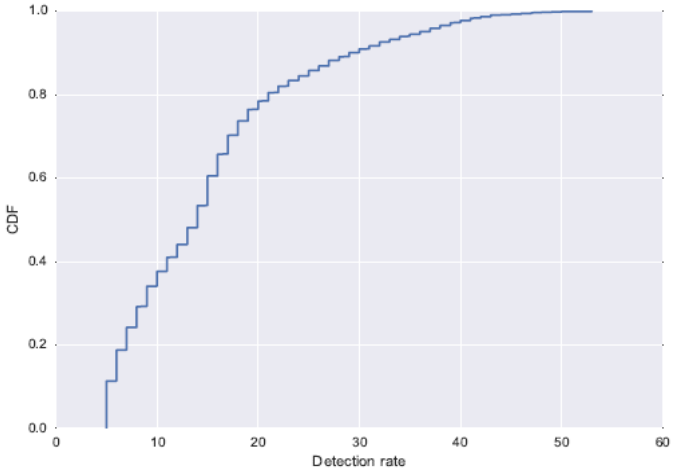


Figure 1. AVT detection CDF for the *malware set*.

5. AVTS SIMILARITY AND COMMUNITIES

Based on the similarity set described in Section 4 we measure the similarity between different AVTs and find existing grouping or communities that share similar decision regarding a given piece of malware. Thus, we first construct the similarity network $G^1 = (V, E, W^1)$ in order to characterize the similarity between different AVTs based on the shared files they label as malware. In order to get relevant results, we discarded from the analysis AVTs that detected less than 0.5% of the files from the *similarity set* i.e. less than 120 files. The node set V consists of the 61 AVTs that meet this condition, whereas the undirected edges set E contains the links between the AVTs that have labelled at least one common malicious file, with an edge weight $w_{ij}^1 \in W^1$ being defined as the Jaccard index between the sets of malware files detected by the two AVTs i and j. Next, we define the similarity between V_i and V_j as the co-occurrence strength. Let us assume that F_i and F_j denote two sets of files, labelled as malware by V_i and V_j, then we can define the Jaccard similarity measure (index) as a co-occurrence strength as follows.

$$sim(V_i, V_i) = \frac{|F_i \cap F_j|}{|F_i \cup F_j|} = w_{ij}^1 = w_{ji}^1, \quad (1)$$

where $|F|$ indicates the size of the set F. The value of w_{ij}^1 is between 0 and 1 (where "0" indicates no co-occurrence relationship between two AVTs and "1" indicates a full co-occurrence).

5.1. AVTS SIMILARITY RESULTS

The visualization plot of the adjacency matrix of the similarity network for the malware set is shown in Fig. 2. The similarity between each AVT is depicted as a square with darker colour. The results show high malware detection similarity between certain AVTs.

Some noticeable similarities are observed for McAfee and McAfee-GW-Edition with a similarity $w_{ij}^{1} = 0.78$. The same observation holds for K7AntiVirus and K7GW with $w_{ij}^{1} = 0.87$. This high similarity is to be expected between different tools coming from same vendors i.e. McAfee and K7 Computing. Yet, we see that different versions of tools i.e. standard and gateway editions, have different capabilities and that AV vendors do not use the same technologies in different products to maximize their detection capabilities, but rather propose tailored solutions for different applications. Gateway edition of these products are company solutions while other are basic customer version. Company solutions my implement more refined and customizable engine that explain this small dissimilarity.

While having comparatively high similarity score $w_{ij}^{1} = 0.71$, VIPRE and BluePex AvWare are developed by different vendors. After making searches we found out that BluePex AvWare actually uses VIPRE engine for malware detection explaining the high similarity in detected files. The conclusion is that detection engines integrated in third party solutions seem to be different than the one integrated in homemade product explaining a still significant dissimilarity (0.28) between these two tools. Observing the VIPRE line in Fig. 2 we see that it has quite high similarity with many AV tools e.g. Sophos, McAfee and Comodo, suggesting that this engine may be used in many other tools.

One AVT group showing high similarity is BitDefender, F-Secure, Emsisoft, MicroWorld-eScan and Ad-Aware with a similarity $w_{ij}^{1} > 0.6$ between these AVTs. Ad-Aware, F-Secure, Emsisoft and MicroWorld-eScan actually use BitDefender's detection engine along with other in-house detection solution, which explains the high similarity and small differences between all these tools. Globally BitDefender engine is largely used in several AVTs. G-data while embedding as well BitDefender engines shows less similarity than previously cited tools ($w_{ij}^{1} = 0.53$) suggesting that their in-house detection solution is more prominent than in other tools.

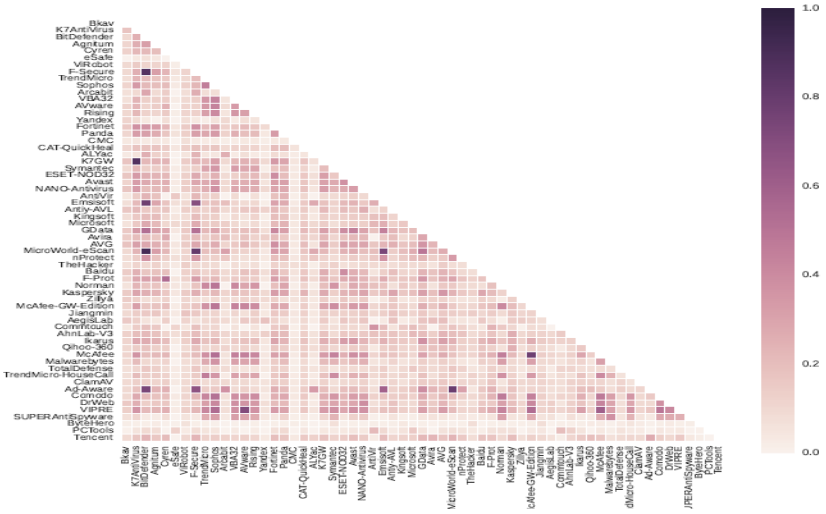


Figure 2. AVT's similarity for the *malware set*.

On the other hand, it is visible that some AVTs must have quite unique detection engine showing low similarity with any other tool with very light colour line in Fig. 2. Some examples are ByteHero with 1.611 files detected and $w_{ij}^1 < 0{:}07$ with any other tool, CMC with 1.439 detected files and $w_{ij}^1 < 0.08$, or Yandex with 639 files and $w_{ij}^1 < 0.19$. ByteHero is a self-developed unknown virus detection software that does not include virus database explaining their uniqueness in detection. Similarly, CMC anti-virus uses its own detection engine. Yandex anti-virus relies partly on Sophos for signature based detection ($w_{ij}^1 = 0.07$). However, our results seem to show that their proprietary anti-virus technology based on behavioural approach is prominent in their product.

5.2. AVTS COMMUNITIES

In this Subsection we detect structural communities, groups and/or modules in the AVT set using modularity-based community-detection algorithm [15]. The structural communities translate into groups of AVTs, which react in a similar manner to a certain malware. However, for a complete functional definition of the detected structural communities [19, 20] we have to know more details about the AVTs, including having an expert knowledge, and the AVTs response to different type of malware for different type of platforms. We underline, that this type of analysis is not part of this work, due to the restricted dataset and the fact that there is no existing effort between the AV companies to have a standardized malware labelling [3], thus, this approach may be used for future analysis.

The modularity-based community-detection algorithm is a simple heuristic method, which extracts community structures in networks, based on modularity optimization. The modularity Q is actually a scalar value (between -1 and 1), which measures the links density inside communities as compared to links between communities and is calculated as:

$$Q = \frac{1}{2m}\sum_{i,j}\left[W^1 - \frac{k_i k_j}{2m}\delta(c_i, c_j)\right], \quad (2)$$

where $k_i = \sum_j W^1$ is is the sum of the weighted degree of node i, c_i is the community to which the node i is assigned, the δ-function $\delta(u,v)$ is 1 if $u = v$, and 0 otherwise, and $m = \frac{1}{2}\sum_{ij} W^1$. In this work in order to find the optimal partitioning, i.e. optimize Q, we use the algorithm presented in [15].

In Fig. 3 we show that the algorithm partition the similarity network in 3 communities, with a highest modularity score of $Q = 0.12$, where each community is represented by a different colour, the size of the node (the AVT) is inversely proportional to its weighted degree ($k_i = \sum_j W^1$), and the width of the edge is proportional to the value w_{ij}^1 (see Eq. 1).

The biggest community, the one with the largest number of AVTs is the violet community, where some of the AVT with a biggest detection rate are Ikarus, ESET-NOD32, K7AntiVirus, K7GW, The Hacker and Baidu. Strong similarity in this community exists between K7GW and K7AntiVirus. The AVTs with the lowest similarity scores (i.e. lowest values for k_i) are ByteHero, CMC, Yandex and TheHacker.

In the orange community some of the AVTs with a highest detection rate are GData, F-Prot, Fortinet, Panda, Agnitium and F-Secure. Strong similarity ties exist between F-Secure, BitDefender, Emsisoft, MicroWorld-eScan and Ad-Aware as previously seen in Fig. 2 as well. BitDefender is used in several AVTs, thus its detected files are a subset of many AVT detected files presenting a smaller node with strong links to many other nodes. The AVTs with the lowest similarity scores are PCTools, eSafe, Commtouch and ALYac. Another remark is that in this

community there are old AVTs that are not maintained anymore. Symantec PCTools line was retired in 2013, VirusBuster was similarly closed in 2012. CommTouch anti-virus company became Cyren in 2014 and eSafe is not a product of Gemalto (previously SafeNet) anymore.

The rest of the AVTs are in the community with the highest detection rate, i.e. the green one. Here the leaders are Symantec, TrendMicro-HouseCall, McAfee, McAfee-GW-Edition, Rising and DrWeb. Strong similarity ties exist between McAfee and McAfee-GW-Edition (w_{ij}^{1} = 0.78), AVware and VIPRE (w_{ij}^{1} = 0.71), McAfee and VIPRE (w_{ij}^{1} = 0.60), Sophos and VIPRE (w_{ij}^{1} = 0.54). The AVTs with the lowest similarity scores are SUPERAntiSpyware, Malwarebytes, TrendMicro-HouseCall and Symantec.

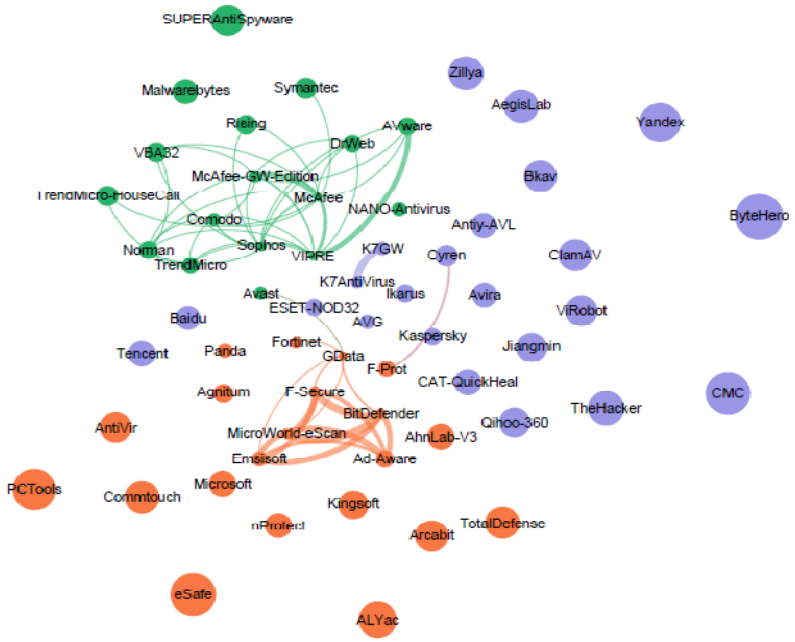


Figure 3. AVTs communities on the *similarity network* based on modularity-based community-detection algorithm. Edges that have Jaccard index below 0.4 are not shown.

6. AVTS COVERAGE AND ADVANTAGE

In order to analyse the AVTs coverage and the advantage of a given AVT compared to the others we define another measure as follows

$$c(V_i, V_j) = w_{i,j}^{2} = \begin{cases} |F_i|, & if\ i = j \\ |F_j \backslash F_i|, & otherwise, \end{cases} \tag{3}$$

when $i \neq j$, higher value of w_{ij}^{2} means that AVT j did found many diverse files as malware, compared to AVT i, i.e. this value defines the AVT advantage, whereas the self-loop weight w_{ii}^{2} shows how many files were found in the file set F_i, i.e. it defines the detection rate.

Now, let us construct a second network, which we call *coverage network* $G^2 = (V, E, W^2)$ in order to characterize the coverage and advantage of different AVTs based on the shared malware. Again, the node set V consists of AVTs that were reported by Virus Total and labelled at least 0.5% from the files in the malware set as malicious (N = 61), whereas the directed edges set E contains the links between the AVTs that have labelled at least one common malicious file with an edge weight $w_{ij}^{2} \in W^{2}$, and self-loops with a weight $w_{ii}^{2} \in W^{2}$.

In Fig. 4 we visualize the *coverage network* for the *malware set*, where the size of the nodes is the in-degree and the colour represents the out-degree. The bigger the node the more unique is its detected malware file set. In a similar way the red colour means lower out-degree, whereas blue means high value for the out-degree. Thus, the AVTs core is actually consisted of big red nodes represented in Fig. 4. Moreover, the colour of the edge represents the direction, i.e. the source AVT, and the width is the value of w_{ij}^2 (for e clearer visualization only the weights above 5.000 are shown). For instance, one can notice that McAfee has a lot of thick blue edges, i.e. incoming edges, which means that it has a great advantage over many AVTs.

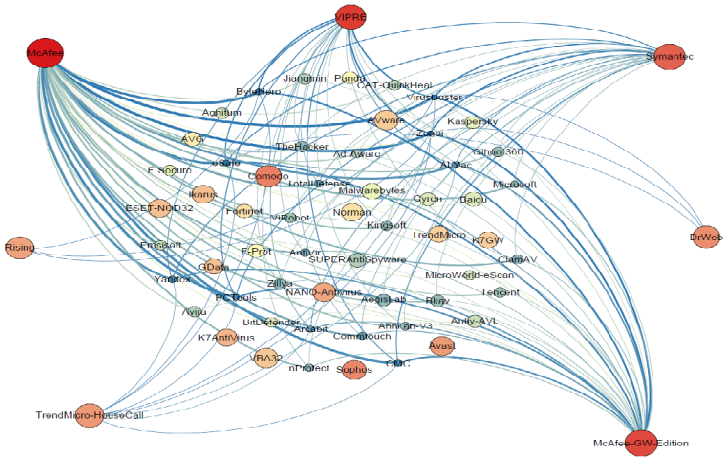


Figure 4. Coverage network for the malware set.

6.1. AVTS COVERAGE RESULTS

Without going in too much details, we observe the detection rate (w_{ii}^2) for the malware set and we find out that the AVTs showing "best" detection rate across all 10.745 files is McAfee, followed by McAfee-GW-Edition, VIPRE, Symantec, TrendMicro-HouseCall, DrWeb, Rising, Comodo, Sophos, etc. (see Fig. 5).

However, these results should not be taken too strict because if we increase the threshold from $k \geq 5$ to $k \geq 30$ then the "best" AVTs are GData and VIPRE, followed by McAfee, Sophos, Avast, Comodo, etc. (the plot is not shown). When the majority of the AVTs "votes" ($k \geq 30$) that a given file is a malware, there is no obvious winner among AVTs, though best results show GData, VIPRE and McAfee. The discrepancies in the results for different thresholds, bring us to one possible conclusion that some of the AVTs might report too many "false positives", i.e. they have a high malware detection rate when the rest of the AVTs disagree, or maybe they have a unique AV engine compared to the other AVTs.

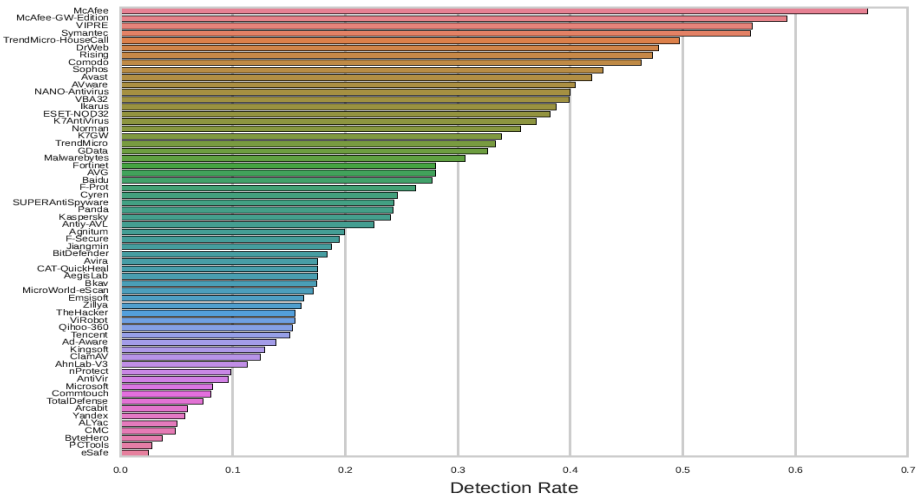


Figure 5 AVE Detection rate for the malware set.

The disagreement between AVTs comes from not having a common definition of what constitutes a malware [16]. For instance, adware can be considered as unwanted software or not by different AV products. As described in Section 4 as well, the labelling of a given file can evolve overtime and performances per AV for a given set of files are only valid at a given time. Finally, Virus Total implements the command line interface of AVTs, which is different from the desktop version that can implement more detection capabilities such as signature matching that could be bypassed in Virus Total [7]. This could lead to apparent performance degradation for a given AV program. Hence, a comparison of AVT detection rate against a given set of files cannot be performed using the VirusTotal interface and is out of the scope of this paper.

Instead, in the following using the *malware set* we focus more on optimizing the protection against malwares in a multi-scanning approach, i.e. ***find an optimal AVT set M, which will have the best malware detection coverage for a given price P***. This problem can be represented as a Mixed Integer Linear Programming (MILP) optimization problem, as following.

$$\begin{aligned} &\max \cup_{i=1}^{|M|} |F_i| \\ &s.t. \sum_{i=1}^{|M|} cost_i \leq P \end{aligned} \tag{4}$$

where $|M|$ is the number of AVTs in the optimal set M, $cost_i$ is the cost needed to buy AVT i and P are the available resources.

Using Eq.4 we show which AVTs to choose in order to have the highest coverage of the detected malware under given price constraint P. Due to the unknown price of the AV software we set **cost = $\mathbf{1}^T$** in Eq. 4. The best malware coverage, both for the *malware set* and the *ground-truth set*, as a function of the number of AVTs is shown in Fig. 6. The coverage follows logarithmic increase as a function of the number of AVTs. For instance, if a company would like to cover 95% of the labelled malware from the *ground-truth set* it would need four AVTs, and six for the *malware set*. In Tables 1 and 2 we give the names of the AVTs and the exact coverage obtained with them. In Table 1 we show the best coverage for the *malware set* for a given resource constraint P, where $P \in [1, 10]$. The best coverage when choosing 3 (three) AVTs is obtained by McAfee, ESET-NOD32 and Trend Micro-House Call with a total coverage of 87.6%.

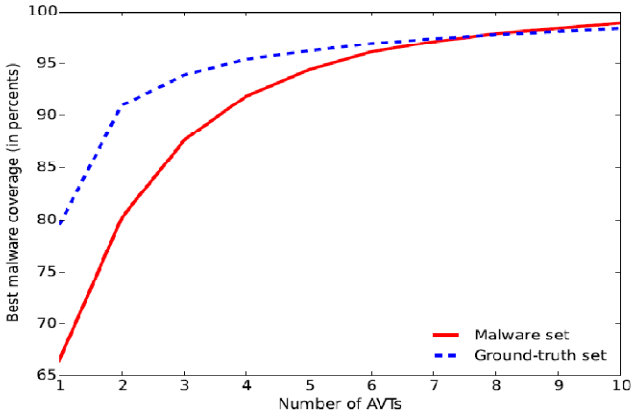


Figure 6 Maximizing malware coverage.

If we map the AVTs that provide best coverage to the community they belong, it is obvious that the best choice is to mix the AVT to be either from the orange or violet community shown in Fig. 3 and the last column in Table 1.

Table 1. Best coverage for a *malware set*

P	**AVT**	**Coverage (%)**	**Community (Color)**
1	McAfee	66.5	green
2	+ ESET-NOD32	80.2	violet
3	+ TrendMicro-HouseCall	87.6	green
4	+ Ikarus	91.9	violet
5	+ NANO-Antivirus	94.4	green
6	+ TheHacker	96.1	violet
7	+ Symantec	97.1	green
8	+ BKAV	97.9	violet
9	+ Antiy-AVL	98.4	violet
10	+ VIPRE	98.9	green

Finally, in Table 2 we show the best coverage for the *ground-truth set* for a given resource constraint *P*. The best coverage when choosing 3 (three) AVTs is obtained by McAfee, F-Secure and Ikarus with a total coverage of 93.9%. However, we must mention that these results are biased towards F-Secure because they have evaluated the *ground-truth set* of malwares.

Table 2. Best coverage for the *ground-truth set.*

P	**AVT**	**Coverage (%)**
1	McAfee	79.5
2	F-Secure + Ikarus	91.0
3	McAfee + F-Secure + Ikarus	93.9
4	+ Cyren	95.4
5	+ Symantec	96.2
6	+ Zillya	96.9
7	+ SUPERAntiSpyware	97.4
8	+ AegisLab	97.8
9	+ CAT-QuickHeal	98.1
10	+ Rising	98.4

6.2. AVTS ADVANTAGE RESULTS

Using the *coverage network* we can define the *advantage* of a given AVT by implementing a random walk on G^2. The directed network with self-loops G^2 consists of information about the *coverage* of AVT i, represented by the weight w_{ii}^2 showing how large is the covered file set F_i, and on the other hand we have the uniqueness of AVT i, compared to the other AVTs, represented by the weight w_{ji}^2. Thus, by running random walk on G^2 we should obtain information about the advantage of the AVT i which depends i) on its coverage (w_{ii}^2), i.e., how many files were detected by the AVT as a malware and ii) recursively on the weights of the incoming links from other AVTs $j \neq i$, i.e. how much unique is the AVT i.

Before presenting the results for the *malware set* and some results for the *ground-truth set*, let us show a simple scenario for the *coverage network*. Let us assume that there are 4 AVTs (x,y,w,z), i.e. $N = 4$, and there are $L = 13$ malware files, named {1, 2,… , 13}, and the following possible scenario. AVT x has detected $F_x = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ as malware, AVT y has detected $F_y = \{1, 3, 11\}$ as malware, AVT z has detected $F_z = \{1, 3, 7, 9, 12, 13\}$ as malware, and AVT w has detected $F_w = \{11, 12, 13\}$ as malware then using Eq. 3 we obtain

$$W^2 = \begin{bmatrix} 10 & 1 & 2 & 3 \\ 8 & 3 & 4 & 2 \\ 6 & 1 & 6 & 1 \\ 10 & 2 & 4 & 3 \end{bmatrix}. \tag{5}$$

In the case shown in Eq. 5 the stationary distribution vector equals to $\pi = [0.56, 0.08, 0.21, 0.15]$, showing that the AVT with the biggest advantage is x, whereas the least advantageous is y, even though it has the same weight self-loop as w, which is due to the fact that w has detected 3 new files that were not in F_x.

Using the weighted adjacency matrix $\mathbf{W}^2$ of the *coverage network*, and the random walk process we obtain different results for the AVT's advantage for the *malware set* and the *ground-truth set*. For instance, for the *malware set* the most advantageous AVT is McAfee, followed by McAfee-GW-Edition, Symantec, TrendMicro-HouseCall, Rising and DrWeb (see Fig. 7). For the *ground-truth set* the most advantageous AVT is again McAfee, this time followed by GData, Ikarus, F-Secure, VIPRE and Avast (see Fig. 8).

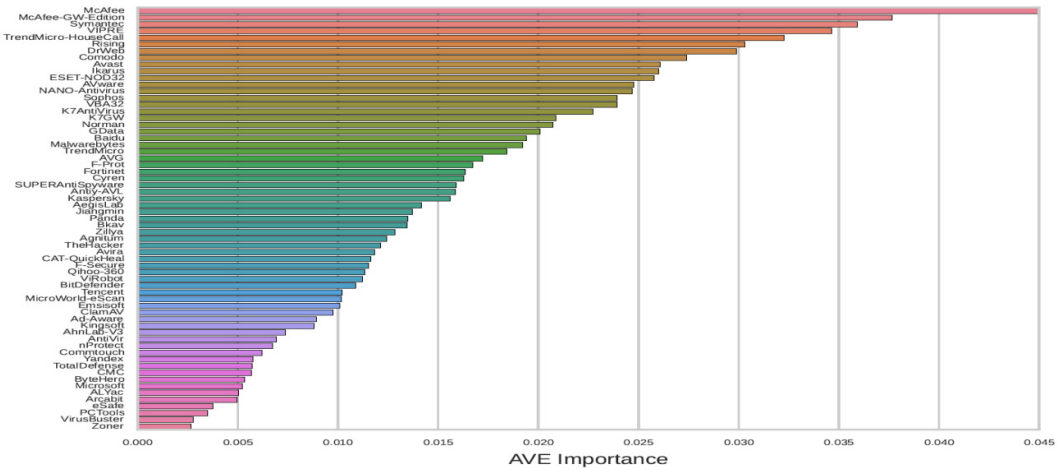


Figure 7. AVT advantage for the *malware set*.

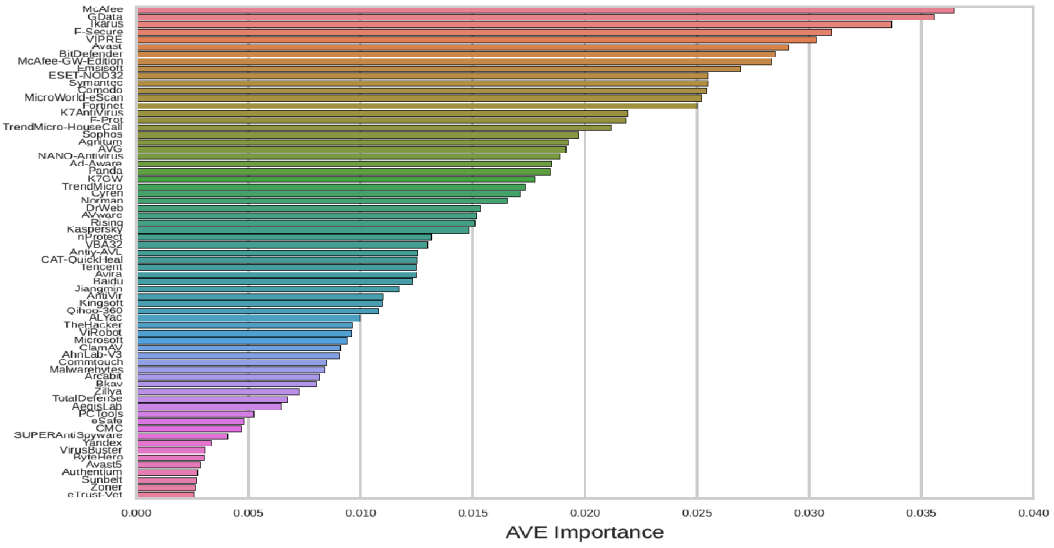


Figure 8. AVT advantage for the *ground-truth set.*

7. CONCLUSIONS

In this work we presented an anti-virus tools analysis using Deep Web malware dataset. The analysis was done on large malware dataset that was crawled by the F-Secure company, using state-of-the-art data analysis techniques, visualizations and graph theory tools, such as community detection algorithm. The analysis was done in order to i) detect common detection capabilities between different anti-virus tools (AVTs), ii) optimize the protection against the largest number of malicious program in a multi-scanning approach and iii) find which AVTs present capability to detect malicious files that other AVTs were not able to detect. The results showed that a lot of the AVTs share similar detection capabilities, due to the fact that they use same detection engine. However, there are some discrepancies between them, such as between gateway and standard AVTs edition, or two AVTs that use same detection engine (due to some in-house solutions). On the other hand, the AVTs that use behavioural approach in detecting malware showed quite unique detection capabilities. The similarity/dissimilarity between AVTs was also shown using community detection algorithm, where three larger AVTs communities were found.

When using a multi-scanning approach, the best solution for the company is to use the most advantageous AVTs, in combination with AVTs from different communities. The MILP approach proposed in the paper, can be used in the future by any company that uses a multi-scanning approach in detecting malware on their mail gateways.

As future work, it remains to analyse the capabilities of different AVT to detect files coming from different sources i.e. downloaded from different domain names. This study could show that some AVTs are more amenable than others to detect several files coming from a given source. The results can denote detection ability for a given malware family (distributed with a domain name specialized for it), which may be due to the crawling of suspicious domain by AV companies to analyse suspicious files in a proactive manner and improve the detection capabilities against new malware distributed by known malicious domains.

ACKNOWLEDGEMENTS

Authors gratefully acknowledge the CyberTrust research project and F-Secure for their support. I.M. work was partially financed by the Faculty of Computer Science and Engineering at the University 'Ss. Cyril and Methodius'.

REFERENCES

[1] M. Lindorfer, M. Neugschwandtner, L. Weichselbaum, Y. Fratantonio, V. v. d. Veen, and C. Platzer, "Andrubis - 1,000,000 apps later: A view on current android malware behaviors," in Proceedings of the Third International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BAD-GERS), 2014, pp. 3-17.

[2] M. K. Bergman, "White paper: the deep web: surfacing hidden value," Journal of electronic publishing, vol. 7, no. 1, 2001.

[3] A. Mohaisen and O. Alrawi, "Av-meter: An evaluation of antivirus scans and labels," in Proceedings of the 11th International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment, ser. DIMVA '14. Springer International Publishing, 2014, pp. 112-131.

[4] "VirusTotal: Free service to analyze suspicious files and URLs," https://www.virustotal.com/en/, online; accessed 14 July 2016.

[5] I. Gashi, V. Stankovic, C. Leita, and O. Thonnard, "An experimental study of diversity with off -the-shelf antiVirus engines," in Proceedings of the 8th IEEE International Symposium on Network Computing and Applications, 2009.

[6] I. Gashi, B. Sobesto, V. Stankovic, and M. Cukier, "Does malware detection improve with diverse antivirus products? an empirical study," in Proceedings of the 32nd International Conference on Computer Safety, Reliability, and Security, ser. SAFECOMP '13. Springer Berlin Heidelberg, 2013, pp. 94-105.

[7] J. Canto, M. Dacier, E. Kirda, and C. Leita, "Large scale malware collection: lessons learned," in Proceedings of the 27th International Symposium on Reliable Distributed Systems, ser. SRDS '08, 2008.

[8] Q. Jerome, K. Allix, R. State, and T. Engel, "Using opcode-sequences to detect malicious android applications," in 2014 IEEE International Conference on Communications (ICC), 2014, pp. 914-919.

[9] M. Zheng, P. P. Lee, and J. C. Lui, "Adam: an automatic and extensible platform to stress test android anti-virus systems," in International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment, Springer, 2012, pp. 82-101.

[10] F. Maggi, A. Bellini, G. Salvaneschi, and S. Zanero, "Finding non-trivial malware naming inconsistencies," in Proceedings of the 7th International Conference on Information Systems Security, ser. ICISS '11. Springer Berlin Heidelberg, 2011, pp. 144-159.

[11] A. Kantchelian, M. C. Tschantz, S. Afroz, B. Miller, V. Shankar, R. Bachwani, A. D. Joseph, and J. D. Tygar, "Better malware ground truth: Techniques for weighting anti-virus vendor labels," in Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security, ser. AISec '15. ACM, 2015, pp. 45-56.

[12] A. Kantchelian, M. C. Tschantz, S. Afroz, B. Miller, V. Shankar, R. Bachwani, A. D. Joseph, and J. Tygar, "Better malware ground truth: Techniques for weighting anti-virus vendor labels," in Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security. ACM, 2015, pp. 45-56.

[13] J. Chang, K. K. Venkatasubramanian, A. G. West, and I. Lee, “Analyzing and defending against web-based malware," ACM Computing Surveys (CSUR), vol. 45, no. 4, p. 49, 2013.

[14] H. S. S.L., “Virustotal public api."

[15] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks," Journal of statistical mechanics: theory and experiment, vol. 2008, no. 10, p. P10008, 2008.

[16] H. T. T. Truong, E. Lagerspetz, P. Nurmi, A. J. Oliner, S. Tarkoma, N. Asokan, and S. Bhattacharya, “The company you keep: Mobile malware infection rates and inexpensive risk indicators," in Proceedings of the 23rd International Conference on World Wide Web, ser. WWW '14. ACM, 2014, pp. 39-50.

AUTHORS

Igor Mishkovski was born in Skopje, Macedonia, in 1981. He graduated and received the master degree in computer science and engineering at the University Ss. Cyril and Methodius, Skopje and the Ph.D. degree from Politecnico di Torino, Torino, Italy, in 2008 and 2012, respectively. After he received the Ph.D. degree in 2012 he was elected as an assistant professor at the Faculty of Computer Science and Engineering in Skopje. His research interests include complex networks and modelling dynamical processes, network science, computer networks, semantic web, operating systems.

Sanja Scepanovic received PhD degree at Aalto Univ. in Helsinki on Big Data with focus to Network Analysis applications to socio-technical systems. She holds diploma in Mathematics from Univ. of Montenegro and MSc in Computer Science from Aalto Univ., Finland and Univ. of Tartu, Estonia. Having spent two years in industry and six in research, Sanja aims to work in the applications of research to space industry; in particular by applying her data science skills to analyzing vast amounts of astronomical and other space data. She is an ISU SSP12 alum and serves as a National Point of Contact (NPoC) for Montenegro for the Space Generation Advisory Council (SGAC)

Miroslav Mirchev received his B.S. in computer engineering and M.S. in computer networks in 2008 and 2009 respectively from the Ss. Cyril and Methodius University in Skopje (UKIM), Macedonia. During 2010 he had a research stay at the City University, Hong Kong. In 2014 he defended his Ph.D. thesis at the Polytechnic University of Turin, Italy, and as part of his studies he spent a period at the BioCircuits Institute, UCSD, USA. His areas of interest include network science, computer networks, nonlinear systems, optimization and machine learning. Currently he is an Assistant Professor at the Faculty of Computer Science and Engineering at UKIM.

Sasho Gramatikov received a Bachelor degree in Computer science, information technologies and automation in 2005 and a Master degree in Computer Science and computer engineering degree in 2009, both from the University of Ss. Cyril and Methodius in Skopje, Macedonia. In 2013 he received a PhD degree at the Universidad Politecnica de Madrid (UPM), Madrid, Spain. He is currently working as an Assistant Professor at the Faculty of Computer Science and Engineering at the University of Ss. Cyril and Methodius in Skopje, Macedonia. His research interests are distribution and streaming of video contents in networks.

INTENTIONAL BLANK

ENHANCING COMPUTER NETWORK SECURITY ENVIRONMENT BY IMPLEMENTING THE SIX-WARE NETWORK SECURITY FRAMEWORK (SWNSF)

Rudy Agus Gemilang Gultom , Tatan Kustana and
Romie Oktovianus Bura

Indonesia Defense University, Bogor, Indonesia

ABSTRACT

This paper proposes a network security framework concept, so called the Six-Ware Network Security Framework (SWNSF). The SWNSF aim is to increase a Local Area Network (LAN) security readiness or awareness in a network security environment. This SWNSF proposal is proposed in order to enhance an organization's network security environment based on cyber protect simulation experiences. Strategic thoughts can be implemented during cyber protect simulation exercise. Brilliant ideas in simulating an network security network environment become good lesson learned. The implementation for proper security strategy could secure an organization LAN from various threats, attack and vulnerabilities in concrete and abstract levels. Countermeasure strategy, which is implemented in this simulation exercise is presented as well. At the end of this paper, an initial network security framework proposal, so called the Six-Ware Network Security Framework has been introduced.

KEYWORDS

Network security environment; cyber protect simulation; cyber threats, attack and vulnerabilities; countermeasures strategy, LAN, SWNSF framework.

1. INTRODUCTION

In terms of network security environment it cannot be denied that as the cost of information processing and internet accessibility falls, civilian, military and government organizations security environments are becoming increasingly vulnerable from cyber threats or attack, e.g., network intrusions, DoS, phishing, spoofing, viruses, flooding, etc. At this point, a LAN security manager might allocate budget, spreading it for network security tools, e.g., anti-virus software, firewalls, intelligent routers or expensive modeling and simulation (M&S) tools. M&S is an effective technique to support better understanding for LAN security managers in concrete and abstract levels [1]. M&S can be used to identify weaknesses proactively and it can also provide education and training using "what if" scenarios reactively. Ultimately when new threats appear the ability of an organization to respond is significantly enhanced. One good lesson learned in the context of network security environment issue today is the phenomenon of Panama papers where over 11.5 million files have been leaked including 2.6 terabytes of data. In the case of Panama papers leak, E-mail is the most of affected records (4,804,618 files), followed by database format (3,047,306 files), PDF document (2,154,264 files), image file (1,117,026 files), text documents

Natarajan Meghanathan et al. (Eds) : SPPR, SCAI, CSIA, WiMoA, ICCSEA, InWeS, NECO, GridCom - 2018
pp. 153–166, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.81714

(320,166) and others file (2,242 files) (see Fig. 1). At this point it is still unclear whether the 11.5 million files were obtained through hacking (data breach) or leaked from someone inside of the Panamanian law firm (insider leak). But from a cyber protect perspective, the lessons are nearly identical either way [2],[3],[4],[5].

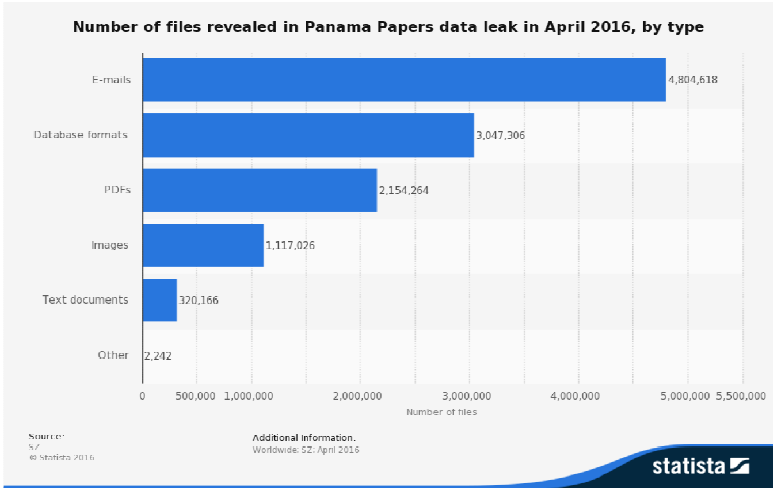


Figure 1. Number of files revealed in Panama Papers data leak in April 2016 by type, Source: Statista.com. [5]

Another good lesson learned in the context of network security environment issue today is the phenomenon of WannaCry Ransomware attack which affected companies and individuals in more than 150 countries, including government agencies and multiple large organizations globally. It was a cyber attack outbreak that started on 12 May 2017 that targeting machines running the Microsoft Windows operating systems. In Indonesia, two major hospitals were affected by this type of computer virus, they are the Harapan Kita Hospital and the Dharmais Cancer Hospital, which halted health information systems services in both hospitals as a result [6]. The affected systems had all data encrypted and a message from the attacker demanding payment of a ransom within 3 days using bitcoins or else the cost would increase. Anyone who refused to pay would eventually lose access to their files and information stored in them. WannaCry Ransomware attack is often delivered via emails which trick the recipient into opening attachments and releasing malware onto their system in a technique known as phishing.

Figure 2. Screenshot of the ransom note left on an infected system, Source: Wikipedia.com. [7]

Based on above two good lessons learned in the context of computer security, the purpose of this paper is to enhance computer network security awareness environment within an organization in order to overcome the various security threats, attack and vulnerabilities through empowering modeling and simulation strategy based on network security framework models. It also meets the demands of the countermeasures strategy and policy of an organization. This paper was inspired by the NIST network security platform version 1.0, 12 February 2014.

The rest of the paper is structures as follows. Section 2 presents the cyber protect simulation tool. Section 3 presents simulation lesson learned. Section 4 explains countermeasures strategy. Section 5 discusses why an organization needs to adopt an appropriate network security framework model to enhance its network security environment. Section 6 describes contribution of this paper by proposing an initial proposal, called The Six-Ware Network Security Framework (The SWNSF), this contribution is an early concept inspired by cyber protect simulation experiences. Section 7 contains concluding remarks and future work for the SWNSF development.

2. CYBER PROTECT SIMULATION TOOL

The Cyber Protect is a software for network security simulation tool designed by the DISA [8]. It is a dynamic learning model environment for information security countermeasures in a Local Area Network (LAN) environment. Cyber Protect has four quarters simulation steps. The user is challenged to make crucial security decision steps about what resources/ countermeasures to purchase and then try to run it [9]. Then, the simulation steps is set in motion and repeated four times where the user faces a various network attack:

- **First step,** choose computer network security resources, e.g., user training, redundant systems, access control, virus protection, backup, disconnection, encryption, firewalls, and intrusion detection.

- **Second step,** applies/installs resources by drag and drop to a specific location on the cyber protect simulation dashboard.

- **Third step,** experiencing a variety of attack. There are nine possible forms of attack, e.g., packet sniffers, viruses, jamming, flooding, imitation (spoofing) and social engineering attack. The attack might come from outside and inside a company.

- **Fourth step,** receiving report indicating performance level. The user receives a final score report based on how well he did in purchasing also applying simulation resources to tackle the variety of attack.

In cyber protect simulation exercise, the user acts as an information leader within an organization. The user has full responsibility to protect or to defend his LAN department. Moreover, by utilizing cyber protect simulation dashboard, the user can freely setup the best and appropriate strategies of a LAN configurations which are expected to be immune from various types of threat, attack or data breach [10]. In order to successfully complete the simulation, meeting a "commanders" goal, the user needs to score 90 or above. But in the real world situation, the information security officers (CISO, etc.) also need a good fortune as well in order to tackle various attakcs. Even with perfect "known" security, the enemy may still find a security hole (see Figure 3).

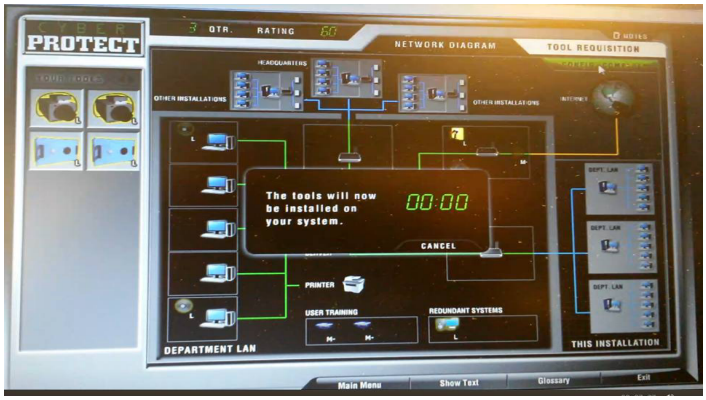


Figure 3. Cyber protect simulation dashboard

3. SIMULATION LESSON LEARNED

During the process of cyber protect simulation exercise, the user will experience several types of threats, attack and vulnerabilities, i.e.:

- **Flooding,** from Internet (external), where the symptom on incident report stating "Network server and/or Router function seriously impaired, degraded or crashed".
- **Viruses,** from internal network stating "Network users report odd characters, noises, tunes, and/or messages appearing on work station screens. Network operations are unusual, degraded or crashed".
- **Packet Sniffer**, from Headquarters (HQ) stating "Slight degradation in time required for network information transference".
- **Jamming,** from HQ stating "Network Transmissions become unreliable or unreadable due to interfering signals".
- **Social engineering attack,** from internal network stating "Report of suspicious attempts by outside individuals to gain access to information".

To deal with those threats, attack and vulnerabilities cyber protect simulation exercise was divided into four quarter tasks, each quarter consist of at least two threat types, attack and vulnerabilities. Every result obtained in each quarter task is displayed into a form of quarter summary reports. Useful experiences during cyber protect simulation process whereby the user can investigate any failures in his network security at the previous quarter. The user determines why controls in place did not prevent threats, attack & vulnerabilities, while making attempts to improve the network security system at the sub-sequent quarter.

4. COUNTERMEASURES STRATEGY

Countermeasure strategy and methodology were needed during cyber protect simulation exercise. The user was asked to design a secure process, technology and personnel of the computer network systems, effectively and efficiently. The user can also identify residual risks of the modelled LAN. At this point, it was found that most of threats and attack came from internal network; these are more difficult to tackle than the external ones (outsiders). From the threat-driven approach perspective, most threats that came from insiders and outsiders (internet) can be

handled effectively through a proper methodology e.g., placing proper security and adequate peripherals, such as, firewalls, IDS and encryption, etc. The threat-driven approach is a methodology, a set of practices and a mindset. In Wikipidia, threat modeling is a process by which potential threats can be identified, enumerated, and prioritized – all from a hypothetical attacker's point of view. Therefore, based on the behavior network model of intended functions, the user identifies and build formal models of security threats, which are potential misuses and anomalies of the intended functions that violate network security aims. [11],[12],[13].

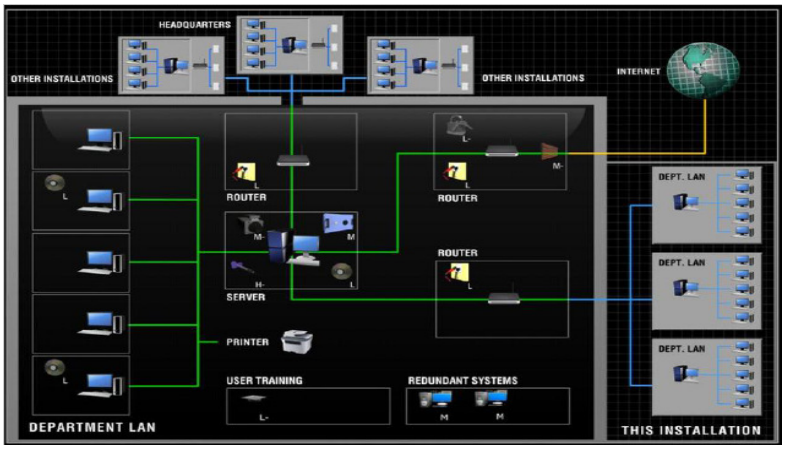


Figure 4. Design of secured LAN

Figure 4 shows a success countermeasure strategy for a LAN modeled configuration by developing appropriate security strategy, effectively and efficiently. It was found that the proper security strategy worked very well in the proposed modeled network security system; the strategy works as follows:

- **First,** configure one medium firewall and one low encryption system at the main router that is connected directly to the internet. The aim is to anticipate threats or attack from outside the network. A good firewall system configuration can anticipate a variety attack from the Internet.

- **Second,** configure three low level of access control units at the entrance and exit of data communication lanes in the network to make it sure that there is no communication path that is not observed in the network. These access control units work as an early warning control system for the network administrator and it has the capability of monitoring all data transmission in the network.

- **Third,** complete system security in every servers with proper security equipments, e.g., high antivirus system, low level backup system, one medium Intruder Detection System (IDS) and one medium redundant system. This strategy can be applied to secure server from various attack.

- **Fourth,** configure two low level backup systems on a particular client who has a high risk job in order to avoid from internal threats or breaches, especially via social engineering attack.

It was found that the proper implementation of countermeasure strategy is a crucial point in cyber protect simulation exercise. The countermeasure strategy might be implemented in various LAN departments, but it depends on its information security and risk management policies. On the

other hand, several countermeasure strategies, e.g., Security-In-Depth Strategy by the US Homeland security or Pro Curve-Pro Active Security Strategy by the Hewlett-Packard innovation centre can be found on internet..

4.1. SECURITY-IN-DEPTH STRATEGY

In October 2009, the US Homeland security developed a security-in-depth strategy as a recommended practice in order to improve Industrial Control Systems (ICS) network security [14]. This strategy is not just about deploying specific technologies to counter certain risks, but it depends on how effective security program for an organization in terms of accepting network security as a constant constraint on all cyber activities in the organization. Figure 5 shows an overview on the key elements of a security-in-depth strategic framework. The basic principles of this framework are as follows:

- First, to know the security risks that an organization faces.
- Second, to quantify and qualify those risks.
- Third, to use key resources to mitigate security risks.
- Fourth, to define each resource's core competency and identify any overlapping areas.
- Fifth, to abide by existing or emerging security standards for specific controls.
- Sixth, to create and customize specific controls that are unique to an organization.

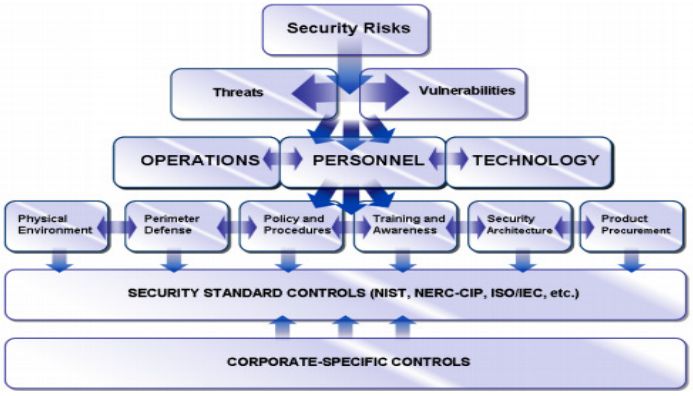


Figure 5. The strategic framework for network security-in-depth

An organization needs to understand its information security risks. It is necessary to understand and improve organizational security awareness as an integral part in implementing the strategy security protection against its sensitive information. Understanding potential threats and vulnerabilities risks is the basic security policy of an organization. The organization should undergo a rigorous risk assessment that covers all aspects to understand risk. Risk assessments are very crucial steps in defining, understanding, and planning remedial efforts against specific threats and vulnerabilities. All level areas and levels in the organization, including executives, must support the valuable risk assessments which are constantly updated at timely intervals.

4.2. PROCURVE-PROACTIVE SECURITY STRATEGY

In February 2007, the Hewlett-Packard (HP) innovation proposed a new comprehensive network security strategy based upon the revolutionary Pro Curve Adaptive EDGE Architecture™ (AEA) [15]. This security strategy embraces distributed intelligence at the network edge and takes a holistic approach to an organization's or company's networking. The HP innovation declared a new security vision, called Pro Curve-Pro Active Security strategy, which is expected to change

dramatically how network security is deployed from now on. Pro Curve-Pro Active security strategy delivers a trusted network infrastructure that is immune to a variety of threats/attack. It has three main pillars:

- **Access Control,** ProCurve ProActive Security strategy proactively prevent network security breaches by controlling which users have appropriate access to network systems.
- **Network Immunity,** ProCurve ProActive Security strategy is able to detect and respond to internal network threats, i.e. viruses, worms, etc., as well as to monitor the behavior and applies security information intelligence in order to assist network security officers in maintaining a high level of network availability.

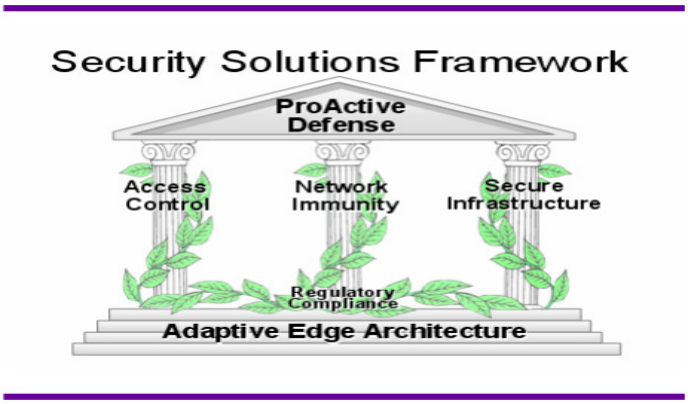


Figure 6. The Three Pillars of Access Control, Network Immunity and Secure Infrastructure

- **Secure Infrastructure.** ProCurve ProActive Security strategy secures the network for policy automation from unauthorized extension or attack to the control plane; includes protection of network components and prevention of unauthorized managers from overriding mandated security provisions. It ensures the integrity and confidentiality of sensitive data, also to protect valuable data from data manipulation or eavesdropping, end-to-end VPN support for remote access or site-to-site privacy, and wireless data privacy (see Figure 6).

One of unique aspects of the ProCurve-ProActive security vision and strategy is that it combines both the security offense and security at the same time and, most importantly, at the network edge. This combined offense and security is possible only because ProActive security is based on AEA principles, which drive intelligence to the network edge while retaining centralized control and management. ProActive security has specific strategy such as identity driven manager, network immunity manager with Network Behavior Anomaly Detection (NBAD) capabilities, policy control at the edge (clientless endpoint integrity web authentication), trusted agent access for LANs, WANs and WLANs as a Standards-based endpoint integrity.

5. NETWORK SECURITY FRAMEWORK COMPARATIVE MODEL

Based on cyber protect simulation experience, organizations need to adopt an appropriate security policy as well as planning and deployment in order to enhance its network security. Every personnel within the organization, from senior level management down to the staff level, must be fully aware of the importance of enterprise information security. All employees should understand the underlying significance of security policy, planning and deployment of the organization. There are several models providing security framework or security reference model,

available in the market, namely the US National Institute of Standards and Technology (NIST) or the Control Objectives for Information and related Technology (CobiT) security framework, etc.

5.1. THE NIST NETWORK SECURITY FRAMEWORK

In February 2013, the US President issued an Executive Order (EO) 13636, in order to improving national critical infrastructure cybersecurity. The EO states: "It is the policy of the United States to enhance the security and resilience of the Nation's critical infrastructure and to maintain a cybersecurity environment that encourages efficiency, innovation and economic prosperity while promoting safety, security, business confidence, privacy and civil liberties". [16]. The US President EO 13636 ordered NIST to work with stakeholders to develop a voluntary framework based upon existing standards, guidelines, and practices in order to reduce cyber risks to national critical infrastructure. The NIST 2014 framework consists of standards, guidelines, and practices to promote the protection of critical infrastructure [17]. It is composed into five basic cybersecurity activities:

- **Identify,** to develop the organization's understanding to manage cybersecurity risk to systems, assets, data and capabilities.
- **Protect,** to develop and implement the appropriate safeguards to ensure delivery of critical infrastructure services.
- **Detect,** to develop and implement the appropriate activities to identify the occurrence of cybersecurity events.
- **Respond** (to develop and implement the appropriate activities to take action regarding a detected cybersecurity event).
- **Recover** (to develop and implement the appropriate activities to maintain the integrity of the security plan and maintain network resilience while restoring impaired ability or services because of cybersecurity attack.

The five activities above are then divided into categories in order to determine a more specific security practices and capabilities, i.e. asset management, access control, etc. Categories are further divided into sub-categories to explain in more detail or technical controls needed to meet the goals of each category (see Table I).

Table 1. The NIST Network Security Framework

Functions	Categories	Sub-categories	Information References
Identify	• Asset Management • Governance	• Invetory devices, systems and software, etc.	• NIST 800-53 CM-8, CA-2, etc.
Protect	• Access Control, etc.	• Review access periodically • Two-factor authentication	• ISO 27001 A6, A9, A11, A13, etc.
Detect	• Detect & Monitor for anomalies and events	• Review logs for suspicious activity, etc.	• NIST 800-53 AU-6, CA-7, etc.

Respond	• Mitigation of security events, etc.	• Report suspicious events, etc.	• ISO 27001 A6, A16, etc.
Recover	• Recovery planning, improve-ments and communi-cation	• Recovery plan • Manage public relations • Repair reputation	• NIST 800-53 CP-10, IR-4, IR-8, etc. • ISO 27001 A16, etc.

6. THE SIX-WARE NETWORK SECURITY FRAMEWORK PROPOSAL

This paper contributes an initial security framework concept, so called, The Six-Ware Network Security Framework (The SWNSF). The SWNSF concept is a comprehensive network security solution to enhance an organization's network security resilience from various threats, attack and vulnerabilities. This is an operational-level security strategy that enables to figure out the most efficient and effective actions that may lead to the success of network security operation [18]. The idea behind this new concept was inspired by NIST network security platform version 1.0., dated 12 February 2014. The SWNSF concept tries to elaborate NIST network security framework to be more practical for the operational level. The security framework discusion can be found also in mashup web data extraction system [19]. The SWNSF concept contributes a common thought to understanding, managing, and expressing network security risks, both internally and externally.

The SWNSF concept contributes increased security awareness environment within an organization where it requires internal/external risk assessment and also threat analysis policies. All levels employees in the organization, ranging from highest level to lowest level must be actively involved in the SWF concept implementation. Otherwise, they cannot obtain better understanding of how threats or attack can be carried out successfully across the entire organization.

6.1. THE SWNSF ENABLERS

The SWNSF enablers provide a set of activities, which consists of six main variables, sub-variables, indicators and information references (e.g., reference guidance). The SWNSF enablers are not only a set of checklist of actions to perform, but it presents key network security solutions to manage security risk and analysis in an organization computer network [20]. The SWNSF enablers comprises six main aspects, e.g., Brain ware, Hardware, Software, Infrastructure ware, Firmware, Budget ware (see Table 2).

- **Brainware** or **human factor,** is the main aspect in network security environment. This variable becomes top list variable within the SWF concept. From network security perspective, it commonly known that human is the weakest link in information security environment. Human factor plays dominant role to enhance or on the contrary, to disrupt all efforts of existing information security within an organization. Therefore, organizations must have function or position related to information security, e.g., Chief Information Security Officer (CISO). The CISO is a company's top executive who is responsible for security of personnel, physical assets, data and information in both physical and digital form. The CISO position has increased in the era of cyberspace where it becomes easier to steal sensitive company information. One of CISO's responsibilities is to conduct

information security certification programs to all level employees. The intention is to produce "information security awareness employees" related to their position and function.

- **Hardware**, plays dominant role in handling threats, attack and vulnerabilities. CISO has to teach all level employees how to use and treat organization's hardware devices safely and wisely. It is because a high-level hacker is not just relying on a specific technique, but still combined with the conventional attack, e.g., social engineering attack. Combination of internal risk assessment and threat analysis are extremely needed, e.g., controlling individual access into the organization's premises or facilities, locking systems and removing unnecessary CD-ROM or USB thumb drives, or monitoring and protecting the security perimeter of organization's facilities, etc.

- **Software,** relates to utilization of software applications security which are used daily in the office, e.g., email, website, social media and other applications. High security awareness is really required because a high profile attacker will always kept on trying to infect or inject malicious emails and its attachments or invite to visit malware-infected websites. The attackers are also constantly introducing new threats although various network security application tools are available in the market.

- **Infrastructure ware,** has an important role in facilitating secure organization network infrastructure, e.g., monitoring network from various threats, attack and vulnerabilities. Nowadays, most of organizations have been highly dependent on Internet access. On the other hand, not all of employees have a good level understanding about security risks they might face in the office, where this condition is making the organization's network infrastructure more vulnerable.

- **Firmware,** includes documentation of an organization security strategy and policy, standard operating procedures (SOPs), business continuity plans (BCPs), network security frameworks or International Organization for Standardization (ISO), i.e. ISO 27001:2013, etc. [21], NIST network security framework version 1.0, government security policy and strategy [22], etc.

- **Budget ware,** plays important and strategic role in facilitating implementation of the five-ware variables above. It is because an organization is urged to provide big enough money or sufficient budget to purchase e.g., network security application tools, patching systems, software licenses, training and education, certification programs, etc. It is highly recommended top level management must put this matter as a high level priority in order to build information security awareness. Allocating sufficient information security budget could protect the entire network system. Otherwise, they will face organization's significant financial losses, etc.

Table 2. The SWNSF Enablers (Enablers and Components)

Aspects	Variables	Sub-variables	Indicators	Infosec References
Brainware	• CISO, etc.	• Security training, etc.	• Security Aware-ness	• CISSP, CISA, etc.
Hardware	• Server Farms	• USB, etc.	• No compromises	• Bench marking, etc.
Software	• Application	• MS Office, etc.	• No pirated Appl. etc.	• Regular updates,etc
Infrastructureware	• Network Infrastructure	• Firewalls. • IDS. • DMZ, etc.	• No network security breaches, etc.	• Self penetration testing, etc.
Firmware	• Security hand book	• Bussiness Continuity Plan	• Good Bussiness pro-cesses	• NIST. • ISO 27001, etc.
Budgetware	• Sufficient budget	• Buy software licen ses, etc.	• Licences always updated, etc.	• Allocated budget policy, etc.

6.2. THE SWNSF COMPONENTS

The SWNSF components proposed, that will be further developed as a theoretical research framework, work together as follows:

- **Variables,** organize network security fundamental aspects as enablers, e.g., brainware, hardware, software, infrastructureware, firmware and budgetware) at highest level. These variables help an organization in managing its security risk and analysis by organizing or clustering data or information, threats and attack activity. Variables align with security and policy framework to reduced impact to organization quality of services (QoS) e.g., investments in human resources, planning and budgeting exercise or recovery actions, etc.

- **Sub-variables,** are sub-divisions of a variable closely tied to a particular (for example, brainware variable) security awareness activities e.g., “security awareness”, “socialization and training”, “network security certification program”, etc.

- **Indicators,** are sub-divisions of a sub-variable, divided into technical outcomes. Indicators provide a set of results to achieve outcomes for each sub-variable. Indicators example (like security awareness sub-variable) e.g., “conducting security awareness training program”; “socializing and implementing security awareness culture in the company”; or “notifications from any social engineering attack or security breaches that are being investigated", etc.

- **Information References (IR),** consists of network security standards, guidelines, methods and practices to achieve solutions or outcomes associated with each indicator. IR which presented in the SWF concept are illustrative and not complete. Examples of IR (like conducting security awareness training program indicator) e.g., “certified ethical hacking (CEH) course from EC-council”; “DoD information assurance awareness training”; and “Achieving ISO 27001 Certification”; etc.

The SWNSF component provides a set of activities to achieve specific network security outcomes, and references examples of guidance to achieve those outcomes. The SWNSF

component is not a checklist of actions to perform. It presents key cybersecurity outcomes identified by organization as helpful in managing the risk within organization network security environment.

7. CONCLUSION AND FUTURE WORK

In terms of network security exercices, the cyber protect simulation is a very good simulation tool. Positively, the cyber protect simulation provides users with useful experiences of tactical and strategical security situation awareness. The users are given the freedom to model and simulate the best strategy to security his secured LAN configurations efficiently and effectively. The cyber protect simulation needs to be developed to face the growth of new variants of security threats, attack and vulnerabilities. The cyber protect needs to comply with sophisticated security frameworks available. In this paper, the authors propose a new security framework, called the SWNSF concept. At this moment, the SWNSF concept cannot be compared, yet, with the NIST or other security frameworks available on the market. It is because, the SWNSF security concept is just an initial proposal to enhance an organization's network security environment.

In the future, the SWNSF concept needs to be implemented and developed more in-depth through further research on specific areas, e.g., determining more technically and specifically security framework variables, sub-variables, indicators, information references, security index scores, etc. Next step, The SWNSF concept will be implemented into a user friendly GUI (Graphics User Interface) or dashboard which acts as an early warning network security system measurement within organizations, institutions or companies. The SWNSF dashboard will work in multi-tasking environments. i.e. portraying the existing LAN security environment while finding the root cause of network security loopholes and suggest some actions to be taken to manage the security aspects. Nevertheless, at the end of the day, it can be concluded that to achieve a perfect or totally a secure network environment is a very difficult task.

ACKNOWLEDGEMENTS

The authors would like to give high appreciation to the i-College, IRMC, the National Defense University (NDU), Washington, DC., USA., for giving a valuable chance to attend the Network Security for Information Leaders (CSIL) course in March 2015. The authors would like to thank also to the Rector of the Indonesia Defense University (IDU) for supporting this strategic paper submission to the NECO 2018.

REFERENCES

[1] J. H. Saunders, "The Case for Modeling and Simulation of Information Security," National Defense University. http://www.johnsaunders.com/papers/securityimulation.htm, last accessed May 2018.

[2] Sara Peters, "7 Lessons From The Panama Papers Leak," vulnerabilities/ threats, http://www.darkreading.com/vulnerabilities---threats/7-lessons-from-the-panama-papers-leak/d/d-id/1324976, last accessed June 2018.

[3] Swati Khandelwal, The Panama papers-Biggest leaks in History Exposes Global Corruption, The Hacker News, http://thehackernews.com/2016/04/panama-paper-corruption. html, May 3, 2016.

[4] Statista.com, "Number of files revealed in Panama Papers data leak in April 2016, by type," http://www.statista.com/statistics/531286/ panama-papers-data-type/, last accessed May 2018.

[5] Statista.com, "Number of files revealed in Panama Papers data leak in April 2016 by type", http://www.statista.com, last accessed May 2018.

[6] Two Major Hospitals in Jakarta had a massive Ransomware WannaCry Attack, https://www.cnnindonesia.com/teknologi/20170513191519-192-214642/dua-rumah-sakit-di-jakarta-kena-serangan-ransomware-wannacry, last accessed May 2018.

[7] Wikipedia.com, "WannaCry ransomware attack," https://en.wikipedia.org/wiki/WannaCry_ransomware_attack, last accessed May 2018.

[8] The i-college, Cyber Cecurity for Information Leaders course, "Cyber Protect Simulation Exercises," National Defense University (NDU), Washington, DC., USA, March 2015.

[9] Ann O'Brien, "Effective Learning Strategies: Cyber Protect – Learning About System Security", Wisconsin School of Bussiness, adapted from Jim Mensching, Chicago State University, USA.

[10] Cyber Protect Network Security Simulation Tool, https://ndu.blackboard.com and http://iatraining.disa.mil/eta/cyber-protect/launchpage.htm, the i-college, NDU, Washington, DC, USA, March 2015. last accessed on June 2018.

[11] Michael Muckin, Scott C. Fitch, "A Threat-Driven Approach to Network Security: Methodologies, Practices and Tools to Enable a Functionally Integrated Network Security Organization," Lockheed Martin Corporation, http://lockheedmartin.com/content/dam/ lockheed/data/ isgs/documents/Threat-Driven%20Approach%20 whitepaper.pdf, last accessed April 2018.

[12] Vicente Pastor, Gabriel Díaz and Manuel Castro, "State-of-the-art Simulation Systems for Information Security Education, Training and Awareness," IEEE EDUCON Education Engineering 2010, The Future of Global Learning Engineering Education, 978-1-4244-6571-2/10, April 14-16, 2010, Madrid, Spain.

[13] Network Security for Information Leaders course, "Information Security and Risk Management," CISSP Textbook Reading, Chapter 3, the i-college, NDU, Washington, DC, USA, March 2015.

[14] The US Homeland Security, Recommended Practice: Improving Industrial Control Systems Cybersecurity with Security-In-Depth Strategies, https://ics-cert.us-cert.gov/sites/ default/files/recommended practices/Security_in_Depth_Oct09.pdf, October 2009, last accessed April 2018.

[15] The Hewlett-Packard (HP) innovation, "ProCurve-ProActive Security: A Comprehensive Network Security Strategy," Pro Curve Networking, February 2007, http://www.hp.com/rnd/pdfs/ProCurve _Security_paper_022107.pdf, last accessed June 2018.

[16] The US White House, Executive Order, "Improving Critical Infrastructure Cybersecurity", 12 February 2013, https://www.whitehouse.gov/the-press-office/2013/02/12/executive-order-improving-critical-infrastructure-cybersecurity, last accessed June 2018.

[17] The National Institute of Standards and Technology (NIST), "Framework for Improving Critical Infrastructure Cybersecurity Version 1.0.," http://www.nist.gov/cyberframework/upload/network security-framework-021214-final.pdf, February 12, 2014, last accessed May 2018.

[18] Chen, J., and Duvall, G., "On Operational-Level Cybersecurity Strategy Formation," Journal of Information Warfare: 13.3: 79-87. SSN 1445-3312 print/ISSN 1445-3347 online, 2014.

[19] Rudy AG Gultom, "Proposing the new Algorithm and Technique Development for Integrating Web Table Extraction and Building a Mashup," Journal of Computer science, Science Publication, NY, USA, DOI: 10.3844/jcssp.2011.129.142, http://www.thescipub.com/ issue-jcs/7/2, 25 February 2011. Download PDF version, http://thescipub.com/PDF/jcssp.2011.129.142.pdf, last accessed April 2018.

[20] Rudy AG Gultom, "The Six-Ware Framework Proposal: A New Comprehensive Network Security Framework To Defend Your Network From Social Engineering Attack," Final Paper, i-college, IRMC, National Defense University (NDU), Washington, DC., USA, 19 March 2015.

[21] ISO, "ISO/IEC 27001: 2013, Information Technology-Security Techniques-Information Security Management Systems-Requirements,"http://www.iso.org/iso/catalogue_detail?csnumber=54534, last accessed May 2018.

[22] Adam Quinn, "Obama's National Security Strategy Predicting US Policy in the Context of Changing Worldviews,"US Research Paper, Project 2015, https://www.chathamhouse.org/sites/files/chathamhouse/field/field_document/20150109 Obama National Security Quinn.pdf, last accessed April 2018.

AUTHORS

Rudy Agus Gemilang Gultom as the author is a researcher and also a senior lecturer at Faculty of Defense Technology, the Indonesia Defense University (IDU), Bogor, Indonesia. He finished his Under Graduate study (Ir.) from the Gunadarma University in Indonesia in 1991, majoring in Information Technology. He finished his Master degree (M.Sc.) in Telematics from the Department of Computer Science, University of Sheffield, United Kingdom in 1999 with scholarship from the British Chevening Award. In 2012, He finished his Doctoral degree (Dr.) in Information Technology from the University of Indonesia, Indonesia with scholarship from Indonesian Government. He can be contacted by mobile phone: +62-81380695525 or at office: 62-21-8795155562-21-87951555 ext. 7152; fax: 62-21-29618766;
e-mail: rudygultom@idu.ac.id.

Tatan Kustana as the co-author is also a researcher and also a senior lecturer at Faculty of Defense Management, the Indonesia Defense University (IDU), Bogor, Indonesia. He finished his Master Degree (M.Bus) from RMIT University of Melbourne, Australia in 1997. He also finished another Master Degree (M.A) from Deakin University, Melbourne, Australia in 2010. He can be contacted by mobile phone: +62-812943406609 or at office: 62-21-8795155562-21-87951555 ext.7001; fax: 62-21-29618766;
e-mail: tatankustana@idu.ac.id.

Romie Oktovianus Bura as the co-author is also a researcher and also a senior lecturer at Faculty of Defense Technology, the Indonesia Defense University (IDU), Bogor, Indonesia. He finished his Master and his Ph.D. studies from the Southampton University, United Kingdom in 1997. He can be contacted by mobile phone: +62-81219588063 or at office: 62-21-8795155562-21-87951555 ext.7001; fax: 62-21-29618766;
e-mail: romieobura@idu.ac.id.

SCALABLE DYNAMIC LOCALITY-SENSITIVE HASHING FOR STRUCTURED DATASET ON MAIN MEMORY AND GPGPU MEMORY

Toan Nguyen Mau and Yasushi Inoguchi

Inoguchi Laboratory, School of Information Science, JAIST

ABSTRACT

Locality-sensitive hashing(LSH) is a significant algorithm for big-data hashing. The original LSH uses a static hash-table as a reduce mapping for the data. Which make LSH challenging to apply on real-time information retrieval system. The database of a realtime system needs to be scalably updated over time. In this research, we concentrate on increasing the accuracy, searching speed and throughput of the nearest neighbor searching problem on big dynamic database. The dynamic Locality-sensitive hashing(DLSH) is proposed for facing the static problem of original LSH. DLSH is targeted for deploying on main memory or GPGPU's global memory, which can increase the throughput searching by parallel processing on multiple cores. We analyzed the efficiency of DLSH by building the big dataset of structured audio fingerprint and comparing the performance with original LSH. To achieve the dynamics, DLSH requires more memory space and takes slightly slower than the LSH. With DLSH's advantages, it can be improved and fully applied in practice in a real-life information retrieval system.

KEYWORDS

Locality-sensitive hashing, Structured dataset, GPGPU Memory, Similarity Searching, Parallel Processing

1. INTRODUCTION

Big-data with high dimensions is becoming a severe problem in analyzing and processing. A high dimensional dataset like audio fingerprint, images featuring or text requires a lot of storage space and takes long time to retrieval. With the optimized storing data and supporting for searching the similarity data, people can deploy many recommender systems likes recommendations of music/video/new, providing the advertising with context or detecting the illegal/similar digital content on the Internet [1].

Because the requirements of fast searching of an information retrieval system, many algorithms are proposed for indexing the big data such as Dimensionality reduction, clustering, classification or hashing algorithm [2{4]. With the principle of the hashing algorithm, the data can be easily divided by the similarity factors. the data that similar to each other by these factors can be

Natarajan Meghanathan et al. (Eds) : SPPR, SCAI, CSIA, WiMoA, ICCSEA, InWeS, NECO, GridCom - 2018
pp. 167–180, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.81715

grouped and store as same location or device [2]. The hashing algorithm can deploy on multiple levels or multiple hash processes that can overlap with each other [5].

Locality-sensitive hashing is a popular hashing algorithm that can group the data into multiple "buckets", the same bucket value will index the similar data. For this characteristic, we can simpler find the nearest neighbors of a query by this hashing value [6].

LSH is an efficient algorithm for approximate nearest neighbor searching. However, the hash-table of LSH is only desirable for the static dataset. Make it not yet widely used for the real-time dynamic database.

It is necessary to improve the structure of LSH to meet the requirements of new generated big dataset system. In this paper, we proposed and analyzed the new approach of storing hash-table targeting for main memory and GPGPU's memory. At this time, DLSH can show the significant result on parallel dynamic and factor on a single memory device. For the further purpose, DLSH can be optimized for use on distributed GPPGU memory.

2. RESEARCH BACKGROUND

2.1 Locality-sensitive hashing (LSH)

Neighborhood-based methods are the method that sorts the similar items/data by its characteristics [7]. LSH algorithm can compute the similarity weights and compare the differences of data/items for detecting the nearest neighbor of an item [8]. Nearest Neighbor Search(NNS) is a typical problem that searches the most similar item for the query [7]. With NNS, there is a difficulty for the guaranty of the output without comparing the output will all feasible data n the dataset. The ε-Nearest Neighbor Search (ε-NNS) is an is a variation of NNS that approximate the nearest neighbor by a threshold function.

Theorem 1 (Nearest Neighbor Search (NNS)). *Given a set P of objects represented as points in a normed space* l_p^d, *preprocess P so as to efficiently answer queries by finding the point in P closest to query point p [9-11].*

NNS is a famous problem in many fields of science and engineering. There are many algorithms already proposed to handle the NNS for every species of the database. However, the complexity of algorithms grows exponentially with the dimensions (curse of dimension), which is a significant difficulty for the real-time system with high dimensions. By a simple trade-off, we can deal with the curse of dimension by using a technique for approximating the NNS [12].

Theorem 2 (ε-Nearest Neighbor Search (ε-NNS)) *Given a set P of objects that are represented as points in a normed space* l_p^d, *pre-process P in order to return efficiently a point p* $\in$ P *for any given query point q in such a way that* $d(q,p) \leq (1+\varepsilon)d(q, P)$ *where d(q, P) is the distance from q to its closest point in P [9,13].*

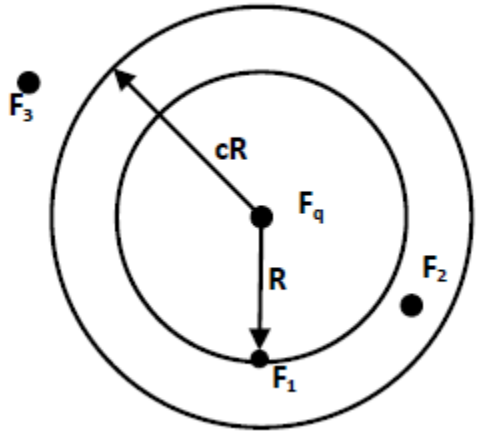


Fig. 1. An illustration of nearest neighbor and approximate nearest neighbor [9, p.2]

In Fig. 1, we can see there are three points in the database F_1, F_2, F_3 and a query point F_q. For the nearest neighbor problem, F_1 should be the chosen one. However, when we consider approximate nearest neighbor problem on this database, suppose that the distance between F_q and F_1 is R and the approximate factor is c, there are two points F_1, F_2 will meet the requirement $|F_q - F_i| \leq cR$. In this case, we can also return F_1 or F_2 both of which are fine.

LSH is a well-known algorithm to handle the ε-NNS problem which uses approximate nearest neighbor. LSH divides the data into multiple buckets, the number of hash function in hash family function will indicate the number of buckets in system. Vectors/points in the same buckets will be similar to each other because of the continuity of the selection of hash functions. Therefore, instead of computing the similarity of the input vector with all of the vectors in database, we need to compare the query with the vectors in several buckets.

With LSH, hash functions will be used for choosing l subnets I_1, I_2,..I_l of database vectors. Let p_I be the projection of vector F_j on the coordinate positions. Then denoting $g_j(p) = P_{I,}$ we store every $F_j \in F$ in the bucket $g_j(F_p)$. Since the number of buckets is probably large or the numbers of points in every bucket differ considerably, so another table is also needed to save the map of buckets.

As to the searching problem within LSH, the same hash functions are also used for each query. As for the query F_q, we define all $g1(q)$, $g2(q)$,…$gl(q)$, and then let $F1$, $F2$,…Ft be the points in bucket on the current process. We have to calculate the distance $l_1(Fp, Fq)$ for each point in this bucket. As for KNN problem, we do not stop until we reach K points in the same or different buckets. However, with the audio fingerprint, it can be returned at the first F_p having the $l_1(Fp, Fq) < P_1$ to attain a good result, along with a better performance. Note that in case the two points are close to each other, P1 is a threshold of the maximum distance.

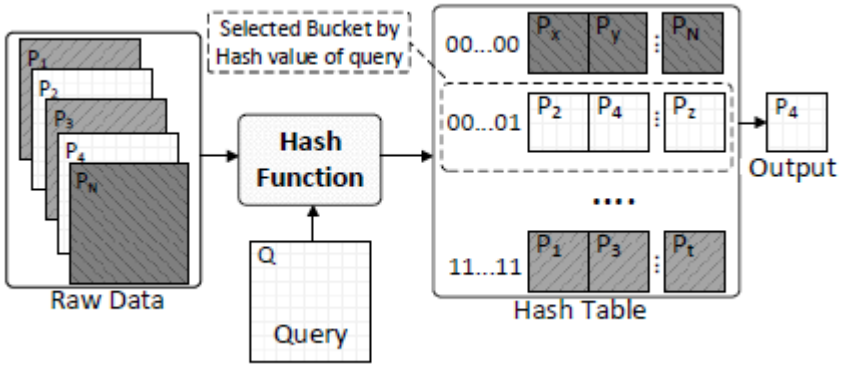


Fig. 2. An illustration of locality-sensitive hashing

In Fig. 2, LSH chooses a family of hash function for handling database and query also. In the preprocessing stage, hash function is used for dividing the database into buckets. There are three buckets with the different colors in the figure. Each bucket will have its hash value by the principle of hash function. In term of the Searching stage, the query also needs calculating the hash value by the previous hash function. This value of hash function will indicate to a bucket that holds the similar points to the query (light box). Next, there is another step for comparing the distance from query to all points in the purple buckets and returning the closest one.

3. DYNAMIC LOCALITY-SENSITIVE HASHING (DLSH)

In this paper, we proposed the DLSH that have the dynamic structure for storing the dataset on heap-memory or GPGPU's global memory.

3.1 Memory Pools Allocation

The memory manager of main memory and GPGPU's memory is very similar by using the address pointer. However, the pointer in CUDA has the restriction for dynamic allocation on the kernel. We choose to use the memory pools allocation for both kind of memory for reducing the number of fragment memory and also the reusable of memory pools.

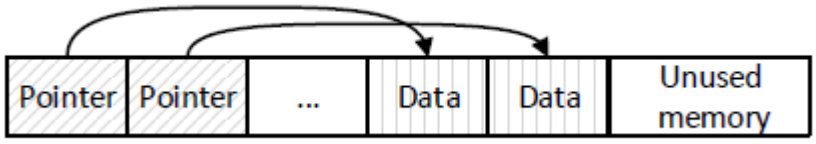


Fig. 3. An example of Memory pools allocation

In Fig. 3, the allocated array have the fixed size and can hold multiple user-defined pointers and data on it. In the last part, there an unused memory for storing the new dynamic data/item.

It is clear that the hash table DLSH will take advantages of memory pools when storing both mapping key and value on the same array.

3.2 Linked-list of bucket

Data will be stored as a sequence that indicated by an hash-value on LSH. We proposed to used linked-list structure to track all the data in a bucket. An important part of hashing table of DLSH is the pointers of all available buckets. To handle this issues, in the first part of hash-table array, we store the static pointers for every bucket by the hash values.

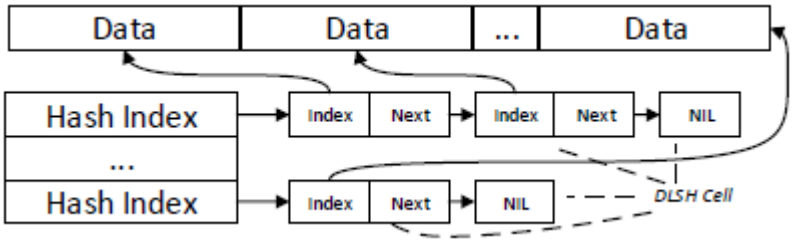


Fig. 4. Linked-list bucket structure of DLSH

Fig. 4 shows the static Hashing Index part correspondings with hashing values of family of hash function. The value of each Hash Index will point to its first index of data. the cell of bucket linked-list contains two important values the address of the data/item and the pointer of the continues cell. The NIL cell is a special cell that denotes the ending of a bucket. In case of a bucket is empty the pointer to the first cell will be set to NIL.

By using the linked-list, the hash-table is compacted and stored on a single array. This help the system can easily allocate the array for both main memory and GPGPU's global memory. Besides that, the linked-list has the significant advantage in modifying the hash-table. However, compare to LSH, each cell of the list required one more addressing space for index the next cell on the list, which make the size of DLSH bigger the size of LSH's hash-table.

3.3 Remove item from DLSH's hash-table

For the real-time information retrieval system, the system needs to support to remove items from the hash-table, which is an essential factor that makes the DLSH become the dynamic hashing system. To delete an item from hash-table, we need to find the previous cell that point directly to the deleting item, then the next pointer of the previous cell will point to the address of next-point of deleting time.

In Fig. 5, the red link will be changed to the blue link, and the deleting item will be unreached by using hash-table's linked-list. That is the reason we proposed to use another pointer-list that point to the deleted cell to keep them on track.

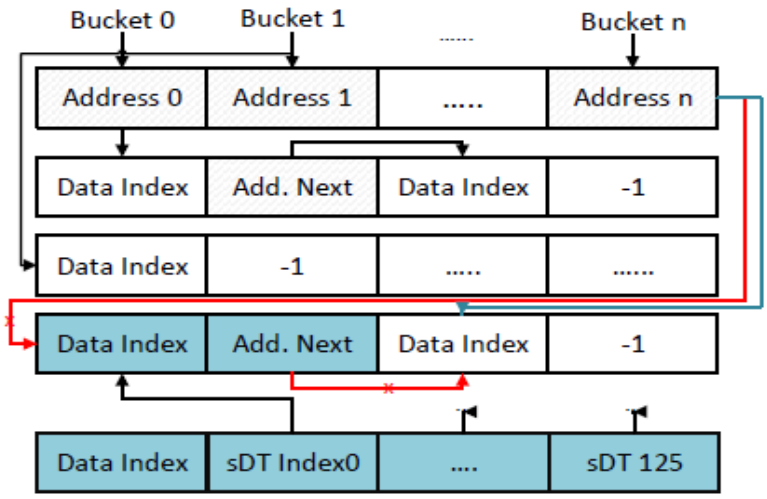


Fig. 5. Scenario of deleting item from DLSH's hash-table

In Fig. 1, because there is no data addressing pointer in static pointer for buckets, the system has to check the first element of linked-list. The previous-pointers of these specific cells are assigned directly to its memory location.

3.4 Add item to DLSH's hash-table

We highly recommend that the allocation of the extra empty space for the hash-table of DLSH. This unused space can be used when the hash-table is full and need more space for the new cell for the new item. In Fig .6, the empty space is used for creating new linked-list cell, the data index points to the address of new item/data on the data array. In this example, we need to find the last cell of the list that indicated by new hash value and assign the next pointer to the newly created cell. However, it will be faster when we assign the new cell as the first element of its linked-list. In this case, the next-pointer of new cell will point to the ex-first cell of its bucket.

Algorithm 1 Removing item/data from DLSH's hash-table

```
1: hash ← HashFunction(q)
2: hash_pre ← hash
3: hash ← HT[hash]
4: is_first_element ← TRUE
5: while hash ≠ NIL do
6:     if Match(q) and GetData(hash) then
7:         if is_first_element then
8:             HT[hash_pre] ← HT[hash + 1]
9:         else
10:            HT[hash_pre + 1] ← HT[hash + 1]
11:        end if
12:    end if
13:    hash_pre ← hash
14:    hash ← HT[hash + 1]
15:    is_first_element ← FALSE
16: end while
```

Algorithm 2 Adding item/data to DLSH's hash-table

```
1: hash ← HashFunction(q)
2: hash ← HT[hash]; hash_pre ← NIL
3: if Exist empty-cell in HT then
4:     hash_restored ← FindEmptyPointer
5: else
6:     hash_restored ← Extend(HT)
7: end if
8: while hash ≠ NIL do
9:     hash_pre ← hash
10:    hash ← HT[hash + 1]
11: end while
12: if hash_pre ≡ NIL then
13:    hash_pre ← hash ← HashFunction(q) -1
14: end if
15: HT[hash_pre + 1] ← hash_restored
16: HT[hash_restored + 1] ← NIL
```

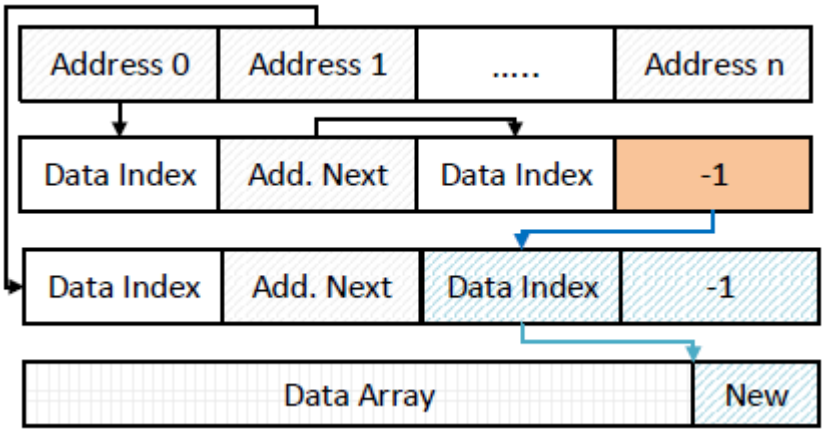


Fig. 6. Scenario of adding item of DLSH's hash-table when the hash-table is full

In Algorithm 2, If the DLSH's hash-table is full, there is required to extend the size of array of HT. We can use the unused space when allocated the memory pool for the HT array. Otherwise, there is an empty linked-list cell, we have to reuse it to reduce the memory size and increase the performance. In Fig. 7, the restored cell (yellow) has been reused, that become the last element in the linked-list of first bucket. The memory space of data/item also can be reused in this figure.

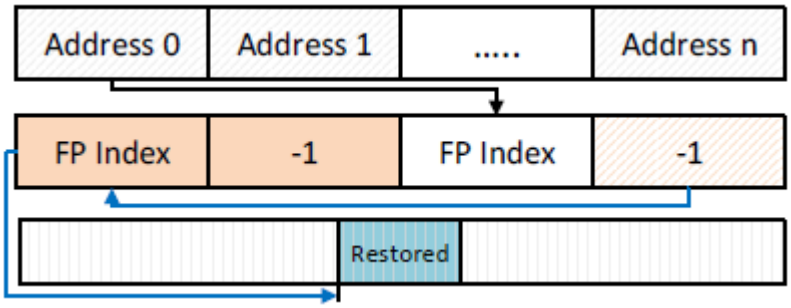


Fig. 7. Scenario of adding item with empty-cell on hash-table

4. EXPERIMENTAL SETUP

In this research, we are targeting on using DLSH for handling the database with millions of structured audio fingerprint for the information retrieval system. For Collaborative Filtering (CF) problem, we experimented ε-NNS problem for the dataset of audio fingerprint to find the approximate most similar song on the database with the queries.

4.1 HiFP2.0: An Binary Sequence Audio Fingerprint

An audio fingerprint is a digital vector that extracted from the audio/song waveform and able to standardize the content of audio/song source. The audio fingerprint can easily help for comparing the similarities and differences of songs. In addition, using audio fingerprint for storing can reduce the size of original audio/song with the standard structure. In our system's database, we store the audio fingerprints and its meta information for every songs/track we considered instead of storage the real waveform of the song [14, 15].

Fig. 8. A example of 4096-bits fingerprint extracted by HiFP2.0 audio fingerprint extraction algorithm

In Fig. 8, there is an audio fingerprint that is represented by a binary sequence extracted by HiFP2.0 fingerprint extraction algorithms. This audio fingerprint is extracted from first 2.97 seconds of a song, and that can reduce the size of the song by 512 times [16].

For description the whole content of an audio input, with the difference in audio length, we will have different size of audio fingerprint [17]. With the different length of audio fingerprint, there is some problem with storage for fast searching [18].

Using audio waveform for comparing have many difficulties because that is un-normalized, so we favor to using an audio fingerprint for storing and processing in our system.

4.2 Dataset and Computer Memory Architecture

We aimed to examine our proposed system on larger memory with the enormous number of data on the database. With the typical size of an HiFP2.0 feature is 512 bytes, we generated a set of 10 million HiFP2.0 with the size of 5 GB for testing. To analyze the accuracy of both LSH and DLSH system, we created numerous of testing queries with different distortion from the dataset and examined with different number of hash function on hash function family.

The Staged-LSH was proposed by [16], that can significantly increase the accuracy of ε-NNS for similarity song searching and that was verified in [16]. Staged-LSH group the data/item into multiple sub-sequences and computes the hash value for each sub-sequence. With the size of HiFP2.0, authors recommended dividing the audio fingerprint to 126 sub-fingerprints, which make a single audio fingerprint can have up to 126 different hash values and indexed by 126 different buckets.

Table 1. The dataset's and hash-table's size for the database of 1 million HiFP2.0 features

Database size (MB)	Num Hash function	Staged-LSH HT Size(MB)	Staged-DLSH HT Size(MB)
488.28	17	480.78	961.43
488.28	18	480.90	961.55
488.28	19	481.15	961.80
488.28	20	481.65	962.30
488.28	21	482.65	963.30
488.28	22	484.65	965.30
488.28	23	488.65	969.30

Table 1 shows the actual memory size for Staged-LSH and Staged-DLSH. Because Staged-LSH increases the number of indexing-cell to 126 times, that make the size of hash-table nearly equal to the size of the dataset. The DLSH take more memory space than LSH by the linked-list pointer, which makes the serious problem of DLSH when trade-off with the dynamic structure.

The specifications of the testing computer are shown in Table 2. We examined and compared the performance of LSH and DLSH using the same computer and same conditions.

5. RESULTS AND COMPARISON

The dynamic feature is the most important factor of DLSH, the dynamic changes of DLSH hash-table (adding/removing) are mainly analyzed. We also compare the efficiencies of DLSH to this main competitor LSH to very the feasibility of DLSH. The principle of DLSH is inherited from LSH, which make the accuracy of ε-NNS of both methods are similar. Therefore, we only showed the results related to performance and memory usage. The performance of DLSH over

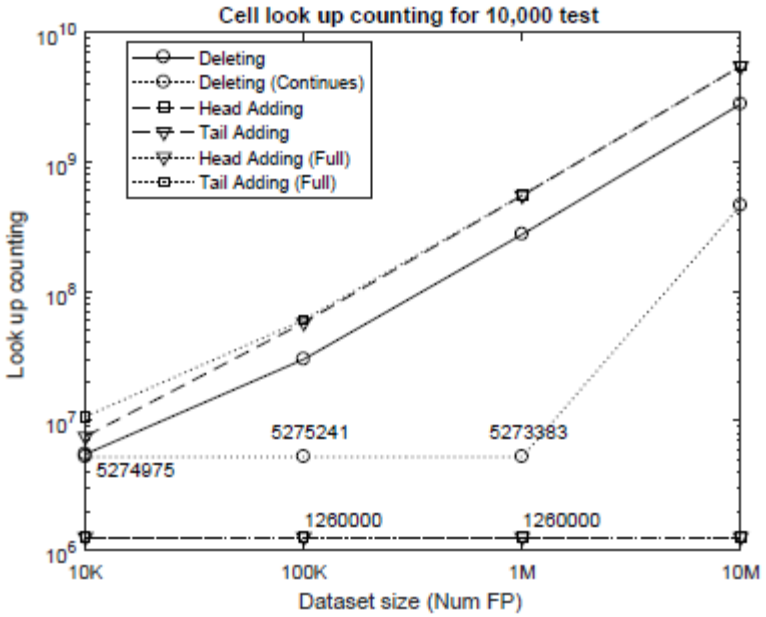


Fig. 9. Number of lookup in hash-table of DLSH; (Full): Adding the data/item when the hash-table is full; (Continues):

the data size is shown in Fig. 9. We deployed tests on various dataset size with 19 hash functions. The number of hash-table lookup of deleting and adding commands is depended to the average number of cell in each bucket. When continues deleting data/item from the dataset, the number of hash-table lookups will be reduced. Adding the data/item at the head of linked-list is recommended for getting performance. However, this approach is easier to make the database fragment.

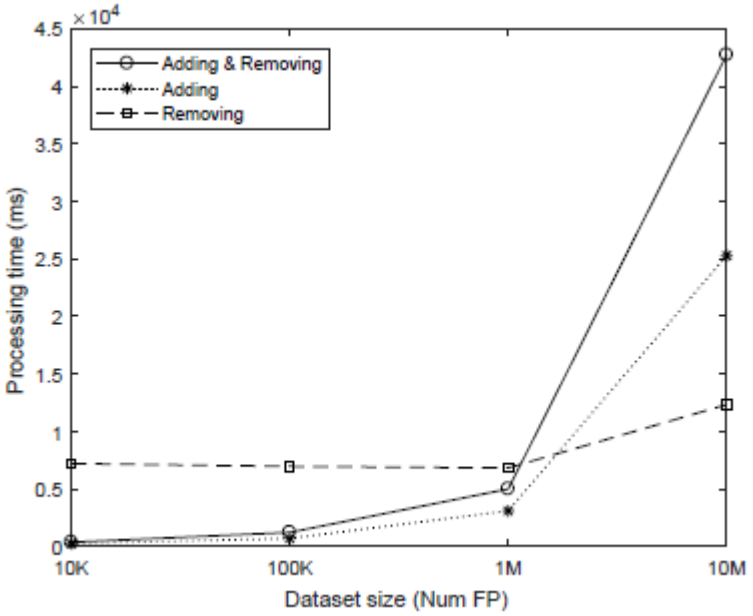


Fig. 10. Processing time of adding/removing 10,000 queries over different dataset size with 19 hashing functions

In Fig. 10, when the dataset size is small, the adding data/item tasks takes less time than removing by using the empty slots in DLSH's hash-table. With 10 million audio fingerprint dataset, the adding/removing tasks requires more to find the suitable position for deleting or adding.

For the scalability factor, we conducted experiments on various dataset size by turning the number of audio fingerprint on the dataset. Fig. 11 compares the searching time of the original LSH with our proposed Dynamic LSH. With numerous additional pointers, our method has an overhead around 10% of the original method. The DLSH's hash-table structure is well developed for supporting parallel processing, especially for multiple core processor by using fix-size pointers of buckets on same 'page-locked' memory.

In Fig. 12, with different number of hash function, the searching of our method have slightly higher than original LSH on finding the approximate nearest neighbor of 10,000 queries. Choosing hashing number is very sensitive, with the low number of hash function we can quickly achieve highly accurate but the high density of buckets will reduce the searching time. Using GPGPU can increase the speed of searching into nearly 2 times than using CPU. In other hand, the performance of GPGPU can get better when we testing on higher throughput of queries.

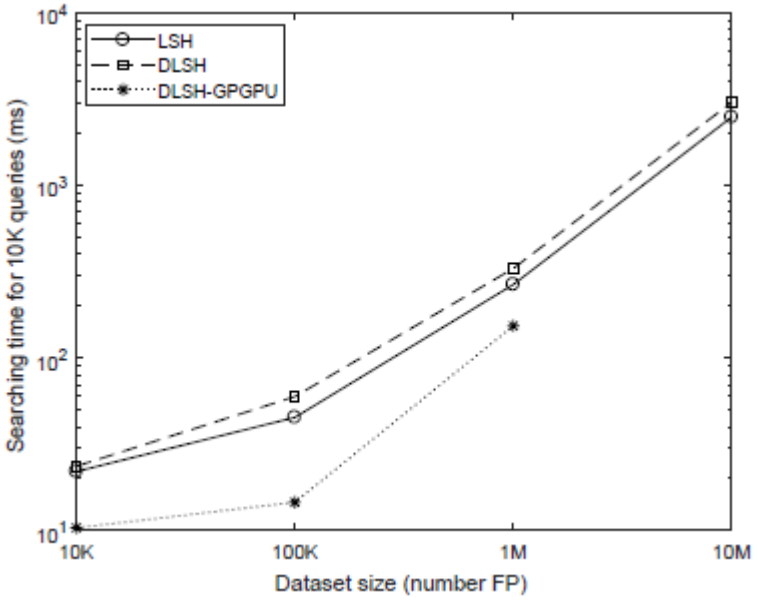


Fig. 11. The ε-NNS processing time comparing between LSH and DLSH on 10,000 queries with 10% distortion queries. Both methods use 19 hash functions on hash function family.

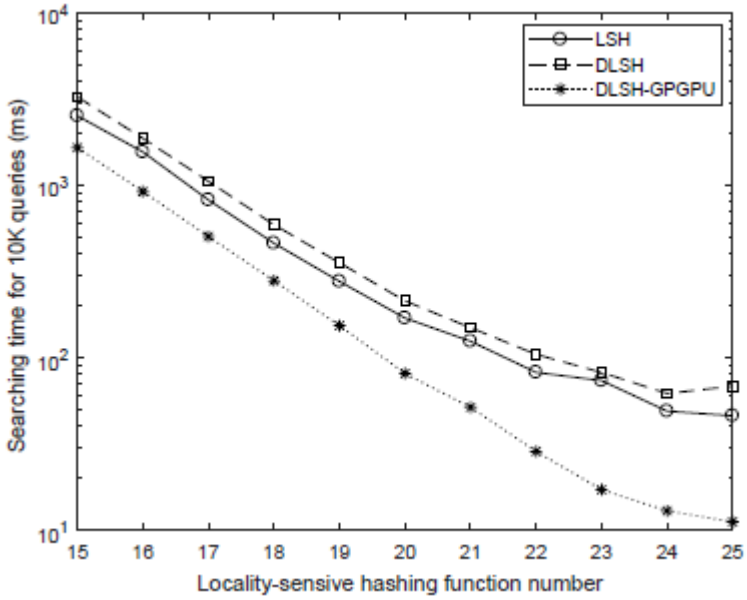


Fig. 12. The ε-NNS processing time comparing between LSH and DLSH on 10,000 queries with 10% distortion queries on dataset with 1 million HiFP2.0 features.

In Fig. 13, when there is no distortion on testing queries, the DLSH of takes more memory accessing by the static pointer. Besides that, the LSH can directly point to the first data of the buckets, which make the searching time of DLSH nearly double the LSH's.

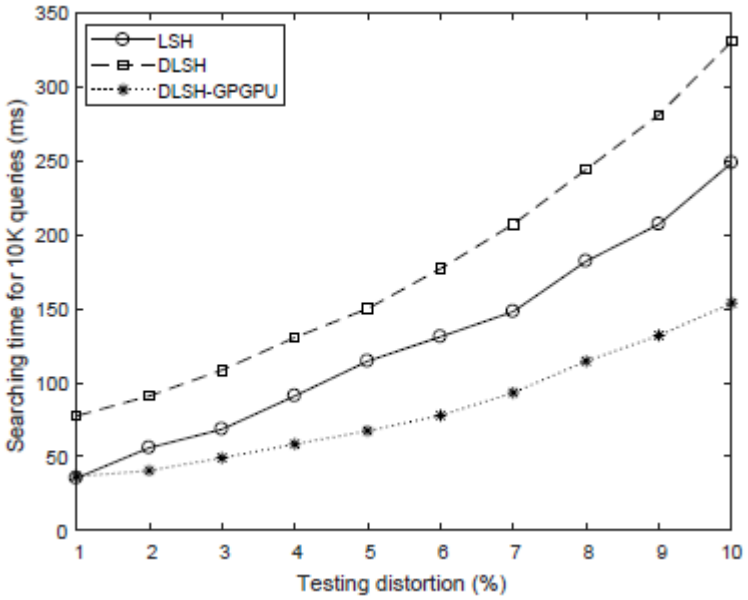


Fig. 13. Comparing performance between LSH and DLSH with different testing distortion

6. CONCLUSION

In this paper, we proposed DLSH a new approach for dynamic hashing dataset that inherited the advantages of locality-sensitive hashing. The DLSH aimed for handling the big real-time dataset for information retrieval system which requires quick response time and high throughput.

The dynamic structure needs extra memory space for holding the pointers of hash-table's cell and addressing to the dataset. However, the performance of DLSH is acceptable for a dynamic system comparing to the performance of the original LSH. The DLSH can reuse the space in main memory or GPGPU memory. Although it still takes time to find the exact memory address.

For the future work, we concentrate on optimizing the DLSH for getting better performance and stabilization. The memory size of main memory and GPGPU is limited, we also target to extend the DLSH for scaling on distributed GPGPU computer system.

REFERENCES

[1] Gema Bello-Orgaz, Jason J Jung, and David Camacho. Social big data: Recent achievements and new challenges. Information Fusion, 28:45-59, 2016.

[2] Jong Soo Park, Ming-Syan Chen, and Philip S Yu. An effective hash-based algorithm for mining association rules, volume 24. ACM, 1995.

[3] Min Chen, Shiwen Mao, and Yunhao Liu. Big data: A survey. Mobile networks and applications,19(2):171-209, 2014.

[4] William Bruce Frakes and Ricardo Baeza-Yates. Information retrieval: Data structures & algorithms, volume 331. prentice Hall Englewood Cliffs, NJ, 1992.

[5] Spyros Blanas, Yinan Li, and Jignesh M Patel. Design and evaluation of main memory hash join algorithms for multi-core cpus. In Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, pages 37-48. ACM, 2011.

[6] Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In Proceedings of the twentieth annual symposium on Computational geometry, pages 253-262. ACM, 2004.

[7] Christian Desrosiers and George Karypis. A comprehensive survey of neighborhood-based recommendation methods. In Recommender systems handbook, pages 107{144. Springer, 2011.

[8] Jia Pan and Dinesh Manocha. Fast gpu-based locality sensitive hashing for k-nearest neighbor computation. In Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems, pages 211{220. ACM, 2011.

[9] Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on, pages 459{468. IEEE, 2006.

[10] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In Proceedings of the thirtieth annual ACM symposium on Theory of computing, pages 604-613. ACM, 1998.

[11] Wei Cen and Kehua Miao. An improved algorithm for locality-sensitive hashing. In Computer Science & Education (ICCSE), 2015 10th International Conference on, pages 61-64. IEEE, 2015.

[12] Anirban Dasgupta, Ravi Kumar, and Tamas Sarlos. Fast locality-sensitive hashing. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1073-1081. ACM, 2011.

[13] Edward Y Chang. Approximate high-dimensional indexing with kernel. In Foundations of Large-Scale Multimedia Information Management and Retrieval, pages 231-258. Springer, 2011.

[14] Jaap Haitsma and Ton Kalker. A highly robust audio fingerprinting system. In Ismir, volume 2002, pages 107-115, 2002.

[15] Kouichi Araki, Yukinori Sato, Vijay K Jain, and Yasushi Inoguchi. Performance evaluation of audio fingerprint generation using haar wavelet transform. In Proceedings of the International Workshop on Nonlinear Circuits, Communications and Signal Processing, Tianjin, China, pages 380-383, 2011.

[16] Fan Yang, Yukinori Sato, Yiyu Tan, and Yasushi Inoguchi. Searching acceleration for audio fingerprinting system. In Joint Conference of Hokuriku Chapters of Electrical Societies, 2012.

[17] Toan Nguyen Mau and Yasushi Inoguchi. Audio fingerprint hierarchy searching strategies on gpgpu massively parallel computer. Journal of Information and Telecommunication, pages 1-26, 2018.

[18] Toan Nguyen Mau and Yasushi Inoguchi. Audio fingerprint hierarchy searching on massively parallel with multi-gpgpus using k-modes and lsh. In Eighth International Conference on Knowledge and Systems Engineering (KSE), pages 49-54. IEEE, 2016.

AUTHORS

Toan Nguyen Mau received the BS degree (Honors Program) from University of Science - Ho Chi Minh Nation University (VNU) , Vietnam, in 2013 and the MS degree in information science from the Graduate School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), Ishikawa, in 2016. He is currently a phD student in Inoguchi laboratory at the Graduate School of Information Science at JAIST. His research interests are parallel processing using GPGPU, massively parallel process and parallel audio fingerprint extraction algorithms.

Yasushi Inoguchi received the BE degree from the Department of Mechanical Engineering, Tohoku University in 1991, and received the MS and PhD degrees from Japan Advanced Institute of Science and Technology (JAIST) in 1994 and 1997, respectively. He is currently a professor of RCAsCI at JAIST. He was a research fellow of the Japan Society for the promotion of Science from 1994 to 1997. He is also a researcher of PRESTO program of Japan Science and Technology Agency from 2002 to 2006. His research interest has been mainly concerned with parallel computer architecture, interconnection networks, GRID architecture, and high-performance computing on parallel machines. Dr Inoguchi is a member of IEEE and IPS of Japan.

A NETWORK OF INTELLIGENT PROXIMITY IOT DEVICES FOR OBJECT LOCALIZATION, INFORMATION COMMUNICATION, AND DATA ANALYTICS BASED ON CROWDSOURCING

Mike Qu[1] and Yu Sun[2]

[1]Northwood High School, Irvine, CA 92602
[2]California State Polytechnic University, Pomona, CA 91768

ABSTRACT

While the advancements of technology have benefited the society in many ways, certain problems remain, and one such problem is the issue of lost people and pets. Current technology has offered many solutions to this problem, yet none is able to encompass all the core aspects to bring an end to the problem. My research proposes a solution that is practical, durable and reliable -- a proximity sensor device powered by a crowd of people, by using their mobile devices as receiving stations of the service, extensively increasing the effectiveness of this service in especially urban and suburban areas where there is a high population density.

KEYWORDS

Beacon, Device Network, Crowdsourcing, Double-blind, Artificial Intelligence.

1. INTRODUCTION

Many pet owners have once come across the trouble of losing their pets. While some animals are eventually retrieved by their owners, others are not so fortunate. In fact, 6.5 million companion animals enter U.S. animal shelters every year. Among them, only 710,000 are returned safely to their owners while more than 1.5 million are euthanized due to a lack of housing capacity in animal shelters [1].

To combat this growing problem, pet owners and vendors have developed a variety of services, including GPS location [2], microchips and written tags. While each type of service has their own merits, there are definitely shortcomings to them as well. GPS location is often expensive as it requires a variety of infrastructure support, including ground stations and positioning satellites to function properly. GPS signals are also very limited in indoor locations, underground areas, and places with large buildings and structures. Aside from that, GPS devices require charging on a regular due to their rapid consumption of energy, both posing a hassle to pet owners and significantly hindering the effectiveness of such a device. On the other hand, the implementation

Natarajan Meghanathan et al. (Eds) : SPPR, SCAI, CSIA, WiMoA, ICCSEA, InWeS, NECO, GridCom - 2018
pp. 181–187, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.81716

of microchips and written tags are a lot less costly, and that they are able to accurately identify a pet to animal control services makes these methods seem practical. However, when the fact that only 6.5 million of the approximately 70 million stray animals living are taken into animal shelters in the U.S. is taken into consideration, the aforementioned methods seem obsolete as there is no absolute guarantee that one's lost pet would eventually end up in one of the 3,500 animal shelters throughout the country.

Missing companion animals, a seemingly small problem is just the tip of an iceberg of a much more important issue. Many other groups in the population suffering from specific medical conditions, such as Alzheimer's and dementia face the same challenges in the society. According to the Alzheimer's association, six in ten patients would wander and oftentimes, they may not remember their home address, and can become disoriented even in familiar places [3]. Aside from that, lost children who are usually unaware of their surroundings also account for a large portion of this problem. According to the FBI, there were 464,324 [4] entries for missing children in 2017, and by providing reliable and accurate locations of potential runaway children, this number can be greatly decreased.

A sprawling population in urban and suburban areas brings various social problems, including the increasing numbers of runaway people and pets call for a more accurate, reliable and practical approach to combat the problem. An increasing population for companion animals has overloaded the current system of animal shelters, and the growing population density also poses an increasing threat to particular groups, especially those who suffer from conditions that affect their capabilities of living independently.

A low-cost, reliable, and crowd-based technology using individuals' mobile devices as receiving stations to map and locate lost items used in accordance the iBeacon technology [5-7]. The iBeacon technology is able to emit Bluetooth 4.0 [15-18] signals and is based on physical proximity sensor devices known as beacons. These devices are small, durable and reliable. This particular technology requires no pairing, which means that mobile devices, such as cell phones and hardware devices could all receive the signal and prompt a specific function, from retrieving the satellite location of that device to sending a statement or message to that device. A number of such mobile devices in the same area, simultaneously sensing a beacon, could provide a well-positioned and accurate satellite location for that purpose.

The rest of the paper is organized as follows: Section 2 details the challenges in this research project; one solution is presented in Section 3, followed by showing a discussion in Section 4; we compare the related works in Section 5, and Section 6 offers the conclusion remarks and future work directions.

2. CHALLENGES

Due to the complexity of this solution, A few challenges are encountered during the development of this system.

2.1 Challenge 1: Potential Lack of Coverage

Without a dense network propagated from a number of devices located in a small area, the system would be not as effective because of the nature of this system, which requires a number of mobile

devices to function. Furthermore, it calls for the density of mobile devices, instead of dispersity. Having multiple devices in a small and close-knit community could benefit its residents more than having the same number of devices across a much larger city. Therefore, we need to overcome the challenge by finding a solution to increase the coverage and make it as effective as other methods.

2.2 Challenge 2: Lack of Appealing Features to the Public

The ultimate usage of this device could be somehow limited as the ultimate beneficiaries are those who need this service the most, such as pet owners to track their pets, concerned family members to know the whereabouts of potential wandering patients, or parents who have young children who are unaware of their surroundings. Due to this, what this service has to offer could be slightly limited to those who do not require this service. Thus, we need to provide flexibility for users when they are using such devices.

2.3 Challenge 3: Risk of Privacy

Since any device could be used as a receiving station for a beacon, privacy could be an important issue in the development of this product. The fact that all beacon-owners could modify the database of information storage could have a negative effect on the practicality and reliability of this system. Therefore, it is necessary to solve the problem and avoid privacy issues in usage of such devices.

3. METHODOLOGY

3.1 Overview of the Solution

The underlying concept of this solution is the use of the powerful iBeacon technology on compact "beacons", which could emit a consistent Bluetooth 4.0 signal without the need to pair with a mobile device. When such a signal is received by either a hardware receiver or a mobile device, the device or receiver would log its own location, and publish that location in terms of longitude and latitude to a database, updating the location to indicate the approximate position of that beacon. This data can be downloaded from the database and displayed on the owner's device. Due to concerns for privacy, the users would have the option to turn on or turn off their abilities to give away their location to potential beacons in its range.

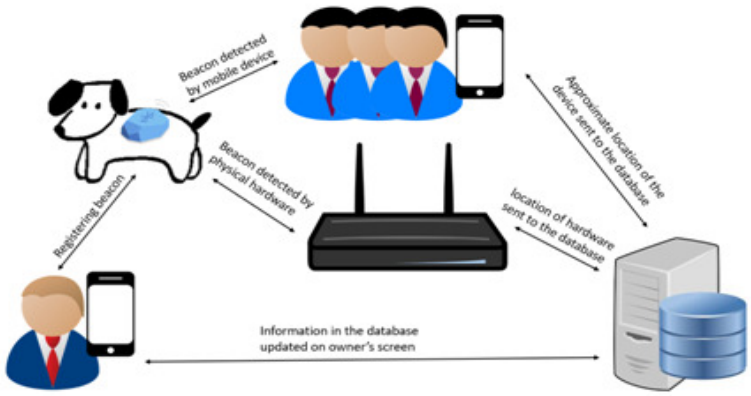


Figure 1. The overview of the solution

3.2. Crowd Analysis

As mentioned above, this system functions best when used under a dense coverage of mobile devices. The more devices there are in one area, the more accurate the provided location will be. This situation can best be modelled by a Venn Diagram -- The three circles (in Figure 2) represent three different devices, either a physical signal receiver or a mobile device, that are of certain distances among one another. Having only one device is the least accurate, as the potential location of that beacon could be anywhere in the whole circle. Having two devices significantly improves the amount of error in locating that beacon, but its effectiveness could greatly vary. While having three or more devices could limit the beacon's location in a small, triangular area, which would be the most accurate.

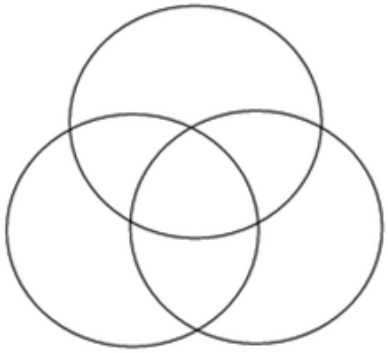


Figure 2. The range of three different devices

3.3 Physical Capacities of a Beacon

As mentioned above, Beacons are usually small, about an inch in diameter, and around a quarter of an inch in height. This attribute makes the beacons very light and durable. These beacons require coin batteries and require replacements once every one to two years, making them reliable under most circumstances. Beacons also vary in coverage area, ranging from around 20 meters to more than 350 meters for advanced models [4], but sensors could distinguish different strengths of iBeacon signal, which can be used in calculations to determine the approximate location of that beacon.

3.4 Signal Receiver Hardware

The signal receiver hardware is a mini-computer that requires both a power outlet and a Wi-Fi connection. This device is able to scan the surrounding area once every one tenth of a second for Bluetooth signals. Although this mini-computer is unable to locate itself due to the lack of a GPS location service, it usually stays stationary in at key locations, such as community gates, shopping centers, street intersections etc. to maximize how effective it is.

3.5 Mobile Application

The mobile application contains three screens: One for navigation, one for monitoring the registered beacons that belong to a user and one for initial registration of a beacon. It is able to pinpoint the exact location of the mobile device or the hardware signal receiver that last came across a signal from that iBeacon. There is also a drop-down menu that allows for one to track

multiple beacons simultaneously, and each of them have a unique Beacon ID, a user-set device name and a password that can be established through registering the beacon. There is also a manual switch that allows one to enable or disable their device's ability to detect signals.

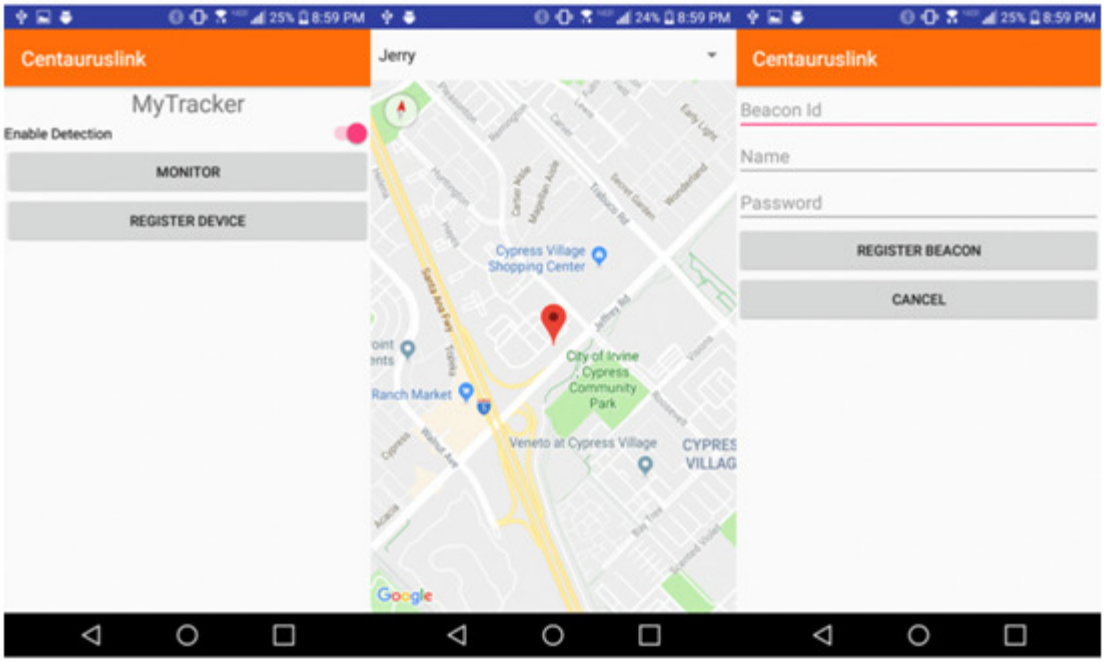


Figure 3. The Screenshots of the mobile application

4. DISCUSSION

To both increase the popularity of this service and improve signal coverage, a rewards mechanism can be put in this system. Since every mobile device that has the app installed could act as a sensor to any nearby beacon, anybody could contribute to locating service "accidentally" even if they are just moving along their daily routine. Everyone who contributes to the ultimate finding of a lost pet, for example, could receive some benefits issued by the pet's owner, such as financial rewards, or even words of gratitude. By incorporating this element of surprise to find something and potentially receive rewards even when one is minding their own business, the public can be encouraged to use this service regardless of whether this technology is of personal importance to them. To improve the security of this system, a sign-in mechanism could be implemented that incorporates registering a device, people would be limited to viewing their own beacons only. And receiving the location of a beacon would be a double-blind process. The owner wouldn't know who contributed to the location, and the locator wouldn't know whose beacon it is that is being located by his or her device. This both improves the security of the exchange of information and reinforces the element of surprise mentioned above.

5. RELATED WORK

iBeacon technology refers to a new generation of low-cost devices which is allowing marketers to track the exact location of consumers via their mobile devices [5-7]. BLE beacons [8] emit a signal that can be received by a BLE-enabled device within a close range. Apps can be built to

cause events to be triggered within an instant of a device coming within the detectable range of the beacon. Moreover, the device can calculate how near or far away it is from the beacon, meaning that different events can be triggered depending on whether a device is within a range. A device can identify numerous beacons simultaneously and, by calculating its relative distance from each of the beacons, the device can gain an element of location awareness.

BLE beacons are important because they overcome challenges, such as secure, proximity-based communication, indoor geo-location, and wide-reaching distribution. It has a variety of application [9-14]. In this paper, we use iBeacon technology based on the technique of crowdsourcing. to solve the problem of lost pets, runaway children or wandering Alzheimer's patients, and similar needs in application.

6. CONCLUSION AND FUTURE WORK

In this project, we proposed a this system is designed to solve the problem of lost pets, runaway children or wandering Alzheimer's patients as well as anyone else who are struggling in a similar condition. By using mobile devices as receiving stations for signals in addition to having physical and designated receivers, the coverage of this service can be greatly improved to benefit the society and people as a whole.

This project also has the potential to become something much more sophisticated. Using such a proximity sensor device, one could request information based on their proximity with a beacon. This specific implementation could be used in areas such as museums for guided tours, showrooms for pricing and item specifics and billboards for further information. Furthermore, analysis could also be done on collected data. For example, one could determine what is a popular amusement park ride, how many customers did a department store have in one hour and so on. The future to this technology is endless and it is important to make it so that it can best benefit our society as a whole.

As for the future work, we will investigate accuracy and efficiency of the system. We also would like to explore the possibility of applying the system on other domains and providing convenience for human being.

REFERENCES

[1] Scarborough, Pet al, Prachi Bhatnagar, Kremlin Wickramasinghe, Kate Smolina, Colin Mitchell, and M. Rayner. "Coronary heart disease statistics 2010 edition." British Health Foundation Health Promotion research group, Department of Public Health, University of Oxford (2010).

[2] Bajaj, Rashmi, Samantha Lalinda Ranaweera, and Dharma P. Agrawal. "GPS: location-tracking technology." Computer 4 (2002): 92-94.

[3] Glenner, George G., and Caine W. Wong. "Alzheimer's disease: initial report of the purification and characterization of a novel cerebrovascular amyloid protein." Biochemical and biophysical research communications 120, no. 3 (1984): 885-890.

[4] Person, NCIC Missing. "Unidentified Person Statistics for 2009." National Crime Information Center, Federal Bureau of Investigation (2014).

[5] Newman, Nic. "Apple iBeacon technology briefing." Journal of Direct, Data and Digital Marketing Practice 15, no. 3 (2014): 222-225.

[6] Koühne, Markus, and Jürgen Sieck. "Location-based services with iBeacon technology." In Artificial Intelligence, Modelling and Simulation (AIMS), 2014 2nd International Conference on, pp. 315-321. IEEE, 2014.

[7] Lin, Xin-Yu, Te-Wei Ho, Cheng-Chung Fang, Zui-Shen Yen, Bey-Jing Yang, and Feipei Lai. "A mobile indoor positioning system based on iBeacon technology." In Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE, pp. 4970-4973. IEEE, 2015.

[8] Ji, Myungin, Jooyoung Kim, Juil Jeon, and Youngsu Cho. "Analysis of positioning accuracy corresponding to the number of BLE beacons in indoor positioning system." In Advanced Communication Technology (ICACT), 2015 17th International Conference on, pp. 92-95. IEEE, 2015.

[9] Yang, Jingjing, Zhihui Wang, and Xiao Zhang. "An ibeacon-based indoor positioning systems for hospitals." International Journal of Smart Home 9, no. 7 (2015): 161-168.

[10] Corna, Andrea, L. Fontana, A. A. Nacci, and Donatella Sciuto. "Occupancy detection via iBeacon on Android devices for smart building management." In Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition, pp. 629-632. EDA Consortium, 2015.

[11] Burzacca, P., M. Mircoli, S. Mitolo, and A. Polzonetti. "“iBeacon” technology that will make possible Internet of Things." (2014): 159-165.

[12] Conte, Giorgio, Massimo De Marchi, Alessandro Antonio Nacci, Vincenzo Rana, and Donatella Sciuto. "BlueSentinel: a first approach using iBeacon for an energy efficient occupancy detection system." In BuildSys@ SenSys, pp. 11-19. 2014.

[13] Akinsiku, Adegboyega, and Divyesh Jadav. "BeaSmart: A beacon enabled smarter workplace." In Network Operations and Management Symposium (NOMS), 2016 IEEE/IFIP, pp. 1269-1272. IEEE, 2016.

[14] Fard, Hadis Kakanejadi, Yuanzhu Chen, and Kyung Kook Son. "Indoor positioning of mobile devices with agile iBeacon deployment." In Electrical and Computer Engineering (CCECE), 2015 IEEE 28th Canadian Conference on, pp. 275-279. IEEE, 2015.

[15] Bhagwat, Pravin. "Bluetooth: technology for short-range wireless apps." IEEE Internet Computing 5, no. 3 (2001): 96-103.

[16] McCurdy, Roger A. "Emergency assistance system using bluetooth technology." U.S. Patent 6,340,928, issued January 22, 2002.

[17] Erasala, Naveen, and David C. Yen. "Bluetooth technology: a strategic analysis of its role in global 3G wireless communication era." Computer Standards & Interfaces 24, no. 3 (2002): 193-206.

[18] Tajika, Yosuke, Takeshi Saito, Keiichi Teramoto, Naohisa Oosaka, and Masao Isshiki. "Networked home appliance system using Bluetooth technology integrating appliance control/monitoring with Internet service." IEEE Transactions on Consumer Electronics 49, no. 4 (2003): 1043-1048.

AUTHOR INDEX