Arslan Brömme, Christoph Busch, Naser Damer, Antitza Dantcheva, Marta Gomez-Barrero,
Kiran Raja, Christian Rathgeb, Ana F. Sequeira, Andreas Uhl (Eds.):
BIOSIG 2021 - 20th International Conference of the Biometrics Special Interest Group

GI-Edition

Lecture Notes
in Informatics

Arslan Brömme, Christoph Busch,
Naser Damer, Antitza Dantcheva,
Marta Gomez-Barrero, Kiran Raja,
Christian Rathgeb, Ana F. Sequeira,
Andreas Uhl (Eds.)

BIOSIG 2021

Proceedings of the 20th International Conference of the Biometrics Special Interest Group

15.–17. September 2021
International Digital Conference

315

Proceedings

GESELLSCHAFT
FÜR INFORMATIK

Arslan Brömme, Christoph Busch, Naser Damer,
Antitza Dantcheva, Marta Gomez-Barrero, Kiran Raja,
Christian Rathgeb, Ana F. Sequeira, Andreas Uhl (Eds.)

BIOSIG 2021

Proceedings of the 20th International Conference of the Biometrics Special Interest Group

15.-17. September 2021
International Digital Conference

Gesellschaft für Informatik e.V. (GI)

Lecture Notes in Informatics (LNI) - Proceedings
Series of the Gesellschaft für Informatik (GI)

Volume P-315
ISBN 978-3-88579-709-8
ISSN 1617-5468

Volume Editors

Arslan Brömme
GI BIOSIG
Gesellschaft für Informatik e.V.
Ahrstraße 45, D-53175 Bonn
arslan.broemme@aviomatik.de

Christoph Busch
Hochschule Darmstadt
Haardtring 100,
D-64295 Darmstadt
christoph.busch@h-da.de

Naser Damer
Fraunhofer IGD
Fraunhoferstraße 5,
D-64283 Darmstadt
naser.damer@igd.fraunhoder.de

Antitza Dantcheva
INRIA Sophia Antipolis
2004 Rue des Lucioles,
F-06902 Sophia Antipolis Cedex
antitza.dantcheva@inria.fr

Marta Gomez-Barrero
Hochschule Ansbach
Residenzstraße 8,
D-91522 Ansbach
marta.gomez-barrero@hs.ansbach.de

Kiran Raja
Norwegian University of Science
and Technology NTNU
NO-7491 Trondheim
kiran.raja@ntnu.no

Christian Rathgeb
Hochschule Darmstadt
Haardtring 100
D-64295 Darmstadt
christian.rathgeb@h-da.de

Ana F. Sequeira
INESC TEC, Campus da Feup
Rua Dr. Roberto Frias,
PT-4200-465 Porto
ana.f.sequeira@inesctec.pt

Andreas Uhl
University of Salzburg
Jakob-Haringer Str. 2,
A-5020 Salzburg
uhl@cosy.sbg.ac.at

printed by Köllen Druck+Verlag GmbH, Bonn

Chairs' Message - 20th anniversary of BIOSIG

Welcome to the annual international conference of the Biometrics Special Interest Group (BIOSIG) of the Gesellschaft für Informatik (GI) e.V.

GI BIOSIG was founded in 2002 as an experts' group for the topics of biometric person identification/authentication and electronic signatures and its applications. For almost two decades the annual conference in strong partnership with the Competence Center for Applied Security Technology (CAST) established a very well known forum for biometrics and security professionals from industry, science, representatives of the national governmental bodies and European institutions who are working in these areas.

The BIOSIG 2021 international digital conference is jointly organized by the Biometrics Special Interest Group (BIOSIG) of the Gesellschaft für Informatik e.V., the Competence Center for Applied Security Technology e.V. (CAST), the German Federal Office for Information Security (BSI), the European Association for Biometrics (EAB), the TeleTrusT Deutschland e.V. (TeleTrusT), the Norwegian Biometrics Laboratory (NBL), the National Research Center for Applied Cybersecurity (ATHENE), the Hochschule Ansbach (HAB), the Institution of Engineering and Technology Biometrics Journal (IET Biometrics), and the Fraunhofer Institute for Computer Graphics Research (IGD). This year's international conference BIOSIG 2020 is once again technically co-sponsored by the Institute of Electrical and Electronics Engineers (IEEE) and is enriched with digital satellite workshops by the TeleTrust Biometric Working Group and the European Association for Biometrics. BIOSIG 2021 is held the the second time as a pure international digital conference due to the global pandemic situation.

The international program committee accepted full scientific papers strongly according to the LNI guidelines (**acceptance rate ~27%**) within a scientific double-blinded review process of at minimum five reviews per paper. All papers were formally restricted for the digital proceedings up to 12 pages for regular research contributions including an oral presentation and up to 8 pages for further conference contributions.

Furthermore, the program committee has created a program including selected contributions of strong interest (further conference contributions) for the outlined scope of this conference. All paper contributions for BIOSIG 2021 will be published additionally in the IEEE Xplore Digital Library.

We would like to thank all authors for their contributions and the numerous reviewers for their work in the program committee.

Darmstadt, 15th September 2021

Arslan Brömme *(GI BIOSIG, GI e.V.)*, **Christoph Busch** *(Hochschule Darmstadt)*, **Naser Damer** *(Fraunhofer IGD)*,
Antitza Dantcheva (*INRIA Méditerranée)*, **Marta Gomez-Barrero** *(Hochschule Ansbach)*, **Kiran Raja** *(NTNU)*,
Christian Rathgeb *(Hochschule Darmstadt)*, **Ana F. Sequeira** *(INESC TEC),* **Andreas Uhl** *(University of Salzburg)*

Chairs

General Chair

Christoph Busch, *Hochschule Darmstadt, Germany*

Marta Gomez-Barrero, *Hochschule Ansbach, Germany*

Program Chairs

Antitza Dantcheva, *INRIA Méditerranée, Sophia Antipolis, France*

Kiran Raja, *Norwegian University of Science and Technology NTNU, Jøvik, Norway*

Christian Rathgeb, *Hochschule Darmstadt, Germany*

Andreas Uhl, *University of Salzburg, Austria*

Naser Damer, *Fraunhofer IGD, Germany*

Publication Chair

Arslan Brömme, *GI BIOSIG, GI e.V., Bonn, Germany*

Publicity Chairs

Victor Philipp Busch, *Sybuca GmbH, Hamburg, Germany*

Ana Filipa Sequeira, *INESC TEC, Porto, Portugal*

Local Chairs

Alexander Nouak, *Fraunhofer IUK, Darmstadt, Germany*

Claudia Prediger, *CAST e.V., Darmstadt, Germany*

Program Committee

Fernando Alonso-Fernandez (HaU, SE)
Patrick Bours (NTNU, NO)
Fadi Boutros (FHG IGD, DE)
Julien Bringer (Smart Valor, FR)
Arslan Brömme (GI BIOSIG, DE)
Christoph Busch (CAST-Forum, DE)
Victor-Philipp Busch (Sybuca, DE)
Patrizio Campisi (RomaIIIU, IT)
Cunjian Chen (MSU, US)
Adam Czajka (WUT, PL)
Naser Damer (FHG IGD, DE)
Antitza Dantcheva (INRIA, FR)
Maria De Marsico (SUoR, IT)
Abdenour Hadid (UoV, FR)
Olaf Henniger (FHG IGD, DE)
Ulrike Korte (BSI, DE)
Ajay Kumar (HKPU, HK)
Davide Maltoni (UoB, IT)
Gian Luca Marcialis (UoC, IT)
Johannes Merkle (secunet, DE)
Amir Mohammadi (Idiap, CH)
Aythami Morales (UAM, ES)
Emilio Mordini (RT, FR)
Abe Narishige (Fujitsu, JP)
Mark Nixon (UoS, UK)
Alexander Nouak (FHG IUK, DE)

Max Fermann (BoNZ, NZ)
Pawel Drozdowski (HDA, DE)
Matteo Ferrara (UoB, IT)
Julian Fierrez (UAM, ES)
Lothar Fritsch (OMU, SE)
Steven Furnell (UoN, UK)
Sonia Garcia (TSP, FR)
Marta Gomez-Barrero (HSAN, DE)
Ester Gonzalez-Sosa (Nokia Bell Labs, ES)
Markus Nuppeney (BSI, DE)
Javier Ortega-Garcia (UAM, ES)
Dijana Petrovska-Delacretaz (TSP, FR)
Kiran Raja (NTNU, NO)
Raghavendra Ramachandra (NTNU, NO)
Kai Rannenberg (GUF, DE)
Christian Rathgeb (HDA, DE)
Heiko Roßnagel (FHG IAO, DE)
Stephanie Schuckers (CU, US)
Günter Schumacher (JRC, IT)
Ana F. Sequeira (ITEC, PT)
Takashi Shinzaki (Fujitsu, JP)
Luis Soares (ISCTE-IUL, PT)
Luuk Spreeuwers (UoT, NL)
Juan Tapia (HDA, DE)
Philipp Terhörst (FHG IGD, DE)
Ruben Tolosana (UAM, ES)
Udo Mahlmeister (TDIS, US)
Andreas Uhl (PLUS, AT)
Markus Ullmann (BSI, DE)
Narayan Vetrekar (GU, IN)
Daishi Watabe (SIT, JP)
James Wayman (SJSU, US)
Andreas Wolf (BDR, DE)
Bian Yang (NTNU, NO)

Hosts

Biometrics Special Interest Group (**BIOSIG**) of the Gesellschaft für Informatik (GI) e.V.
http://www.biosig.org

Competence Center for Applied Security Technology e.V. (**CAST**)
http://www.cast-forum.de

Bundesamt für Sicherheit in der Informationstechnik (**BSI**)
http://www.bsi.bund.de

European Association for Biometrics (**EAB**)
http://www.eab.org

TeleTrusT Deutschland e.V (**TeleTrust**)
http://www.teletrust.de

Norwegian Biometrics Laboratory (**NBL**)
https://www.ntnu.edu/nbl

National Research Center for Applied Cybersecurity (**ATHENE**)
https://www.athene-center.de/

Hochschule Ansbach (**HAB**)
https://www.hs-ansbach.de/en/home

Institution of Engineering and Technology Biometrics Journal (**IET Biometrics**)
http://www.theiet.org/

Fraunhofer-Institut für Graphische Datenverarbeitung (**IGD**)
http://www.igd.fraunhofer.de

BIOSIG 2021 – Biometrics Special Interest Group

“**2021 International Conference of the Biometrics Special Interest Group**”
15th -17th September 2021

Biometrics provides efficient and reliable solutions to recognize individuals. With increasing number of identity theft and misuse incidents we do observe a significant fraud in e-commerce and thus growing interests on trustworthiness of person authentication.

Nowadays we find biometric applications in areas like border control, national ID cards, e-banking, e-commerce, e-health etc. Large-scale applications such as the European Union Smart-Border Concept, the Visa Information System (VIS) and Unique Identification (UID) in India require high accuracy and also reliability, interoperability, scalability and usability. Many of these are joint requirements also for forensic applications.

Multimodal biometrics combined with fusion techniques can improve recognition performance. Efficient searching or indexing methods can accelerate identification efficiency. Additionally, quality of captured biometric samples can strongly influence the performance.

Moreover, mobile biometrics is an emerging area and biometrics based smartphones can support deployment and acceptance of biometric systems. However, concerns about security and privacy cannot be neglected. The relevant techniques in the area of presentation attack detection (liveness detection) and template protection are about to supplement biometric systems, in order to improve fake resistance, prevent potential attacks such as cross matching, identity theft etc.

BIOSIG 2021 addresses these issues and will present innovations and best practices that can be transferred into future applications. Once again a platform for international experts’ discussions on biometrics research and the full range of security applications is offered to you.

Table of Contents

BIOSIG 2021 – Regular Research Papers

BIOSIG 2021 – Further Conference Contributions

BIOSIG 2021

Regular Research Papers

Identical Twins as a Facial Similarity Benchmark for Human Facial Recognition

John McCauley[1], Sobhan Soleymani[1], Brady Williams[1], John Dando[1],
Nasser Nasrabadi[2], Jeremy Dawson[2]

Abstract: The problem of distinguishing identical twins and non-twin look-alikes in automated facial recognition (FR) applications has become increasingly important with the widespread adoption of facial biometrics. This work presents an application of one of the largest twin datasets compiled to date to address two FR challenges: 1) determining a baseline measure of facial similarity between identical twins and 2) applying this similarity measure to determine the impact of doppelgangers, or look-alikes, on FR performance for large face datasets. The facial similarity measure is determined via a deep convolutional neural network. The proposed network provides a quantitative similarity score for any two given faces and has been applied to large-scale face datasets to identify similar face pairs.

Keywords: Facial Similarity, Facial Recognition, Identical Twins, Look-alikes.

1 Introduction

Identical, or monozygotic, twins continue to pose significant challenges to facial recognition (FR) systems. Paone et al. [Pa14] found that the average face recognition system has a significantly higher equal error rate (EER) when presented with a population of identical twins versus a non-twin population. As any facial dataset becomes large, the probability of look-alikes becomes larger, giving a higher likelihood of false matches. However, for identical twins and look-alikes, a high comparison score may not be directly correlated to human-perceived facial similarity due to transformations applied and features extracted by the FR algorithm.

The primary purpose of this work is to develop and apply a similarity measure to evaluate baseline facial similarity between identical twins as a worst-case scenario of similarity in face comparison. This work is motivated by two factors 1) the need to better understand the relationship between facial similarity and the comparison score returned by a FR system (i.e., the difference between FR and facial similarity) and 2) the need for isolating potential look-alikes in large face datasets to estimate the frequency of look-alike occurrence in any dataset.

[1] West Virginia University, Lane Department of Computer Science and Electrical Engineering, Morgantown, West Virginia, USA, {jamccauley, ssoleyma, bwwilliams, jmdando}@mix.wvu.edu

[2] West Virginia University, Lane Department of Computer Science and Electrical Engineering, Morgantown, West Virginia, USA, {nasser.nasrabadi, jeremy.dawson}@mail.wvu.edu

This work will present: 1) an analysis of the performance of one commercial off the shelf (COTS) FR tool and one academic FR algorithm on one of the largest identical twin databases to date, as well as two large non-twin datasets, 2) a convolutional neural network (CNN) based measure for quantifying facial similarity that can inform how facial similarity is related to comparison score, and 3) an application of this similarity measure to determine the baseline similarity of identical twin pairs and identify non-twin look-alikes in a large face dataset.

2 Background

In previous studies of twin recognition, the challenge of distinguishing between two identical twins has been explored for face, fingerprint, and iris modalities [Ne12], [RM13], [BF16], [Su10]. While face recognition is one of the most widely used biometric modalities, it also faces the biggest challenge when presented with identical twins. Two of the earliest studies on the face recognition of twins [Ph11] & [Pr11] found that several COTS face matchers could identify twins when imaging conditions were ideal (i.e., studio lighting, neutral expression), but performance was measurably decreased as imaging conditions were varied [Ph11]. The goal of the work presented herein is not intended to explore the performance of FR systems on identical twins. Instead, this work investigates the related topic of facial similarity using identical twin pairs as a worst-case baseline of facial similarity. Previous studies of facial similarity have drawn an important distinction between face recognition and facial similarity. An early study identified the topic of facial similarity to be distinct from that of face recognition and developed a facial similarity measure based on an Eigenface framework [RCC04]. A recent work from Sadovnik et al. [Sa18] explored a CNN approach to rank similar faces within a dataset, showing evidence that facial similarity is highly related to, but distinct from, facial recognition. This work used hand selected similar face pairs to train a neural network to accomplish the distinct task of facial similarity determination.

Another application of facial similarity determination is the selection of faces for morph generation. Röttcher et al. [RSB20] present a method of determining facial similarity using a variety of factors, showing that their intelligent morph pair selection produces better morphs than random selection.

Accurate identification of look-alikes poses a similar challenge to that of identical twin recognition. Work from Kosmerlj et al. [Ko05] on the effect of look-alikes on a border control FR application found that the system evaluated would not be robust to the occurrence of look-alikes based on their estimated frequency of look-alike individuals. Studies presented in [La11] & [STN15] found FR performance on look-alikes to be very low for feature-based and deep learning algorithms. These works motivate need for the ability to identify potential look-alikes in any given dataset to better evaluate the hardest cases presented to FR systems. Here, we seek to expand upon previous research in this sphere by determining a baseline facial similarity measure for any two faces based on the similarity of identical twins.

2.1 Datasets & Match Performance Analysis

The face image data used in this work comes from multiple sources. The first of these is a twin dataset[1] that contains 2,269 unique identities, 1,438 of which are identical twins. The remaining portion of the dataset is comprised of fraternal twins, relatives to the twin pairs, and non-twin participants. The second dataset is comprised of face images of participants from the general public (i.e., non-twins) with 5,295 unique identities. A third dataset was constructed using the second, non-twin dataset combined with the CelebA dataset [Li15], resulting in a dataset that contains a total of 15,455 unique identities.

The initial task of the work presented here was to analyze the performance of two facial recognition systems on the datasets described above. Two FR tools were used in this experiment, the first of which was a COTS "black box" matcher, and the second was the FaceNet matcher [SKP15], which is based on the Inception-ResNet v1 architecture. The first experiment was a baseline analysis of the matchers when presented with only the identical twin pairs from the twin dataset to determine the effect of highly similar faces on the non-mated distribution of a FR experiment. In addition, a mated comparison was made for each identity to show the relationship between the identical twin non-mated distribution and mated distribution. The mean comparison score of the identical twins in this baseline experiment is used as the threshold for each of the remaining comparison experiments. This score represents an experimental threshold for individuals with high facial similarity, and is used later in this paper to extract potential look-alikes in our dataset for further analysis by the proposed similarity network. For the remaining comparison experiments, the approach presented in Howard et al. [HSV19] was used, wherein all-to-all matching was performed on each of the face datasets retaining only the non-mated, or impostor, distributions. The remaining comparison experiments correspond to each of the face datasets used in this work and are as follows: twin dataset, non-twin dataset, and the combined non-twin and CelebA dataset.

2.2 Similarity Network

To determine a quantitative measure of similarity between identical twins, a deep CNN was implemented. This network was designed with a twin architecture (also known as Siamese architecture) to directly compare two faces. Each half of this network was comprised of a FaceNet architecture, with the weights of the network shared between the two halves, as shown in Figure 1.

[1] This is the first use of the twin dataset; it is available upon request.

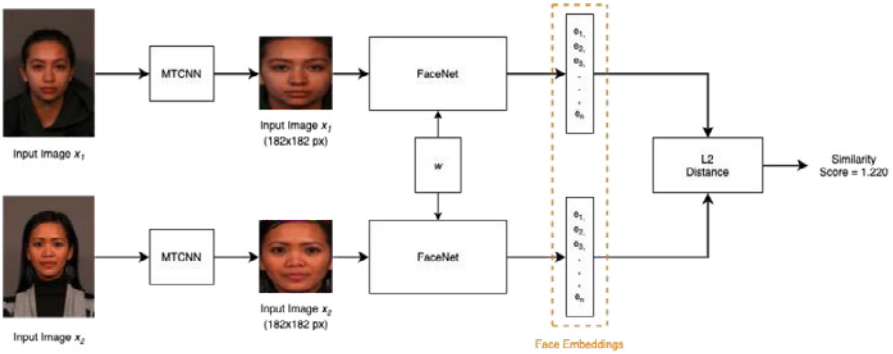

Input Image x_1
MTCNN
Input Image x_1 (182x182 px)
FaceNet
w
e_1, e_2, e_3, ..., e_n
L2 Distance
Similarity Score = 1.220
Input Image x_2
MTCNN
Input Image x_2 (182x182 px)
FaceNet
e_1, e_2, e_3, ..., e_n
Face Embeddings

Fig. 1: Similarity network diagram.

The FaceNet architecture was chosen as the foundation of this network for its high accuracy on typically difficult, similar face images, as well as its ability to generate highly representative face embeddings. In [SKP15], the network is trained to generate embeddings which directly correspond to facial similarity in the embedded feature space. This is advantageous, as our work seeks to quantify the similarity of identical twin pairs rather than simply generating a comparison score. The network was optimized using the contrastive loss function [HSL06] to minimize the L2 distance between similar samples who reside close to one another in the feature space while maximizing the L2 distance between the dissimilar samples. The output of this network consists of the L2 distance between two samples in the feature space. As this calculation gives similar samples a low score and dissimilar samples a high score, for clarity, the scores are inverted such that similar samples have a high score and dissimilar samples a low score. This was achieved by subtracting each resultant similarity score from the maximum similarity score in a given set of scores. The training phase of the network consisted of fine tuning the weights with data from the twin dataset. Starting with the network pre-trained on the VGGFace2 dataset, the network was fine-tuned on a subset of the twin database. This fine tuning was performed on a tailored verification task where a pair of identical twin images represents the positive case, and a pair of unrelated look-alikes represents the negative case. This training encouraged the network to group together those samples with the facial similarity of identical twins in the embedded feature space, and inversely pushed apart those samples not as similar as identical twins.

The training and testing dataset for this network was comprised of a subset of the twin dataset. This dataset contained images of identical twin pairs and non-mated look-alikes to the twin pairs sorted into an equal number of mated and non-mated pairs, where a mated pair consists of (Twin A vs. Twin B), and a non-mated pair consists of (Twin A vs. look-alike). The look-alikes for each identity were found by selecting the identities with the highest FaceNet comparison score to each twin identity. This training schema was chosen because the network should learn to determine facial similarity from the most similar face pairs available (i.e., identical twins). It is expected that an individual's identical twin will be more similar than any potential look-alike, as such, the network is trained to identify the face pairs with the highest facial similarity. The dataset contains 645 identical twin identities, with a total of 3,203 images, split 80/20 for training and testing.

3 Results

3.1 Match Performance Analysis

Figures 2 illustrates the results of the identical twin baseline experiments, indicating that the average comparison score for identical twins trends higher than the comparison score for non-twin matches. The mean comparison score for identical twins in this baseline represents the experimental twin threshold *T*, and, when compared to the mated score distribution, this threshold approximates the left tail of the mated distribution for both matchers tested. All-to-all impostor or non-mated matching was performed using both matchers. The comparison scores for each of these experiments were analyzed to extract the scores falling at and above the experimental twin threshold, *T*, presented in Table 1. In each of the matching experiments, it is shown that an overwhelming majority of comparison scores fall below the experimental twin threshold, indicating that non-mated look-alikes are a rare occurrence in the population used in this study. Due to the relatively small number of identities in the datasets used for this evaluation, it is not possible to accurately predict the frequency of look-alike occurrence in general from these results. However, the similarity measure described in the next section provides a method of finding highly similar faces in any given dataset.

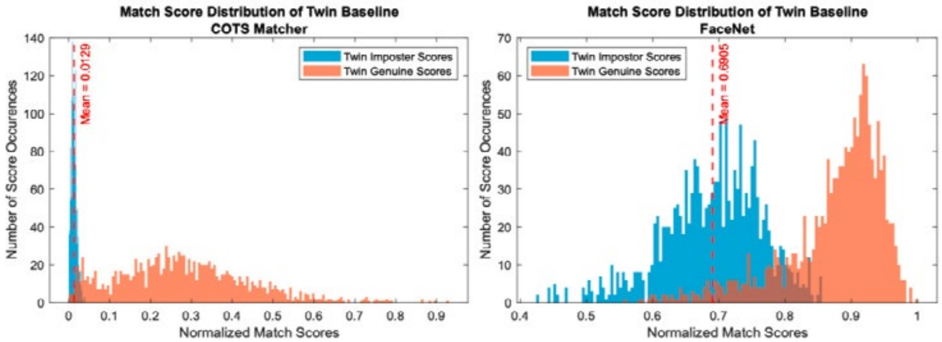


Fig. 2: Twin baseline match experiment results from COTS (left) and FaceNet (right) matchers. The red line shows the mean comparison score of identical twins.

Dataset	Relationship	# Scores >= *T*	Avg. Score	Score Range	% of Matches
Twin Dataset – COTS Matcher	Ident. Twin	713	0.0188	[0.0129-0.0431]	0.0139%
	Family Member	50	0.0197	[0.0135-0.0344]	0.00097%
	No Relation	199	0.0137	[0.0129-0.0189]	0.0038%
Twin Dataset – FaceNet Matcher	Ident. Twin	868	0.746	[0.6905-0.856]	0.0168%
	Family Member	50	0.753	[0.6905-0.83]	0.00097%
	No Relation	6	0.702	[0.694-0.7177]	0.00012%
Non-twin dataset – COTS Matcher	No Relation	16274	0.0144	[0.0129-0.0283]	0.0580%
Non-twin dataset – FaceNet Matcher	No Relation	97	0.704	[0.6905-0.76]	0.000346%
Large Scale Non-twin dataset – FaceNet* Matcher	No Relation	792	0.71	[0.6905-0.76]	0.000331%
	* the large-scale non-twin dataset experiment was performed exclusively on the FaceNet matcher				

Tab. 1: Matching analysis experiments, comparison scores above the twin threshold.

3.2 Similarity Network

After training and testing, the proposed network achieved a train AUC of 0.917, and test AUC of 0.979 in the classification of a pair of face images as a twin pair or look-alike pair. While the end goal of this network is not verification, the accuracy of the network on the tailored verification task shows that the network can accurately identify similar face pairs. This similarity network was then applied to both the twin dataset and large-scale non-twin dataset to observe the general similarity of twin and non-twin individuals. Initially, the similarity score of only the identical twin pairs was calculated (Figure 3). This distribution of similarity scores for identical twin pairs is the foundation of the worst-case baseline measure of similarity. As identical twins exist on a spectrum of similarity, two measurements of the baseline similarity between identical twin pairs are reported. The mean similarity score between identical twin pairs, 1.09, captures the similarity of both highly similar and dissimilar twins, while the fourth quartile of the similarity score distribution, ≥1.29, represents only the most similar twin pairs. In this experiment, the fourth quartile score of the distribution may more accurately represent the worst case of similarity presented to FR systems.

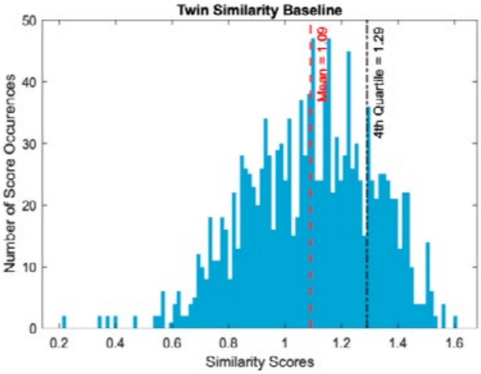


Fig. 3: Identical twin only similarity baseline. The red line shows the mean of the twin similarity distribution, and the black line the fourth quartile of the distribution.

After determining the worst-case baseline for facial similarity, this measure was used to set the threshold for the similarity scores of the large-scale non-twin dataset. Since the network was fine-tuned using only ideal face images, the similarity score returned for "in-the-wild" face images may not be as robust as the similarity score returned for controlled images. Several examples of identical twin pairs and non-mated pairs with similarity scores exceeding the baseline measurements are shown in Figure 4.

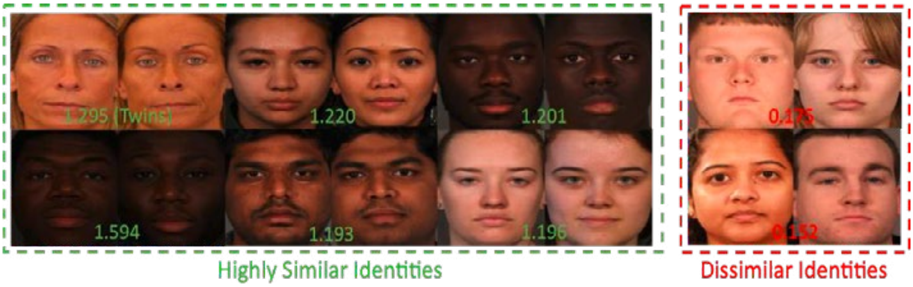


Fig. 4: Examples of identical twin pairs, non-twin look-alikes, and dissimilar face comparisons as determined by the similarity scores from the similarity network.

An additional analysis was performed to correlate the comparison score results of the COTS matcher to the similarity score obtained from our similarity network. Using the non-mated pairs whose comparison scores exceeded the experimental twin threshold in the matching experiments detailed above, a comparison was made to the similarity score calculated for the same pairs. Examples of individuals with high COTS comparison scores and the corresponding similarity score are highlighted below (see Figure 5).

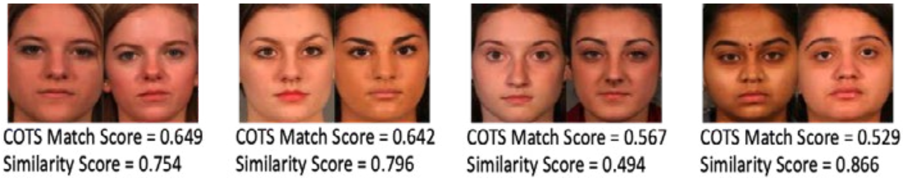


Fig. 5: High COTS comparison score, non-mated pairs and their corresponding similarity score.

As Figure 5 indicates, the COTS comparison score for each of these face pairs was high; however, none of the pairs' similarity scores are above the twin similarity threshold. This indicates that the comparison score returned by the COTS matcher is not directly correlated with facial similarity, and instead, may rely on other features of the image in its comparison score determination process.

Finally, an investigation into the number of potential look-alike pairs returned by the network while varying the similarity threshold was performed to further understand the occurrence of look-alikes in a given population of unrelated individuals (Fig. 6).

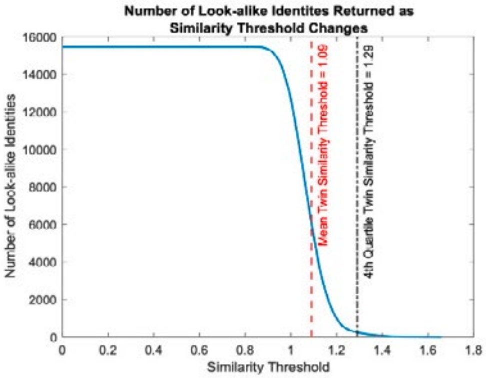


Fig. 6: Number of look-alike identities versus a range of similarity thresholds – large scale non-twin dataset.

Given the mean twin similarity threshold of 1.09, 6,153 of the total 15,455 identities in the large-scale non-twin dataset have at least one similarity comparison at or above the threshold. This means 39.8% of the identities have one or more potential look-alike at this level of similarity. At and above the fourth quartile threshold, only 228 identities have one or more potential look-alike, or 1.475% of identities in the dataset.

4 Conclusion

This work presents an application of one of the largest twin databases to date to better understand the challenges that look-alikes pose to biometric face recognition. Using this dataset, a baseline measure of the worst-case scenario of facial similarity in FR was calculated using a deep CNN. Additionally a performance analysis of two FR tools presented with highly similar faces was carried out, to demonstrate the impact of highly similar faces on FR tools. Using an experimental twin threshold, potential look-alikes were extracted from the datasets for further analysis.

The similarity measure presented here has several applications in FR at large. First, this measure is one way to compare facial similarity to a comparison score from a FR system in order to better understand the impact that facial similarity has on FR. Second, this measure can be directly applied to large face datasets to identify potential look-alikes. Face pairs with high facial similarity can then be identified as difficult cases for a FR system or be used for other applications such as the selection of suitably similar faces for intelligent morph pair generation.

Future work in this area could further explore the relationship between the comparison score returned by a FR tool and the similarity score returned by the proposed similarity network. Another topic of interest in this sphere is a translation of the so called ‘birthday paradox’ to large facial datasets. Much like the birthday paradox seeks to calculate the probability of two people in a given population sharing a birthday, calculating the probability of two unrelated individuals having high facial similarity based on the number of identities in a dataset would be a useful measure as face datasets continue to grow in size. This measure could be used to estimate the number of look-alikes in the population at large, or determine the difficulty of large face datasets.

This material is based upon work supported by the Center for Identification Technology Research and the National Science Foundation under Grant No. 1650474.

References

[Pa14] Paone, J. R. et al.: Double Trouble: Differentiating Identical Twins by Face Recognition. IEEE Transactions on Information Forensics and Security vol. 9 no. 2, S. 285–295, 2014.

[Ne12] Nejati, H. et al.: Wonder ears: Identification of identical twins from ear images. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012). S. 1201–1204, 2012.

[RM13] Ricanek, K.; Mahalingam, G.: Biometrically, How Identical Are Identical Twins?. Computer vol. 46 no. 3, S. 94–96, 2013.

[BF16] Bowyer, K. W.; Flynn, P. J.: Biometric identification of identical twins: A survey. In: 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS). S. 1–8, 2016.

[Su10] Sun, Z. et al.: A study of multibiometric traits of identical twins. Biometric Technology for Human Identification VII vol. 7667, p. 76670T, 2010.

[Ph11] Phillips, P. J. et al.: Distinguishing identical twins by face recognition. In: 2011 IEEE International Conference on Automatic Face Gesture Recognition (FG). S. 185–192, 2011.

[Pr11] Pruitt, M. T. et al.: Facial recognition of identical twins. In: 2011 International Joint

Conference on Biometrics (IJCB). S. 1–8, 2011.

[RCC04] Ramanathan, N.; Chellappa, R.; Chowdhury, A. K. R.: Facial similarity across age, disguise, illumination and pose. In: 2004 International Conference on Image Processing (ICIP). vol. 3, S. 1999-2002, 2004.

[Sa18] Sadovnik, A. et al.: Finding your Lookalike: Measuring Face Similarity Rather than Face Identity. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). S. 2408–24088, 2018.

[RSB20] Röttcher, A.; Scherhag, U.; Busch, C.: Finding the Suitable Doppelgänger for a Face Morphing Attack. In: 2020 IEEE International Joint Conference on Biometrics (IJCB). S. 1–7, 2020.

[Ko05] Kosmerlj, M. et al.: Face recognition issues in a border control environment. Advances in Biometrics - Lecture Notes in Computer Science vol. 3832, S. 33-39, 2005.

[La11] Lamba, H. et al.: Face recognition for look-alikes: A preliminary study. In: 2011 International Joint Conference on Biometrics (IJCB). S. 1–6, 2011.

[STN15] Sun, X.; Torfi, A.; and Nasrabadi, N.: Deep Siamese Convolutional Neural Networks for Identical Twins and look-alike Identification. in Deep Learning in Biometrics, 2015.

[Li15] Liu, Z. et al.: Deep Learning Face Attributes in the Wild. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 3730–3738, 2015.

[SKP15] Schroff, F.; Kalenichenko, D.; Philbin, J.: FaceNet: A Unified Embedding for Face Recognition and Clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). S. 815–823, 2015.

[HSV19] Howard, J. J. ; Sirotin, Y. B.; Vemury, A. R.: The Effect of Broad and Specific Demographic Homogeneity on the Impostor Distributions and False Match Rates in Face Recognition Algorithm Performance. 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS). S. 1-8, 2019.

[HSL06] Hadsell, R.; Chopra, S.; LeCun, Y.: Dimensionality Reduction by Learning an Invariant Mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06). S. 1735–1742, 2006.

Impact of Doppelgängers on Face Recognition: Database and Evaluation

Christian Rathgeb[1], Pawel Drozdowski[1], Marcel Obel[1], André Dörsch[1],
Fabian Stockhardt[1], Nathania E. Haryanto[1], Kevin Bernardo[1], Christoph Busch[1]

Abstract: Lookalikes, a.k.a. doppelgängers, increase the probability of false matches in a facial recognition system, in contrast to random face image pairs selected for non-mated comparison trials. In order to analyse and improve the robustness of automated face recognition, datasets of doppelgänger face image pairs are needed. In this work, we present a new face database consisting of 400 pairs of doppelgänger images. Subsequently, two state-of-the-art face recognition systems are evaluated on said database and other public datasets, including the Disguised Faces in The Wild (DFW) database. It is found that the collected image pairs yield very high similarity scores resulting in a significant increase of false match rates. To facilitate reproducible research and future experiments in this field, the dataset is made available.

Keywords: Biometrics, face recognition, doppelgänger, lookalike, database.

1 Introduction

Face recognition technologies are used in numerous personal, commercial, and governmental identity management systems worldwide. Recent developments in convolutional neural networks have led to remarkable improvements in facial recognition accuracy, surpassing human-level performance [GZ19, Ta14, Ra18]. In particular, state-of-the-art deep recognition systems turn out to be robust against a variety of covariates which may lead to false rejections, such as facial expression [LD20], ageing [BRJ18], or beautification [RDB19].

The improved robustness of said deep face recognition systems may, however, increase the vulnerability against impostors. This has for instance been shown for presentation attacks where an attacker aims at impersonating a target subject by using some attack instrument [MBM18]. In contrast, in a zero-effort impostor attempt, an individual submits their own biometric characteristic while attempting to obtain a successful verification against another subject [IS21]. Previous works reported high success chances for zero-effort impostor attempts in the presence of kin-relationship, in particular for monozygotic, *i.e.* identical, twins [Pr11]. Specific efforts have been devoted to differentiate monozygotic twins in the framework of a facial recognition system, *e.g.* through the analysis of facial marks [Sr12]. It is worth noting that the mentioned effect is far less pronounced for other popular biometric characteristics, *e.g.* fingerprint [Ta12] or iris [DD20].

[1] Hochschule Darmstadt, Germany, contact: `christian.rathgeb@h-da.de`

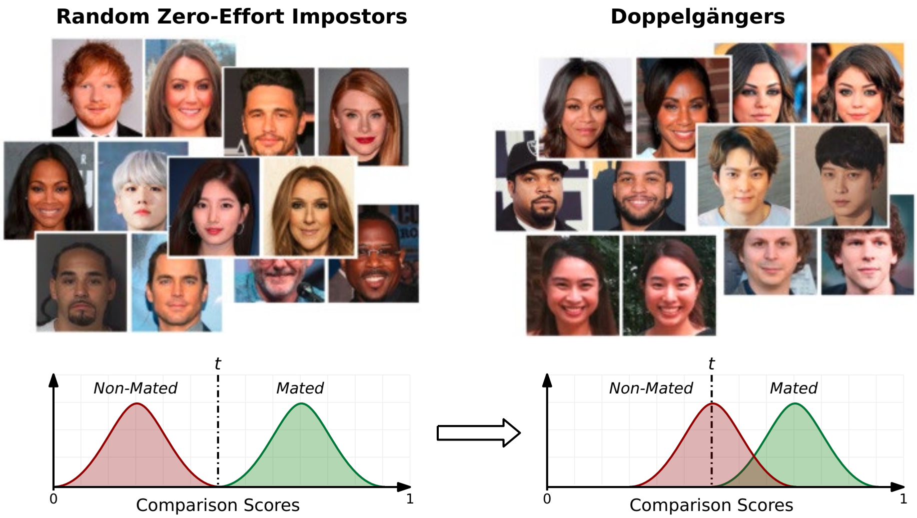


Fig. 1: Random zero-effort impostors (left) achieve low non-mated comparison scores while doppelgängers (right) achieve high non-mated comparison scores and may, if above the decision threshold t, be falsely matched.

In contrast to monozygotic twins, doppelgängers usually refer to biologically unrelated lookalikes. Apart from demographic attributes, doppelgängers also share facial properties such as facial shape. Additionally, some facial properties may further be altered to obtain even higher similarity to a target subject, *e.g.* through the use of makeup [RDB20]. Similar to identical twins, doppelgängers were found to yield high success probabilities compared to random zero-effort impostor attempts, see figure 1. This may lead to serious risks in various scenarios, *e.g.* blacklist checks, where innocent subjects may have a higher chance to match to a lookalike in the list. Lamba *et al.* [La11] presented a preliminary study on the ability of humans and automated face recognition to distinguish lookalikes. Their analysis showed that neither humans nor automatic face recognition algorithms were able to correctly recognise lookalikes. The authors proposed a comparison of facial regions to distinguish lookalikes. Moeini *et al.* [Mo17] suggested to employ 3D reconstruction methods in order to differentiate lookalike faces. To learn highly discriminative facial representations which should also allow to distinguish doppelgängers, Smirnov *et al.* [Sm17] refined the mini-batch selection of a general-purpose face recognition model using a list of lookalikes. Deng *et al.* [De17] introduced the Similar-looking LFW (SLLFW) database, a subset of the Labeled Faces in the Wild (LFW) database, which was selected by human crowdsourcing. It is worth noting that the facial images of LFW are generally unconstrained and of low sample quality. In their Disguised Faces in the Wild (DFW) dataset, Singh *et al.* [Si19] collected facial images which represent challenging face recognition scenarios, including lookalike pairs. More recently, Swearingen and Ross [SR20] presented an approach to improve facial identification performance by re-ranking candidate lists using a lookalike disambiguator which is specifically trained to distinguish between lookalike face images.

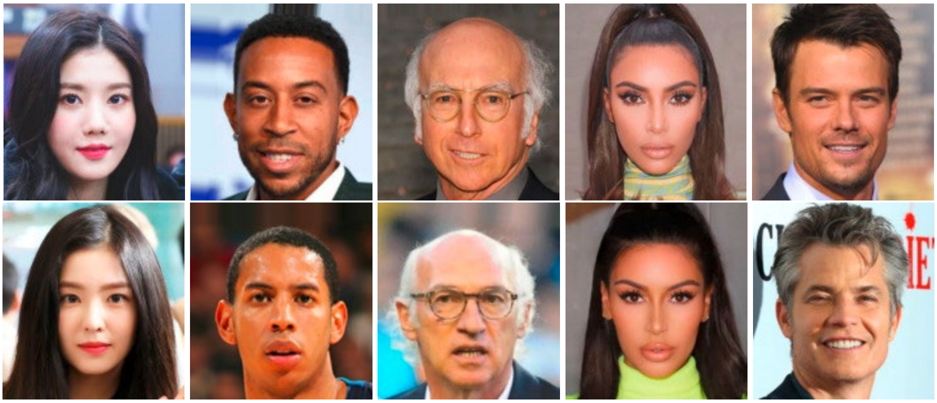


Fig. 2: Example doppelgänger image pairs (column-wise) from the collected database.

In this work, we introduce the *HDA Doppelgänger Face Database* consisting of 400 high quality image pairs (with gender parity), which is made publicly available for the research community upon request[3]. Two face recognition systems are evaluated on this newly collected dataset: the well-known open-source ArcFace system and a Commercial-of-the-Shelf (COTS) system. In experiments, the results obtained on the collected dataset are compared with those achieved for lookalikes in the DFW dataset. The rest of this paper is organised as follows: section 2 describes the collected database. Experiments are presented in section 3. Finally, conclusions are given in section 4.

2 Database

The database introduced in this work was collected from the web using search terms like “lookalike” or “doppelgänger”. A total number of 400 mostly frontal doppelgänger image pairs was collected and manually checked. During the collection, gender parity as well as diversity in other demographic attributes was assured, resulting in 200 male and female image pairs of various age groups and skin colours. Example image pairs of the collected dataset are shown in figure 2. Similarly to the DFW dataset, the majority of facial images are of celebrities.

3 Experiments

In the experiments, we used the newly collected database described in the previous section as well as a subset of the DFW database [Si19] which contains lookalike face image pairs to investigate the success probability of zero-effort impostor attempts of doppelgängers. In addition to these datasets, mated and non-mated comparison trials were obtained from the FRGCv2 face database [Ph05].

[3] HDA Doppelgänger Face Database:
`https://dasec.h-da.de/research/biometrics/hda-doppelgaenger-face-database/`

Tab. 1: Number of comparisons for the used databases and face recognition systems.

Comparisons	**Ours**		**DFW**	
	ArcFace	COTS	ArcFace	COTS
Doppelgänger	397	389	4,353	4,305
Mated	8,883	6,375	894	893
Non-mated	4,998,147	3,664,320	493,521	496,506

Tab. 2: Descriptive statistics of the used databases and face recognition systems.

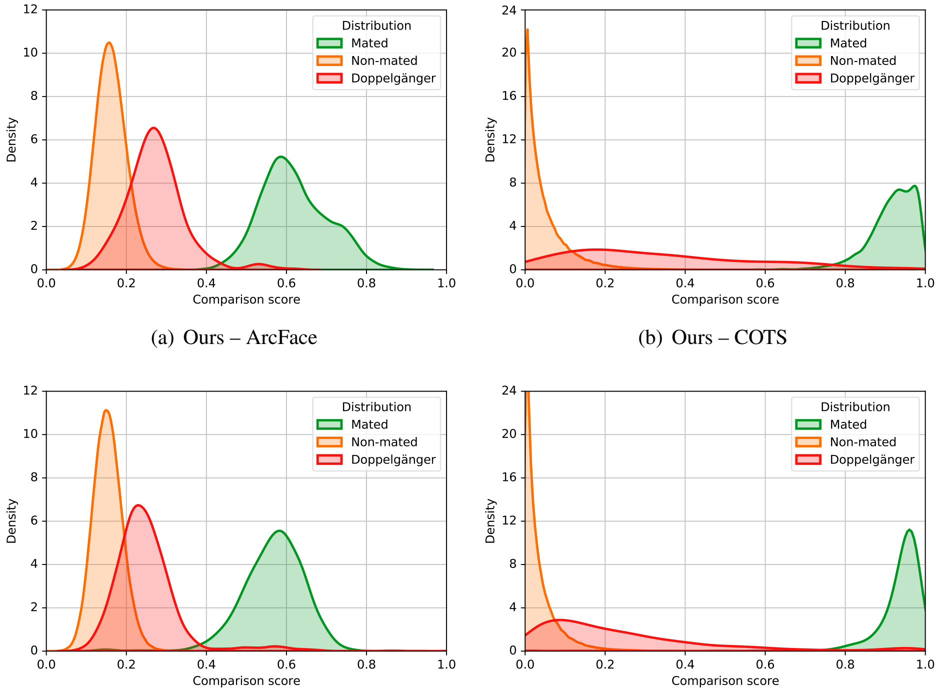
System	**Comparisons**	**Ours**				**DFW**			
		Mean	Std. dev.	Skew.	Ex. kurt.	Mean	Std. dev.	Skew.	Ex. kurt.
ArcFace	Doppelgänger	0.27	0.07	1.12	3.33	0.25	0.08	2.25	9.00
	Mated	0.62	0.08	0.36	-0.23	0.57	0.08	-0.98	3.87
	Non-mated	0.16	0.04	0.36	0.46	0.15	0.04	0.27	0.38
COTS	Doppelgänger	0.34	0.23	0.71	-0.32	0.24	0.22	1.58	2.41
	Mated	0.92	0.06	-1.68	6.13	0.93	0.09	-6.59	57.19
	Non-mated	0.05	0.06	2.63	9.88	0.04	0.05	3.23	17.85

For face recognition, we use a strong open-source system (ArcFace [De19]) with a pre-trained model provided by its authors. ArcFace produces feature vectors of 512 floating-point elements, whose dissimilarity can be computed using the Euclidean distance. For the purposes of visualisation of the results, those dissimilarity scores were mapped into the range $[0,1]$ using min-max normalisation and converted into similarity scores. While the use of this publicly available and well-known tool facilitates reproducibility, an evaluation with a state-of-the-art commercial off-the-shelf (COTS) system was additionally conducted to increase the practical relevance of the obtained results.

Table 1 summarises the number of comparisons (mated, non-mated doppelgänger, and non-mated) for our dataset and the DFW database for both of the employed face recognition systems. For the COTS system, the number of comparisons tends to be smaller since it failed more often in extracting the face embeddings.

Biometric performance is evaluated using metrics standardised by ISO/IEC [IS21, IS17]. Specifically, biometric recognition performance is reported using false match rate (FMR) and false non-match rate (FNMR); the efficacy of doppelgänger impostor attacks is reported by the impostor attack presentation match rate (IAPMR), *i.e.* the fraction of non-mated doppelgänger comparisons resulting in a false match.

Table 2 lists descriptive statistics of the resulting score distributions which are plotted in figure 3. It can be observed that the comparison scores obtained from lookalike face image pairs are generally higher compared to the non-mated scores. Further, it can be seen that for both face recognition systems, the doppelgängers of the collected dataset tend to yield higher comparison scores than those of the DFW database. Moreover, we observe that doppelgänger score distributions exhibit high standard deviations and longer tails, in



(a) Ours – ArcFace

(b) Ours – COTS

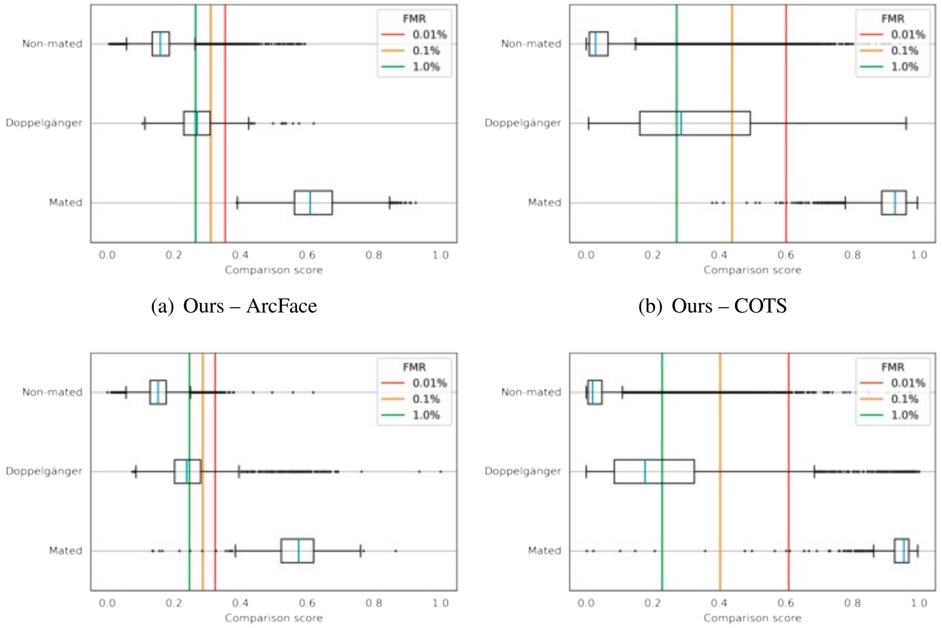
(c) DFW – ArcFace

(d) DFW – COTS

Fig. 3: Probability density functions of scores for both databases and face recognition systems.

particular for the COTS system. That is, some doppelgänger image pairs yield very high comparison scores while the overall distribution is skewed towards the non-mated score distribution. This is further pronounced in the corresponding comparison score boxplots in figure 4 which additionally include decision thresholds obtained from the FMRs. Examples of doppelgängers achieving high comparison scores are shown in figure 5.

Table 3 summarises the performance obtained on both databases in the absence of lookalikes. Here, it can be observed that both face recognition systems obtain competitive recognition performances on both datasets (across the considered, practically relevant [eu15], decision thresholds). The IAPMRs, *i.e.* success chances for doppelgängers, at corresponding decision thresholds are shown in table 4. For a conservative decision threshold, *i.e.* FMR of 0.01%, IAPMRs range from 9.5% to 17% for the collected database for ArcFace and COTS, respectively. As expected based on the analysis of the score distributions, IAPMRs on the DFW database are a bit lower – 6.8% for COTS and 9.6% for ArcFace. For more liberal decision thresholds, *e.g.* FMR of 0.1% or 1%, IAPMRs quickly raise above approximately 25% to 52% for the collected dataset and approximately 17% to 40% on the DFW database. These IAPMR values are alarmingly high and show that the employed face recognition systems are not capable of reliably distinguishing lookalikes. On both



(a) Ours – ArcFace

(b) Ours – COTS

(c) DFW – ArcFace

(d) DFW – COTS

Fig. 4: Boxplots of scores for both databases and face recognition systems.

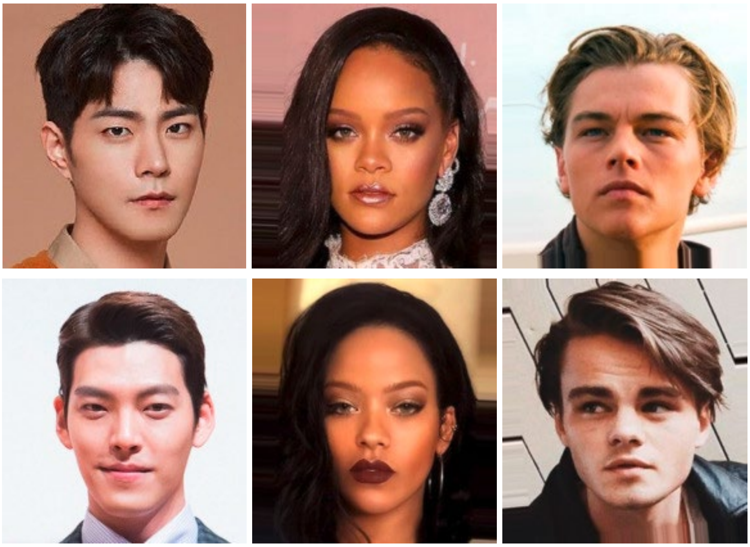


Fig. 5: Example doppelgänger image pairs (column-wise) achieving high comparison scores.

datasets, the obtained IAPMR values are significantly higher than the FMRs expected for random non-mated comparisons.

Tab. 3: Performance rates for both databases and face recognition systems.

Database	**System**	**FNMR at FMR of**		
		1.00%	0.10%	0.01%
Ours	ArcFace	0.00%	0.00%	0.00%
	COTS	0.00%	0.05%	0.17%
DFW	ArcFace	0.56%	0.78%	0.78%
	COTS	0.56%	0.67%	1.12%

Tab. 4: Attack success chance of doppelgängers for both databases and face recognition systems.

Database	**System**	**IAPMR at FMR of**		
		1.00%	0.10%	0.01%
Ours	ArcFace	54.16%	24.94%	9.57%
	COTS	52.44%	29.82%	17.22%
DFW	ArcFace	45.26%	21.59%	9.65%
	COTS	39.70%	17.12%	6.85%

4 Conclusion

Many face recognition evaluation protocols randomly pair face images to obtain non-mated comparisons. Obtained non-mated comparison score distribution may then be used to set up decision thresholds at fixed FMRs. It may be concluded that FMRs (and decision thresholds) obtained in such a way overestimate the security of the underlying face recognition system. Furthermore, one may reasonably argue that zero-effort impostor attacks are less likely to be launched by attackers that look very different from the attacked target subject.

The database of doppelgänger image pairs collected in this work allows for a better estimation of face recognition security w.r.t. zero-effort impostor attacks. It was shown, that a large proportion of doppelgängers contained in our dataset falsely results in a match decision for different state-of-the-art face recognition systems. The collected database is made available to the interested researchers upon request. We believe that this may facilitate improvements in face recognition towards a reliable distinction of lookalikes. Further, we would expect that such improvements would enhance the security of face recognition in general as well as against attacks, *e.g.* presentation attacks.

Acknowledgements

This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[BRJ18] Best-Rowden, L.; Jain, A. K.: Longitudinal Study of Automatic Face Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(1):148–162, 2018.

[DD20] Daugman, J.; Downing, C.: Broken Symmetries, Random Morphogenesis, and Biometric Distance. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2(3):271–278, 2020.

[De17] Deng, W.; Hu, J.; Zhang, N.; Chen, B.; Guo, J.: Fine-grained face verification: FGLFW database, baselines, and human-DCMN partnership. Pattern Recognition, 66:63–73, 2017.

[De19] Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S.: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 4685–4694, 2019.

[eu15] eu-LISA: Best Practice Technical Guidelines for Automated Border Control ABC Systems. Technical Report TT-02-16-152-EN-N, European Agency for the Management of Operational Cooperation at the External Borders of the Member States of the European Union, September 2015.

[GZ19] Guo, G.; Zhang, N.: A survey on deep learning based face recognition. Computer Vision and Image Understanding, 189:102805, 2019.

[IS17] ISO/IEC JTC1 SC37 Biometrics: . ISO/IEC 30107-3. Information Technology – Biometric presentation attack detection – Part 3: Testing and Reporting, September 2017.

[IS21] ISO/IEC JTC1 SC37 Biometrics: . ISO/IEC 19795-1:2021. Information Technology – Biometric Performance Testing and Reporting – Part 1: Principles and Framework, June 2021.

[La11] Lamba, H.; Sarkar, A.; Vatsa, M.; Singh, R.; Noore, A.: Face recognition for look-alikes: A preliminary study. In: International Joint Conference on Biometrics (IJCB). pp. 1–6, 2011.

[LD20] Li, S.; Deng, W.: Deep Facial Expression Recognition: A Survey. IEEE Transactions on Affective Computing, pp. 1–1, 2020.

[MBM18] Mohammadi, A.; Bhattacharjee, S.; Marcel, S.: Deeply vulnerable: A study of the robustness of face recognition to presentation attacks. IET Biometrics, 7(1):15–26, January 2018.

[Mo17] Moeini, A.; Faez, K.; Moeini, H.; Safai, A. M.: Open-set face recognition across look-alike faces in real-world scenarios. Image and Vision Computing, 57:1–14, 2017.

[Ph05] Phillips, P. J.; Flynn, P. J.; Scruggs, T.; Bowyer, K. W.; Chang, J.; Hoffman, K.; Marques, J.; Min, J.; Worek, W.: Overview of the face recognition grand challenge. In: Conference on Computer Vision and Pattern Recognition (CVPR). volume 1. IEEE, pp. 947–954, June 2005.

[Pr11] Pruitt, M. T.; Grant, J. M.; Paone, J. R.; Flynn, P. J.; Bruegge, R. W. Vorder: Facial recognition of identical twins. In: Int'l Joint Conf. on Biometrics (IJCB). pp. 1–8, 2011.

[Ra18] Ranjan, R.; Sankaranarayanan, S.; Bansal, A.; Bodla, N.; Chen, J.; Patel, V. M.; Castillo, C. D.; Chellappa, R.: Deep Learning for Understanding Faces: Machines May Be Just as Good, or Better, than Humans. IEEE Signal Processing Magazine, 35(1):66–83, 2018.

[RDB19] Rathgeb, C.; Dantcheva, A.; Busch, C.: Impact and Detection of Facial Beautification in Face Recognition: An Overview. IEEE Access, 7:152667–152678, October 2019.

[RDB20] Rathgeb, C.; Drozdowski, P.; Busch, C.: Makeup Presentation Attacks: Review and Detection Performance Benchmark. IEEE Access, 8:224958–224973, December 2020.

[Si19] Singh, M.; Singh, R.; Vatsa, M.; Ratha, N. K.; Chellappa, R.: Recognizing Disguised Faces in the Wild. Transactions on Biometrics, Behavior, and Identity Science (TBIOM), 1(2):97–108, March 2019.

[Sm17] Smirnov, E.; Melnikov, A.; Novoselov, S.; Luckyanets, E.; Lavrentyeva, G.: Doppelganger Mining for Face Representation Learning. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). pp. 1916–1923, 2017.

[Sr12] Srinivas, N.; Aggarwal, G.; Flynn, P. J.; Vorder Bruegge, R. W.: Analysis of Facial Marks to Distinguish Between Identical Twins. IEEE Transactions on Information Forensics and Security, 7(5):1536–1550, 2012.

[SR20] Swearingen, T.; Ross, A.: Lookalike Disambiguation: Improving Face Identification Performance at Top Ranks. In: 25th International Conference on Pattern Recognition (ICPR. pp. 1–6, 2020.

[Ta12] Tao, X.; Chen, X.; Yang, X.; Tian, J.: Fingerprint Recognition with Identical Twin Fingerprints. PLOS ONE, 7(4):1–7, 04 2012.

[Ta14] Taigman, Y.; Yang, M.; Ranzato, M.; Wolf, L.: DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In: Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 1701–1708, 2014.

My Eyes Are Up Here: Promoting Focus on Uncovered Regions in Masked Face Recognition

Pedro C. Neto[1,2], Fadi Boutros[3,4], João Ribeiro Pinto[1,2], Mohsen Saffari[1,2], Naser Damer[3,4] , Ana F. Sequeira [1], Jaime S. Cardoso[1,2]

Abstract: The recent Covid-19 pandemic and the fact that wearing masks in public is now mandatory in several countries, created challenges in the use of face recognition systems (FRS). In this work, we address the challenge of masked face recognition (MFR) and focus on evaluating the verification performance in FRS when verifying masked vs unmasked faces compared to verifying only unmasked faces. We propose a methodology that combines the traditional triplet loss and the mean squared error (MSE) intending to improve the robustness of an MFR system in the masked-unmasked comparison mode. The results obtained by our proposed method show improvements in a detailed step-wise ablation study. The conducted study showed significant performance gains induced by our proposed training paradigm and modified triplet loss on two evaluation databases.

Keywords: Face recognition, masked face recognition, Covid-19, triplet loss, vggface2.

1 Introduction

Computer vision tools have been successfully applied to face recognition (FR) in the past [SKP15]. New challenging conditions, such as the face occlusion caused by the use of face masks in public, mandatory during the SarsCov2 pandemic, raised limitations for well-performing and established FR methods. The pandemic has also stressed the importance of hygienic and contactless biometrics [Go21], such as FR. Recently, the National Institute of Standards and Technology (NIST), in the scope of the ongoing Face Recognition Vendor Test (FRVT), published a study on the effect of face masks on the performance of vendor's FR systems (FRVT -Part 6A). The NIST study concluded that the algorithm accuracy with masked faces declined substantially. The Department of Homeland Security has conducted an evaluation with similar goals, however on more realistic data[5]. They also observed the significant negative effect of wearing masks on the accuracy of automatic FR methods.

The lack of robustness of current systems to perform masked face recognition (MFR) fostered an interest in the research community to address this challenge [Di20; Ge20; Ho20; Li21]. Damer *et al.* [Da20] evaluated the verification performance drop in three

[1] INESC TEC, Porto, Portugal, pedro.d.carneiro@inesctec.pt

[2] Faculdade de Engenharia da Universidade do Porto, Porto, Portugal

[3] Fraunhofer Institute for Computer Graphics Research IGD, Germany

[4] TU Darmstadt, Germany

[5] https://mdtf.org/ Rally2020/Results2020

face biometric systems when verifying masked vs not-masked faces compared to verifying not-masked faces to each other. This study was extended [Da21a] to both synthetic and real masks, pointing out the questionable use of simulated masks to represent the real mask effect on face recognition. Furthermore, the performance of human inspectors in face verification has been shown to have a drop, consistent with automatic FR systems, when faces are masked [Da21b]. The effect of facial masks extends to other components of biometric systems as it has been shown to largely change the behaviour of face presentation attack detection [Fa21]. Recently, Boutros *et al.* [Bo21a] proposed a template unmasking approach that can be adapted on top of any face recognition network aiming at creating unmasked-like templates from masked faces by the proposed self-restrained triplet loss. Other initiatives were also incited by the lack of systems capable of handling this task, such as the "Competition on Masked Face Recognition" (IJCB-MFR-2021 [Bo21b]) and "The International Workshop on Face and Gesture Analysis for COVID-19" (FG4COVID19)[6].

The work proposed in this paper comprises the construction of a synthetic masked face dataset based on the VGGFace2 [Ca18] and proposes a solution to address the challenge of MFR. This solution is based on a proposed loss that combines the traditional triplet loss and the mean squared error (MSE) intending to improve robustness in the comparison between masked and unmasked samples.

The contributions of this work are: 1) A cascaded training paradigm that leverages the benefits of both a conventional identity classification learning in the first stage and the subsequent embedding optimization fine-tuning stage. 2) A specifically modified triplet loss function (for the embedding optimization) that incorporates a mean square error measurement to control the training process in a weighted manner. 3) A thorough ablation study on multiple databases (including our own created database), showing, in a step-wise manner, the benefit of our training paradigm and the specifically designed loss.

This paper is organised as follows. In this introductory section, we contextualise the challenge addressed within the related work and detail the contributions of the paper; and in the conclusion, we reflect on the findings and future work possibilities. In Section 2 we present the methodology used. Section 3 details the metrics, the datasets used (the creation of the synthetic face masks and the dataset with real masks), the implementation details, and finally, presents the results and its discussion.

2 Methodology

Performing the direct optimization of embedding predictions is often trickier than learning a model capable of performing classification. Hence, we propose a constrained triplet loss, specially crafted for masked face recognition (MFR), to be used after the classification optimization. Our approach redirects the focus of learned embeddings towards unmasked areas. In the embedding learning stage, the training focuses on comparing two images against a reference (anchor) and distinguishing between images of the same person (positives) and from a different person (negatives). Together, the anchor, the positive, and the

[6] `https://fg4covid19.github.io/index.html`

negative form a triplet. This will enable the model to capture the benefits of both strategies and therefore be more flexible to inter-class variations, as we will experimentally illustrate.

In this section, we start by describing the approach followed to train these models for classification. Afterwards, we expose our proposed change to the triplet loss that improves the representations learned by these models for MFR.

Classification Training: In our scenario, the number of classes and the universe of possible inputs is unknown once the model is deployed. Thus, supervised classification methods can only guide us up to a certain point. Nonetheless, they have been used as a technique to improve the convergence speed of the model for other tasks. Therefore, initially, the problem is approached as a closed-set recognition. Afterwards, the pre-trained model is fine-tuned towards learning meaningful embeddings.

The approach to train the classification model was based on minimizing the cross-entropy (CE). This loss function is frequently used for the classification of the input as a single class, which suits our use case since a picture can only belong to one subject. It attempts to minimize the confidence of the model on erroneous classes while maximizing its confidence in the correct class. The validation occurred after each epoch and it evaluated the accuracy of the model in the classification of masked pictures unseen during the training. And while these images were unseen, the network already knew the subject from past pictures. To separate validation and training sets, we followed an 80%/20% data split. This training process is similar to the one designed by Cao *et al.* [Ca18].

Embedding Optimization: Embedding optimization is a task that requires the network to learn the representations of the inputs instead of classifying them. This is no trivial task, and the hyperparameters search of this process has to be done carefully since this is an expensive task when compared to the use of classification losses [Yu20]. For this, the fully connected layer is removed, and another untrained embedding layer is added. Besides this last layer, all the weights of the network are frozen, and thus, further training does not update them. To train the embedding layer, which outputs an embedding vector with a size of 512, the triplet loss is used, based on Equation 1.

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

$$x_i = W' \frac{\phi(l_i))}{||\phi(l_i)||_2} \tag{2}$$

Equation 2 describes the embedding layer added after the last convolutional layer. It receives as input the output of the convolutional layer (represented by l on the equation) and normalizes it in the euclidean space.

The triplet loss has some aspects that serve our goals, for instance, it relies on three inputs, referred to as the “anchor”, the “positive” and the “negative”, which suits the structure of our evaluation method. Moreover, this loss verifies the distances between the anchor and the positive and between the anchor and the negative. It penalizes the network if the last one is smaller. The formulation of the triplet loss is given by Equation 1, where it is possible to see that it penalizes the model if the distance between the negative and the anchor is shorter than the one of the positive and the anchor. Moreover, the equation includes a term α, which is the margin. The margin, which in this case is set to 0.2, helps the model to define some separability between positives and negatives.

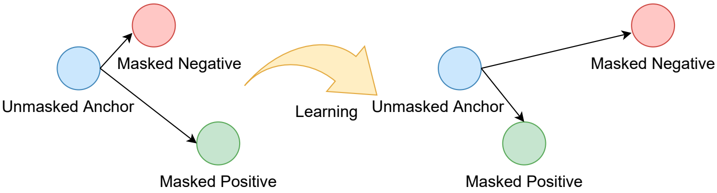


Fig. 1: Triplet loss effect in the euclidean space

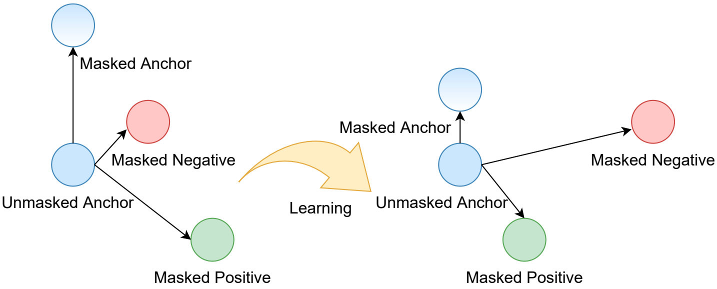


Fig. 2: Proposed Triplet loss effect in the euclidean space

It is possible to see on Figure 1 the effect that optimizing with triplet loss has on the embeddings. The anchor is a randomly selected image (without mask) to be used as a comparison point. The positive image, is a masked image from the same identity as the anchor image, whereas the negative is from a different identity. Our proposed approach, $TripletLoss_{Prop}$, constraints the loss of the original triplets, through the minimization of the distance between the masked and unmasked anchor embeddings. Our loss is formulated in Equation 3 and the effects of its optimization are visible on Figure 2.

$$TripletLoss_{Prop} = \sum_{a,am,p,n} TripletLoss(a,p,n) + MSE(am,a) \tag{3}$$

Since the mask occlusions are always on the same facial area, it is known that the network should focus its attention on areas that will not have occlusions. And thus, as seen on Equation 3, a mean squared error term is added to the loss. This way, we introduce more information to the model so that embedding optimization can be done more effectively.

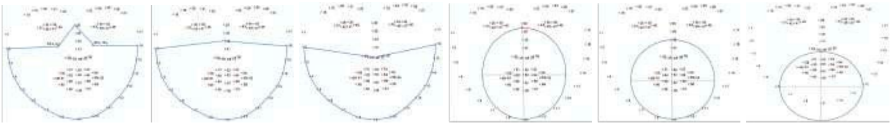
3 Experimental Setup and Results

Evaluation Metrics: To report the results we present the *false non-match rate (FNMR)*; the *FMR100* and *FMR10* which are the lowest *FNMR* for a *false match rate (FMR)* $< 1.0\%$

and $<$ 10.0%, respectively; the *equal error rate (EER)*; and the *area under the receiver operating characteristic curve (AUC)*. We also report the genuine mean (GMean) and impostors mean (IMean), which represent the mean distances between the embeddings of the same individual and from different people respectively.

Face Data: The development of face recognition methods requires large and diverse datasets. When the Covid-19 pandemic started and it became evident that the use of face masks had a negative impact on FR systems there was no ready-to-use data for research. The creation of synthetic data allows leveraging from existing data so that it fits the problem. Still, using real data is crucial as the ultimate test to the models and the community started to also collect face samples of individuals using face masks. The creation of these two types of data used in our method is described as follows.

Synthetic masked face data (SMFD): Here we describe the synthetic masked face data (SMFD) creation process. Adding a facial mask requires information regarding facial landmarks. Moreover, the position and inclination of the face affect the positioning of the mask. Due to the lack of large-scale pairs of masked and unmasked identities, in this work, we synthetically generate different types of masks and adjust them on the unmasked samples of the VGGFace2 dataset [Ca18]. The dataset includes 3,310,000 face images from 8,631 train identities and 500 disjoint identities for the test, and includes a diverse set of samples regarding the various poses, ages, and ethnicity. Mask generation is carried out using the proposed algorithm by NIST [NGH20]. The algorithm exploits the Dlib C++ toolkit to obtain 68 facial landmarks for each image; afterward, using the extracted facial landmarks and interpolation between the points, various synthetic masks are generated. The details of the landmark extraction and mask generation are described in [Ki09; NGH20]. Due to variability regarding shape (Wide vs. Round) and face coverage (High, Medium, Low) we obtained six possible combinations. Figure 3 shows the result of applying the mask generation algorithm on a randomly selected sample from VGGFace2. Both the shape of the mask and its colour are randomly selected, for each image, while generating the masks.

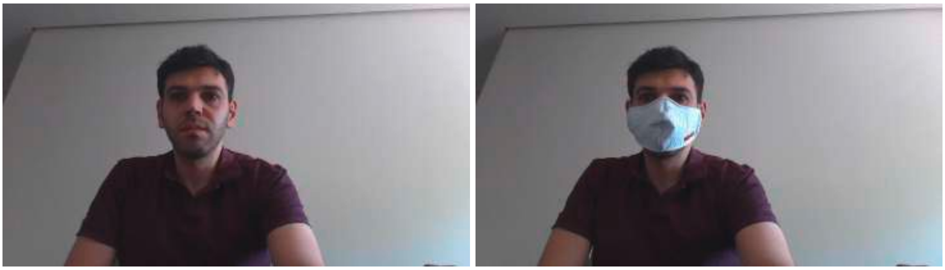


(a) Wide, high (b) Wide, medium (c) Wide, low (d) Round, high (e) Round, medium (f) Round, low

Fig. 3: Examples of face landmarks obtained from one image with different types of masks added. These masks vary in shape and face coverage.

Real masked face dataset: We evaluated our proposed solution using masked face recognition competition dataset (MFRC-21) [Bo21b]. MFRC-21 dataset was collected on 3 different days from 47 subjects. The first day is considered as a reference session, while the second and third sessions are considered probe sessions. In each session, 3 (two with mask and one without) videos are recorded using a webcam, while the subjects are requested to look directly into their camera. An overlapping database, and the same capture and frame selection procedure is describe in [Da20; Da21a]. In total, the references contain 470 unmasked images and 940 masked images. The probes contain 940 unmasked images and

1880 masked images. We evaluate our proposed solution under two evaluation scenarios. The first is between unmasked references and masked probes (U-M) and the second is between masked references and masked probes (M-M).



(a) Example from the test set of a reference image without a mask

(b) Example from the test set of a probe image with a mask

Fig. 4: Examples of images from the real masked faces dataset: images of the same individual with (a) and without a mask (b) (from [Da20]).

Implementation Details: To implement the mask FR model, we exploit ResNet50 architecture [He16] to extract the features from masked/unmasked facial samples. We trained the model by making use of cross-entropy (CE). The training of this model required around 150 thousand iterations. It trained with an initial learning rate of 0.1. It was decreased by a factor of 10 whenever the validation accuracy decreased. Stochastic gradient descent was used, with a batch size of 400, momentum of 0.9, and 0.0005 weight decay. We did not use any face alignment of the VGGFace2 dataset, giving as input of the model 224x224 images. After achieving convergence, it was fine-tuned with triplet loss and the combination of triplet loss with the mean squared error. Triplets were randomly generated at training time, and thus, it was unlikely to have them seen by the network more than once. Furthermore, we did not use any triplet mining. The models trained for 65 thousand iterations with a batch size of 200. And thus, 13 million triplets were created and used for the weight updates. The margin hyper-parameter α in Triplet loss is empirically determined as 0.2.

Results and discussion: We evaluated our method on two distinct datasets. One evaluation used synthetic masked face data (SMFD), with all the identities used for testing being disjoint from the training identities. We also evaluated the model with real masked face data (RMFD). Evaluating on these datasets allows us to infer the generalization capabilities of the model for unknown identities and images with different characteristics from the gallery images (e.g.real masks). The results are provided using the already mentioned metrics: GMean, IMean, AUC, EER, FMR100, and FMR10.

Our method is evaluated through a detailed step-wise ablation study that allowed us to understand the impact of the proposed modification of the triplet loss. Hence, we evaluate the model, in both datasets, with different training frameworks. This allows us to capture information regarding the impact of individual components of the model, such as, training with triplet loss, or not optimizing the embeddings. Besides, we also included the results of the method referred to in the tables as “VGG Face” [Ca18; SKP15] consisting of an Inception-ResNet pre-trained on the original VGGFace2 datasets.

In Table 1, it can be observed that good performance is achieved with just cross-entropy training (CE). Nevertheless, optimizing the produced embeddings with triplet loss (CE+TL) led to significant improvements in performance, lower distances for impostors, and higher distances for genuines. Moreover, our proposed adapted triplet loss resulting from the addition of the MSE constraint (CE+TL+MSE) lead to even more significant improvements, for example, approaching the AUC to 0.99, besides improvements in all the other metrics.

Tab. 1: Results obtained for the synthetic masked face data (EER, FMR100, FMR10 in %).

Method	GMean	IMean	AUC	EER	FMR100	FMR10
VGG Face[Ca18; SKP15]	0.505	0.325	0.951	11.8	38.2	13.5
CE Loss	0.528	0.426	0.941	13.2	38.5	21.5
CE + TL	0.601	0.320	0.977	7.8	28.9	11.9
CE + TL + MSE **(Ours)**	0.596	0.319	**0.985**	**6.2**	**18.5**	**4.1**

In Table 2, can be observed that, the model's performance degrades when compared to the previous table. Regardless of that, for the targeted comparison mode - U versus M - the results show the superior performance of our proposed loss. The distance of impostors increases as the distance of genuines increases too. Furthermore, it is possible to conclude that the model is competent in the task despite being trained only on synthetic data.

Tab. 2: Results obtained for the real masked face data (in the column "Mode": U and M stands for unmasked and masked data, respectively; EER, FMR100, FMR10 in %).

Method	Mode	GMean	IMean	AUC	EER	FMR100	FMR10
VGG Face[Ca18; SKP15]	U-M	0.523	0.426	0.769	29.419	90.587	58.959
	M-M	0.616	0.461	0.847	23.552	68.979	38.159
CE Loss	U-M	0.610	0.475	0.931	11.687	32.041	12.852
	M-M	0.702	0.503	0.936	**9.002**	**16.628**	**8.791**
CE + TL	U-M	0.647	0.396	0.943	11.213	34.744	11.874
	M-M	0.699	0.414	0.945	10.806	26.457	11.249
CE + TL + MSE **(Ours)**	U-M	0.649	0.383	**0.957**	**9.799**	**28.252**	**9.678**
	M-M	0.699	0.390	**0.959**	9.292	23,507	9.035

It should be noted that, the M-M mode experimental results do not keep up with the U-M one, in order words while our method improves the U-M, it offers no improvements for M-M recognition. This can be due to the fact that the embedding optimization process is made in a way that the model is trained to minimise the distance between masked and unmasked (U-M) genuine pairs, thus aiming at making it greater than the distance between imposter pairs. However, this was performed because the main application scenario commonly would contain unmasked references, such as automatic border control with an unmasked passport-stored reference and a possible masked live probe.

Besides the quantitative evaluation of the proposed method, a more qualitative approach was also studied. In order to infer the importance of each pixel for the overall embedding produced we used the Smooth Grad-CAM++method. The output of this gradient-based method was computed for each of the 512 features. Afterwards, all the map outputs were summed and divided by 512, thus generating a final map with the average importance of

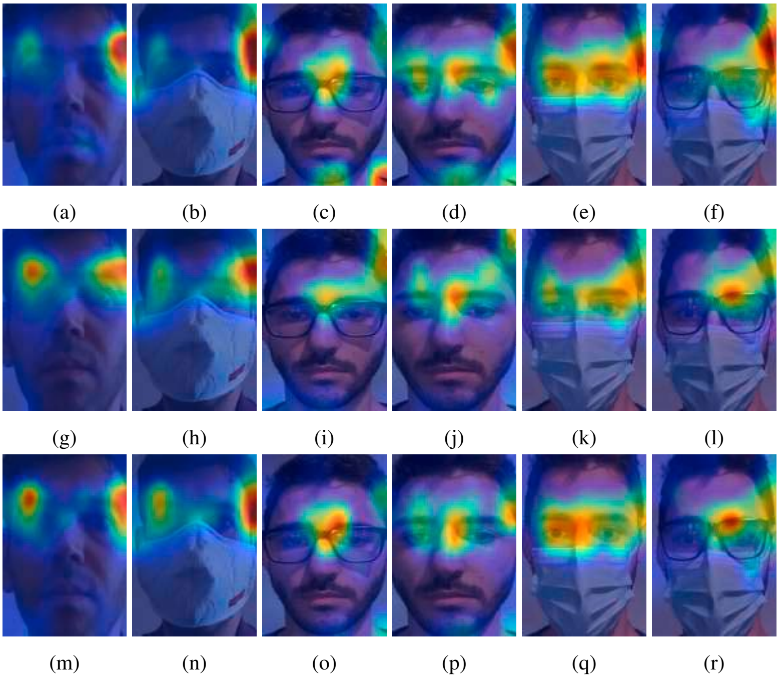


Fig. 5: Output of the Smooth Grad-CAM++ computed for each of the 512 features of the feature vector and normalized. Subfigures (a)-(f) computed from the cross-entropy model; (g)-(l) computed from the triplet loss model; and (m)-(r) computed from the mean squared error and triplet loss model.

the pixels to the overall embedding. Figure 5 displays the outputs for three of the studied methods, from the top to the bottom we have the CE, the TL and the TL+MSE methods. While the first is already capable of ignoring the masks, it still constructs the embedding of the unmasked images based on the chin area. Between the other two models, the main difference seems to be that the model with the MSE uses a wider area of pixels to construct the embedding, thus, capturing more information.

4 Conclusion and Future Work

In this work we addressed the challenge of masked face recognition motivated by the recent Covid-19 pandemic causing that wearing masks is now essential to prevent the spread of contagious diseases and has been currently forced in public places in many countries. However, recent research has shown that the performance, and thus the trust in contactless identity verification through face recognition, can be impacted by the presence of a mask. The scenario addressed is the evaluation of the verification performance in face recognition systems when verifying masked vs not-masked faces compared to verifying not-masked faces to each other. It was already noted in the literature, that the effect of masks was stronger on genuine pairs decisions in comparison to imposter pairs decisions. In this work, we proposed a methodology that targeted that observation and aimed at improving the performance of MFR systems in the comparison of unmasked versus masked faces. The results obtained by our proposed method showed consistent improvements in a detailed step-wise ablation study. The ablation studies performed showed that our proposed triplet loss modification improved the performance of the models in the addressed scenario.

Acknowledgements

This work was financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project UIDB/50014/2020, and within the PhD grants "2021.06872.BD" and "SFRH/BD/137720/2018". This research work has been also funded by the German Federal Ministry of Education and Research and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[Bo21a] Boutros, F.; Damer, N.; Kirchbuchner, F.; Kuijper, A.: Unmasking Face Embeddings by Self-restrained Triplet Loss for Accurate Masked Face Recognition. arXiv preprint arXiv:2103.01716, 2021.

[Bo21b] Boutros, F.; Damer, N.; Kolf, J. N.; Raja, K.; Kirchbuchner, F.; Ramachandra, R.; Kuijper, A.; Fang, P.; Zhang, C.; Wang, F.; Montero, D.; Aginako, N.; Sierra, B.; Nieto, M.; Erakin, M. E.; Demir, U.; Ekenel, H. K.; Kataoka, A.; Ichikawa, K.; Kubo, S.; Zhang, J.; He, M.; Han, D.; Shan, S.; Grm, K.; Štruc, V.; Seneviratne, S.; Kasthuriarachchi, N.; Rasnayaka, S.; Neto, P. C.; Sequeira, A. F.; Pinto, J. R.; Saffari, M.; Cardoso, J. S.: MFR 2021: Masked Face Recognition Competition. In: 2021 IEEE International Joint Conference on Biometrics (IJCB). Pp. 1–10, 2021.

[Ca18] Cao, Q.; Shen, L.; Xie, W.; Parkhi, O.; Zisserman, A.: VGGFace2: A Dataset for Recognising Faces across Pose and Age. In. Pp. 67–74, 2018.

[Da20] Damer, N.; Grebe, J. H.; Chen, C.; Boutros, F.; Kirchbuchner, F.; Kuijper, A.: The Effect of Wearing a Mask on Face Recognition Performance: an Exploratory Study. In: BIOSIG 2020 - Proceedings of the 19th International Conference of the Biometrics Special Interest Group, online, 16.-18. September 2020. Vol. P-306. LNI, Gesellschaft für Informatik e.V., pp. 1–10, 2020.

[Da21a] Damer, N.; Boutros, F.; Süßmilch, M.; Kirchbuchner, F.; Kuijper, A.: An Extended Evaluation of the Effect of Real and Simulated Masks n Face Recognition Performance. IET Biometrics, 2021.

[Da21b] Damer, N.; Boutros, F.; Süßmilch, M.; Fang, M.; Kirchbuchner, F.; Kuijper, A.: Masked Face Recognition: Human vs. Machine.
arXiv preprint arXiv:2103.01924, 2021.

[Di20] Ding, F.; Peng, P.; Huang, Y.; Geng, M.; Tian, Y.: Masked Face Recognition with Latent Part Detection. In: Proceedings of the 28th ACM International Conference on Multimedia. Pp. 2281–2289, 2020.

[Fa21] Fang, M.; Damer, N.; Kirchbuchner, F.; Kuijper, A.: Real Masks and Spoof Faces: On the Masked Face Presentation Attack Detection. CoRR abs/2103.01546, 2021.

[Ge20] Geng, M.; Peng, P.; Huang, Y.; Tian, Y.: Masked Face Recognition with Generative Data Augmentation and Domain Constrained Ranking. In: Proceedings of the 28th ACM Int. Conference on Multimedia. Pp. 2246–2254, 2020.

[Go21] Gomez-Barrero, M.; Drozdowski, P.; Rathgeb, C.; Patino, J.; Todisco, M.; Nautsch, A.; Damer, N.; Priesnitz, J.; Evans, N. W. D.; Busch, C.: Biometrics in the Era of COVID-19: Challenges and Opportunities. CoRR abs/2102.09258, 2021.

[He16] He, K.; Zhang, X.; Ren, S.; Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Pp. 770–778, 2016.

[Ho20] Hong, Q.; Wang, Z.; He, Z.; Wang, N.; Tian, X.; Lu, T.: Masked Face Recognition with Identification Association. In: 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, pp. 731–735, 2020.

[Ki09] King, D. E.: Dlib-ml: A machine learning toolkit. The Journal of Machine Learning Research 10, pp. 1755–1758, 2009.

[Li21] Li, Y.; Guo, K.; Lu, Y.; Liu, L.: Cropping and attention based approach for masked face recognition. Applied Intelligence 515, pp. 3012–3025, 2021.

[NGH20] Ngan, M. L.; Grother, P. J.; Hanaoka, K. K.: Ongoing Face Recognition Vendor Test (FRVT) Part 6B: Face recognition accuracy with face masks using post-COVID-19 algorithms. 2020.

[SKP15] Schroff, F.; Kalenichenko, D.; Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Pp. 815–823, 2015.

[Yu20] Yuan, Y.; Chen, W.; Yang, Y.; Wang, Z.: In Defense of the Triplet Loss Again: Learning Robust Person Re-Identification with Fast Approximated Triplet Loss and Label Distillation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Pp. 1454–1463, 2020.

Interoperability of Contact and Contactless Fingerprints Across Multiple Fingerprint Sensors

Brady Williams[1], John McCauley[1], John Dando[1], Nasser Nasrabadi[2], Jeremy Dawson[2]

Abstract: Contactless fingerprinting devices have grown in popularity in recent years due to speed and convenience of capture. Also, due to the global COID-19 pandemic, the need for safe and hygienic options for fingerprint capture are more pressing than ever. However, contactless systems face challenges in the areas of interoperability and matching performance as shown in other works. In this paper, we present a contactless vs. contact interoperability assessment of several contactless devices, including cellphone fingerphoto capture. In addition to evaluating the match performance of each contactless sensor, this paper presents an analysis of the impact of finger size and skin melanin content on contactless match performance. AUC results indicate that contactless match performance of the newest contactless devices is reaching that of contact fingerprints. In addition, match scores indicate that, while not as sensitive to melanin content, contactless fingerprint matching may be impacted by finger size.

Keywords: Fingerprint Interoperability, Contactless Fingerprint, Finger Size, Palm Color.

1 Introduction

The use of fingerprints for identification and verification has been commonplace for many years in commercial, consumer, and government applications. As technology has advanced, so have the methods for fingerprint collection. From inked fingerprints on paper, to contact-based livescan fingerprinting, to contactless fingerprint imaging, while the image capture process may be different, the resulting fingerprint must still be interoperable in matching against legacy contact galleries. Traditional contact-based digital fingerprints impart some degree of elastic deformation on the finger, and consequently, to the ridges of the fingerprint. Contactless fingerprints pose an interoperability problem as they lack the elastic deformation caused by pressing the finger against the capture device [Li18]. In addition, because they are essentially created from fingerphotos, contactless fingerprints may contain high degrees of photometric distortion that, in addition to the lack of elastic deformation, may further reduce matching interoperability [Li18][Li20][Pr21]. The ubiquitous nature of smartphone cameras and their use in multibiometric capture, as well as the emergence of COVID-19 as a major health crisis,

[1] West Virginia University, Lane Department of Computer Science and Electrical Engineering, Morgantown, West Virginia, USA, {bwwilliams, jamccauley, jmdando}@mix.wvu.edu

[2] West Virginia University, Lane Department of Computer Science and Electrical Engineering, Morgantown, West Virginia, USA, {nasser.nasrabadi, jeremy.dawson}@mail.wvu.edu

have driven the need for fast, hygienic capture of contactless fingerprints, making studies of contactless fingerprint imaging interoperability even more necessary. The overall goal of the work presented here is to evaluate the interoperability of multiple contactless fingerprints when matched against contact fingerprints collected from the same individuals. In addition to this baseline interoperability analysis, physiological factors such as skin color and finger size is evaluated to determine their impact on contactless fingerphoto matching. The contributions of this research effort are: 1) a quantification of the interoperability of contactless fingerprints from two contactless devices and one cellphone-based fingerprint collection method against a traditional contact-based digital fingerprinting device, 2) a measurement of the effect of hand size on the overall matching performance of fingerprints, and 3) an exploration of the effect of skin color measured by skin reflectance on the overall matching interoperability and matching performance of contactless-based fingerprints. The results presented here provide critical insight into the application of contactless fingerprinting systems in a variety of biometric scenarios.

2 Background

Two forms of contactless fingerprints were examined in this effort. The first form is contactless fingerprints captured from a standalone kiosk-type sensor that images the finger when in the field of view of the device (see, e.g., [Li18], [Li20], [Th21], [Id21], [Tb21]). The second form is contactless fingerprints that are captured using a cellphone app that employs the built-in camera to capture fingerphotos [Li18]. The images from the cellphone undergo processing to create a binarized or grayscale fingerprint image that are representative of the original fingerphoto captured from the cellphone camera. To evaluate the interoperability of these fingerprints, two commercial 'black-box' fingerprint matchers will be used, along with one open-source matcher. These three solutions rely on minutiae correspondence as the primary method for matching [Ma14], [JRP04].

While the use and capture of contactless fingerprints are relatively new developments, there has been work done to evaluate and use this form of capture with contact-based fingerprint galleries. NIST has provided recommendations on evaluation of contactless fingerprint devices [Li18]. This study outlines the considerations necessary for proper capture of contactless fingerprints, and how these differ from traditional fingerprints.

In addition to best practices for contactless fingerprint applications, there have been other studies into the interoperability of contactless and contact-based fingerprints [Li20], [Bi17] , [De18]. These studies have shown the challenges and variability issues that are common when collecting contactless fingerprints. To close the interoperability gap between contact and contactless fingerprints, convolutional neural networks (CNN) that use preprocessed versions of both the contact and contactless prints to perform the matching were demonstrated in [LK19]. An alternative CNN-based method presented in [Da19] uses a pair of CNNs to first find the amount of warp on the contactless fingerprint image, and then use that warp parameter to generate a new version of the contactless fingerprint that is representative of a contact-based fingerprint of the same finger.

Because of the nascent nature and methodology of contactless fingerprinting via photo-based capture, physiological features that have little to no impact on contact fingerprint collection, such as finger size and skin color, may negatively impact contactless fingerprint interoperability. However, these features have received little evaluation in the literature in this context. Hand geometry features have been used in biometric verification applications. Hand geometry biometrics rely on the hand shape and various parameters of the hand's size as the features to be extracted and compared [SSG00]. Relating to contactless fingerprints, the variation in finger sizes from person to person may have an impact on contactless matching performance when compared to a gallery of contactless images.

Skin tone, also referred to as skin reflectance, is an important factor to consider in face detection and recognition [BM00]. Variations in skin reflectance, as well as differences in lighting, in facial imagery can have a major effect on the outcome of facial recognition and matching. This is typically not an issue when it comes to contact fingerprints because the method of acquisition is not photo-based. Contactless fingerprints, however, rely on fingerphotos to obtain the ridge and valley information of the fingerprint. As with facial images, variation in skin reflectance could have a significant effect on the matching accuracy of the fingerprint extracted from fingerphotos.

3 Dataset Details and Matching Experiments

The fingerprints used in these experiments were collected from 215 individuals who each provided fingerprint data across multiple commercial fingerprint capture devices[1]. These devices include one contact device, two kiosk-style contactless devices, and a COTS cellphone-based fingerphoto application. At the request of the sponsor, these devices have been anonymized and will be referred to in this paper as Contact-1, Contactless-1, Contactless-2, and Cellphone-1. Contact-1 is an optical livescan device that captures fingerprints via frustrated total internal reflection (FTIR). Contactless-1 and Contactless 2 are both kiosk-style capture devise. Contactless-1 captures fingerprints using multiple cameras and special illumination while Contactless-2 operates using a single camera and structured light approach. The cellphone devices capture fingerphotos using the integrated cameras and utilize app-specific post processing to convert the fingerphoto to a contact-equivalent image. Sample images from each device are shown in Fig. 1.

[1] This is the first use of this dataset. The dataset is available upon request.

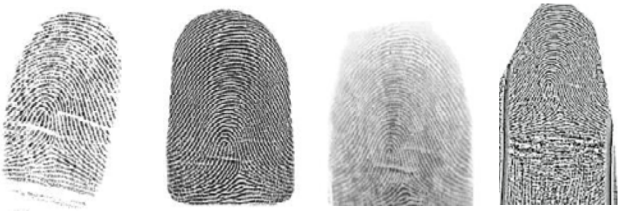


Figure 1: Example of images from contact-1, contactless-1, contactless-2, and cellphone-1

The total number of fingerprint images used in matching experiments was 1,165, consisting of fingerprints from the index, middle, ring, and little fingers only. Thumbs were excluded from the analyses because not every device captured thumbprints. A summary of the number of images from each sensor is provided in Table 1. The dataset also contains finger size data collected from hand geometry images and skin reflectance data measured using the Cortex Technology DSM III sensor [Co21]. Some devices captured images across multiple sessions, with others only capturing one session. In addition, the skin reflectance data collected with the DSM III provided CIEL*a*b* RGB data and a measure of melanin and erythema in the skin [Co21].

Device	Image Type	No. of Samples	No. of Sessions	Total Samples
Contact 1	slaps & rolls	2 slaps 2 thumbs 10 rolls	1	4300
Contactless 1	slaps	2 slaps 2 thumbs	1	2150
Contactless 2	slaps	2 slaps 2 thumbs	2	4300
Cellphone 1	slaps	2 slaps	3	5160

Tab 1: Dataset Description

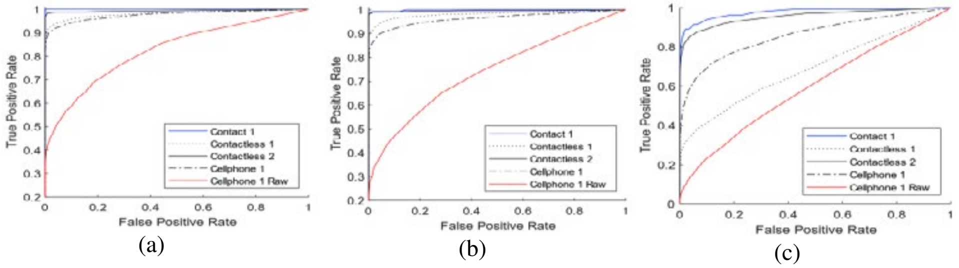
Before matching, preprocessing was performed on the raw versions of the cellphone-based fingerphotos. The photos were converted to grayscale, histogram equalization was applied, and they were inverted so the ridges are shown as the dark regions of the fingerprint to match traditional fingerprinting techniques. These processed photos, referred to as Cellphone-1-Raw, were matched to provide a comparison of the fingerprint processing done by the COTS application in Cellphone-1. Along with the raw photos, the cellphone-based application provided binarized generated prints from the photos that were also used in matching (i.e., Cellphone-1 images).

Using this dataset, matching experiments were performed on two commercial black-box matchers and one open-source matcher with the segmented slap fingerprints from Contact-1 as the gallery for all matches. The two commercial black box matchers and the open-source matcher are referred to as Matcher-1, Matcher-2, and Matcher-3, respectively. All three matchers were used in an ‘out-of-the-box’ configuration, with no optimizations made

for contactless fingerprint images. All matches were performed in a one-to-many fashion so that scores were generated for all probes versus all gallery images. The threshold for all matchers was set to 0 to allow all match results to be extracted. As a baseline for the match scores, the rolled fingerprint data that was collected with Contact-1 was matched against the gallery of segmented slaps used for all other matches. Using the results of these matching experiments, receiver operating characteristic (ROC) curves were generated. This was followed by a statistical analysis of the matching results and a statistical correlation of the finger size and skin reflectance data with the matching results.
The analysis of the impact of finger size on contactless fingerprint match performance was performed using the width of the middle finger of the right hand of all individuals. Using this measurement, the finger sizes were split into equal-sized groups and the mated match scores were sorted into these groups to produce a distribution for analysis. The mated match scores were scores obtained by comparing two fingerprint images collected from the same finger. The analysis for the skin reflectance data involved splitting the data into three equal-sized ranges of melanin value using the melanin value provided by the DSM III. From there, a distribution was generated using the mated match scores to evaluate any effect caused by the amount of melanin on the resulting scores.

4 Results

The results shown in Figure 2 shows ROC curves for the contactless devices compared against Contact-1 as well as the baseline match using Contact-1. Along with the contact baseline, there is a 'worst-case' baseline determined using the preprocessed raw images from Cellphone-1 to show a difference in performance when using the binarized images produced by the cellphone app.



(a) (b) (c)

Figure 2: Receiver Operating Characteristic for contact and contactless fingerprint devices against Contact-1 using (a) Matcher-1, (b) Matcher-2, and (c) Matcher 3.

The results show a clear distinction in match performance between the three devices that is consistent for the three matchers used. As is shown by the area under the curve (AUC) calculated from the ROC curves, shown in Table 2, Cellphone-1 exhibited the worst matching performance out of the three contactless sensors for the first two matchers, but only by a small margin below Contactless-1. Of the contactless images used in Matcher-

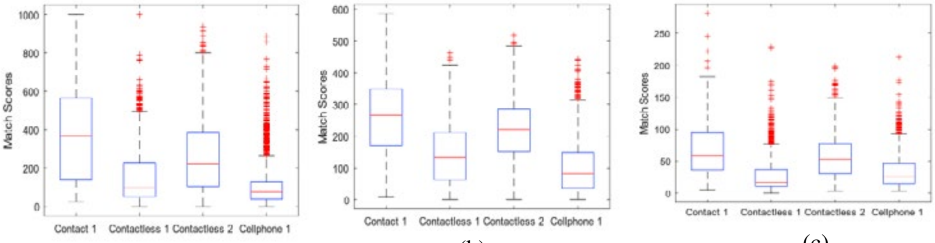
3, Contactless-1 performed the worst with an AUC of only 0.6897.

The match results using Matcher-3 exhibited lower accuracy when compared to the other matchers. All devices performed similarly to the other experiments, except for Contactless-1, which had a much lower matching accuracy, below the performance of images from Cellphone-1. It should be noted that all matchers were used in an 'out of the box' configuration with no optimization for minutiae detection in contactless prints in order to keep the matching results fair.

Device	Matcher-1	Matcher-2	Matcher-3
Contact-1	1.0000	0.9989	0.9765
Contactless-1	0.9818	0.9820	0.6897
Contactless-2	0.9940	0.9955	0.9551
Cellphone-1	0.9764	0.9635	0.8606
Cellphone-1-Raw	0.8252	0.7422	0.5964

Tab 2: AUC of ROC Curves

The results shown in Figure 3 are a comparison of the mated match scores for each of the devices using a specific matcher. In agreement with the ROC curves, the match scores of the two contactless devices trend higher than Cellphone-1, with Contactless-2 achieving the highest match scores.



(a) (b) (c)

Figure 3: Comparison of the distribution of mated match scores for each device using (a) Matcher-1, (b) Matcher-2, and (c) Matcher-3.

The results shown in Figures 4-11 are distributions of mated match scores for each device on all three matchers. Each figure shows the distribution all matchers based on either the melanin values or finger width values. For these distributions, the data is into three bins for each device. These bins separate the data based on the melanin measurement obtained from the skin reflectance data collected from the palm of the subjects or the middle finger width calculated for each hand. The threshold values used for these bins were calculated to split the groups into even ranges of melanin amounts or finger width.

For the melanin distributions, these plots show many outliers; however, the overall average area does not indicate a statistically relevant relationship between melanin content and match score. The plots for the melanin value lower than the first threshold do tend to have more outliers at the top end, however, it is apparent that the vast majority of the match scores fall within a similar range for all of the data. As expected, the contact fingerprint matching data is clearly unaffected by the amount of melanin present. Considering finger width distributions, the middle range of values from 30.99 to 42.17 has the highest-reaching whisker values. In terms of the overall results from this data, Matcher-1 was most affected by finger size for Contactless-1 and Contactless-2, with larger sizes producing higher match scores. For images captured from the other devices, and all images on Matcher-2, there was no noticeable effect of finger width on match scores. For Matcher-3 there was no noticeable effect of the finger width on the matching performance. Along with the width analysis focused on the middle finger, an experiment was also performed using width data for the little finger of the right hand of all participants. The resulting match score distributions showed similar results to the middle finger values, and thus, were not included here.

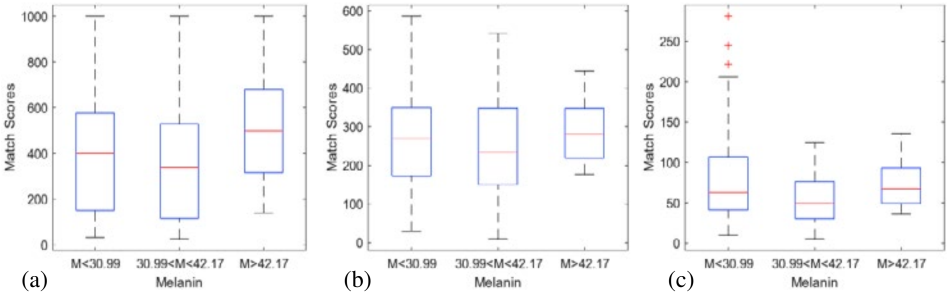


Figure 4: Comparison of the distribution of mated match scores based on melanin amount using probes from Contact-1 and (a) Matcher-1, (b) Matcher-2, and (c) Matcher-3.

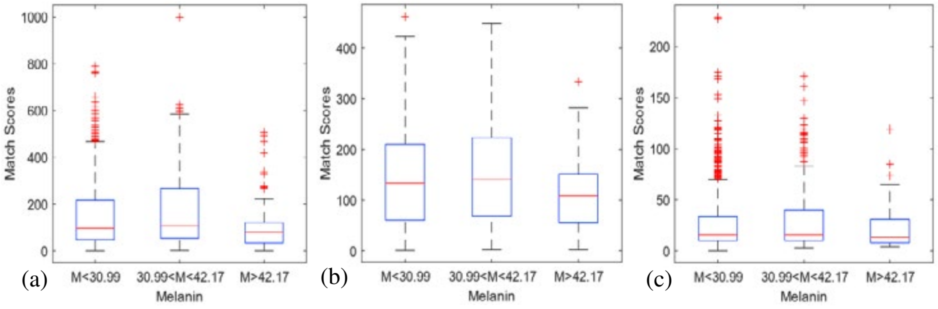


Figure 5: Comparison of the distribution of mated match scores based on melanin amount using probes from Contactless-1 and (a) Matcher-1, (b) Matcher-2, and (c) Matcher-3.

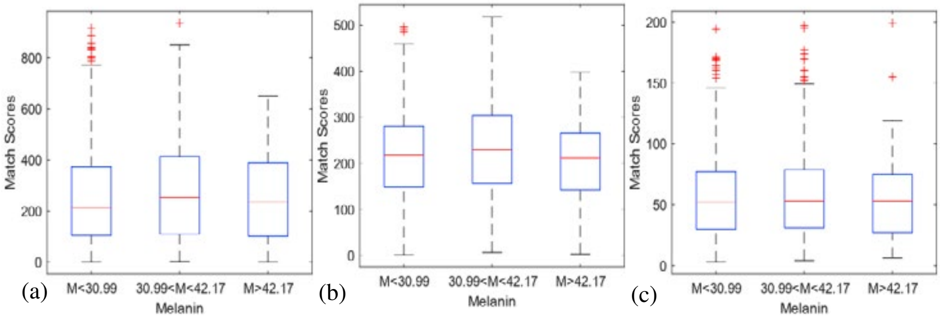


Figure 6: Comparison of the distribution of mated match scores based on melanin amount using probes from Contactless-2 and (a) Matcher-1, (b) Matcher-2, and (c) Matcher-3.

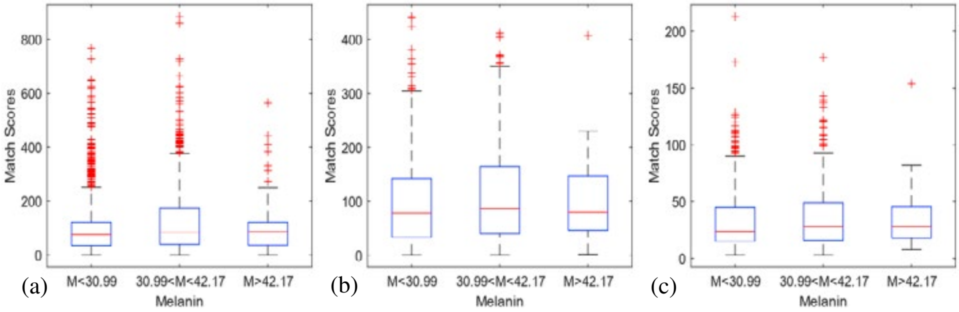


Figure 7: Comparison of the distribution of mated match scores based on melanin amount using probes from Cellphone-1 and (a) Matcher-1, (b) Matcher-2, and (c) Matcher-3.

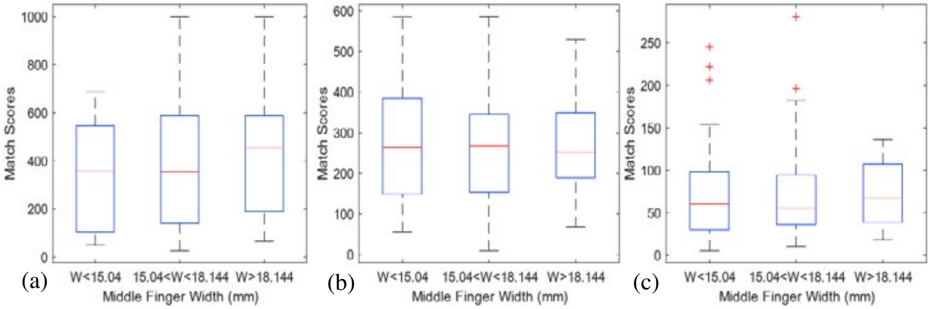


Figure 8: Comparison of the distribution of mated match scores based on middle finger width using probes from Contact-1 and (a) Matcher-1, (b) Matcher-2, and (c) Matcher-3.

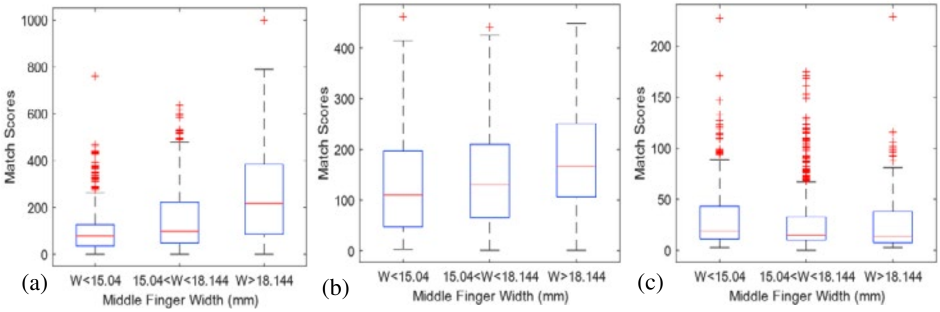


Figure 9: Comparison of the distribution of mated match scores based on middle finger width using probes from Contactless-1 and (a) Matcher-1, (b) Matcher-2, and (c) Matcher-3.

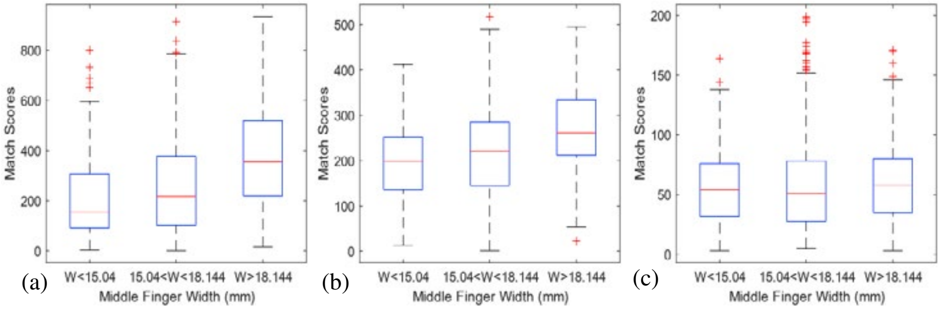


Figure 10: Comparison of the distribution of mated match scores based on middle finger width using probes from Contactless-2 and (a) Matcher-1, (b) Matcher-2, and (c) Matcher-3.

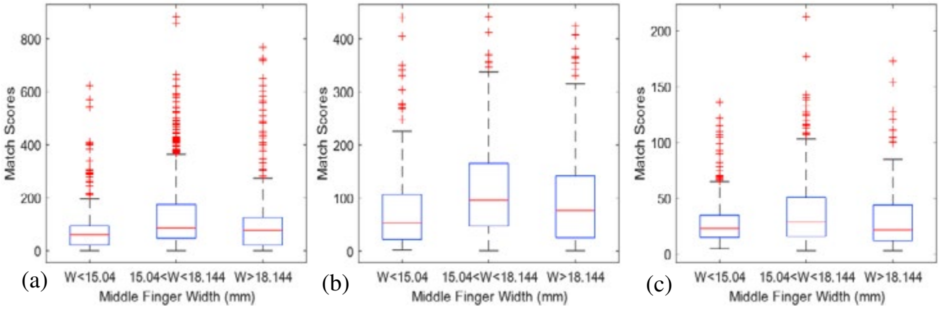


Figure 11: Comparison of the distribution of mated match scores based on middle finger width using probes from Cellphone-1 and (a) Matcher-1, (b) Matcher-2, and (c) Matcher-3.

5 Conclusion

This work explored the interoperability of fingerprints captured from multiple contactless fingerprint devices matched against a gallery of fingerprints captured using a contact-based fingerprint device. Based on the results shown, the Contactless-2 device outperformed both Contactless-1 and Cellphone-1, with an AUC of 0.9818, 0.9955, and 0.9551 for Matcher-1, Matcher-2, and Matcher-3, respectively, with the latter two performing within 0.0176 and 0.0185 of each other for Matcher-1 and Matcher-2, respectively. Using Matcher-3, Contactless-1 fell below Cellphone-1 by a margin of 0.1709. Again, this performance is likely due to the lack of optimization done for Matcher-3. The Cellphone-1 images outperformed the baseline Cellphone-1-raw images by a margin of 0.1512, 0.2213, 0.2642 based on the AUC, as expected.

After the matching analysis was completed, an evaluation of the impact of skin color, collected via skin reflectometer, on match performance was conducted. From this skin reflectance data, a measure of the melanin present in the palm of the subjects was used to split the match scores into groups. This was used to generate new distributions to show the performance for each group. Based on these distributions of the results, there was no perceivable impact across all of the experiments based on statistical significance. This also shows that the contact-based fingerprints were unaffected by melanin content, as was the expected outcome.

A similar analysis was performed for finger size using the width of the middle finger from the right hand of each participant. Again, the data was split into groups based on finger width data, and the match results were used to generate a distribution to convey the performance of the matching based on the various widths. In this case, there was a noticeable effect on the match scores of Contactless-1 and Contactless-2 when using Matcher-1. This effect was not present in either Contact-1 or Cellphone-1 images used as probes to the same matcher, nor was it observed with probes images from any of the devices matched by Matcher-2 or Matcher-3.

Based on the results of this work, it has been shown that contactless fingerprint devices, such as Contactless-2, can achieve a match performance approaching that of contact fingerprints. In comparison to previous work from [Li20], Cellphone-1 with an AUC of 0.9764, 0.9635, and 0.8606 from Matcher-1, Matcher-2, and Matcher-3 respectively outperforms similar cellphone-based device performance. As well, Contactless-1, while not matching the results of Contactless-2, exceeds the match performance results of many of the devices from [Li20] as well.

This material is based upon work supported by the Center for Identification Technology Research and the National Science Foundation under Grant No. 1650474.

6 References

[Bi17] Biller, E.: Interoperability Analysis of Non-Contact Fingerprinting Devices vs. Contact-Based Fingerprinting Devices. Graduate Theses, Dissertations, and Problem Reports, 2017.

[BM00] Brand, J.; Mason, J. S.: A Comparative Assessment of Three Approaches to Pixel-Level Human Skin-Detection. 15th International Conference on Pattern Recognition, Barcelona, Spain, 2000.

[Da19] Dabouei, A. *et al.*: Deep Contactless Fingerprint Unwarping. International Conference on Biometrics, Crete, Greece, 2019.

[De18] Deb, D. *et al.*: Matching Fingerphotos to Slap Fingerprint Image. arXiv preprint arXiv:1804.08122, 2018.

[Th21] *Identification and Authentication within Reach*. (kein Datum). (Thales Group) Abgerufen am March 2021.

[Tb21] *TBS 3D Terminal*. (kein Datum). (TBS) Abgerufen am March 2021.

[JRP04] Jain, A. K.; Ross, A.; Prabhakar, S.: An Introduction to Biometric Recognition. IEEE Transactions on Circuits and System for Video Technology, 14, 2004.

[Li20] Libert, J. *et al.*: Interoperability Assessment 2019: Contactless-to-Contact Fingerprint Capture, National Institute of Standards and Technology.

[Li18] Libert, J. *et al.*: Guidance for Evaluating Contactless Fingerprint Acquisition Devices. National Institute of Standards and Technology, 2018.

[LK19] Lin, C.; Kumar, A.: A CNN-Based Framework for Comparison of Contactless to Contact-Based Fingerprints. IEEE Transactions on Information Forensics and Security, 2019.

[Ma14] Maltoni, D. *et al.*: Handbook of Fingerprint Recognition. Springer, 2014.

[Id21] *MorphoWave Compact*. (kein Datum). (Idemia) Abgerufen am March 2021

[Pr21] Priesnitz, J. *et al.*: An Overview of Touchless 2D Fingerprint Recognition. EURASIP Journal on Imageand Video Processing, 2021.

[SSG00] Sanchez-Reillo, R.; Sanchez-Avila, C.; Gonzalez-Marcos, A.: Biometric Identification Through Hand Geometry Measurements. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(10), 1168-1171, 2000.

[Co21] *Skin Color Meter DSM III*. (kein Datum). (Cortex Technology) Abgerufen am March 2021 *TBS 3D Terminal*. (kein Datum). (TBS) Abgerufen am March 2021

Curricular SincNet: Towards Robust Deep Speaker Recognition by Emphasizing Hard Samples in Latent Space

Labib Chowdhury[1], Mustafa Kamal[2], Najia Hasan[3], Nabeel Mohammed[4]

Abstract: Deep learning models have become an increasingly preferred option for biometric recognition systems, such as speaker recognition. SincNet, a deep neural network architecture, gained popularity in speaker recognition tasks due to its parameterized sinc functions that allow it to work directly on the speech signal. The original SincNet architecture uses the softmax loss, which may not be the most suitable choice for recognition-based tasks. Such loss functions do not impose inter-class margins nor differentiate between easy and hard training samples. Curriculum learning, particularly those leveraging angular margin-based losses, has proven very successful in other biometric applications such as face recognition. The advantage of such a curriculum learning-based techniques is that it will impose inter-class margins as well as taking to account easy and hard samples. In this paper, we propose Curricular SincNet (CL-SincNet), an improved SincNet model where we use a curricular loss function to train the SincNet architecture. The proposed model is evaluated on multiple datasets using intra-dataset and inter-dataset evaluation protocols. In both settings, the model performs competitively with other previously published work. In the case of inter-dataset testing, it achieves the best overall results with a reduction of 4% error rate compare to SincNet and other published work.

Keywords: Biometric Authentication, Speaker Recognition, Angular Margin Loss, Curriculum Learning.

1 Introduction

Speaker Recognition(SR) is widely adopted in real-life scenarios as it has brought remarkable changes in security systems, authentication programs, automated identifications, and forensics. SR is divided into two subtasks: - Speaker Verification (SV) and Speaker Identification (SI). SV involves the comparison of two speech signals and determining whether they belong to the same person. It is simply a validation task where the system is required to indicate whether a speech signal given matches the subject who is being considered. Unlike SV, SI is not a validation task but instead can be considered as a search problem, where given a voice sample of a person, the system attempts to identify the speaker from a list of previously registered speakers.

[1] Department of Electrical & Computer Engineering, North South University, Dhaka, Bangladesh, labib.chowdhury@northsouth.edu

[2] Department of Electrical & Computer Engineering, North South University, Dhaka, Bangladesh, mustafa.kamal@northsouth.edu

[3] Department of Electrical & Computer Engineering, North South University, Dhaka, Bangladesh, najia.tasnim@northsouth.edu

[4] Department of Electrical & Computer Engineering, North South University, Dhaka, Bangladesh, nabeel.mohammed@northsouth.edu

Before the emergence of deep learning in this field, the popular method included the i-vector method [De11], where the features were extracted from MFCC coefficients and Filter-bank Features [Va14], [RD15], [Sn17]. These features are then used in a variety of classifiers, including Probabilistic Linear Discriminant Analysis (PLDA) [PE07] and heavy-tailed PLDA [Ma11]. Numerous recent SR tasks have been based on the popular SincNet [RB18] architecture and as can be appreciated. SR is a very challenging task due to audio signals having a high dimension. Unlike other methods, SincNet can work directly on audio signals because it leverages the parameterized sinc function, which extracts features from audio signals. The deeper layers of the network later process these features.

Biometric systems such as SR and Facial Recognition (FR) can be considered as open-set problems, where the number of classes is not fixed [CM20]. The original SincNet model was trained using the softmax loss [RB18]. Following studies have incorporated various angular margin-based loss functions with SincNet, to achieve better results in both inter-dataset and intra-dataset testing [NZ19], [CM20]. While existing models achieve excellent performance on standard datasets [NZ19], [CM20], the study performed in [CM20] demonstrated that these results do not carry over when performing the inter-dataset evaluation, raising a question about the generalizability of these models. To address this, this study proposes the use of a curriculum learning based loss function and incorporates it with the SincNet architecture. Previously curriculum learning based loss function [Hu20] obtained outstanding performance on biometric tasks such as FR. Influenced by such findings, in this study, we propose Curricular SincNet (CL-SincNet), where we use SincNet architecture as a feature extractor and incorporate curricular loss with it. The contributions of our paper are as follows:

- To the best of our knowledge, we are the first one to introduce curriculum learning applied in the angular space to the speaker recognition task.
- We conducted extensive experiments on two popular speaker recognition datasets, TIMIT and LibriSpeech, and achieve competitive performance on both. In the case of LibriSpeech, we do better than previously published studies. In fact, our approach reduces the frame error rate by 17% in intra-dataset testing.
- Most significantly, we find our proposed approach achieves a lower Classification Error Rate (CER), compared to previously published models in inter-dataset testing. In fact, our proposed approach reduces the CER by 4% when trained on LibriSpeech and tested on TIMIT, thus indicating the better generalizability of our approach.

2 Background Study

This section includes a brief discussion of some background of the softmax loss function and the later part of the discussion includes the SincNet architecture.

Softmax loss is usually defined as the pipeline combination of the last fully connected layer, softmax function, and cross-entropy loss [Wa18]. Softmax loss can be formulated as:

$$L_{softmax} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{W_k^T f_i + b_k}}{\sum_{c=1}^{C} e^{W_c^T f_i + b_c}} \tag{1}$$

Here, f_i denotes the feature vector from last fully connected layer, W_k represents the kth row of weight matrix W and b_k, b_c are the bias scalar value of respective index value k and c. C is the total number of classes and the number of training samples in a mini-batch is N. By setting bias, $b_k, b_c = 0$ and ensuring W_k^T and f_i are unit norm, equation 1 can be rewritten equation 2

$$L_{softmax} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{s\cdot\cos\theta_k}}{e^{s\cdot\cos\theta_k} + \sum_{c=1,c\neq k}^{C} e^{s\cdot\cos\theta_c}} \tag{2}$$

Here s is rescaling parameter and θ_k is angle between weight vector W_k and feature vector f_i. Softmax loss in equations 1and 2 result in a decision boundary between two classes without having any margin being imposed [Wa18]. However for open-set problems, particularly in biometric recognition areas, margin-based loss functions in particular angular margin-based loss functions have obtained superior and encouraging results [De19], [Hu20].

To this end, authors of [De19] proposed arcface loss function that mitigates the issue with softmax loss by imposing a margin in angular space, thus creating more robust and larger decision boundaries between classes. The formulation of [De19] is as follows:

$$L_{ArcFace} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{s\cos(\theta_{k,i}+m)}}{e^{s\cos(\theta_{k,i}+m)} + \sum_{c=1,c\neq k}^{C} e^{s\cos(\theta_c,i)}} \tag{3}$$

Where authors added an additional margin with the angle between the target weight vector and the feature vector and then rescale the feature by multiplying with s. Although this loss function is verified to obtain good performance [De19] it does not consider each sample's difficulties into consideration [Hu20].

The authors of [Hu20] proposed a new loss function where they leverage curriculum learning and introduce a modulation coefficient in the negative cosine similarity. Authors defined positive cosine similarity as $\cos(\theta_k + m)$, which is same as [De19] but they changed the representation of negative cosine similarity from $\cos\theta_j$ to $N(t, \cos\theta_j)$. The loss is defines as follows:

$$L_{CurricularLoss} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{s(\cos(\theta_{k,i}+m)}}{e^{s(\cos(\theta_{k,i}+m)} + \sum_{c=1,c\neq k}^{C} e^{sN(t,\cos\theta_c)}} \tag{4}$$

The modulation coefficient function $N(t, \cos\theta_c)$ is defined as in [Hu20]

$$N(t, \cos\theta_j) = \begin{cases} \cos\theta_j, & \cos(\theta_k + m) > \cos\theta_j \\ \cos\theta_j(t + \cos\theta_j), & \cos(\theta_k + m) < \cos\theta_j \end{cases} \tag{5}$$

According to equation 5 a sample is considered to be easy if the angle between the embedding vector and the target weight vector plus the margin is still smaller than the angle between the embedding vector and the weight vector of non-ground truth classes. At the beginning of the training, the hyper-parameter t should be closed to zero so that the model can emphasize the easy samples first, gradually t will increase and the model will focus on the hard example. Since t will increase, the hard sample will be emphasized with larger weights in the later part of the training. the value of t is adaptive in the loss function, as the training goes the estimate of t is formulated as:

$$t^{(k)} = \alpha r^{(k)} + (1-\alpha)t^{(k-1)} \tag{6}$$

Here, $r^{(k)}$ is the average of positive cosine similarity of k-th batch, defined as $r^{(k)} = \sum_i \cos\theta_{y_i}$, α is a momentum parameter, the author from [Hu20] defined $\alpha = 0.99$.

3 Proposed model Curricular SincNet

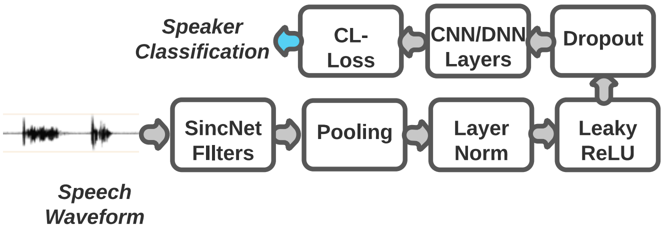


Fig. 1: Portrayal of our proposed architecture

Recent work such as [RB18], [NZ19], [CM20] improved significantly in SI and SV tasks and reports competitive results. Recently [NZ19], [CM20] used the SincNet architecture and incorporated margin-based losses. The initial convolution operation is performed by using parameterized sinc function to extract low-level features from the raw audio waveform and the only learnable parameters of the convolutional filter are the high and low cutoff frequencies. The convolution operation is expressed in equation 7

$$y[n] = x[n] * g[n,\theta] \tag{7}$$

Where $x[n]$ represents a chunk of audio/speaker's signal, $y[n]$ is the filtered output. g is a predefined function and it is defined as equation 8, where a_1 & a_2 represents the low and high cutoff frequencies

$$g[n,a_1,a_2] = 2a_2\frac{\sin(2\pi a_2 n)}{2\pi a_2 n} - 2a_1\frac{\sin(2\pi a_1 n)}{2\pi a_1 n} \tag{8}$$

It has been verified that introducing margin-based losses as a last layer of SincNet, helps to increase the distance between classes and decrease the intra-class distance in the embedding space [CM20], [NZ19]. Although the previously mentioned work showed competitive results, they did not consider each sample's difficulties during the training time.

To address this, in this study we propose Curricular SincNet aka CL-SincNet. We use the same feature extractor as others which is SincNet and we use the curricular loss function as outlined in equation 4 to optimize the network. The motivation is that this approach can allow the model to learn features by allowing the network to learn easy samples first and the harder samples later in the training loop. The graphical representation of our CL-SincNet is shown in figure 1.

4 Experimental Setup

This section describes the details of datasets we use to train and evaluate our model, training and testing procedures, and the metrics we use to measure the model's performance.

4.1 Datasets

We consider TIMIT [Ga93] and LibriSpeech [Pa15] for training and evaluation our model. These two datasets are very popular in SR tasks. TIMIT dataset has 462 classes, or unique speakers, and each class has 8 samples. LibriSpeech has a total 2484 number of classes/unique speakers with a total number of 21933 samples. We used 12-15 seconds of audio for training and 2-6 seconds for testing.

4.2 Training & Testing Procedure

For the training procedure, we use similar settings as [RB18] except for the last layer. We discarded this layer, and instead used the output of the previous layer. We normalized both the feature vector and the row vector of weights by L2 normalization and calculated the cosine similarity of the easy sample and hard sample with corresponding labels. For the two hyperparameters in equation 4 we used the same value as used in [Hu20], which is $m = 0.5$ and $s = 64$. To train the model we use a mini-batch of 128 and the learning rate is set to 10^{-2}. To evaluate our model, we use the same two protocols as [CM20] i.e - Intra dataset test and Inter dataset test. An intra-dataset test is performed to evaluate the Speaker Verification performance and an inter-dataset test is performed to evaluate the Speaker Identification performance. All codes are available at github's project repository[3].

For SV we use Frame Error Rate (FER) and Classification Error Rate (CER) in percentage to demonstrate the performance of our proposed model. These are widely used metrics to measure the performance in SR-based task [RB18], [NZ19], [CM20]. FER is calculated over a window of 200 ms while CER is calculated by averaging the posterior probabilities computed at each frame of the sample and voting for the speaker with the highest average probability. We also use CER(%) in inter-dataset testing for SI task.

The motivation of using the aformentioned metrics is, to demonstrate a fair comparison with recently published works [CM20], [NZ19], [RB18].

[3] `https://github.com/jongli747/Curricular-SincNet`

5 Results

In this section, we discuss our results in two parts. Initially, we speak about speaker verification tasks in the intra-dataset protocol, and then we will talk about speaker identification in an inter-dataset protocol.

5.1 Speaker Verification

	FER(%)		CER(%)	
Configuration	**TIMIT**	**LibriSpeech**	**TIMIT**	**LibriSpeech**
SincNet [RB18]	47.38	45.23	1.08	3.2
AM-SincNet [NZ19]	28.09	44.73	0.36	6.1
AF-SincNet [CM20]	**26.90**	44.65	**0.28**	5.7
Ensemble-SincNet [CM20]	35.98	45.97	0.79	7.2
ALL-SincNet [CM20]	36.08	45.92	0.72	6.4
CL-SincNet (Ours)	37.36	**27.63**	1.08	**0.64**

Tab. 1: Comparison of FER(%) and CER(%) evaluation for both TIMIT and LibriSpeech

As we mentioned earlier in section 4.2 we used FER and CER in percentage as evaluation metrics for SV task(intra- dataset test protocol). Table 1 presents the FER and CER obtained by our model on the TIMIT and LibriSpeech datasets. The performance reported in the previously published models is also shown for comparison. From table 1 we can see that on the LibriSpeech dataset, our proposed model outperforms previously published methods with a significant reduction of FER and CER. In FER, we can see that our proposed model not only outperforms the previously published model, but we have also achieved at least 17% less error rate on the LibriSpeech dataset. Moreover, in terms of CER on the LibriSpeech dataset, our proposed approach has achieved the lowest error rate of 0.64%, reducing the CER by 2.5% in the speaker verification task. Although our model does not show better performance than [NZ19], [CM20] on TIMIT, it is worth mentioning that, TIMIT is a comparatively smaller dataset than LibriSpeech.

5.2 Speaker Identification (Inter-Dataset Evaluation)

For the SI task, we usually compare $x : n$, where x is the given speaker's sample and n is a set of registered lists of speakers. Usually, Cosine Similarity or Euclidean Distance is used for evaluation, this study considered Cosine Similarity.

We have adopted the protocol of inter-dataset testing from [CM20] for the evaluation of the SI task, where the model is trained on one dataset and tested using a different independently collected dataset. This is a good indicator of the generalizability of a model. Table 2 presents the CER obtained by our model on the TIMIT and LibriSpeech datasets. For the sake of simplicity, we refer to the TIMIT trained LibriSpeech tested model as protocol-1 and LibriSpeech trained TIMIT tested model as protocol-2. The first two columns represent the result of protocol-1, and the last two columns represent the results of protocol-2.

Protocol-1		**Protocol-2**	
Configuration	**CER (%)**	**Configuration**	**CER (%)**
SincNet[RB18]	10.09	SincNet[RB18]	10.94
AM-SincNet[NZ19]	9.39	AM-SincNet[NZ19]	13.10
AF-SincNet[CM20]	9.14	AF-SincNet[CM20]	10.83
Ensemble-SincNet[CM20]	8.10	Ensemble-SincNet[CM20]	12.87
ALL-SincNet[CM20]	7.15	ALL-SincNet[CM20]	10.72
CL-SincNet(Ours)	**6.39**	CL-SincNet(Ours)	**6.06**

Tab. 2: Comparison of interdataset evaluation for both TIMIT and LibriSpeech

It is worth mentioning that no samples or classes are overlapped between the two datasets. Our proposed CL-SincNet outperforms the previously published model in both settings. At protocol-1, our proposed model has achieved 0.8% less error rate than compared to previously published work. Most significantly in protocol-2, our proposed CL-SincNet has achieved a 6.06% error rate which is a reduction of more than 4% error rate than compared to previously published works. As we have mentioned earlier, the TIMIT dataset is small than the LibriSpeech dataset, which may be a reason why an improvement in TIMIT is less significant. Tables 1, 2 indicate that our proposed CL-SincNet has the capacity to generalized better than other published models with significant performance improvements.

6 Conclusions

This study has proposed Curricular SincNet (CL-SincNet), where we leverage an angular margin-based curriculum learning loss function on the SincNet architecture for the speaker recognition task. The proposed CL-SincNet has manifested superior results compared to previously published studies [RB18], [NZ19], [CM20]. Our proposed approach reduces the frame error rate by 17% on the LibriSpeech dataset for speaker verification tasks in intra-dataset test protocol and reduces the 4% classification error rate in inter-dataset testing for speaker identification tasks. The results indicate that introducing such a curriculum learning-based loss function on SincNet architecture can have positive outcomes for open-set biometric recognition systems.

References

[CM20] Chowdhury, Labib; Zunair, Hasib; Mohammed, Nabeel: Robust Deep Speaker Recognition: Learning Latent Representation with Joint Angular Margin Loss. Applied Sciences, 10:7522, 2020.

[De11] Dehak, N.; Kenny, P. J.; Dehak, R.; Dumouchel, P.; Ouellet, P.: Front-End Factor Analysis for Speaker Verification. Trans. Audio, Speech and Lang. Proc., 19(4):788?798, May 2011.

[De19] Deng, Jiankang; Guo, Jia; Xue, Niannan; Zafeiriou, Stefanos: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4690–4699, 2019.

[Ga93] Garofolo, John S.; Lamel, Lori; Fisher, William M.; Fiscus, Jonathan G.; Pallett, David S.; Dahlgren, Nancy L.: DARPA TIMIT:: acoustic-phonetic continuous speech corpus CD-ROM, NIST speech disc 1-1.1. 1993.

[Hu20] Huang, Yuge; Wang, Yuhan; Tai, Ying; Liu, Xiaoming; Shen, Pengcheng; Li, Shaoxin; Jilin Li, Feiyue Huang: CurricularFace: Adaptive Curriculum Learning Loss for Deep Face Recognition. pp. 1–8, 2020.

[Ma11] Matejka, P.; Glembek, O.; Castaldo, F.; Alam, M. J.; Plchot, O.; Kenny, P.; Burget, L.; Cernocky, J.: Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4828–4831, 2011.

[NZ19] Nunes, João Antônio Chagas; Macêdo, David; Zanchettin, Cleber: Additive margin sincnet for speaker recognition. In: 2019 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1–5, 2019.

[Pa15] Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S.: Librispeech: An ASR corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5206–5210, 2015.

[PE07] Prince, S. J. D.; Elder, J. H.: Probabilistic Linear Discriminant Analysis for Inferences About Identity. In: 2007 IEEE 11th International Conference on Computer Vision. pp. 1–8, 2007.

[RB18] Ravanelli, Mirco; Bengio, Yoshua: Speaker recognition from raw waveform with sincnet. In: 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, pp. 1021–1028, 2018.

[RD15] Reynolds, Fred Richardson Douglas A.; Dehak, Najim: A Unified Deep Neural Network for Speaker and Language Recognition. CoRR, abs/1504.00923, 2015.

[Sn17] Snyder, David; Garcia-Romero, Daniel; Povey, Daniel; Khudanpur, Sanjeev: Deep Neural Network Embeddings for Text-Independent Speaker Verification. In: INTERSPEECH. 2017.

[Va14] Variani, E.; Lei, X.; McDermott, E.; Moreno, I. L.; Gonzalez-Dominguez, J.: Deep neural networks for small footprint text-dependent speaker verification. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4052–4056, 2014.

[Wa18] Wang, Feng; Cheng, Jian; Liu, Weiyang; Liu, Haijun: Additive margin softmax for face verification. IEEE Signal Processing Letters, 25(7):926–930, 2018.

Gait Authentication based on Spiking Neural Networks

Enrique Argones Rúa[1], Tim Van hamme[2], Davy Preuveneers[2], Wouter Joosen[2]

Abstract: In this paper we address gait authentication using a novel approach based on spiking neural networks (SNNs). This technology has proven advantages regarding energy consumption and it is a perfect match with some proposed neuromorphic hardware chips, which can lead to a broader adoption of user device applications of artificial intelligence technologies. One of the challenges when using this technology is the training of the network itself, since it is not straightforward to apply well-known error backpropagation, massively used in traditional artificial neural networks (ANNs). In this paper we propose a new derivation of error backpropagation for the spiking neural networks that integrates lateral inhibition and provides competitive results when compared to state of the art ANNs in the context of IMU-based gait authentication.

Keywords: Spiking neural network, membrane potential, gait authentication, average spiking rate, normalized membrane potential, synaptic time dependent plasticity, gait recognition system, error backpropagation, surrogate function, angular velocity, variable length gait sequence, continuous authentication, open set biometric authentication, IMU gait authentication.

1 Introduction

Artificial Neural Networks (ANNs) have become the most prevalent pattern recognition tool, being used in a multiplicity of applications. Regarding biometrics, it is already used in most biometric modalities, such as speaker authentication, face recognition, fingerprint recognition, hand-based biometrics, electrocardiogram-based recognition, handwritten online signature recognition, or inertial gait recognition. However, deep learning neural networks often come at the cost of larger complexity and computational requirements in terms of memory and processing power, which may thwart its deployment on constrained user devices, especially for continuous authentication. In recent years there has been an increasing interest in a different type of neural network, the Spiking Neural Networks (SNN) [Ma97a]. This interest has been driven mainly by the possibility to use these networks within ultra-low power consumption specific hardware modules called neuromorphic hardware [Ba20], ideally suited for instance for continuous authentication. However, the drawback of this technology is the lack of mature learning approaches, as opposed to standard ANNs. The most widely used method for SNNs is Synaptic Time-Dependent Plasticity (STDP), which is a biologically inspired unsupervised method based on Hebbian

[1] imec – COSIC, Dept. of Electrical Engineering, KU Leuven. Kasteelpark Arenberg 10, bus 2452, B-3001 Heverlee, Belgium, enrique.argonesrua@esat.kuleuven.be

[2] imec – DistriNet, Dept. of Computer Science, KU Leuven. Celestijnenlaan 200A B-3001 Heverlee, Belgium, {tim.vanhamme, davy.preuveneers, wouter.joosen}@cs.kuleuven.be

rules [Ma97b], reaching limited discrimination performance when compared to gradient descent for ANNs. However, some approximations to gradient descent have been recently proposed, as we will discuss later. In this paper we propose to use SNNs topologically organized in columns, and a novel backpropagation method that allows to perform competitive supervised learning. We demonstrate the potential of this approach in the challenging biometric problem of inertial gait recognition. Inertial measurement unit (IMU) based gait recognition uses the inputs captured by IMU sensors placed somewhere in the subjects' body. Thus, gait is modeled as a six dimensional time series: 3D linear acceleration and 3D angular velocity. We augment this signal and represent it as a 26D time sequence as described in prior work [Va18].

To the best of our knowledge, this constitutes the first work on stream-based biometrics using this technology. We compare the obtained results in terms of authentication performance to ANNs.

2 Previous work on Spiking Neural Networks

Although SNNs are a relatively new neural network paradigm, a lot of research efforts have focused on it lately. Regarding the learning technique, initially most of the works focused on diverse versions of Synaptic Time Dependent Plasticity, a non-supervised technique that produces generative models, which produce spikes related to the most relevant or frequent inputs observed by the network. There are plenty of examples, such as [TM19, DC15, Kh18]. However, all of them suffer from one intrinsic limitation, its generative nature makes them not as accurate when dealing with classification tasks as other state of the art discriminative approaches. The main problem is the non-differentiability of the spikes, the output of the networks, which makes traditional error backpropagation (EBP) not directly applicable. There exist two main approaches to tackle this problem when using spiking neural networks, which consist of converting conventional neural networks trained with traditional supervised EBP-based techniques, or developing a framework where it is possible to perform EBP on the spiking neural network. A comparison of the mentioned approaches can be found in [Ta19]. Although both approaches can provide reasonable performances, specific training can be theoretically more accurate, especially if SNNs have special topological features which do not easily map from ANNs, as the columns used in this paper. This motivated us to use a specific training approach. Developments shown in [Mo18, Pa20, Wu18] are some representative examples of EBP on SNNs. All the backpropagation frameworks made for SNNs so far share that the objective function is either specialized for classification tasks (trying to get a specific label as the average spiking rate at the output layer), or it tries to provide spike trains very similar in different inputs, implying homogeneous dimensionality and duration (if applicable) of input signals. In both cases, it is hard to use these ideas to get a universal (open set, where the same network will be used to process samples from individuals not included in the training set) network for sequence-based biometric authentication, where the length of the sequence is variable (non-homogeneous, making the use of sequence approximation cumbersome).

3 Gait recognition system using SNNs

In this paper we use Leaky Integrate and Fire (LIF) neurons as fundamental computation units. Our networks will be constituted of columns of these neurons, organized in layers. Each layer feeds with the outputs of the columns from previous layers, after an optional max pool layer. Regarding the input layer, it is fed with the min-max normalized 26-D available gait signal and their negative counterparts, since synaptic weights are constrained to be positive in our SNNs.

We train the SNNs in two phases. In the first one, we use the unsupervised method Synaptic Time-Dependent Plasticity to get a good baseline for the second phase, which is a novel supervised Gradient Descent. In the following sections we describe in detail the topological elements of the networks and the learning algorithms.

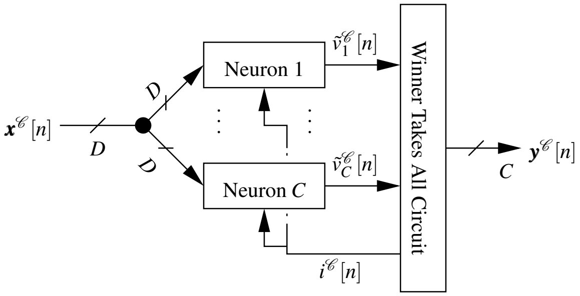

$\boldsymbol{x}^{\mathscr{C}}[n]$
D
D
D
Neuron 1
Neuron C
$\tilde{v}_1^{\mathscr{C}}[n]$
$\tilde{v}_C^{\mathscr{C}}[n]$
Winner Takes All Circuit
C
$\boldsymbol{y}^{\mathscr{C}}[n]$
$i^{\mathscr{C}}[n]$

Fig. 1: Column of neurons with lateral inhibition

Columns of LIF neurons As mentioned above, neurons are organized in columns, as shown in Fig. 1. Neurons in a column share the synaptic field (input space), and only one can emit a spike at a given instant in time. This behaviour is modelled using a Winner Takes All (WTA) circuit, which chooses the neuron with the maximum normalized membrane potential above unity as the spiking one. When a neuron spikes, a lateral inhibition signal $i^{\mathscr{C}}[n]$ is fed back to the neurons in the column, which will put them in a refractory state, as explained below.

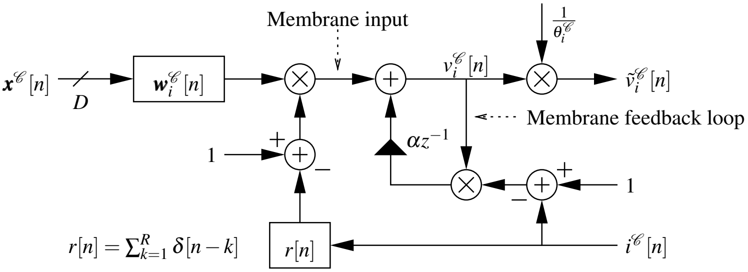

Membrane input
$\frac{1}{\theta_i^{\mathscr{C}}}$
$\boldsymbol{x}^{\mathscr{C}}[n]$
D
$\boldsymbol{w}_i^{\mathscr{C}}[n]$
$v_i^{\mathscr{C}}[n]$
$\tilde{v}_i^{\mathscr{C}}[n]$
Membrane feedback loop
αz^{-1}
+
1
−
+
−
1
$r[n] = \sum_{k=1}^{R} \delta[n-k]$
$r[n]$
$i^{\mathscr{C}}[n]$

Fig. 2: Diagram of neuron i in column $\mathscr{C}$.

LIF neurons The LIF neurons used in this paper, shown in Fig. 2, work in discrete time. They can be in two different states: *active* and *refractory*. During the active state, the membrane potential leaky integrates the inputs filtered by their corresponding synaptic response filters, as shown in the membrane feedback loop. Before the membrane potential is fed into the WTA circuit, it is divided by θ, which can be understood as a neuron-specific activation threshold: if the membrane potential exceeds this value, the normalized membrane potential $\tilde{v}[n]$ will exceed unity. The WTA circuit then performs:

$$y_i^{\mathscr{C}}[n] = \begin{cases} 1 & \Longleftrightarrow \quad i = \mathrm{argmax}_j \left\{ \hat{v}_j^{\mathscr{C}}[n] \mid \hat{v}_j^{\mathscr{C}}[n] \geq 1 \right\} \\ 0 & \Longleftrightarrow \quad \nexists j \mid \hat{v}_j^{\mathscr{C}}[n] \geq 1 \end{cases} \quad , \; i^{\mathscr{C}}[n] = \max_i \left\{ y_i^{\mathscr{C}}[n] \right\} \qquad (1)$$

When any neuron in the column emits a spike, each neuron in the column receives an inhibition impulse from the column WTA circuit (i.e., $i^{\mathscr{C}}[n] = 1$), and goes into refractory state, where the membrane potential is reset and synaptic inputs are ignored, by using $1 - i^{\mathscr{C}}[n]$ and $1 - i^{\mathscr{C}}[n] * r[n]$ as gate signals in the membrane potential feedback loop and input respectively, with $r[n] = \sum_{k=1}^{R} \delta[n-k]$, and R the refractory period duration. It must be noticed that if $R = 0$ the inhibition only affects the feedback loop. Both R and the membrane persistence α (or equivalently are the same in all the neurons in our model, since we understand that these are parameters related to the neurons' physiology in this biologically-inspired system.

Unsupervised initialization using STDP Synaptic weights are randomly initialized from Gaussian distributions, and clamped into the interval $[0,1]$. Then, STDP is used to adapt these weights to the statistics of the gait input sequences. Different update rules for the weights have been proposed for STDP, but the main idea is to implement (i) Long Term Potentiation (LTP) to weights recently contributing to the potential of a neuron that spikes, i.e., increasing this weight; and (ii) Long Term Depression (LTD) to weights that did not contribute to the potential of a neuron that spikes. In other words, neurons that spike reinforce the weights that contribute to the spiking, and weaken those weights that do not contribute. In our case, we implement LTP and LTD by increasing the weights proportionally to the observed input contribution. Given the synaptic filter $\boldsymbol{w}[n] = \sum_{i=0}^{L} w^k \delta[n-k]$, if a spike is emitted at time s, then $\Delta w^k = p \sum_{j=0}^{k} x[s-(k+j)] \alpha^j$ if $\sum_{j=0}^{k} x[s-(k+j)] \alpha^j > \varepsilon$, and $\Delta w = -d$ otherwise. A homeostatic rule is also used to avoid too dominant neurons in a column, decreasing the activation thresholds of neurons that spike below the columns' average spiking rate, and increasing them for neurons with spiking rate higher than the columns' average. These averages are computed using first order average estimators $\bar{r}[n] = \tau^{-1} r[n] + (1 - \tau^{-1}) \bar{r}[n-1]$, with τ equal to 5 times the average gait sequence length. Thresholds are updated at the end of each sequence STDP by $\Delta\theta_i^{\mathscr{C}} = \beta(\bar{r}^{\mathscr{C}} - C\bar{r}_i^{\mathscr{C}})$, where C is the number of neurons in column $\mathscr{C}$, $\bar{r}^{\mathscr{C}}$ and $\bar{r}_i^{\mathscr{C}}$ are the average spiking rate of the column and its i^{th} neuron respectively. After the update, the thresholds are clamped to the interval $[1,10]$.

Supervised training using EBP The main obstacle to perform EBP on SNNs is the non-differentiability of the neuron's output. Its discontinuous (it is either 0 or 1) and homogeneous (all the spikes look the same) nature require adopting a different approach to the one used in traditional ANNs. One of the most common is to define a surrogate function $\boldsymbol{f}$ that is differentiable, and substitute the output of the neuron by that surrogate during gradient computations, i.e., $\delta \boldsymbol{y} \sim \delta \boldsymbol{f}$. One sensible approach is to make this surrogate function model the probability of that neuron firing. A sensible surrogate function for the proposed neuron columns should be based on the normalized membrane potential signals $\tilde{v}_i^{\mathscr{C}}[n]$, which are differentiable with respect to the neurons' input and synaptic filter weights. In this regard, it should also:

1. Be monotonically increasing with its normalized membrane potential, and monotonically decreasing with the normalized membrane potential of the other neurons in the column.
2. Saturate when the membrane potential of the neuron dominates the other neurons' membrane potential in the column.

Taking these into account, we propose to use the softmax function of the normalized membrane potential as surrogate function:

$$\boldsymbol{f}^{\mathscr{C}}[n] = \operatorname{softmax}\left(\tilde{\boldsymbol{v}}^{\mathscr{C}}[n]\right) = \left(\frac{e^{\hat{v}_1^{\mathscr{C}}[n]}}{\sum_{j=1}^{C} e^{\hat{v}_j^{\mathscr{C}}[n]}}, \ldots, \frac{e^{\hat{v}_C^{\mathscr{C}}[n]}}{\sum_{j=1}^{C} e^{\hat{v}_j^{\mathscr{C}}[n]}}\right)^t \tag{2}$$

We evaluate this differentiable surrogate function during the backward phase of EBP *only at the spiking times*, i.e., when one neuron in the column emits a spike. By doing this, we only take into account the spiking events, which are the ones forwarding real information, thus removing influence of instants where the neurons do not get enough evidence to spike, and saving a lot of computation power and training time. Thus, spikes serve as noise removing and energy efficiency mechanisms.

Objective function, optimizer and boosting Although there could be information on time dependencies among spiking events (local information), we only use the average spiking rate at the output layer neurons. We use a siamese architecture, where a reference sequence $\mathscr{S}_R$, belonging to class C_R, and a probe sequence $\mathscr{S}_P$ belonging to class C_P make a forward pass through the network, obtaining two average output average spiking rates $\boldsymbol{r}_R^o$ and $\boldsymbol{r}_P^o$. We then maximize the following cosine similarity based objective function:

$$O\left(\mathscr{S}_R, \mathscr{S}_P\right) = \left(2\delta\left[C_R - C_P\right] - 1\right) \frac{\boldsymbol{r}_R^{o\,t} \boldsymbol{r}_P^o}{|\boldsymbol{r}_R^o|\,|\boldsymbol{r}_P^o|} \tag{3}$$

We use the Nadam optimizer as described in [Do16] to improve convergence, abiding the parameters suggested in this paper. We also incorporated boosting to train the neural network, grouping the pairs of sequences matching a given enrollment sequence in the same minibatch, as in tuplet loss [YT19], and performing the updates based only on the hard negative sequences and the positive sequence (if any of the negative examples is hard, or if the positive sequence is hard itself).

4 Gait recognition system using ANNs

We built a gait recognition system based on ANNs which serves as a baseline to adequately demonstrate the potential of SNNs to model gait. We choose an architecture that led to competitive results for other gait sequence modelling tasks [Va19]. Specifically, we chose temporal convolutional networks (TCN), which are proven to be an effective strategy to model time series data in general [BKK18]. Furthermore, to accommodate the needs of authentication, i.e. an open world assumption with limited amounts of enrollment data, we leverage metric learning to train the system, i.e. we use the triplet loss [SKP15].

As can be seen in Fig. 3, our TCN has a fairly simple architecture with only three layers. The final embedding is a vector of size 128. We trained the network for 500 epochs with the adam optimzer with a learning rate of 0.001. During training we feed fixed length sequences of size 180 to the network, due to the sample rate of $100Hz$ this corresponds to $1.8s$ of data. We chose this value such that at least one gait cycle is captured. Moreover, which part of the full variable length gait sequence is retained is chosen at random. During testing the whole variable length gait sequence is used.

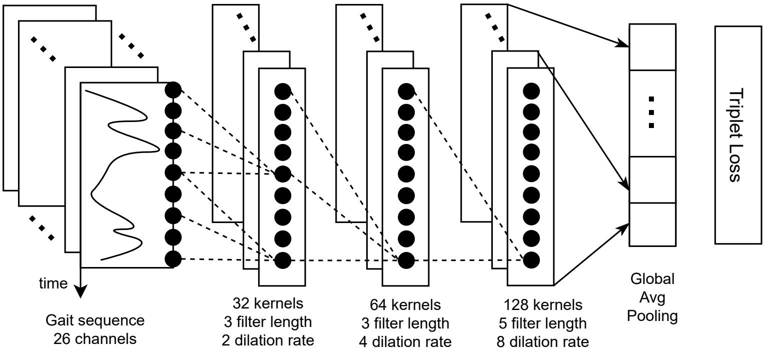


Fig. 3: The architecture of the ANN system which is based on 1 dimensional convolutions with dilation and trained with the triplet loss.

5 Database and experimental setup

To train and evaluate our systems we use the IMU sequences contained in the OU-ISIR dataset labeled as *level walk* and captured by the center sensor. Usually for each user there are two sequences, of which we use one for enrollment and one for testing. We only consider the 483 users with two valid walking sequences for all three sensors, i.e. left, right and center located sensors. Only the center sensor is used for the experiments.

The walk sequences of the OU-ISIR dataset is represented as a six dimensional (6D) time series: 3D linear acceleration and 3D angular velocity. We augment this signal and represent it as a 26D time sequence as described in prior work [Va18]. Eight dimensions are derived from the *gyroscopic dynamics (8D)*: the angular velocity, their first order differences, and the magnitude of both vectors. Six dimensions are related to the *vertical and horizontal components (6D)*, which are an approximation of the vertical and horizontal acceleration in the world pane. They are complemented with their first order differences and jerks. Twelve dimensions are related to *roll and pitch (12D)*, which are approximated from the linear accelerations. We also improve this approximation by fusing it with an estimation of roll and pitch from the angular velocity. We keep both approximations and complement them with their first-order and second-order differences.

We adopt a 5-fold cross-validation protocol, where we randomly split the 483 users in 5 disjoint sets which contain either 96 or 97 users: $S_s = \{U_0^s, U_1^s, \ldots, U_x^s\}$ with $x \in [96,97]$ and $s \in [0,4]$. Thus, we do 5 rounds, where during each round k we select S_k as the test set, and the remaining 4 sets as training set. We report the Equal Error Rate (EER) in the test set, i.e. the threshold at which False Rejection Rate (FRR) is equal to False Acceptance Rate (FAR). Besides, we plot the Detection Error Trade-off curves for a visual comparison.

6 Experimental results

In our experiments we tried different SNN topologies, with one and two layers, using max pooling between them. We present here the performance obtained by the TCN architecture presented in Section 4, together with the flowing SNNs:

N1 1 layer, 4 columns, 32 neurons/column, 1 coeff. filter, 32-D column input.

N2 1 layer, 4 columns, 32 neurons/column, 4 coeff. filter, 32-D column input.

N3 2 layers, each with 4 columns, 32 neurons/column, and 4 coeff. filters. First layer has 32-D and second layer has 96-D column inputs. Max-pooling with window and stride 2 is used between layers.

N4 1 layer, 32 columns, 32 neurons/column, 2 coeff. filter, 32-D column input.

7 Conclusions

In this work we presented a novel column-based SNN architecture and EBP approach for supervised learning. We tested this approach on the challenging task of authenticating persons using IMU gait signals in an open set protocol. Although results obtained by EBP on SNNs are yet behind from the ones shown by state of the art ANN architectures, a clear improvement over STDP is demonstrated. The low power consumption shown by SNN hardware implementations, together with the development of EBP techniques and reduced performance gap with respect to ANNs encourage further research on the application of SNN for biometrics, but also for a wider range of applications.

Label	Algorithm	Test fold 1	2	3	4	5	Average ± Std. dev.
N1	STDP	7.97	4.88	6.19	10.66	6.25	7.19 ± 2.23
	EBP	3.98	2.51	3.09	6.25	5.21	4.21 ± 1.53
N2	STDP	9.28	5.96	8.25	10.71	11.46	9.13 ± 2.17
	EBP	4.16	2.13	4.12	7.27	4.17	4.37 ± 1.84
N3	STDP	10.53	7.22	11.16	15.62	11.46	11.20 ± 3.00
	EBP	7.62	2.80	4.99	6.33	8.33	6.01 ± 2.20
N4	STDP	5.62	4.12	5.15	9.38	6.25	6.10 ± 1.99
	EBP	2.49	1.03	2.06	3.12	2.08	2.16 ± 0.76
TCN	EBP	1.03	1.92	1.92	2.08	1.06	1.60 ± 0.46

Tab. 1: Authentication performance in terms of test EER(%) of the different networks.

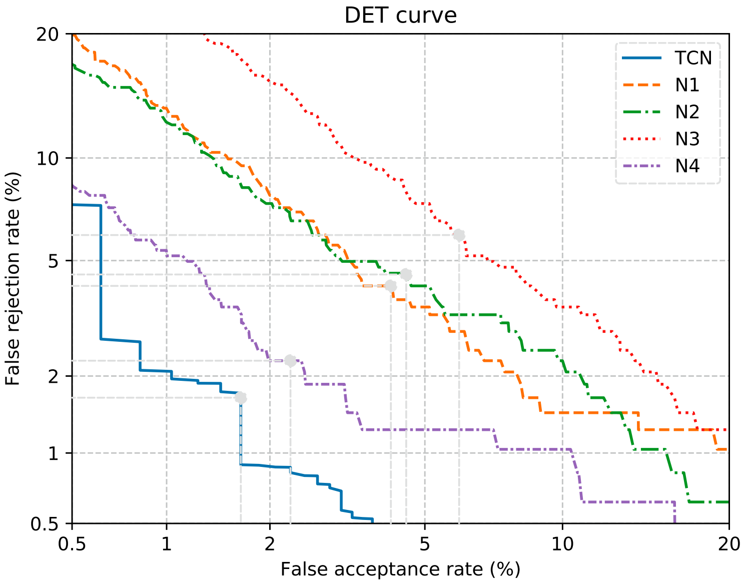


Fig. 4: DET curve for the EBP-trained SNNs and TCN.

References

[Ba20] Balaji, Adarsha; Das, Anup; Wu, Yuefeng; Huynh, Khanh; Dell'Anna, Francesco G.; Indiveri, Giacomo; Krichmar, Jeffrey L.; Dutt, Nikil D.; Schaafsma, Siebren; Catthoor, Francky: Mapping Spiking Neural Networks to Neuromorphic Hardware. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 28(1):76–86, 2020.

[BKK18] Bai, Shaojie; Kolter, J. Zico; Koltun, Vladlen: , An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling, 2018.

[DC15] Diehl, Peter; Cook, Matthew: Unsupervised learning of digit recognition using spike-timing-dependent plasticity. Frontiers in Computational Neuroscience, 9:99, 2015.

[Do16] Dozat, Timothy: Incorporating Nesterov Momentum into Adam. In (Bengio, Yoshua; LeCun, Yann, eds): 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings. 2016.

[Kh18] Kheradpisheh, Saeed Reza; Ganjtabesh, Mohammad; Thorpe, Simon J.; Masquelier, Timothée: STDP-based spiking deep convolutional neural networks for object recognition. Neural Networks, 99:56–67, 2018.

[Ma97a] Maass, Wolfgang: Networks of spiking neurons: The third generation of neural network models. Neural Networks, 10(9):1659–1671, 1997.

[Ma97b] Markram, Henry; Lübke, Joachim; Frotscher, Michael; Sakmann, Bert: Regulation of Synaptic Efficacy by Coincidence of Postsynaptic APs and EPSPs. Science, 275(5297):213–215, 1997.

[Mo18] Mostafa, H.: Supervised Learning Based on Temporal Coding in Spiking Neural Networks. IEEE Transactions on Neural Networks and Learning Systems, 29(7):3227–3235, 2018.

[Pa20] Patiño-Saucedo, Alberto; Rostro-Gonzalez, Horacio; Serrano-Gotarredona, Teresa; Linares-Barranco, Bernabé: Event-driven implementation of deep spiking convolutional neural networks for supervised classification using the SpiNNaker neuromorphic platform. Neural Networks, 121:319–328, 2020.

[SKP15] Schroff, Florian; Kalenichenko, Dmitry; Philbin, James: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823, 2015.

[Ta19] Tavanaei, Amirhossein; Ghodrati, Masoud; Kheradpisheh, Saeed Reza; Masquelier, Timothée; Maida, Anthony: Deep learning in spiking neural networks. Neural Networks, 111:47–63, 2019.

[TM19] Tavanaei, Amirhossein; Maida, Anthony: BP-STDP: Approximating backpropagation using spike timing dependent plasticity. Neurocomputing, 330:39–47, 2019.

[Va18] Van hamme, Tim; Argones Rúa, Enrique; Preuveneers, Davy; Joosen, Wouter: Gait template protection using HMM-UBM. In (Brömme, Arslan; Busch, Christoph; Dantcheva, Antitza; Rathgeb, Christian; Uhl, Andreas, eds): 2018 International Conference of the Biometrics Special Interest Group, BIOSIG 2018, Darmstadt, Germany, September 26-28, 2018. volume P-282 of LNI. GI / IEEE, pp. 1–8, 2018.

[Va19] Van hamme, Tim; Garofalo, Giuseppe; Argones RÃºa, Enrique; Preuveneers, Davy; Joosen, Wouter: A Systematic Comparison of Age and Gender Prediction on IMU Sensor-Based Gait Traces. Sensors, 19(13), 2019.

[Wu18] Wu, Yujie; Deng, Lei; Li, Guoqi; Zhu, Jun; Shi, Luping: Spatio-Temporal Backpropagation for Training High-Performance Spiking Neural Networks. Frontiers in Neuroscience, 12:331, 2018.

[YT19] Yu, B.; Tao, D.: Deep Metric Learning With Tuplet Margin Loss. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6489–6498, 2019.

Shuffled Patch-Wise Supervision for Presentation Attack Detection

Alperen Kantarcı[1], Hasan Dertli[2], Hazım Kemal Ekenel[1]

Abstract: Face anti-spoofing is essential to prevent false facial verification by using a photo, video, mask, or a different substitute for an authorized person's face. Most of the state-of-the-art presentation attack detection (PAD) systems suffer from overfitting, where they achieve near-perfect scores on a single dataset but fail on a different dataset with more realistic data. This problem drives researchers to develop models that perform well under real-world conditions. This is an especially challenging problem for frame-based presentation attack detection systems that use convolutional neural networks (CNN). To this end, we propose a new PAD approach, which combines pixel-wise binary supervision with patch-based CNN. We believe that training a CNN with face patches allows the model to distinguish spoofs without learning background or dataset-specific traces. We tested the proposed method both on the standard benchmark datasets —Replay-Mobile, OULU-NPU— and on a real-world dataset. The proposed approach shows its superiority on challenging experimental setups. Namely, it achieves higher performance on OULU-NPU protocol 3, 4 and on inter-dataset real-world experiments.

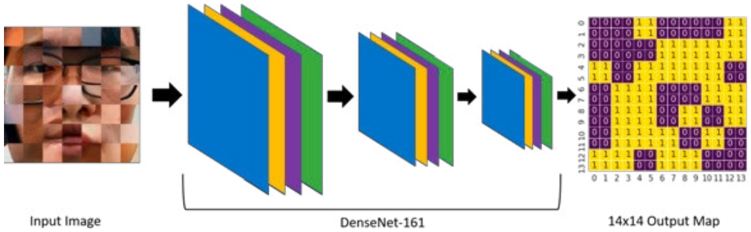
Keywords: Face antispoofing, presentation attack detection, convolutional neural networks, real-world dataset.

1 Introduction

In recent years, facial recognition systems are widely used as they are robust and reliable for common usage. However, these recognition systems have to be careful about the authenticity of a given face input. If the given input is recorded from a video of an authorized user, the recognition system should not recognize the person in the video and give access to the system. Presentation attack detection (PAD) systems aim to prevent this problem by evaluating the liveness of the given person's image.

In recent years, PAD methods improved significantly with the progress in deep learning methods and publicly available large, representative datasets [Bo17, LJL18, Zh20, Co16]. Most of the significant progress has been achieved when researchers found different cues to decide liveness of a face [Li16, MHP11, At17]. These different cues used with complex deep neural networks to create PAD systems that are very successful in intra-dataset benchmark results. However, the real challenge in PAD still remains as an inter-dataset benchmark which shows the real performance of the PAD systems in real-world like scenario. Most of the systems that use CNNs overfit the data easily by memorizing reflection

[1] Department of Computer Engineering, Istanbul Technical University, Turkey, {kantarcia,ekenel}@itu.edu.tr
[2] Sodec Technologies, Istanbul, Turkey, hasan.dertli@sodecapps.com



	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	0	0	0	0	1	1	0	0	0	0	0	0	1	1
1	0	0	0	0	1	1	0	0	0	0	0	0	1	1
2	0	0	0	0	0	0	1	1	1	1	1	1	1	1
3	0	0	0	0	0	0	1	1	1	1	1	1	1	1
4	1	1	0	0	1	1	1	1	1	1	1	1	0	0
5	1	1	0	0	1	1	1	1	1	1	1	1	0	0
6	0	0	1	1	1	1	0	0	0	0	1	1	1	1
7	0	0	1	1	1	1	0	0	0	0	1	1	1	1
8	0	0	1	1	1	1	0	0	1	1	0	0	1	1
9	0	0	1	1	1	1	0	0	1	1	0	0	1	1
10	0	0	1	1	1	1	1	1	0	0	1	1	0	0
11	0	0	1	1	1	1	1	1	0	0	1	1	0	0
12	1	1	1	1	0	0	1	1	1	1	0	0	0	0
13	1	1	1	1	0	0	1	1	1	1	0	0	0	0

Fig. 1: Overview of the proposed method

and illumination effects. To address this problem, in this paper, we propose a new training procedure for face PAD systems. We show that our training method utilizes the pixel-wise binary loss in a better way. Moreover, we show that our proposed method improves model performances on real-world experiments.

2 Related Work

PAD approaches are mainly grouped into two categories; video-based and frame level. While video-based methods use temporal consistency and temporal cues, frame-based methods use subtle cues from the given face images. These cues can be summarized as liveness, texture, and 3D geometry [Mi20]. Among these cue-based methods, liveness cues are applicable for video-based PAD. Therefore, texture and 3D geometry cue-based methods are more popular for detecting liveness from a frame. 3D geometry cue-based methods either use depth or pseudo-depth signals to distinguish attacks from real attempts. Even though devices with depth sensors, time-of-flight cameras, Lidar sensors, etc., are getting popular in daily usage, most mobile phones or video cameras do not have depth sensors. Therefore, methods mostly rely on pseudo-depth maps which are not real data and may not reflect real-world data distribution very well. Therefore, most of these methods might get good results on specific datasets but fail to generalize.

As initial work with deep neural networks, [YLL14] proposes to use a face alignment network for preparing face images to train an AlexNet [KSH12] model. They use the model for extracting features of the face and use an SVM classifier to classify images as artefact or bona fide.

In order to improve PAD performances, researchers search different supervisions, along with the binary classification objective, for training their models. [GM19] proposes an effective model for frame-level PAD. They add additional supervision, which they call pixel-wise binary supervision, to simplify the necessity of complex depth maps and temporal information. Their model creates a 14x14 score map which helps to perform pixel-wise binary supervision. On top of this supervision, their model is guided with binary cross-entropy. We build our model on top of [GM19] by using their pixel-wise binary supervision and model architecture. Instead of using binary cross-entropy, we propose to use only pixel-wise binary supervision. Moreover, we train our models with shuffled face images that are created by multiple patches of different face parts of different subjects.

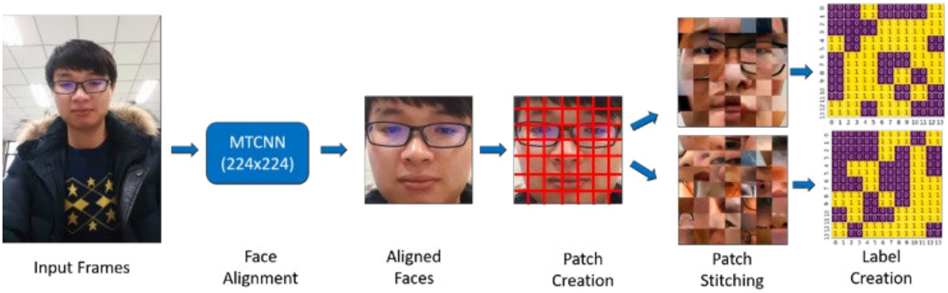


Fig. 2: Our overall pipeline of input and ground truth label creation.

[At17] proposes a two-stream CNN that uses patches and depth maps. They claim that patch-based CNN learns to discriminate artefact patches independent of the spatial face areas whereas depth-based CNN allows the model to learn how a face-like holistic depth map should look like. Our work, which combines pixel-wise binary supervision with patch-based CNN, is inspired by [GM19, At17]. As [At17] showed, we believe that training a CNN with face patches allows the model to distinguish artefacts without learning background or dataset-specific traces. Therefore, it prevents the model to overfit. Effectiveness of similar patch shuffling method is presented in [Ka17]. In [Ka17] it is shown that shuffling pixels within a patch can increase the generalization of the models. Our approach tries to mimic this behaviour from a higher level of view.

Currently, the state-of-the-art frame-level PAD method [Yu20] leverages a novel convolution operation. Authors propose Central Difference Convolution (CDC) for detecting detailed artefact traces. CDC specifically focuses on artefact traces. Their model is trained with pseudo-depth maps which require additional pseudo-depth map creation steps for ground truth. They also use computationally expensive Neural Architecture Search (NAS) to find a better and more efficient model which they call CDCN++. They report the lowest error rates on the OULU-NPU [Bo17] dataset.

3 Methodology

Preprocessing of the images is an important task in PAD systems. The preprocessing pipelines are very similar among different methods. We first detect faces and face landmarks in the given frame with MTCNN [Zh16] face detector. Then by using Bob framework [An12, An17] we align the detected faces according to the eye coordinates and crop these aligned faces in 224x224 resolution. After that, we create patches from each face image by dividing the face into 7x7. Therefore, we get 49 face patches with 32x32 resolutions for each face image. We then combine these patches to create a new 224x224 image. Each face patch corresponds to a 2x2 location in the ground truth 14x14 label map. Bona fide patches have 1 as the label and fake patches have 0 as the label. We call this process patch stitching. We use two different strategies while stitching the patches. In the first approach, which we call random stitching, we completely randomize the patches and do not care about facial structure while combining different patches. Therefore, in this method, we get completely shuffled face images. As we randomly choose patches, the same parts of faces can be found in stitched images. For example, multiple noses or eyes can be seen in the bottom part of the Figure 2. In the second approach, which we call controlled-stitching, we combine patches of different faces while keeping the facial structure as much as possible. Therefore, we actually create an input that resembles a face and consists of multiple subjects' face parts. The input contains both bona fide and attack patches, therefore stitched images must have labels for each patch separately. In our experiments we use the former approach which gives better performance in inter-dataset experiments. The overview of the proposed method is illustrated in the Figure 1.

3.1 Model Architecture and Training

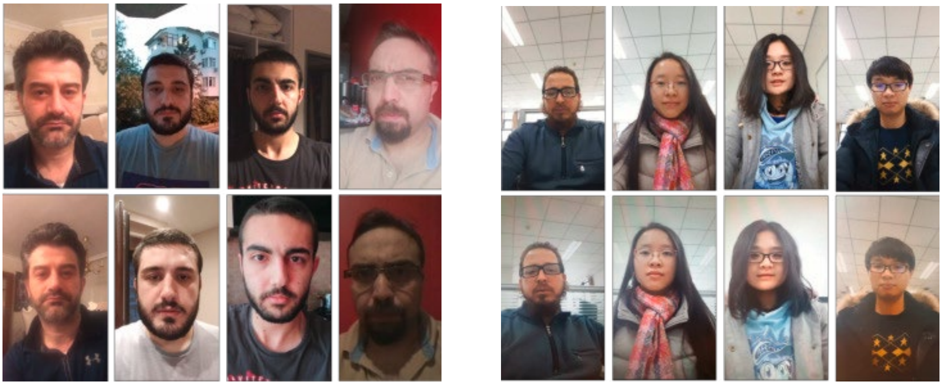
We use a deep CNN network that takes 224x224 input images and creates a 14x14 output map. We use the model that was proposed in [GM19] which is based on DenseNet-161 [Hu17] architecture. However, we do not have the final linear and sigmoid layers. The model contains the first eight layers of the DenseNet [Hu17] and we use pretrained weights. At the end of eight layers, we add a 1x1 convolution to produce a 14x14x1 feature map which is the model output.We use input images that contain multiple patches together, therefore, we do not have a binary label. In training time, we use 14x14 pixel-wise binary labels to train the model with Binary Cross Entropy (BCE) loss. We assign ground truth y=0 for patches that come from attack images and y=1 for patches that come from bona fide images. The equation for pixel-wise loss is shown below.

$$\mathscr{L}_{pixel-wiseBCE} = -(y(log(p)) + (1-y)log(1-p)) \tag{1}$$

In Equation 1 p is the 14x14 model output that contains probability values between 0 and 1. We minimize this loss with Adam [KB15] optimizer. We use 0.001 as the initial learning rate and halve this value at each 10th epoch. We use 32 as batch size and we generally train our methods for 30 epochs. We use horizontal flip and color jitter as data augmentation. In the test time, we give the aligned face image to the model and our model gives 14x14 output. We calculate the mean probability score by using the 14x14 output, then we use this probability as our liveness score. If the score is higher than the predefined threshold, we classify the given input as bona fide, else we classify it as an attack.

4 Experimental Results

This section presents the datasets that have been used to assess the performance of the proposed approach, the experiments carried out, and the obtained results.



(a) Samples from Sodec Real-World dataset

(b) Samples from OULU-NPU dataset

Fig. 3: Sample images from OULU-NPU and Sodec Real-World datasets.

4.1 Datasets and Metrics

Replay Mobile: Replay Mobile dataset [Co16] consists of 1190 video clips of 40 subjects. It contains paper and replay presentation attacks under five different lighting conditions. Each subject has 10 videos under different lighting and background conditions. There are mainly two kinds of attacks; matte screen and print. Matte screen attacks are the replay attack scenario where subject videos are displayed on a 1080p monitor, then recorded off of it. In print attacks, the faces of the subjects are printed on A4 paper then put on a stationary surface. The dataset has *”grandtest”* protocol for evaluating the global performance of an algorithm. We report our results on the *”grandtest”* protocol.

OULU-NPU: OULU-NPU dataset [Bo17] is a high resolution antispoofing dataset. It has over 5900 videos of 55 subjects. The dataset has both print and replay attacks with two printers and two display devices. There are 4 protocols for evaluating the methods’ generalization capabilities. The protocol names were set based on the level of difficulty.

SiW: SiW dataset [LJL18] is one of the largest high-quality antispoofing datasets. It has over 4400 videos of 165 subjects collected over 4 different sessions. All videos have 1080p resolution and contain variations of distance, pose, illumination, and expression. It contains both print and replay attacks. The dataset has 3 protocols to test the generalization of the models. We utilize this dataset for inter-dataset experiments because of the following reasons: it contains one of the highest numbers of subjects among PAD datasets, it contains different poses and expressions, and the acquisition device changes its distance to subjects which is a common behavior in the real world PAD attempts.

Sodec Real-World Dataset: Sodec Real-World dataset has been collected to simulate the real-world presentation attack scenarios. It contains more than 51k frames of 31 different subjects. There are 16 male and 15 female subjects. It contains only replay attacks over 3 different presentation attack instruments (PAI), namely mobile phone, computer display, and television. Unlike controlled datasets, every subject recorded real videos with their mobile phone in their own home. Moreover, subjects were asked to rotate the phone vertically on the spot themselves while holding the phone in their outstretched arms and recording the video of themselves. It allows us to capture different backgrounds, illumination conditions, and pose variations with different input sensors. We show some examples from the dataset in Figure 3. We utilize this dataset in inter-dataset experiments to test models on real-world attacks.

In all of our experiments, we used ISO/IEC 30107-3 [IS16] standards which are standard metrics for the PAD. Attack Presentation Classification Error Rate (APCER), Bona fide Presentation Classification Error Rate (BPCER), Half Total Error Rate (HTER) along with the Average Classification Error Rate (ACER) in the test set are reported in the experiments. For all of our experiments, we use the threshold according to the equal error rate (EER) criterion. Average of False Recognition Rate (FRR) and False Acceptance Rate (FAR) is equal to HTER. We show the calculation of these metrics in Equation 2 and Equation 3.

$$ACER = \frac{APCER + BPCER}{2} \tag{2}$$

$$HTER = \frac{FRR + FAR}{2} \tag{3}$$

4.2 Experiments and Results

We use OULU-NPU and Replay Mobile datasets for our intra-dataset experiments. As explained above, on OULU-NPU we report APCER, BPCER, and ACER performances of our and other models. On Replay Mobile dataset we report EER and HTER performances. We compare our method with CDCN++ [Yu20] and DeepPixBis [GM19] models. CDCN++ is a state-of-the-art method that uses pseudo-depth maps and NAS. DeepPixBis has the same CNN architecture as our model. We differ from DeepPixBis in only our input and label creation where we utilize patch-wise labels. Therefore, our proposed method is directly comparable with DeepPixBis [GM19]. Similar to us, in the Replay-Mobile dataset, most of the newest methods report 0% error. Table 1 shows that our proposed method also achieves 0% error on the dataset. In Table 2 we report our intra-dataset results on the

Model	**EER(%)**	**HTER(%)**
CDCN	**0.0**	**0.0**
DeepPixBis	**0.0**	**0.0**
Ours	**0.0**	**0.0**

Tab. 1: Intra-dataset test results of Replay-Mobile *"grandtest"* protocol. The first column represents the Equal Error Rate (EER) in percentage and the second represents the Half Total Error Rate (HTER) in percentage.

Protocol	Model	APCER(%)	BPCER(%)	ACER(%)
1	CDCN	**0.4**	1.7	1.0
	CDCN++	**0.4**	**0.0**	**0.2**
	DeepPixBis	0.83	**0.0**	0.42
	Ours	2.14	2.14	2.14
2	CDCN	**1.5**	1.4	1.5
	CDCN++	1.8	**0.8**	**1.3**
	DeepPixBis	11.39	0.56	5.97
	Ours	6.22	6.26	6.24
3	CDCN	2.4 ± 1.3	2.2 ± 2.0	2.3 ± 1.4
	CDCN++	**1.7 ± 1.5**	**2.0 ± 1.2**	**1.8 ± 0.7**
	DeepPixBis	11.67 ± 19.57	10.56 ± 14.06	11.11 ± 9.4
	Ours	6.10 ± 2.57	6.30 ± 2.55	6.20 ± 2.55
4	CDCN	4.6 ± 4.6	9.2 ± 8.0	6.9 ± 2.9
	CDCN++	**4.2 ± 3.4**	**5.8 ± 4.9**	**5.0 ± 2.9**
	DeepPixBis	36.67 ± 29.67	13.33 ± 16.75	25.0 ± 12.0
	Ours	11.51 ± 7.86	11.58 ± 7.84	11.54 ± 7.84

Tab. 2: Intra-dataset test results of OULU-NPU dataset.

OULU-NPU dataset. Table 2 shows that our method outperforms DeepPixBis on Protocol 3 and Protocol 4 which are the hardest protocols in the dataset. These protocols have smaller training data and their test data is not very similar to the training data in terms of environment, PAI, PA acquisition device. Our method has a higher error rate on Protocol 1 and Protocol 2, but according to the our experiment results we show that our method is well suited for generalization whereas other methods gain an advantage of similar training and testing sets in these protocols. Our result does not outperform the state-of-the-art PAD model, however, CDCN and CDCN++ use pseudo-depth maps. In the case of CDCN++, it employs computationally expensive NAS operations.

4.3 Real-World Experiments

There are many face antispoofing datasets that are collected in controlled environments. Most of these datasets only consider two or three backgrounds with controlled illumination changes. However, in real-world applications, there are various backgrounds, illumination, poses, and expression changes. Moreover, attackers are more careful when creating a face presentation attack. We have collected a dataset to simulate the real-world use case of the presentation attacks. We trained CDCN [Yu20], DeepPixBis [GM19], and our model on the SiW dataset Protocol 1. We choose the SiW dataset because it is one of the most representative datasets in terms of distance, pose, illumination, and expression changes. From Table 3 we see that state-of-the art CDCN model has achieved the lowest error rate in the SiW dataset, but gets a higher error rate on Sodec Real-World Dataset. Our method is able to outperform DeepPixBis on both the SiW dataset and Sodec Real-World Dataset.

	Trained on SiW	
Model	tested on SiW	tested on Sodec Real-World Dataset
CDCN	**0.12**	12.52
DeepPixBis	3.68	12.05
Ours	2.15	**5.24**

Tab. 3: Experiment results of SiW (Protocol 1) and inter-dataset test results on Sodec Real-World Dataset. Reported metrics in the table represents the ACER values in percentage (%)

The results show that our proposed training method is useful for real-world inter-dataset scenarios which is the hardest task to perform.

5 Conclusion and Future Work

In this paper, we propose a new training method for the PAD models. Our proposed method uses combined face patches instead of one single face image in training time. We show that training models with pixel-wise binary loss and shuffled face patches can improve PAD performance. Our proposed method improves DeepPixBis' [GM19] performance on OULU-NPU Protocol 3 and 4 which are the hardest protocols. Moreover, the proposed method performs much better when we test the models on a real-world dataset. For future work, we are planning to extend the Sodec Real World dataset to print attacks. Furthermore, we are planning to modify the backbone architecture of the model and analyze the effects of different patch creation methods on model behavior.

6 Acknowledgements

This work was partially supported by a Sodec Technologies research grant. We would like to thank Sodec Technologies for their data collection efforts and support for this work.

References

[An12] Anjos, A.; Shafey, L. El; Wallace, R.; Günther, M.; McCool, C.; Marcel, S.: Bob: a free signal processing and machine learning toolbox for researchers. In: 20th ACM Conference on Multimedia Systems (ACMMM), Nara, Japan. October 2012.

[An17] Anjos, A.; Günther, M.; de Freitas Pereira, T.; Korshunov, P.; Mohammadi, A.; Marcel, S.: Continuously Reproducing Toolchains in Pattern Recognition and Machine Learning Experiments. In: Intl. Conference on Machine Learning (ICML). August 2017.

[At17] Atoum, Y.; Liu, Y.; Jourabloo, A.; Liu, X.: Face anti-spoofing using patch and depth-based CNNs. In: IEEE Intl. Joint Conference on Biometrics. 2017.

[Bo17] Boulkenafet, Z.; Komulainen, J.; Li, Lei.; Feng, X.; Hadid, A.: OULU-NPU: A mobile face presentation attack database with real-world variations. In: IEEE Intl. Conference on Automatic Face and Gesture Recognition. May 2017.

[Co16] Costa-Pazo, A.; Bhattacharjee, S.; Vazquez-Fernandez, E.; Marcel, S.: The REPLAY-MOBILE Face Presentation-Attack Database. In: Proceedings of the Intl. Conference on Biometrics Special Interests Group. September 2016.

[GM19] George, A.; Marcel, S.: Deep Pixel-wise Binary Supervision for Face Presentation Attack Detection. In: 12th IAPR Intl. Conference on Biometrics (ICB). 2019.

[Hu17] Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K. Q.: Densely Connected Convolutional Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2261–2269, 2017.

[IS16] ISO/IEC JTC1 SC37: Information Technology - Biometric presentation attack detection - Part 3: Testingand Reporting. Technical report, International Organization for Standardization, Geneva, CH, February 2016.

[Ka17] Kang, Guoliang; Dong, Xuanyi; Zheng, Liang; Yang, Yi: Patchshuffle regularization. 2017.

[KB15] Kingma, Diederik P.; Ba, Jimmy: , Adam: A Method for Stochastic Optimization, 2015.

[KSH12] Krizhevsky, A.; Sutskever, I.; Hinton, G.: ImageNet Classification with Deep Convolutional Neural Networks. In: Advances in Neural Information Processing Systems. volume 25, 2012.

[Li16] Li, X.; K., Jukka; Z., G.; Y., Pong-Chi; Pietikainen, M.: Generalized face anti-spoofing by detecting pulse from face videos. In: 2016 23rd International Conference on Pattern Recognition (ICPR). pp. 4244–4249, 2016.

[LJL18] Liu, Y.; Jourabloo, A.; Liu, X.: Learning Deep Models for Face Anti-Spoofing: Binary or Auxiliary Supervision. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 389–398, 2018.

[MHP11] Maatta, J.; Hadid, A.; Pietikainen, M.: Face spoofing detection from single images using micro-texture analysis. In: 2011 Intl. Joint Conference on Biometrics. pp. 1–7, 2011.

[Mi20] Ming, Z.; Visani, M.; Luqman, M. M.; Burie, J.C.: A Survey On Anti-Spoofing Methods For Face Recognition with RGB Cameras of Generic Consumer Devices. In: arxiv.org/abs/2010.04145. 2020.

[YLL14] Yang, J.; Lei, Z.; Li, S. Z.: Learn Convolutional Neural Network for Face Anti-Spoofing. In: arxiv.org/abs/1408.5601. 2014.

[Yu20] Yu, Z.; Zhao, C.; Wang, Z.; Qin, Y.; Su, Z.; Li, X.; Zhou, F.; Zhao, G.: Searching Central Difference Convolutional Networks for Face Anti-Spoofing. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5294–5304, 2020.

[Zh16] Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y.: Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. IEEE Signal Processing Letters, 23(10):1499–1503, 2016.

[Zh20] Zhang, S.; Liu, A.; Wan, J.; L., Yanyan; Guo, G.; Escalera, S.; Escalante, H. J.; Li, S. Z.: CASIA-SURF: A Large-Scale Multi-Modal Benchmark for Face Anti-Spoofing. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2(2):182–193, 2020.

Towards Generating High Definition Face Images from Deep Templates

Xingbo Dong[1], Zhe Jin[2], Zhenhua Guo[3], Andrew Beng Jin Teoh[4]

Abstract: Face recognition based on deep convolutional neural networks (CNN) has manifested superior accuracy. Despite the high discriminability of deep features generated by CNN, the vulnerability of the deep feature is often overlooked and leads to the security and privacy concerns, particularly the risks of reconstructing face images from the deep templates. In this paper, we propose a method to generate high definition (HD) face images from deep features. To be specific, the deep features extracted from CNN are mapped to the input (latent vector) of the pre-trained StyleGAN2 using a regression model. Subsequently, HD face images can be generated based on the latent vector by the pre-trained StyleGAN2 model. To evaluate our method, we derived the face features from the generated HD face images and compared them against the bona fide face features. In the sense of face image reconstruction, our method is simple, yet the experimental results suggest the effectiveness, which achieves an attack performance as high as SAR=46.08% (18.30%) @ FAR=0.1 threshold under type-I (type-II) attack settings. Besides, experiment results also indicate that 50.7% of the generated HD face images can pass one commercial off-the-shelf (COTS) liveness detection.

Keywords: Face template security, face image reconstruction, deep templates.

1 Introduction

The recent thriving of deep learning technology has succeeded in numerous computer vision applications such as face recognition (FR). In fact, the deep learning enabled approach has become a de-facto standard for face recognition due to the superior recognition performance. In general, a deep learning-based FR system is composed of three main components, i.e., pre-processing, convolution neural network based feature extractor, and a matcher. Despite enjoying decent performance and convenience, security and privacy concerns on FR systems rise among the public because of the inherent linkage between the face data and the owner identity. For example, the disclosure of face data may expose the private and sensitive information of the user (e.g., race, age, gender). Moreover, face templates could be inverted, hence face images can be generated to gain illegal access to the system.

A number of work have been done on the restoration of the face template to the face image [MJ13, Ma18]. One of the latest works, namely a neighborly de-convolutional neural

[1] School of Information Technology, Monash University, Malaysia Campus, Malaysia, xingbo.dong@monash.edu

[2] School of Information Technology, Monash University, Malaysia Campus, Malaysia, jin.zhe@monash.edu

[3] Shenzhen International Graduate School, Tsinghua University, Shenzhen, China, zhenhua.guo@sz.tsinghua.edu.cn

[4] School of Electrical and Electronic Engineering, Yonsei University, Seoul, South Korea, bjteoh@yonsei.ac.kr

network (NbNet) is proposed to reconstruct the face image from its deep feature counterpart [Ma18]. The generated face images can achieve an attack success rate of 95.20% (58.05%) on LFW dataset [Hu08] under type-I (type-II) attacks with a false accept rate of 0.1%. The type-I attack here refers to comparing the generated fake image against the face image from which the template was extracted. In contrast, type-II refers to comparing the generated fake image against the bona fide[3] face images of the same subject that were not used for template creation.

However, the generated face images in [Ma18] are in low-resolution and leaves over the clue of synthetical artifacts, which is easy to be detected by the image based face liveness detection. As opposed to the low-resolution face images, high definition (HD) face images aid adversaries to perform the attack easily and efficiently, e.g., construct a 3D face mask from an HD face image. In this paper, we focus on generating HD face images based on the deep features[4] and attempt to utilize the generated HD face images to access a targeted FR system by launching type-I and type-II attacks. The proposed method to generate an HD face image based on the deep features can be regarded as a partial invertible scheme without precisely recovering the original face data, which may lead to privacy leakage and security compromises. Our work may shed light on the biometric deployment to meet the privacy-preserving, and security policies, such as the EU General Data Protection Regulation (GDPR) [Co18].

In summary, this paper makes the following contributions:

1. By learning a mapping network between the latent vector space of a pre-trained StyleGAN2 model and the feature vector space of a pre-trained CNN extractor, a simple but effective method to generate HD face images from deep features is established.
2. Attacks to compromise the security of the face recognition systems are simulated. Specifically, the CNN-based face feature extractor is regarded as a black-box in the attacks. The generated HD face images are used to compare with the original face features to simulate the compromising of security of the face recognition systems.
3. The proposed method is also evaluated with a Commercial-Off-The-Shelf (COTS) face liveness detector. It shows that the generated HD face images manage to fool the COTS detector.

In the remainder of this paper, some existing related works will be reviewed in Section 2, and the detailed methodology will be discussed in Section 3. The experiments and results are shown in Section 4. Finally, the conclusion is drawn in Section 5.

[3] We refer to Bona fide as genuine face images.

[4] Deep features and deep templates are used interchangeably.

2 Related works

Face image generation can be traced back to face image reconstruction [RdJG98, Tr99, Tr06]. The application of such face image reconstruction can facilitate the witnesses in a crime scene. For example, as presented in [Tr06], generated face images are shown to the witness, and the candidate face images are selected by the witness. The selected candidate face images are further evaluated based on an optimization algorithm by narrowing the eigenface coefficient space iteratively and then generated face images are computed and shown to the user again. To achieve this task, the user's interactive input is always required in such a system.

In [Ad03, FLY14], hill-climbing is utilized to generate a synthetic face image from a corresponding real-valued template. Specifically, a random face image is initiated firstly; then, the face image is perturbed iteratively based on the matching score between the current iteration and the previous iteration-based synthetic face image' features. The iteration is ended when the matching score decreases to a decision threshold. The corresponding synthetic face image is used as the final output. In [MJ13], radial basis function (RBF) regression in the face eigenspace is adopted to reconstruct visually realistic face images from the local pattern features. In [MSK07], a scheme to reconstruct face images from match scores is developed. An affine transformation is utilized to approximate the behavior of the face recognition system. A similarity score matrix is generated firstly based on the target face recognition system, and an affine space is subsequently learned based on the similarity matrix. Given the distances of the targeted subject's template, the template is embedded in the affine space; an affine transformation is applied to retrieve the original template.

In the era of deep learning, FR systems based on deep learning models have been widely deployed. Reconstructing face images from the deep features draws the attention of the public due to the privacy and security concerns. To generate face images from deep templates, two main branches have been proposed in the literature, i.e., white-box based (feature extractor model is known) and black-box based (feature extractor model is unknown) approaches. [ZS16, Co17] are typical white-box based approaches while [Ma18] is a black-box approach.

[ZS16] proposed a method to invert FaceNet face embedding [SKP15] to realistic-looking face images based on convolutional neural networks. Specifically, face image reconstruction is formulated as a minimization problem that attempts to minimize the template difference between original and reconstructed images. However, a regularization function constructed using the intermediate nodes of the target extractor model network is required in the proposed scheme. Hence the detailed parameter of the target template extractor should be known. In reality, however, the extractor model may not usually be available. In [Co17], a method to synthesize a frontal, neutral expression face image from the FaceNet feature [SKP15] is proposed. Firstly, the landmarks and textures of face images are estimated by off-the-shelf landmark detection tools and a warping technique. Face images are then generated based on differentiable image warping by combining landmarks and textures information. In the implementation, however, the last convolution layer instead of the final output of a pre-trained FaceNet model is used; hence the parameters of the extractor model, i.e., FaceNet, should be known.

In [Ma18], a neighborly de-convolutional neural network (NbNet) is designed to reconstruct face images from their deep templates. Unlike the aforementioned models, the knowledge of the target subject and the deep network are not required. Specifically, the NbNet is a cascade of multiple stacked de-convolution blocks and a convolution block. Unlike the conventional convolution operation, de-convolution operations can up-sample the input data to produce a larger output feature map. Subsequently, a convolution operation is applied on the output of the de-convolution output to generate the output face images. In [Ma18], GAN synthesized face images, and two augmented benchmark face datasets are used to train the model. The system is evaluated with type-I and type-II attack settings.

Although, a variety of approaches to reconstruct face images from deep features are reported. The face images generated from the aforementioned approaches are not in HD resolution. For example, the output image size in [Ma18] is 160×160, and the size in [Co17, ZS16] is 224×224. Such low resolution may not meet the requirement of some applications, for example, face liveness detection.

On the other hand, a number of techniques that generate face images based on adversarial models had been proposed. Among various techniques, StyleGAN2 [KLA19, Ka20] is one of the most popular methods to generate high-resolution and realistic face images. In StyleGAN2, a mapping network is used to map points in latent space to an intermediate latent space, then the intermediate latent space is utilized to control the style in the generator model.

In this paper, we propose an alternative way to achieve template inverting tasks by incorporating the StyleGAN2's capabilities to generate face images from the deep templates, and show that generating HD face images may threaten FR systems, especially liveness detection.

3 Generating HD Face Images from deep features

A method to generate HD face images based on deep features is presented in this section. We firstly assume that the stored template $v = f(x) \in \mathscr{V}$ is known to the adversary, $\mathscr{V}$ denotes the feature space. We also assume that the adversary can generate unlimited input-output data pairs of the deep feature extractor. However, the deep feature extractor is regarded as a black-box that is not necessarily known to the adversary. By learning a mapping between latent vector space of the StyleGAN2 and the feature vector space of the face feature extraction model, latent code of the corresponding compromised template can be predicted. Next, a pre-trained StyleGAN2 model [Ka20] is utilized to generate HD face images by exploiting the information originating from the deep features. Next, those generated fake images are used to access (attack) a target system illegally. An overview of the method is shown in Fig. 1.

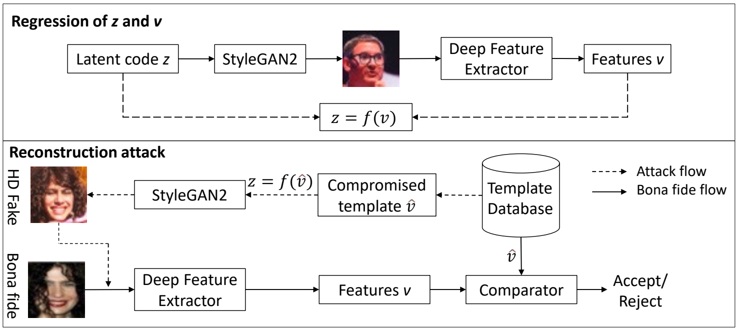


Fig. 1: An overview of the proposed method.

We expect that the fake face images generated based on deep features should preserve some characteristics from the original face image, e.g., gender and age. Besides, we also expect that the generated fake face images can also be utilized to gain illegal access to the system by launching type-I and type-II attacks. To achieve this task, the input of the StyleGAN2, i.e., the latent code, should be determined first. The problem now can be defined as given a deep features v, how to determine the corresponding latent code z.

To solve above problem, a regression model based on neural networks is established to learn the mapping between latent space and feature space and denoted as the mapping network. The mapping network has two hidden layers with 1024 nodes having a relu activation function. The input and output size of the network are both 512, while the linear activation function is used for the output layer as this is a regression task. Finally, the model is trained based on the MSE loss by Adam optimizer. Given an input face features v, the model is expected to output the corresponding latent code z.

To obtain the training data to train the regression model, 3.2 million $< v_i, z_i >$ pairs are generated based on the adopted pre-built models. Specifically, the random latent code vector $z_i \in \mathbb{R}^{512}$ is generated pirorly, then the latent code z_i is supplied as an input to the pre-trained StyleGAN2 model, and a fake face image can be generated. Subsequently, face feature v_i is extracted from the generated image by the InsightFace model. It is worth highlighting that no dataset is needed to train the regression model, as we directly utilize the pre-trained model to generate the training samples.

It is worth highlighting that the proposed method to generate face images from features enjoys advantages compared with existing methods:

1. HD face images are generated in our scheme. Compared with the existing works such as 160×160 in [Ma18] and 224×224 in [Co17, ZS16], the generated face images in this paper is in 1024×1024 resolution. HD face images provide extra advantages to perform the attack, such as liveness detection.
2. Our pipeline is concisely simple and efficient, which only needs to train a regression model. Besides, no extra training dataset is needed. Simultaneously, the attack can still be feasible.

Tab. 1: SARs of type-I and type-II attacks on LFW.

FAR	Normal TAR	Threshold	Type I	Type II
0.0%	91.77%	0.4183	0.49%	0.10%
0.1%	93.80%	0.4008	1.42%	0.46%
1.0%	97.40%	0.3669	10.11%	3.13%
10.0%	98.93%	0.3337	46.08%	18.30%

4 Experiments and Results

In our experiment, a pre-trained StyleGAN2 model[5] is adopted. The pre-trained model is trained on FFHQ dataset [KLA19] at 1024×1024. The 512-D face features were extracted by the InsightFace (ArcFace) with ResNet-100 backbone [De18] [6].

To evaluate the performance of the proposed method, the Labeled Faces in the Wild (LFW) [Hu08] face dataset is adopted in this experiment. The face features of LFW bona fide face images are extracted. Next, the face features are fed into the regression model to compute the corresponding latent code vectors. Then the fake HD face images are generated by the pre-trained StyleGAN2.

To simulate the attack, the generated fake face images are then directly used as the input of the feature extractor to extract the features. Finally, features are compared with the stored template to compute the similarity score.

We quantitatively evaluated the security of the deep features under type-I and type-II attacks. Specifically, the official LFW verification protocol[7] is adopted in this paper to compute true accept rate (TAR) at different false accept rate (FAR) on the deep features. To distinguish with the TAR in a normal situation, the TAR corresponding to specific FAR under attack situations is denoted as Success Attack Rate (SARs), and higher SAR means a high risk of compromising. The results are shown in Table 1.

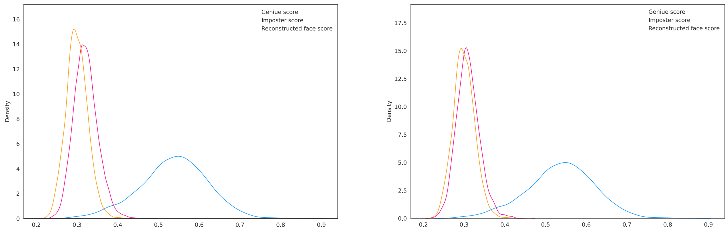
From the table 1, we observe that the generated face images can achieve relatively high SARs under type-I attack settings. The SAR can reach 46% under the threshold at FAR=10%, and the attack SAR is nearly 5 times higher than the false accept attack rate (FAR=10%) by attempting the access with a random imposter sample, which implies high risks of adversary attacks. Under FAR=1%, the attack SAR under type-I can still achieve 10.11%. On the other hand, the type II attack shows weaker performance than type-I, as the attack SAR can only reach 18.30% under the FAR=10% threshold. However, the risks still persist under this setting.

To show the difference between the proposed attack and the false accept attack, the comparing scores (similarity) distribution between genuine pairs, imposter pairs, and attack pairs are shown in Fig. 2. It is seen that the generated HD face images can generate higher similarity scores than a random imposter sample. This suggests that the proposed method

[5] https://nvlabs-fi-cdn.nvidia.com/StyleGAN2/networks/StyleGAN2-ffhq-config-f.pkl

[6] https://www.dropbox.com/s/tj96fsm6t6rq8ye/model-r100-arcface-ms1m-refine-v2.zip

[7] http://vis-www.cs.umass.edu/lfw/#views



(a) Type-I

(b) Type-II

Fig. 2: Scores (similarity) distribution between genuine pairs, imposter pairs and attack pairs.

Tab. 2: Comparison with the state-of-the-art methods

	Deep feature extractor	Resolution	Liveness pass rate
Mai et al. [Ma18] (VGG-NbA-P)	Facenet	160×160	1.77%
Ours	InsightFace	1024×1024	50.70%

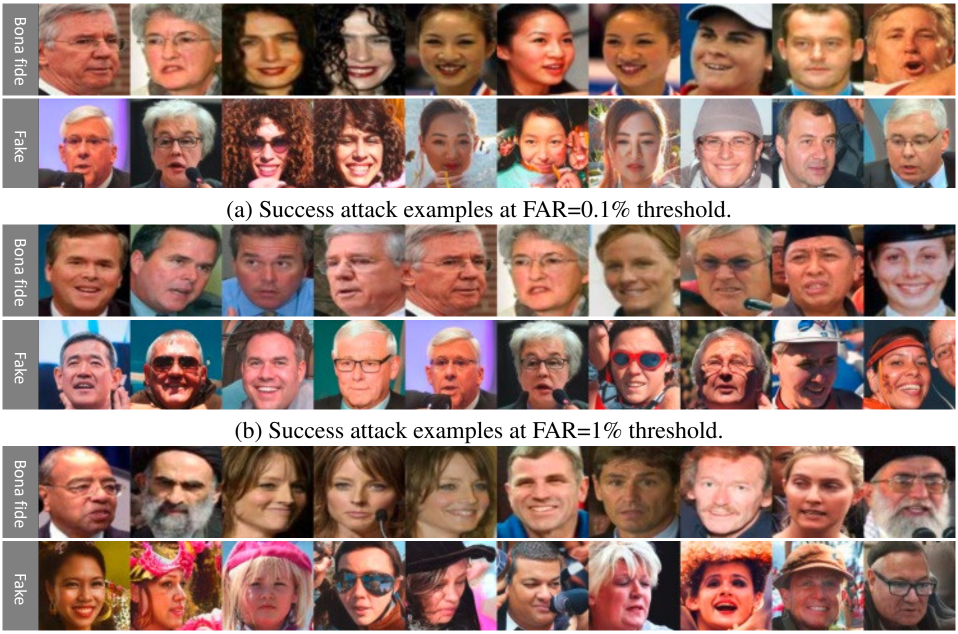
is not a false acceptance attack and more vulnerable in the practical biometric systems deployment.

Examples of the generated HD face images based on randomly selected subjects in LFW are shown in Fig. 3. We can find that the succeeded fake face instances show high similarity visually compared with the corresponding bona fide images. Failure cases are also shown in Fig. 3 (c). The results suggest that the proposed method could be a real threat to practical FR systems.

To further validate the advantage of the HD face images, a COTS liveness detection cloud computing API is used to evaluate the generated face images. The COTS returns three suggestions for each image, i.e., block, review, and pass. If the image is detected as block class, then this face image will be regarded as non-live. The detailed results are shown in Fig. 4.

As shown in Fig. 4, 50.7% of our generated HD face images manage to fool the face liveness detector, while 60.4% images from [Ma18] under VGG-NbA-P setting are blocked. This is because the generated face images from [Ma18] are in low-resolution, and also contain artifacts. Examples of blocked and passed face images generated by our model can be found in Fig. 5.

Table 2 shows a comparison with one of the current state-of-the-art schemes. [Ma18] shows superior performance due to the specific designed neighborly de-convolutional neural network (NbNet). But surprisingly, the combination of StyleGAN2 and a simple regression can still achieve 10.11% SAR at the FAR=1%.



(a) Success attack examples at FAR=0.1% threshold.

(b) Success attack examples at FAR=1% threshold.

(c) Failed attack examples at FAR=1% threshold.

Fig. 3: Generated HD face images from the randomly selected subjects in LFW. (a) and (b) shows the success attack examples at the FAR=0.1% and FAR=1% threshold, respectively. (c) shows the failure attack examples at the FAR=0.1% threshold (Best view in color and zoom in).

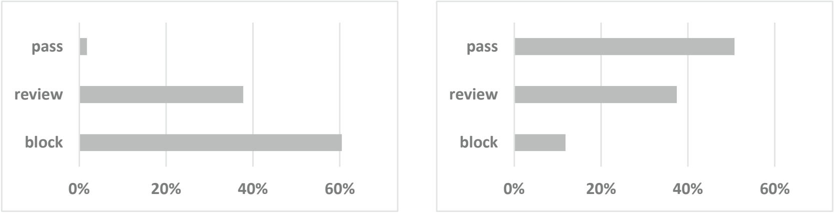
5 Conclusion

In this paper, a method to generate HD face images from the deep features has been presented. By finding a mapping between the latent space and the feature space, the method allows HD face images generation from deep face features with StyleGAN2. The forged HD face images do not need to resemble exactly the bona fide face images. However, it could be exploited by the adversary to gain illegal access to the face recognition systems.

We also simulate type-I and type-II attacks on the LFW dataset, and the quantitative results show that the proposed method can achieve comparable performance. Compared with state-of-the-artwork in [Ma18], our method is simple, and higher resolution images can be attained.

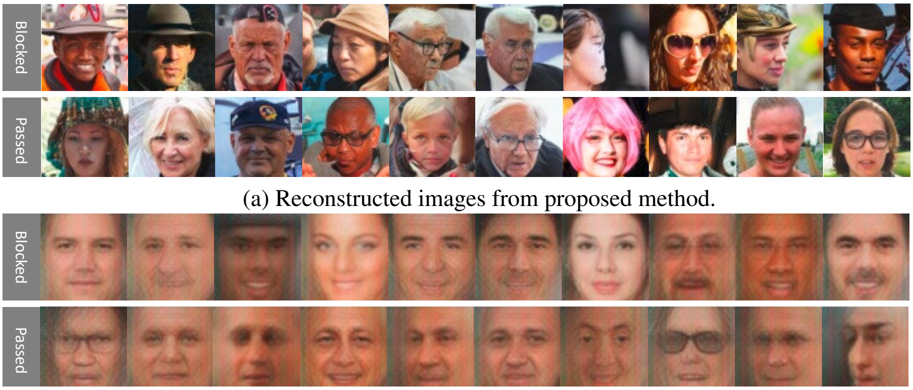
The generated HD face images are also evaluated by a COTS liveness detection API, and it shows that 50% of samples can pass the liveness detection system. This indicates that the current usage of COTS photo-based liveness detection API is at risk and still needs to be improved.

It is interesting to extend the proposed approach into the investigation of existing biometric template protection algorithms. For example, by finding a mapping between the protected (or transformed) template space and feature space, is it possible to generate HD face im-



(a) Mai et al. [Ma18] (b) Ours

Fig. 4: Liveness detection results based on COTS .



(a) Reconstructed images from proposed method.

(b) Reconstructed images from [Ma18]. (Source images provided by Mai.)

Fig. 5: Blocked and passed examples.

ages to compromise the security? On the other hand, StyleGAN2 model shows impressive performance in generating HD face images, but the modeling capability of StyleGAN2 model is limited in our work as it is pre-trained. The attack SAR could be improved by fine-tuning the model with specific training data. In addition, utilizing the reconstructed images to attack different face recognition systems is also an interesting future investigation.

Acknowledgement

This work was supported by grants from Ministry of Higher Education (MOHE) Malaysia through Fundamental Research Grant Scheme (FRGS/1/2018/ICT02/ MUSM/03/3). The authors would like to thank Dr. Mai Guangcan for his gratitude in offering the reconstructed face images.

References

[Ad03] Adler, Andy: Sample images can be independently restored from face recognition templates. In: CCECE 2003-Canadian Conference on Electrical and Computer Engineering.

Toward a Caring and Humane Technology (Cat. No. 03CH37436). volume 2. IEEE, pp. 1163–1166, 2003.

[Co17] Cole, Forrester; Belanger, David; Krishnan, Dilip; Sarna, Aaron; Mosseri, Inbar; Freeman, William T: Synthesizing normalized faces from facial identity features. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3703–3712, 2017.

[Co18] Commission, European: , 2018 reform of EU data protection rules, 2018.

[De18] Deng, Jiankang; Guo, Jia; Xue, Niannan; Zafeiriou, Stefanos: Arcface: Additive angular margin loss for deep face recognition. arXiv preprint arXiv:1801.07698, 2018.

[FLY14] Feng, Yi C; Lim, Meng-Hui; Yuen, Pong C: Masquerade attack on transform-based binary-template protection based on perceptron learning. Pattern Recognition, 47(9):3019–3033, 2014.

[Hu08] Huang, Gary B; Mattar, Marwan; Berg, Tamara; Learned-Miller, Eric: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: Workshop on Faces in 'Real-Life' Images: Detection, alignment, and recognition. 2008.

[Ka20] Karras, Tero; Laine, Samuli; Aittala, Miika; Hellsten, Janne; Lehtinen, Jaakko; Aila, Timo: Analyzing and Improving the Image Quality of StyleGAN. In: Proc. CVPR. 2020.

[KLA19] Karras, Tero; Laine, Samuli; Aila, Timo: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410, 2019.

[Ma18] Mai, Guangcan; Cao, Kai; Yuen, Pong C; Jain, Anil K: On the reconstruction of face images from deep face templates. IEEE transactions on pattern analysis and machine intelligence, 41(5):1188–1202, 2018.

[MJ13] Mignon, Alexis; Jurie, Frédéric: Reconstructing faces from their signatures using RBF regression. In: British Machine Vision Conference 2013. pp. 103–1, 2013.

[MSK07] Mohanty, Pranab; Sarkar, Sudeep; Kasturi, Rangachar: From scores to face templates: a model-based approach. IEEE transactions on pattern analysis and machine intelligence, 29(12):2065–2078, 2007.

[RdJG98] Rosenthal, Y; de Jager, G; Greene, J: A computerised face recall system using eigenfaces. In: Proceedings of the Eighth Annual South African Workshop on Pattern Recognition. pp. 53–57, 1998.

[SKP15] Schroff, Florian; Kalenichenko, Dmitry; Philbin, James: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823, 2015.

[Tr99] Tredoux, Colin; Rosenthal, Yon; da Costa, Lisa; Nuenz, D: Face reconstruction using a configural, eigenface-based composite system. SARMAC III, 1999.

[Tr06] Tredoux, Colin; Nunez, David; Oxtoby, Oliver; Prag, Bhavesh: An evaluation of ID: an eigenface based construction system: reviewed article. South African Computer Journal, 2006(37):90–97, 2006.

[ZS16] Zhmoginov, Andrey; Sandler, Mark: Inverting face embeddings with convolutional neural networks. arXiv preprint arXiv:1606.04189, 2016.

Emerging biometric modalities and their use: Loopholes in the terminology of the GDPR and resulting privacy risks

Tamas Bisztray[1], Nils Gruschka[2], Thirimachos Bourlai[3], Lothar Fritsch [4]

Abstract: Technological advancements allow biometric applications to be more omnipresent than in any other time before. This paper argues that in the current EU data protection regulation, classification applications using biometric data receive less protection compared to biometric recognition. We analyse preconditions in the regulatory language and explore how this has the potential to be the source of unique privacy risks for processing operations classifying individuals based on soft traits like emotions. This can have high impact on personal freedoms and human rights and, therefore, should be subject to data protection impact assessment.

Keywords: biometric data, data protection impact assessment, GDPR, taxonomy, profiling, privacy, digital identity.

1 Introduction

The General Data Protection Regulation (GDPR) [Eu16] is the data protection regulation of the European Union (incl. the European Economic Area). All processing of personal data of EU citizens must comply with this regulation. In particular the processing of biometric data has the potential to impose high risks to the rights and freedoms of individuals. The exact definition of biometric data is not always consistent across legal and scientific publications, or standards [Ar12, Me08, JNR16, Or10]. Thus, it can be difficult to interpret the resulting guidelines, implement appropriate privacy controls, or to conduct a data protection impact assessment (DPIA) as defined in Article 35 of the GDPR (which is required upon processing of biometric data according to [Ki01]).

As per the GDPR, data defined as biometric does not receive any special protection when compared to other personal data, such as name, email addresses, usernames, etc. Only when the processing of biometric data is done "*for the purpose of uniquely identifying a natural person*", the data falls under the protection of Article 9. This article prohibits processing unless special conditions are fulfilled (e.g, the data subject has given explicit consent). This means that, other processing purposes using biometric data (unlike identification) only require the same level of protection as processing "ordinary" personal data. As an example emotional reactions are measurable physiological processes, and in a study

[1] University of Oslo, Department of Informatics, Gaustadalléen 23B, 0373 Oslo, Norway, tamasbi@ifi.uio.no

[2] University of Oslo, Department of Informatics, Gaustadalléen 23B, 0373 Oslo, Norway, nilsgrus@ifi.uio.no

[3] University of Georgia, 115 Boyd Graduate Studies Research Center 200 D.W. Brooks Drive, Athens, Georgia 30602, United States, thirimachos.bourlai@uga.edu

[4] Oslo Metropolitan University, Pilestredet 35, 0166 Oslo, Norway, lothar.fritsch@oslomet.no

researchers showed how seeing positive or negative social media posts can manipulate the emotions of people, experimenting on 689,003 individuals without their knowledge [KGH14]. In such situations, the rights and freedoms of individuals can be threatened and, yet, Article 9 would not apply as the purpose of processing is not identification. Secondly, whether such data can classify as biometric under the interpretation of the GDPR must be examined, as Article 9 is only applicable to special categories, not to personal data in general. In this paper we analyse the definition of biometric data in the GDPR and compare it to other existing definitions. We show how preconditions in the GDPR are excluding towards certain biometric modalities and modes of operations namely, soft biometrics and classification. Finally, we will show risks resulting from the limited protection for these modes of operation.

The GDPR doesn't point to other standards such as the ISO/IEC 2382-37:2017(E) (ISO standard), to interpret the language it uses, therefore, in the remainder of this paper the default interpretation of the terminology should be according to the taxonomy of the GDPR unless specified otherwise. Section 2 presents the Methodology for our analysis, Section 3 examines preconditions in the legal taxonomy, Section 4 discusses classification as a mode of operation, while Section 5 presents unique risks to rights and freedoms using soft biometrics for classification purposes. Section 6 summarises the results.

2 Methodology

This paper utilises the WPR (*What's the problem represented to be*) approach developed by Bacchi with the aim of policy analysis [Ba12]. It constitutes of the following seven steps: (i) WPR apprehends that policies can contain an implicit representation of the problem they are aiming to solve, by constructing a representation of reality which the policy responds to. (ii) This representation is based on (iii) presuppositions and assumptions that often (iv) omit and silence other aspects of reality, (v) and as such it produces a series of undesirable effects for the subjects in question. (vi) This requires the representation to be analysed and altered, leading to a new problem representation which then needs to be (vii) analysed again by the WPR approach.

In our analysis this approach translates to: (i) examining how the GDPR defines and protects biometric data (ii) where the description of reality only considers recognition as the purpose of use, (iii) requiring the notions of "specific technical processing" and "uniqueness" (Art. 4 GDPR). (iv) Tying the definition and protection of data to preconditions, involuntarily silences classification purposes and the processing of non-unique biometric information. The main focus of this paper is to (v) highlight a series of undesirable effects stemming from these silences, in the form of privacy risks that are impacting, rights and freedoms of natural persons. Based on this the following (vi) problem representations will be examined: 1. The inclusion of specific technical processing in the definition silences and excludes certain forms of biometric information from the category of biometric data. 2. Requiring that biometric data must allow or confirm unique identity can exclude forms of processed biometric information such as soft biometric templates. 3. Tying the protection provided by Article 9 to "purpose of use" is excluding towards other modes operations such as verification and classification, based on biometric traits and characteristics.

We use discourse analysis as a methodology scrutinising legal text, where our focus is to highlight risks to rights and freedoms of individuals. The arrangement of this paper and the presentation of the results of our analysis will follow a thematic line of thought instead of the repetitive steps of the WPR approach. Note, that in this paper our main focus is to outline the problem and the resulting risks, not to propose modifications to the GDPR. Related work regarding legal analysis, privacy evaluation, and biometric technologies will be referenced at the relevant sections respectively. In [Ki01] Kindt argued that the artificial distinction the GDPR makes between categories of biometric data based on "specific technical processing" and "purpose of use" is unnatural and calls for further discussion about this definition. In the following we highlight how these distinctions can translate to unique privacy risks. We further expand the context of the discussion by the inclusion of the silenced notions of classification and soft biometrics.

3 Taxonomy adapted by regulators and resulting risks

Defining biometric data is not that simple as it seems on the first glance. One can find different definitions in relevant sources, like legal regulations, technical standards and scientific literature. The term biometrics comes from the ancient Greek words 'βιος' (bios) for life and 'μετρον' (metron) for measure, it is the measurement of traits and characteristics of living beings [Sc09]. Most sources in literature acknowledge physical, physiological and behavioural traits, while others use categories such as biological, physiographic, motoric or biochemical [BT20].

The GDPR defines biometric data in Article 4.14 as: "*personal data resulting from 'specific technical processing' relating to the physical, physiological or behavioural characteristics of a natural person, which 'allow or confirm the unique identification' of that natural person, such as facial images or dactyloscopic data*". As shown in Fig. 1, this puts presuppositions on personal data containing measured traits and characteristics of human beings, to be recognised legally as biometric data. This together with preconditions in Article 9 can dampen the protection of such personal data.

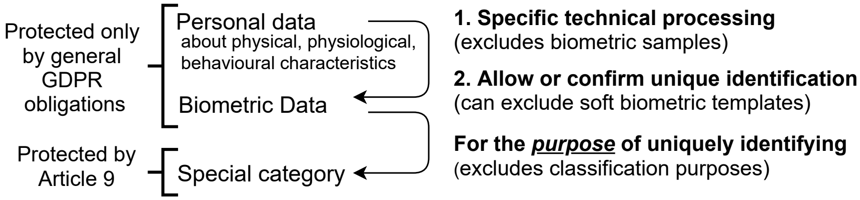

Protected only by general GDPR obligations
Personal data
about physical, physiological, behavioural characteristics
Biometric Data
Protected by Article 9
Special category
1. Specific technical processing
(excludes biometric samples)
2. Allow or confirm unique identification
(can exclude soft biometric templates)
For the *purpose* of uniquely identifying
(excludes classification purposes)

Fig. 1: Conditions in the GDPR

Personal data in question must relate to physical, physiological, and/or behavioural characteristics. This has to go through specific technical processing, where the resulting data allows or confirms identity, for such personal data to qualify as biometric. The order of the steps is not interchangeable. Article 9 only applies for data produced this way when its used for identification purposes. In contrast a proposed definition by the European Commission which was not adapted, defined biometric data as: "*Any personal data relating to the physical, physiological or behavioural characteristics of an individual which allow their unique identification, such as facial images, or, dactyloscopic data.*" [Ja15]. This definition doesn't require specific technical processing.

Analysing the notion "allow or confirm the unique identification" Jasserand argues that "allow" refers to establishing the identity (biometric *identification*), whereas "confirm" refers to *verifying* identity, where together these correspond to the notion of *recognition* [Ja15, Ja16]. As Article 9 only mentions uniquely identifying, Bygrave and Tosoni concludes that verification purposes are excluded [BT20]. Considering recognition as the only mode of operation is a shortcoming not unique to the GDPR, but shared by several legal articles, scientific papers and standards, aiming to give a technological overview, including the ISO standard.

Article 29 Working Party (WP29) defines biometric data as: "*biological properties, behavioural aspects, physiological characteristics, living traits or repeatable actions where those features and/or actions are both unique to that individual and measurable, even if the patterns used in practice to technically measure them involve a certain degree of probability.*" This only requires the measurable features to be unique regardless of use, thus neither tying to "specific technical processing" nor to "recognition". However, requiring uniqueness can be silencing towards soft biometric traits. Soft biometric traits are defined by Dantcheva et al. as: "*physical, behavioural, or material accessories, which are associated with an individual, and which can be useful for recognising an individual. These attributes are typically gleaned from primary biometric data, are classifiable in pre-defined human understandable categories, and can be extracted in an automated manner.*" The authors further note that the use of soft biometrics is possible beyond recognition purposes [DER16]. Using the notion of "*characteristics*" in Article 4.14 instead of trait or feature, further reinforces that the goal is "*unique identification*", as a characteristic is a unique/distinctive trait.

The GDPR provides no interpretation for "specific technical processing" although it is a pivotal moment for understanding and managing potential risks. A biometric system performing recognition has two phases: enrolment and recognition. In [BT20] the authors note that such processes consist of multiple steps, listing seven higher level points which can be considered as specific technical processing. Point (a) in their list states: "*acquiring a reference measure of one or more physical, physiological or behavioural characteristics of a person (often termed 'enrolment')*". Enrolment can be further broken down to technical steps [JNR16]. These steps can be different in other modes of operation. For recognition these include: (1) capture the biometric trait of the individual (2) create a biometric sample, (3) feature extraction (4) creation of templates (5) storing templates as biometric reference data. Collecting and storing facial images and fingerprints (biometric samples) are not considered as biometric data or processing of biometric data under the GDPR (unlike in the ISO standard) as reflected by Recital 51 [Eu16] and also pointed out by Kindt [Ki01]. This means that "specific technical processing" only starts at the step of (3) feature extraction, but even that and step (4),(5) is tied to “purpose of use” whether it can produce data legally considered as biometric, as the result must allow recognition. This view of specific technical processing where biometric systems extract unique traits, to be used solely during the enrolment or the recognition phase, completely neglects other potential applications of biometric systems.

4 Classification as a mode of operation

WP29 notes that biometric systems in addition to recognition can be also used for other purposes: *“The categorisation/segregation of an individual by a biometric system is typically the process of establishing whether the biometric data of an individual belongs to a group with some predefined characteristic in order to take a specific action [...]* ”. While unique biometric traits are most suitable for recognition purposes, both unique and soft measurable traits are classifiable. For example a fingerprint can be used in all three discussed modes of operation, it can identify, verify or classify an individual. An example for the latter can be gender classification from fingerprint ridge count and fingertip size [GV19].

In multi-modal systems soft traits can increase performance, and even using only soft traits for identification is possible [Da11, NB20], but their most common use is for classification purposes [Ki13, vdP11, NB16]. A general system performing classification, instead of “enrolment and recongition” performs “training and classification”, usually based on machine learning (ML) algorithms. There are several approaches for training these classifiers, the discussion of which is outside the scope of this paper. Training can be performed on a data set that is independent from the data subject. Potential biases introduced during training such as sampling bias, over-fitting data, or temporal bias are among the few problems that could have a negative impact on rights and freedoms of the individuals. Under certain circumstances user data can be reused for training and correction, but for most individuals the first interaction with the biometric system will happen in the classification phase.

The primary classification tasks in ML are: binary classification where data is segregated into two sets, like classifying faces into female/male, thereby gendering biometric data on the way [OR15]. In multi-class classification there are more than two pre-established sets. Multi-label classification are typical in video surveillance scenarios, where age, gender, hair style, etc. can be determined from a single image. In imbalanced classification the number of cases for each class is not equally distributed.

The phase of classification contains: (1) capture of the biometric feature(s), (2) sample creation, (3) feature extraction. (4) Create biometric probe(s) (5) labelling by trained ML classifier (6) produce classification decision(s). For certain applications there can be a step in conjunction with the capture process, which is linking the measured signals to a certain activity, like presenting a stimulus, and observing the performed action of the individual. Such processing is not subject to Article 9 but can have far reaching consequences in profiling applications, and combined with certain modalities can contribute to increased risks.

5 Discussion

Reflecting on the example from the introduction, presenting a stimulus and measuring the emotional reactions of individuals for the purpose of determining their emotional state, is not under the protection of Article 9. We present the following non-exhaustive list of risks in connection to classification purposes considering both unique and soft traits:

Risk 1: Data that researchers and other definitions would already call biometric, needs to satisfy condition 1. and 2. from Fig. 1, to reach this category in the GDPR. If condition 1. (specific technical processing) is related to classification steps/purposes (i.e. either the training or the classification phase), resulting data in most cases won't qualify as biometric.

Risk 2: Due to the data minimisation principle, if the intended purpose is not identification, the extracted feature should not identify the data subject. Still if recognition is not the purpose, and strong anonymization or obfuscation would render the data unusable, the controller is not obliged to anonymise, and could as a result handle biometric data that confirms identity but doesn't get the protection of Article 9.

Risk 3: Classifying biometric traits of individuals, while they are identified by non-biometric means like a username, is not prohibited by Article 9, as usernames are not biometric data and the purpose of processing of the biometric data in questions is not identification. This can lead to individual profiling using physical, physiological or behavioural traits and characteristics of the data subject.

While determining if someone has a moustache might not appear to be a high risk scenario, McStay points out far reaching consequences for the use of emotional surveillance in AI applications [Mc20]. Psychological or cognitive biometrics relies on the measurement of cognitive or emotional states of an individual linked to certain activities [OTW19]. These states of mind are deducted from physical or behavioral actions/reactions, or bio-signals such as the electroencephalogram (EEG), electrocardiogram (ECG), or electrodermal response (EDR) of the individual in response to the presentation of a certain stimulus, e.g., viewing an image portraying a memorable event. These can be unique like EEG suitable for biometric recognition, or soft (non-unique) like measuring the emotional reaction to certain videos from watch time, pupil dilation or a survey.

Risk 4: Technological advancements will allow such psychological-based techniques to be more accurate, continuously present and immersive. This can change the impact or the level of risks such techniques can impose to the rights and freedoms of individuals. Therefore, consequences and harms may change as well as likelihood.

Psychological states such as emotions are measurable through several modalities. If such data is not considered to be biometric by the GDPR, another way to be eligible for special protection is through other special categories of data from Article 9, like racial or ethnic origin, data concerning health etc. As Kindt points out the GDPR doesn't confirm or reject this interpretation [Ki01]. For certain physiological characteristics like cardiac signals it is more obvious that they qualify as health data, but this can't be generalised.

Risk 5: If high quality data about physical, physiological or behavioural traits and characteristics are collected for classification purposes in large amounts, the controller might acquire data that reveals information about health, ethnicity, sexual orientation or other special categories of data, thus accumulating a large amount of sensitive personal data that will need special attention.

Risk 6: Personal data about physical, physiological, behavioural traits and characteristics, whether or not they can satisfy the conditions to be regarded as biometric data, their processing for classification purposes escapes the protection of Article 9.

Risk 7: According to Recital 26 [Eu16], when data is anonymized and the data controller is certain the data subject is not possible to be re-identified by any means and can demonstrate that, the processing of such anonymous data, including statistical or research purposes is not subject to the GDPR. Such data can still be used to evaluate even sensitive aspects of groups of people, leading to group profiling using sensitive information, the collection of which escapes Article 9 for classification purposes, in the context of biometric data. For data to remain unlikable, a sufficient anonymity set of similar data must be present, which is not always guaranteed [Pf07].

The definition of classification is not synonymous with profiling but it can "evaluate certain personal aspects" (Article 4.4 GDPR), and can have similar consequences to profiling which is a serious privacy risk discussed. If such processing enters the category of Article 22 automated decision making including profiling, paragraph 1 states: "*The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.*" However, this is a weaker protection compared to Article 9. It is not prohibited by default and it is not obvious how data subjects can exercise their rights. Similarly to Article 9 explicit consent gives away the right, but more importantly Article 22.2.a states paragraph 1 shall not apply: "if is necessary for entering into, or performance of, a contract between the data subject and a data controller". For example a legitimate legal bases can be direct marketing.

Risk 8: If Article 22.2 is fulfilled it causes the data subject to lose the right not to be subject to such processing operations, and explicit consent is not required anymore, even for classification purposes using biometrics.

While Article 22, Recital 70 and 71 [Eu16] reinforce that the data subject shall have the right to object to direct marketing, automated decision making and profiling, the interest of the controller can override the interests or the fundamental rights and freedoms of the data subjects, as pointed out by Veale et al. [VBA18] and reflected in Recital 69 [Eu16]. Paragraph 4 of Article 22 states the "Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1)", but as we established certain biometric modalities and modes of operations will escape that protection.

6 Conclusion

In this paper we have shown high risks impacting data subject rights in connection with processing their physical, physiological, behavioural traits and characteristics for classification purposes. Unfortunately, as our problem representation showed, certain types of data are regarded as biometric by researchers or in standards but not by the GDPR. Even for data which qualifies as biometric, but used for purposes other than identification, the GDPR gives no more protection than general obligations. Processing certain types of soft biometrics including but not limited to ones from psychological-based techniques for classification, can present threats that rival those posed by the processing of unique traits for identification purposes. These risks and issues we brought forward in the discussion are further aiming to assist in the risk assessment step of data protection impact assessment, which we specifically recommend to use for biometric applications. Technological advancements can change risks associated with the processing of personal data related to physical, physiological, or behavioural traits. Therefore, a more inclusive systematisation, and a more technologically neutral definition by the legislator would be beneficial.

References

[Ar12] Article 29 Working Party: , WP193: Opinion 3/2012 on developments in biometric technologies, 2012.

[Ba12] Bacchi, Carol: Why Study Problematizations? Making Politics Visible. Open Journal of Political Science, 2012.

[BT20] Bygrave, Lee A.; Tosoni, Luca: Article 4(14). Biometric data. In: The EU General Data Protection Regulation (GDPR). Oxford University Press, February 2020.

[Da11] Dantcheva, Antitza; Velardo, Carmelo; D'Angelo, Angela; Dugelay, Jean-Luc: Bag of soft biometrics for person identification. Multimedia Tools and Applications, 51(2):739–777, 2011.

[DER16] Dantcheva, Antitza; Elia, Petros; Ross, Arun: What Else Does Your Biometric Data Reveal? A Survey on Soft Biometrics. 11, March 2016.

[Eu16] European Parliament and Council: , Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance), 2016.

[GV19] Gnanasivam, P.; Vijayarajan, R.: Gender classification from fingerprint ridge count and fingertip size using optimal score assignment. Complex & Intelligent Systems, 5(3):343–352, October 2019.

[Ja15] Jasserand, Catherine A.: Avoiding terminological confusion between the notions of 'biometrics' and 'biometric data': an investigation into the meanings of the terms from a European data protection and a scientific perspective. International Data Privacy Law, p. ipv020, September 2015.

[Ja16] Jasserand, C.: Legal Nature of Biometric Data: From 'Generic' Personal Data to Sensitive Data. European Data Protection Law Review, 2, 2016.

[JNR16] Jain, Anil K.; Nandakumar, Karthik; Ross, Arun: 50 years of biometric research: Accomplishments, challenges, and opportunities. Pattern Recognition Letters, 79:80–105, August 2016.

[KGH14] Kramer, A. D. I.; Guillory, J. E.; Hancock, J. T.: Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National Academy of Sciences, (24), 2014.

[Ki13] Kindt, Els J.: Privacy and Data Protection Issues of Biometric Applications: A Comparative Legal Analysis. Springer Netherlands, 2013.

[Ki01] Kindt, E. J.: Having yes, using no? About the new legal regime for biometric data. Computer Law & Security Review, 34(3):523–538, 2018-06-01.

[Mc20] McStay, Andrew: Emotional AI, soft biometrics and the surveillance of emotional life: An unusual consensus on privacy. Big Data & Society, SAGE Publications Ltd, 7(1):2053951720904386, January 2020.

[Me08] Meints, Martin et al.: Biometric Systems and Data Protection Legislation in Germany. In: 2008 International Conference on Intelligent Information Hiding and Multimedia Signal Processing. pp. 1088–1093, 2008.

[NB16] Narang, Neeru; Bourlai, Thirimachos: Gender and ethnicity classification using deep learning in heterogeneous face recognition. In: 2016 International Conference on Biometrics. 2016.

[NB20] Narang, Neeru; Bourlai, Thirimachos: Classification of Soft Biometric Traits When Matching Near-Infrared Long-Range Face Images Against Their Visible Counterparts. In (Bourlai, Thirimachos; Karampelas, Panagiotis; Patel, Vishal M., eds): Securing Social Identity in Mobile Platforms. Springer International Publishing, Cham, 2020.

[Or10] Ortega-Garcia, Javier et al.: The Multiscenario Multienvironment BioSecure Multimodal Database (BMDB). IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(6):1097–1111, June 2010.

[OR15] Othman, Asem; Ross, Arun: Privacy of Facial Soft Biometrics: Suppressing Gender But Retaining Identity. In (Agapito, Lourdes; Bronstein, Michael M.; Rother, Carsten, eds): Computer Vision - ECCV 2014 Workshops, volume 8926, pp. 682–696. Springer International Publishing, Cham, 2015. Series Title: Lecture Notes in Computer Science.

[OTW19] Obaidat, Mohammad S.; Traore, Issa; Woungang, Isaac, eds. Biometric-Based Physical and Cybersecurity Systems. Springer International Publishing, 2019.

[Pf07] Pfitzmann, Andreas; Dresden, TU; Hansen, Marit; Kiel, ULD: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management – A Consolidated Proposal for Terminology. p. 84, 2007.

[Sc09] Schatten, Markus: , Towards a General Definition of Biometric Systems, 2009.

[VBA18] Veale, Michael; Binns, Reuben; Ausloos, Jef: When data protection by design and data subject rights clash. International Data Privacy Law, 8(2):105–122, 2018.

[vdP11] van der Ploeg, Irma: Normative Assumptions in Biometrics: On Bodily Differences and Automated Classifications. In (van der Hof, Simone; Groothuis, Marga M., eds): Innovating Government, volume 20, pp. 29–40. T. M. C. Asser Press, The Hague, The Netherlands, 2011. Series Title: Information Technology and Law Series.

Improved Post-quantum-secure Face Template Protection System Based on Packed Homomorphic Encryption

Hiroto Tamiya[1], Toshiyuki Isshiki[1], Kengo Mori[1], Satoshi Obana[2], Tetsushi Ohki[3]

Abstract: This paper proposes an efficient face template protection system based on homomorphic encryption. By developing a message packing method suitable for the calculation of the squared Euclidean distance, the proposed system computes the squared Euclidean distance between facial features by a single homomorphic multiplication. Our experimental results show the transaction time of the proposed system is about 14 times faster than that of the existing face template protection system based on homomorphic encryption presented in BIOSIG2020.

Keywords: Face recognition, biometric template protection, post-quantum cryptography, homomorphic encryption, packing method, and squared Euclidean distance.

1 Introduction

Biometric authentication is widely used in many applications ranging from unlocking smartphones to immigration controls at airports due to its convenience of being free from memory and possession. On the other hand, since biometric characteristics are unique and cannot be changed, they are classified as sensitive data in many regulations and laws, such as the EU General Data Protection Regulation (GDPR). Therefore, it is essential to prevent the leakage of biometric information in the authentication process.

To this end, various biometric template protection (BTP) systems, which compare biometric information while keeping it protected, have been presented [BDL15, MK19, RU11]. One of the main types is the BTP systems based on homomorphic encryption (HE), which directly compute similarities between encrypted biometric features. Many HE-based BTP systems have been presented for a wide range of modalities [Bo18, Dr19, Ko19, Ya13].

In recent years, there has been a remarkable development of lattice-based HE schemes, which are considered to be secure even for quantum computers, and a lot of practical schemes has been presented [BGV12, Ch17, FV12]. In BIOSIG2020, Kolberg et al. report biometric performance and running time of a face template protection system using lattice-based HE schemes for different feature representations [Ko20]. Their experimental results show that the transaction time for integer representation features is about 700 ms. This could be a bottleneck when the scheme is deployed in large-scale systems.

[1] NEC Corporation, Japan, {htamiya, toshiyuki-isshiki, ke-mori.bx}@nec.com
[2] Hosei University, Japan, obana@hosei.ac.jp
[3] Shizuoka University, Japan, ohki@inf.shizuoka.ac.jp

This paper proposes an efficient face template protection system based on homomorphic encryption for integer representation features. Our experimental results show the transaction time of the proposed system is about 50 ms and about 14 times faster than that of the system presented in BIOSIG2020 [Ko20].

2 Homomorphic Encryption

Homomorphic encryption (HE) is public key encryption that allows operations on messages without decrypting the ciphertexts. HE is defined as follow:

Definition 1 (Homomorphic Encryption)**.** *A homomorphic encryption scheme* HE*for message space $\mathcal{M}$ consists of five algorithms* KeyGen*,* Enc*,* Dec*,* HomAdd*, and* HomMul*.*

$(pk, sk) \leftarrow \mathrm{KeyGen}(1^\lambda)$ *: The key generation algorithm* KeyGen *is a probabilistic algorithm. On input a security parameter λ, it outputs a key pair (pk, sk).*

$c \leftarrow \mathrm{Enc}(pk, m)$ *: The encryption algorithm* Enc *is a probabilistic algorithm. On input a public key pk and a message $m \in \mathcal{M}$, it outputs a ciphertext c.*

$m \leftarrow \mathrm{Dec}(sk, c)$ *: The decryption algorithm* Dec *is a deterministic algorithm. On input a secret key sk and a ciphertext c, it outputs a message $m \in \mathcal{M}$.*

$c \leftarrow \mathrm{HomAdd}(pk, c_1, c_2)$ *: The homomorphic addition algorithm* HomAdd *is a deterministic algorithm. On input a public key pk, a ciphertext c_1 of a message $m_1 \in \mathcal{M}$, and a ciphertext c_2 of a message $m_2 \in \mathcal{M}$, it outputs a ciphertext c of the message $m_1 + m_2 \in \mathcal{M}$.*

$c \leftarrow \mathrm{HomMul}(pk, c_1, c_2)$ *: The homomorphic multiplication algorithm* HomMul *is a deterministic algorithm. On input a public key pk, a ciphertext c_1 of a message $m_1 \in \mathcal{M}$, and a ciphertext c_2 of a message $m_2 \in \mathcal{M}$, it outputs a ciphertext c of the message $m_1 \times m_2 \in \mathcal{M}$.*

Efficient lattice-based HE schemes, such as [BGV12, BV11, SS11, Ya13], use polynomial ring $R_t = (\mathbb{Z}/t\mathbb{Z})[x]/(x^N + 1)$ as the message space, where t is some integer and N is a power of two. In other words, the HE schemes can encrypt N integers by packing them into a single polynomial of degree $N - 1$ and perform homomorphic operations for polynomials.

For such HE schemes, Yasuda et al. presented a message packing method for computing the inner product with a single homomorphic multiplication [Ya14]. Specifically, for vectors $\boldsymbol{a} = (a_0, a_1, \ldots, a_{n-1})$ and $\boldsymbol{b} = (b_0, b_1, \ldots, b_{n-1})$ of length $n \leq N$, let $\mathrm{Enc}(\boldsymbol{a})$ and $\mathrm{Enc2}(\boldsymbol{b})$ denote the encryption of packed messages $\mathrm{pm}(\boldsymbol{a}) = \Sigma_{i=0}^{n-1} a_i x^i$ and $\mathrm{pm2}(\boldsymbol{b}) = -\Sigma_{i=0}^{n-1} b_i x^{n-i}$. Let c_a and c_b denote ciphertext given by $\mathrm{Enc}(\boldsymbol{a})$ and $\mathrm{Enc2}(\boldsymbol{b})$, respectively. Let c denote the ciphertext given by $\mathrm{HomMul}(c_a, c_b)$. Let m_0 denote the constant term of the decryption result $\mathrm{Dec}(sk, c) \in R_t$. Then we have $m_0 = \langle \boldsymbol{a}, \boldsymbol{b} \rangle \bmod t$, where $\langle \boldsymbol{a}, \boldsymbol{b} \rangle$

denotes the inner product of vectors. In other words, the constant term of the decryption result gives the inner product of two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ for sufficiently large t.

HE schemes are desire to be IND-CPA secure, which implies the schemes have onewayness and indistinguishability against attackers with access only to a public key. Onewayness means it is infeasible to obtain information about a message from the corresponding ciphertext. Indistinguishability means it is infeasible to identify the message corresponding to a received ciphertext.

3 Face Template Protection System

3.1 Security Requirements for Biometric Template Protection System

ISO/IEC 24745 [IS11] defines three security requirements for biometric template protection systems: irreversibility, unlinkability, and renewability. Irreversibility means it is impossible to retrieve original samples from templates. Unlinkability means two templates cannot be linked to the same subject. Renewability means new templates can be created and old templates can be revoked.

3.2 Proposed Face Template Protection System

The proposed face template protection system employs a facial feature extraction algorithm based on deep learning and a HE scheme which encrypts integers described in Section 2. Since facial feature extraction algorithms based on deep learning output floating-point features, the output features need to be converted to integer features. In this section, we first explain a quantization method to convert floating-point features into integer features. Then, we explain the behavior and security of the proposed system.

3.2.1 Quantization

Floating-point values are converted to integer values of arbitrary precision as follows: Firstly, a floating-point value x_{float} is normalized to the range $0 \leq x_{norm} \leq 1$ using the minimum x_{min} and maximum x_{max} values of features. Since it is generally difficult to obtain the true values of x_{min} and x_{max}, the values obtained from pre-collected features are used instead. Outliers are rounded off to 0 or 1. Finally, x_{norm} is converted to a m-bit integer value $x_{int} = \lfloor x_{norm} \times C \rfloor$, where $C = 2^m - 1$ and $\lfloor \cdot \rfloor$ denotes the floor function. We call the number of bits m quantization precision. Integer features are obtained by applying the above method to all dimensions of floating-point features.

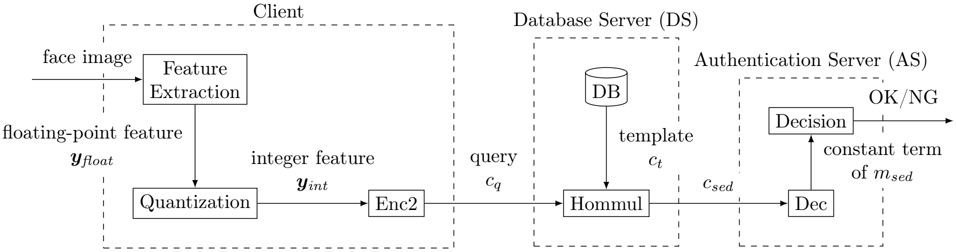


Fig. 1: Overview of the proposed system.

3.2.2 System Behavior

The proposed system computes the squared Euclidean distance (SED) as the similarity of features. The SED between two features $\boldsymbol{x}$ and $\boldsymbol{y}$ is calculated using a single inner product as $\mathrm{SED}(\boldsymbol{x},\boldsymbol{y}) = \big\langle (\|\boldsymbol{x}\|^2, 1, -2\boldsymbol{x}), (1, \|\boldsymbol{y}\|^2, \boldsymbol{y}) \big\rangle$, where $\|\cdot\|$ denotes Euclidean norm of a vector. By utilizing the packing method to compute the inner product by a single homomorphic multiplication described in Section 2, the proposed system computes the encrypted SED by a single homomorphic multiplication.

Figure 1 shows the overview of the proposed system. The proposed system consists of three parties: a client, a database server (DS), and an authentication server (AS). The client and DS own a public key of the HE scheme and AS owns a secret key of the HE scheme. The enrollment and the verification flow of the proposed system works as follows:

Enrollment

1. The client extracts a floating-point feature $\boldsymbol{x}_{float}$ from a captured face image.
2. The client converts the floating-point feature $\boldsymbol{x}_{float}$ to an integer feature $\boldsymbol{x}_{int}$ using the quantization method described in Section 3.2.1.
3. The client computes a template $c_t \leftarrow \mathrm{Enc}\big(pk, (\|\boldsymbol{x}_{int}\|^2, 1, -2\boldsymbol{x}_{int})\big)$ and sends it to DS.
4. DS stores the template c_t to the database.

Verification

1. The client generates an integer feature $\boldsymbol{y}_{int}$ from a captured image in the same way as the enrollment process.
2. The client computes a query $c_q \leftarrow \mathrm{Enc2}\big(pk, (1, \|\boldsymbol{y}_{int}\|^2, \boldsymbol{y}_{int})\big)$ and sends it to DS.
3. DS computes $c_{sed} \leftarrow \mathrm{HomMul}(pk, c_t, c_q)$ and sends it to AS.

4. AS computes $m_{sed} \leftarrow \mathrm{Dec}(sk, c_{sed})$ and decides the result based on the constant term of m_{sed}.

As described in Section 2, the constant term of m_{sed} becomes $\left\langle (\|\boldsymbol{x}_{int}\|^2, 1, -2\boldsymbol{x}_{int}), (1, \|\boldsymbol{y}_{int}\|^2, \boldsymbol{y}_{int}) \right\rangle$ mod $t = \mathrm{SED}(\boldsymbol{x}_{int}, \boldsymbol{y}_{int})$ mod t. Thus, AS decides the result based on the SED between $\boldsymbol{x}_{int}$ and $\boldsymbol{y}_{int}$.

3.2.3 Security

The security of the proposed system is same as that of the HE-based face template protection system presented in BIOSIG2020 [Ko20]. The system assumes the honest-but-curious model, where each party behaves according to the protocol, but tries to learn information about the secret. This model does not take into account obvious attacks in which DS and AS collude to decrypt templates or queries.

Thorem 1. *If a HE scheme used in the proposed face template protection system is IND-CPA secure, the system achieves Irreversibility, unlinkability, and renewability in the honest-but-curious model.*

Proof sketch. *Irreversibility and unlinkability are achieved directly from one-wayness and indistinguishability of the IND-CPA secure HE scheme, respectively. Renewability can be achieved by having users re-enroll their templates due to randomness of the encryption algorithm of the HE scheme.*

Note that brute force attacks and false accept attacks are not taken care of in the proposed system. To prevent these attacks, other measures such as rate limiting and liveness detection are necessary.

4 Experimental Results

Selecting a large quantization precision basically improves biometric performance. However, it slows down running time of HE schemes since the size of messages to be encrypted becomes larger. Therefore, we evaluated both biometric performance and running time for different quantization precisions.

Experimental Setting The experiments were executed on a frontal image subset of the FERET dataset [Ph00] as in the previous work [Ko20]. FaceNet [SKP15] and ArcFace (RESNET34) [De19] with pre-trained models published on [SO20] and VGGFace2 (RESNET50) [Ca18] with the model pre-trained on MS-Celeb-1M were employed for feature extraction methods. The dimensions of their features are 128, 512, and 2048, respectively. The minimum and maximum values of enrolled features are used as x_{min} and x_{max} for quantization, respectively.

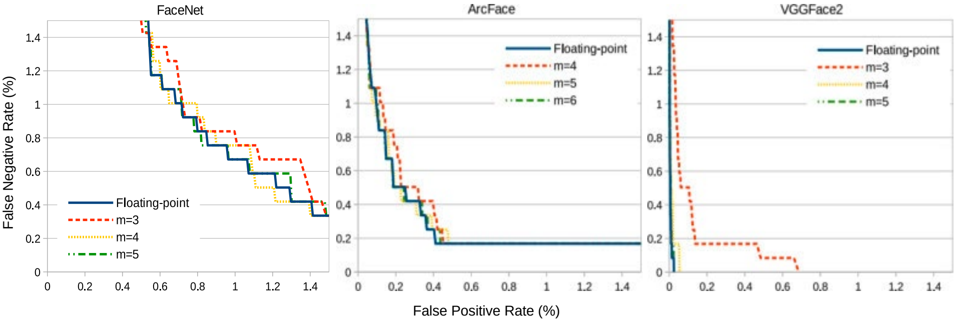


Fig. 2: DET curves for different quantization precisions.

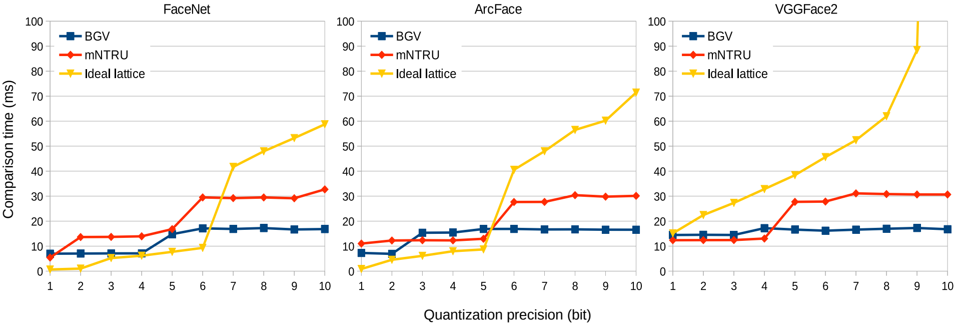


Fig. 3: Comparison time for different quantization precisions.

mNTRU [SS11], BGV [BGV12], and a ideal lattice-based scheme [Ya13] were implemented in the C language for HE schemes. The OpenSSL BIGNUM library [Th] was utilized for multi-precision integer operations to implement HE schemes. The number theoretic transformation (NTT) [CT65] was implemented to speed up polynomial multiplication. Running time was measured on an Ubuntu 18.04 machine with Intel Core i7-8700 3.2 GHz CPU and 16 GB DDR RAM. The parameter t of HE schemes was set the range of the SED between features. That is, for the dimension of features n and the quantization precision m, $t = n(2^m - 1)^2 + 1$. The parameter N of HE schemes was set to achieve a security level of 128 bits under the determined t.

Biometric Performance Analysis Figure 2 shows DET curves of FaceNet, ArcFace, and VGGFace2 for different quantization precisions. The DET curves implies that the biometric performances with quantization precision $m = 5, 6$, and 5 are close enough to those of floating-point features for FaceNet, ArcFace, and VGGFace2, respectively.

Efficiency Analysis Figure 3 shows comparison time, which is the total time of homomorphic multiplication, decryption, and decision, for quantization precisions up to 10 bits. The rapid increases of comparison time occur since the parameter N of HE schemes, which

Tab. 1: Running time and data size.

Feature extraction method	HE scheme	Quantization precision [bit]	Enrollment/ query time [ms]	Comparison time [ms]	Template/ query size [KB]
FaceNet (128 dim.)	ideal lattice	5	4.00	7.73	10.3
ArcFace	BGV	6	33.2	16.3	70
VGGFace2	BGV	5	33.4	16.5	71
FaceNet (512 dim.), ArcFace [Ko20] [1]	BFV [FV12]	2	76	618	132

[1] Measured on a virtualised (single-core) Linux machine with Intel Core i7 2.7 GHz CPU and 16 GB DDR RAM.

determines degree of polynomials that HE schemes encrypt, is doubled in order to maintain the security level. The difference in the timing of the rapid increases among feature extraction methods comes from the size of the parameter t of HE schemes differs due to their different dimensions of features.

Summary Table 1 shows running time and data size of the optimal combinations of feature extraction method and HE scheme. For all three feature extraction methods, the transaction time is less than 50 ms, where the transaction time is the total time of query time and the comparison time.

Focusing on ArcFace, the transaction time of the proposed system, which is 49.5 ms, is 14.0 times faster than that of the HE-based template protection system presented in [Ko20], which is 694 ms. In addition, the template size and the query size of the proposed system, which is 70 KB, is 1.89 times smaller than that of the existing system, which is 132 KB. The integer features encrypted in the proposed system have the same size of dimension and the larger quantization precision as those handled in the existing system. Thus, the speedup is due to the use of efficient HE schemes and the SED computation method.

5 Conclusion

This paper proposed an efficient face template protection system based on HE for integer representation features. By utilizing the message packing method to compute the inner product by a single homomorphic multiplication, the proposed system computes the SED of facial features by a single homomorphic multiplication.

At the quantization precision carefully chosen to have negligible impact on biometric performance, the transaction time of the proposed system is less than 50 ms. It is 14.0 times faster than that of the HE-based face template protection system presented in BIOSIG2020. In addition, our experimental results suggest that the HE scheme used in the proposed system should be selected for each feature extraction method and quantization precision in order to optimize the running time.

References

[BDL15] Barni, M; Droandi, G; Lazzeretti, R: Privacy Protection in Biometric-Based Recognition Systems: A marriage between cryptography and signal processing. IEEE Signal Processing Magazine, 32(5):66–76, sep 2015.

[BGV12] Brakerski, Z; Gentry, C; Vaikuntanathan, V: (Leveled) Fully Homomorphic Encryption without Bootstrapping. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. ITCS'12, Association for Computing Machinery, New York, NY, USA, pp. 309–325, 2012.

[Bo18] Boddeti, V. N: Secure Face Matching Using Fully Homomorphic Encryption. In: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS). IEEE, oct 2018.

[BV11] Brakerski, Z; Vaikuntanathan, V: Fully Homomorphic Encryption from Ring-LWE and Security for Key Dependent Messages. In: Proceedings of the 31st Annual Conference on Advances in Cryptology. CRYPTO'11, Springer-Verlag, Berlin, Heidelberg, pp. 505–524, 2011.

[Ca18] Cao, Q; Shen, L; Xie, W; Parkhi, O. M; Zisserman, A: VGGFace2: A Dataset for Recognising Faces across Pose and Age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, may 2018.

[Ch17] Cheon, J. H; Kim, A; Kim, M; Song, Y: Homomorphic Encryption for Arithmetic of Approximate Numbers. In (Takagi, T; Peyrin, T, eds): Advances in Cryptology – ASIACRYPT 2017. Springer International Publishing, Cham, pp. 409–437, 2017.

[CT65] Cooley, J. W; Tukey, J. W: An algorithm for the machine calculation of complex Fourier series. Mathematics of Computation, 19(90):297–297, may 1965.

[De19] Deng, J; Guo, J; Xue, N; Zafeiriou, S: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, jun 2019.

[Dr19] Drozdowski, P; Buchmann, N; Rathgeb, C; Margraf, M; Busch, C: On the Application of Homomorphic Encryption to Face Identification. In: 2019 International Conference of the Biometrics Special Interest Group (BIOSIG). pp. 1–5, 2019.

[FV12] Fan, J; Vercauteren, F: , Somewhat Practical Fully Homomorphic Encryption. Cryptology ePrint Archive, Report 2012/144, 2012. `https://eprint.iacr.org/2012/144`.

[IS11] ISO/IEC JTC1/SC27: , ISO/IEC 24745:2011 Information technology – Security techniques – Biometric information protection, 2011.

[Ko19] Kolberg, J; Bauspies, P; Gomez-Barrero, M; Rathgeb, C; Durmuth, M; Busch, C: Template Protection based on Homomorphic Encryption: Computationally Efficient Application to Iris-Biometric Verification and Identification. In: 2019 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, dec 2019.

[Ko20] Kolberg, J; Drozdowski, P; Gomez-Barrero, M; Rathgeb, C; Busch, C: Efficiency Analysis of Post-quantum-secure Face Template Protection Schemes based on Homomorphic Encryption. In (Brmme, A; Busch, C; Dantcheva, A; Raja, K; Rathgeb, C; Uhl, A, eds): BIOSIG 2020 - Proceedings of the 19th International Conference of the Biometrics Special Interest Group. Gesellschaft fr Informatik e.V., Bonn, pp. 175–182, 2020.

[MK19] Manisha; Kumar, N: Cancelable Biometrics: a comprehensive survey. Artificial Intelligence Review, 53(5):3403–3446, oct 2019.

[Ph00] Phillips, P; Moon, H; Rizvi, S; Rauss, P: The FERET evaluation methodology for face-recognition algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(10):1090–1104, 2000.

[RU11] Rathgeb, C; Uhl, A: A survey on biometric cryptosystems and cancelable biometrics. EURASIP Journal on Information Security, 2011(1), sep 2011.

[SKP15] Schroff, F; Kalenichenko, D; Philbin, J: FaceNet: A Unified Embedding for Face Recognition and Clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). jun 2015.

[SO20] Serengil, S. I; Ozpinar, A: LightFace: A Hybrid Deep Face Recognition Framework. In: 2020 Innovations in Intelligent Systems and Applications Conference (ASYU). IEEE, pp. 23–27, 2020. `https://github.com/serengil/deepface`.

[SS11] Stehlé, D; Steinfeld, R: Making NTRU as Secure as Worst-Case Problems over Ideal Lattices. In: Advances in Cryptology – EUROCRYPT 2011, pp. 27–47. Springer Berlin Heidelberg, 2011.

[Th] The OpenSSL Project: OpenSSL: The Open Source toolkit for SSL/TLS. `https://www.openssl.org`.

[Ya13] Yasuda, M; Shimoyama, T; Kogure, J; Yokoyama, K; Koshiba, T: Packed Homomorphic Encryption Based on Ideal Lattices and Its Application to Biometrics. In: Security Engineering and Intelligence Informatics, pp. 55–74. Springer Berlin Heidelberg, 2013.

[Ya14] Yasuda, M; Shimoyama, T; Kogure, J; Yokoyama, K; Koshiba, T: Practical Packing Method in Somewhat Homomorphic Encryption. In (Garcia-Alfaro, J; Lioudakis, G; Cuppens-Boulahia, N; Foley, S; Fitzgerald, W. M, eds): Data Privacy Management and Autonomous Spontaneous Security. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 34–50, 2014.

Bloom Filter vs Homomorphic Encryption: Which approach protects the biometric data and satisfies ISO/IEC 24745?

Amina Bassit[1], Florian Hahn[2], Chris Zeinstra[3], Raymond Veldhuis[4], Andreas Peter[5]

Abstract: Bloom filter (BF) and homomorphic encryption (HE) are popular modern techniques used to design biometric template protection (BTP) schemes that aim to protect the sensitive biometric information during storage and the comparison process. However, in practice, many BTP schemes based on BF or HE violate at least one of the privacy requirements of the international standard ISO/IEC 24745: irreversibility, unlinkability and confidentiality. In this paper, we investigate the state-of-the-art BTP schemes based on these two approaches and assess their relative strengths and weaknesses with respect to the three requirements of ISO/IEC 24745. The results of our investigation showed that the choice between BF and HE depends on the setting where the BTP scheme will be deployed and the level of trustworthiness of the parties involved in processing the protected template. As a result, HE enhanced by verifiable computation techniques can satisfy the privacy requirements of ISO/IEC 24745 in a trustless setting.

Keywords: Bloom filter, homomorphic encryption, biometric template protection, ISO/IEC 24745.

1 Introduction

A biometric template is a compact representation of a physiological or a behavioral biometric characteristic such as face, iris, voice, etc. The biometric characteristic itself is not a secret as, in a human-to-human interaction, humans recognize each other from their actual characteristics. However, in a human-to-machine interaction, a biometric template becomes a numerical equivalent of the human characteristic understandable by a machine. Thus, a biometric template reflects the identity of an individual that allows him to be recognized by the system. Given the fact that systems are subject to various types of security threats, a biometric template must be well-protected.

The literature [JNN08, SP17] defines *biometric template protection* (BTP) schemes as the branch of biometrics that tackles the problem of persevering biometric templates while maintaining the recognition performance. There exist different approaches to design BTP schemes that try to satisfy the privacy requirements of the international standard ISO/IEC 24745 [Se11]: irreversibility, unlinkability and confidentiality. Among those approaches, *Bloom filter* (BF) based BTPs, process the template in the transformed domain, while *homomorphic encryption* (HE) based BTPs, process the template in the encrypted domain. Both approaches have common and exclusive interesting properties that deal with the BTP challenges and the tradeoffs.

[1] University of Twente, DMB Group and SCS Group, Enschede, The Netherlands, a.bassit@utwente.nl
[2] University of Twente, SCS Group, Enschede, The Netherlands, f.w.hahn@utwente.nl
[3] University of Twente, DMB Group, Enschede, The Netherlands, c.g.zeinstra@utwente.nl
[4] University of Twente, DMB Group, Enschede, The Netherlands, r.n.j.veldhuis@utwente.nl
[5] University of Twente, SCS Group, Enschede, The Netherlands, a.peter@utwente.nl

There are several surveys that investigate either Bloom filter [BM04, GA13] or homomorphic encryption [MSM17, Ac18, WNK20] and their applications in general. However, none of them focuses on examining these two approaches from a biometrics point of view.

In this paper, we investigate the differences between BF-based BTP schemes and HE-based BTP schemes. We analyze the state-of-the-art in both approaches by studying their core functionalities and how they are exploited in the design of BTP schemes. As both approaches seem promising, we compare their advantages and disadvantages with respect to different levels: fulfillment of the privacy requirements of ISO/IEC 24745, application usability, protected template flexibility, template size and runtime efficiency. We conclude by reflecting on which of BF or HE has the potential to satisfy the three requirements of ISO/IEC 24745 in a trustless setting.

2 Background

In this section, we discuss Bloom filter and homomorphic encryption as technologies we are about to investigate in the context of biometric recognition. We also provide the privacy requirements recommended by ISO/IEC 24745 [Se11].

2.1 Bloom Filter

A standard Bloom filter (BF) is an efficient data structure that is used to verify whether an element belongs to a set or not. Let us denote $S=\{x_1,\cdots,x_n\}$ where $x_i \in \{0,1\}^*$ [3] a set of n elements to-be-represented. A BF consists of an m bits array initially set to zero. The filter uses k independent hash functions $h_1,\cdots,h_k$, where $h_i:\{0,1\}^* \rightarrow \{0,1,\cdots,m-1\}$, that are assumed to be uniformly random. To insert an element $x \in S$ in the BF, the bit at index $h_i(x)$ is set to one for all $1 \leq i \leq k$. To verify whether an element y belongs to S, for all $i \in [1,k]$ the bit at index $h_i(y)$ must be activated [4]. Hence, if at least one index is not activated then with certainty y does not belong to S otherwise y probably belongs to S since the indexes could have been activated by some elements of S distinct from y. [KM08] provides an extensive study on the selection of optimal parameters $(k,n$ and $m)$ of a BF and [Hu09] provides a tool to estimate them and observe parameters variation.

BF is used in biometrics not only for being a space-efficient data structure but also for its invariant property with respect to element insertion since the BF of a set of elements S is identical to the BF of any permutation of S. This property is important for disposing of the inconvenient features alignment, and thus to allow an alignment-free technique. The BFs used in biometrics differ from the standard ones in the number of hash functions, they use a single hash function that is binary-to-integer, and the verification of element membership, instead they calculate the weighed Hamming distance between the BFs of two sets. BFs are close if the distance is small and thus their corresponding sets are likely to overlap.

[3] The set $\{0,1\}^*$ refers to the binary set of arbitrary length

[4] BF is activated at index j means it is set to one at index j.

2.2 Homomorphic Encryption

Homomorphic encryption (HE) allows computation over encrypted data without decryption; $E(x)*E(y)=E(x\circ y)$ where $E(x)$ (resp. $E(y)$) is an encryption of x (resp. y), $*$ operation in the encrypted domain and $\circ$ operation in the plaintext domain. The operations $*$ and $\circ$ can be either an addition, a multiplication or both; depending on HE scheme type. There are three types of HE schemes: partially HE (PHE), somewhat HE (SWHE) and fully HE (FHE). PHE schemes (e.g. Paillier [Pa99], ElGamal [El85]) support only one operation unlimited number of times with a plaintext space either binary or integer. SWHE schemes (e.g. BGN [BGN05]) support a limited [5] number of operations, usually a limited number of multiplications and an arbitrary number of additions, and operate also on a binary or integer plaintext space. FHE schemes (e.g. BFV [Br12,FV12], BGV [BGV14], CKKS [Ch17]) support an unlimited number of both operations and are fundamentally based on Gentry's construction [Ge09] that enables refreshing ciphertexts to prevent them from reaching the allowed limit in each operation, and thus they remain decryptable. Unlike the classical PHEs and SWHEs, that have a limited choice of the plaintext, the state-of-the-art FHEs support binary (e.g. BFV), integers (e.g. BGV), real numbers and complex numbers (e.g. CKKS). Moreover, they offer a new style of operations, called *single-instruction multiple-data* (SIMD), that significantly contributes to speeding up FHEs. For instance, they allow encryption of a vector of plaintexts, packing of a vector of ciphertexts into a single ciphertext, permutations within the same ciphertext and automorphisms of a ciphertext. Although, the practical improvements on accelerating FHE schemes are considerable, it is still an active area of research.

HE offers flexibility in processing encrypted data, however it comes with a significant cost that impacts the storage as well as the runtime. The HE ciphertexts have a large size which implies that the biometric encrypted templates have a large size as well. The biometric recognition performed in the plaintext domain is significantly faster than the biometric recognition performed in the encrypted domain since they require several multiplications which are resource demanding operations under HE. The impact that HE has on the memory space and the runtime is undesirable in biometric recognition systems that try to minimize both of them to meet the usability requirement. However, this optimization should not be at the expense of their security.

2.3 Privacy Requirements of ISO/IEC 24745

The international standard ISO/IEC 24745 [Se11] establishes requirements and guidelines on how the biometric information should be protected throughout its entire lifecycle: storage, transfer and processing. The standard highlights the importance of binding a biometric reference with the corresponding subject identity as well as the privacy protection of subjects' biometric information during the processing. In this work, we focus on the ISO/IEC 24745 privacy requirements that are: *Irreversibility:* for a fixed pre-defined usage (such as recognition), the raw biometric data must be transformed into an irreversible representation that precisely fits the task of the pre-defined usage. *Unlinkability:* there must be no relationship between the stored biometric templates neither across applications nor databases. *Confidentiality:* the biometric template must be preserved and not exposed to unauthorized parties trying to gain unauthorized accesses.

[5] SWHE schemes produce noisy ciphertexts where the noise grows along with each homomorphic operation until it reaches its limit. Subsequently, the resulted ciphertext can no longer be decryptable.

3 Bloom Filter based BTP Schemes

Cancelable biometric systems [TKL08, SS18, Ku20], that apply non-invertible transformations to preserve the biometric template, suffer from significant degradation in their recognition performance due to the use of non-invertible transformations (such as cryptographic hash functions) that hurt the biometric accuracy. BF-based BTP schemes overcome this drawback by taking advantage from the invariant property of BFs to conceal a distorted version of the raw biometric sample in a BF-based template and thus achieve diffusion of the statistical properties of biometric features while maintaining their distinctiveness.

First Category BF-based BTPs: [RBB13] introduced the first BF-based BTP scheme (which we call *first category BF-based BTP scheme* and illustrate in Figure 1), as a form of cancelable biometric system that preserves the recognition performance by circumventing the feature alignment problem during the comparison process. This is achieved since BFs are invariant with respect to the insertion of elements as the BF of a set of elements S is identical to the BF of any permutation of S. This first category of BF-based BTPs was tested on irises [RBB13, Ra14, RB14, Ra15], faces [Go14] and fingerprints [Li15] to demonstrate the diversity of this approach with respect to the biometric modalities as long as they can be expressed as binary feature vectors.

The early security assessment of the first category of BF-based BTPs was studied by [HMP14] who confirmed the irreversibility of their templates but questioned their unlinkability. In particular, the authors showed that for $T_1 = \{\mathrm{BF}^{M}_{B_i}(K_1)\}_1^k$ and $T_2 = \{\mathrm{BF}^{M}_{B_i}(K_2)\}_1^k$ two BF-based templates generated from the same iriscode M using different keys $K_1 \neq K_2$ are determined to conceal the same iriscode with a probability of 96% assuming that the biometric samples are uniformly random. Later, [BMR17] extended the unlinkability analysis and considered the non-uniformity of biometric samples inherited from the acquisition noise to determine whether $\tilde{T}_1 = \{\mathrm{BF}^{M_1}_{B_i}(K_1^i)\}_1^k$ and $\tilde{T}_2 = \{\mathrm{BF}^{M_2}_{B_i}(K_2^i)\}_1^k$, with different iriscodes and different keys, are from the same iris. Their attack is a brute force over the possible keys K per block that saves the key with the lowest dissimilarity score. In other terms, for each block B_i it searches for

$$K = \operatorname*{argmin}_{\hat{K}\in[0,2^n-1]} DS\big(\mathrm{BF}^{M_1}_{B_i}(K_1^i),\mathrm{BF}^{M_2}_{B_i}(K_2^i \oplus \hat{K})\big)$$

where $\mathrm{BF}^{M_2}_{B_i}(K_2^i \oplus \hat{K})$ is computed only from $\mathrm{BF}^{M_2}_{B_i}(K_2^i)$ and $\hat{K}$ by activating the BF at index $j \oplus \hat{K}$ if and only if BF at index j is activated. Hence, the distribution of the dissimilarity scores of the original BF-based templates $DS\big(\mathrm{BF}^{M_1}_{B_i}(K_1^i),\mathrm{BF}^{M_2}_{B_i}(K_2^i)\big)$ and the distribution of the attacked templates $DS\big(\mathrm{BF}^{M_1}_{B_i}(K_1^i),\mathrm{BF}^{M_2}_{B_i}(K_2^i \oplus K)\big)$, where the key K has been chosen from the lowest dissimilarity score, overlap and have a slightly similar error rate. Then, [BMR17] analyzed the irreversibility of a 1st category BF-based template without key $K = 0$ and proposed two attacks that try to reconstruct an approximation of the unprotected template only by extracting some partial information from the protected template. The first attack consists of reconstructing a block by replacing all its columns with the same column computed from averaging the activated indexes of the BF of the protected template. The second attack requires a training set of the form (M_{ID}, T_{ID}) where T_{ID} is the protected template concealing the iriscode M_{ID}. The attack consists of reconstructing the iriscode of a protected template from the test set by replacing each block with the block corresponding to the nearest BF belonging to the protected templates of the training set. This attack assumes that $K=0$ which implies that it does not take into account neither the variability of the key among different subjects nor the effect of the key for the same subject. As reported by the authors, the experimental results of both attacks are ineffective.

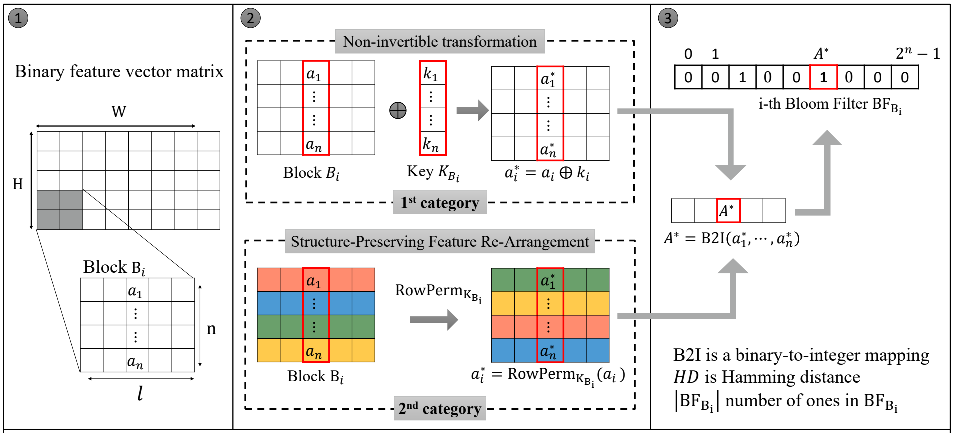
Second Category BF-based BTPs: In order to address the linkability vulnerability of 1st category BF-based BTPs, [Go16b] proposed a technique called *structure-preserving feature re-arrangement* to replace the XOR with the key before computing the BF, and thus the *second category BF-based BTP scheme* that we illustrate in Figure 1. This technique permutes the rows of a feature block according to a keyed random permutation to diffuse the statistical properties of a biometric feature vector and at the same time to preserve the biometric performance. Later, [Ma17] uses the same technique with a minor addition, that is, after a row-wise permutation there is a circular shift within each column. However, this circular shifting does not contribute to the dissipation of the biometric information but rather might lead to some accuracy loss since different columns after shifting might result in the same column.

[Go17a] studied the unlinkability of any BTP scheme from an information theory perspective and proposed a linkability evaluation procedure (Section 5 in [Go17a]). This procedure helps to assess whether two protected templates of a given BTP scheme are concealing the same or different biometric instances. This is determined only by observing the score resulted from the BTP's comparison measure and comparing it with the prior mated score distribution and the prior unmated score distribution. The same work defined three degrees of unlinkability that are: *fully unlinkable*, *semi unlinkable*, and *fully linkable* templates. [Go17a] tested their framework analysis on a HE-based BTP that uses Euclidean distance and reported that it is fully unlinkable while the BF-based BTP in [Go16b] lies between fully unlinkable and semi unlinkable. Note that this procedure works only if the comparison score is known, however for an HE-base BTPs this score can be hidden [PPV17] and only the comparison outcome is revealed. Hence, this procedure studied the unlinkability of the underlying unprotected template instead of the one protected by HE.

4 Homomorphic Encryption based BTP Schemes

Homomorphic Encryption (HE) has been the centerpiece of many privacy-preserving schemes, in particular biometric recognition in the encrypted domain [AM14, Ka15, IJL20] as it allows processing of encrypted templates without decryption. The use of an IND-CPA[6] secure HE scheme guarantees unlinkability, irreversibility and confidentiality under the constraint of the hardness of the underlying mathematical problem. Unlike classical BTP schemes, HE-based BTPs provide template protection even for a remote biometric recognition since an encrypted template can be sent over an unprotected public channel as only the party holding the private key is able to decrypt, and thus the importance of key management in the design of HE-based BTPs. Hence, HE allows a distributed comparison between the client and the server where only the party with the disclosure right is entitled to learn the recognition outcome. Therefore, in this survey,

[6] Indistinguishability under Chosen Plaintext Attack ensures that the encryption of the same plaintext twice yields two different ciphertexts. This property contributes to the dynamism of the protected template.


1
Binary feature vector matrix
W
H
Block B_i
a_1
⋮
a_n
n
l
2
Non-invertible transformation
a_1
a_n
k_1
k_n
a_1^*
a_n^*
Block B_i
Key K_{B_i}
$a_i^* = a_i \oplus k_i$
1st category
Structure-Preserving Feature Re-Arrangement
a_1
a_n
$\text{RowPerm}_{K_{B_i}}$
a_1^*
a_n^*
Block B_i
$a_i^* = \text{RowPerm}_{K_{B_i}}(a_i)$
2nd category
3
0 1 A^* $2^n - 1$
0 0 1 0 0 1 0 0 0
i-th Bloom Filter BF_{B_i}
A^*
$A^* = \text{B2I}(a_1^*, \cdots, a_n^*)$
B2I is a binary-to-integer mapping
HD is Hamming distance
$|BF_{B_i}|$ number of ones in BF_{B_i}

① Extract binary feature vectors and partition them into k feature blocks of dimension n × l

② Each block B_i is transformed using one of the transformations below:

- **1st category** uses the non-invertible transformation based on a XOR with a key
- **2nd category** uses the Structure-Preserving Feature Re-Arrangement based on a row-wise random permutation

③ Compute a Bloom filter BF_{B_i} for each transformed block B_i.
The BF-based template T is of the set of those BFs $T = \{BF_{B_i}\}_1^k$
The dissimilarity between two templates T_1 and T_2 is given by

$$DS(T_1, T_2) = \sum_{i=1}^{k} DS\left(BF_{B_i}^1, BF_{B_i}^2\right) \text{ where } DS\left(BF_{B_i}^1, BF_{B_i}^2\right) = \frac{HD\left(BF_{B_i}^1, BF_{B_i}^2\right)}{\left|BF_{B_i}^1\right| + \left|BF_{B_i}^2\right|}$$

Fig. 1: Overview of the 1st category and the 2nd category BF-based BTP schemes. Step 1 and Step 3 are common to both categories. In Step 2, the 1st category (resp. 2nd category) each block is transformed via a XOR with user's key (resp. row-wise random permutation). Note that the key is user-specific and should be different from an application to another to avoid cross-matching over databases. The original scheme [RBB13] uses the same key for all blocks while [BMR17], who assessed its security, proposed to use a different key per block, as depicted in this figure.

we classify HE-based BTPs according to their key management approach: either a single key HE[7], where the template is encrypted with the public key of one of the parties and is decryptable with its private key, or threshold HE where the template is encrypted using a joint public key between the client and the server and is decryptable using their both partial private keys.

Single key HE-based BTPs: The choice of a suitable HE scheme for designing a HE-based BTP scheme depends on the comparison measure that produces either a similarity score or a dissimilarity score. Some comparison measures (such as Hamming distance) can be efficiently implemented under encryption using only a PHE scheme while others that consume more multiplications (such as Cosine similarity) can benefit from SIMD operations of a SWHE scheme or a FHE scheme to improve their efficiency under encryption. The design of a HE-based BTP

[7] Here, single key means that there is one single private key (decryption key) that is retained by one single party unlike in threshold HE where the private key is divided between more than one single party or in multi-key HE where each party holds its own private key.

scheme also depends on the recognition protocol architecture, the parties involved (such as client, authentication server and database server, where the two later are sometimes combined as a single server), which party has the right to learn the recognition outcome based on which the key management is handled.

For applications such as access control, the client is entitled to learn the recognition outcome. For instance, schemes such as [Ba10, SSNS15, Ch16] encrypt the template with the client's public key and stores the encrypted template on the server's database who computes the comparison measure under encryption and sends the final score encrypted to the client. While in other applications such as remote authentication to a service, the authentication server is entitled to learn the recognition outcome. For example, schemes such as [Še14, Go16a, GBFG16, Go17b, Ko20] differentiate between an authentication server and a database server with the assumption that both do not collude. In these schemes, the template is encrypted with the authentication server's public key and stored on the database server. This time the database server performs the comparison under encryption and sends the encrypted final score to the authentication server. In both cases, the party, entitled to learn the recognition outcome, decrypts the encrypted final score and then compares it with the system's threshold, if the score exceeds the threshold then the party counts it as a match otherwise a no match. Hence, the comparison is not fully in the encrypted domain as the comparison with the threshold is performed after the decryption and the entitled party learns more than what it needs to learn, the final score and the recognition outcome.

In some schemes, such as [Up10, IJL20], the template is encrypted with the client's public key although the authentication server is the entitled party. For the comparison measure, [Up10] uses the support vector machine (SVM) classifier while [IJL20] uses the squared Euclidean distance (SED). During the enrollment of a given individual, in [Up10] the classifier is trained on several biometric samples of that individual and the encrypted template is formed by encrypting the classifier's parameters using the client's public key while in [IJL20] the encrypted template is simply the encrypted feature vector.

During the comparison, in [Up10] the client sends an encrypted freshly extracted feature vector to the authentication server who multiples them feature-wise with the encrypted template and a random value in order to blind the individual products. Subsequently, the server sends these blinded products to the client who decrypts and adds them and then sends back the result to the server so that it cancels out the blinding to learn the final score based on which it makes its decision. Similarly, in [IJL20] the server computes a blinded SED under encryption, sends the encrypted blinded final score to the client who decrypts it and sends it back. Then, the server removes the blinding from the blinded final score and performs the comparison with the threshold. Again in these cases the final score is revealed to the server and thus the comparison with the threshold is performed outside the encrypted domain.

Threshold HE-based BTPs: The encryption of the template with the authentication server's public key, even if the encrypted template is stored on the database server, is unsafe since in case the authentication server intercepts the communication between the client and the database server or illegally obtains the encrypted template, the authentication server is able to decrypt the encrypted template and learns the clear template that is supposed to be protected. HE-based BTP schemes such as [Ka15, PPV17] use a threshold variant of HE to encrypt the template in order to address the above mentioned limitation introduced by the use of a single key HE scheme. Hence, a threshold HE encrypted template cannot be decrypted by neither the client nor the server on his own but instead both of them need to participate in the decryption process, and thus a better control of the biometric data flow from both parties.

In general, the exposure of the final score, whether to the client or to the server, leaks the closeness between a freshly processed biometric data (probe) and the static previously processed biometric data (template) as well as the quality of a user's biometric modality. Taking advantage from HE that allows processing under encryption, [PPV17] shows that the final score can be hidden. Moreover, [PPV17] performs the comparison with the threshold under encryption and then reveals only the recognition outcome, *match* or *no match*, at moment of decryption.

Tab. 1: Comparison Table Showing the Advantages and Disadvantages of Each Approach

BTP approaches	**BF-based BTP**		**HE-based BTP**	
Categories	**1st Category**	**2nd Category**	**Single Key HE**	**Threshold HE**
Schemes	[RBB13, Ra14, RB14] [Go14, Ra15, Li15]	[Go16b, Ma17]	[Up10, Ch16, SSNS15] [Ba10, Še14, IJL20, Ko20]	[Ka15, PPV17]
Irreversibility	✓	✓	✓	✓
Unlinkability	✗[1]	✓[2]	✓	✓
Confidentiality	✓	✓	✓	✓
Supported modalities	All	All	All	All
Supported features	Binary and integer	Binary and integer	Binary, integer and float	Binary, integer and float
Feature alignment	Not needed [3]	Needed [3,4]	Needed	Needed
Comparison	Centralized	Centralized	Centralized and distributed	Centralized and distributed
Malleability	Malleable	Malleable	Malleable	Malleable
Final score exposure	Exposed	Exposed	Can be hidden	Can be hidden
Template dynamism	Static [5]	Static [5]	Refreshable and Randomizable	Refreshable and Randomizable
Template size	Linear in #feature blocks and BF size	Linear in #feature blocks and BF size	Linear in #features and ciphertext size	Linear in #features and ciphertext size
Runtime Efficiency	Fast	Fast	Practical to slow [6]	Practical to slow [6]
Recognition Accuracy	No accuracy loss	No accuracy loss	No accuracy loss	No accuracy loss

[1] Shown by [HMP14] and [BMR17]. [2] [Go17a] reports that it is slightly linkable. [3] However, it compares BFs generated from the same block of features. [4] For faces, it assumes pre-aligned images. [5] Once it is generated, it cannot be refreshed. [6] Depends on HE scheme security level.

5 BF-based BTP Schemes vs HE-based BTP Schemes

Both approaches present pros and cons and differently satisfy the tradeoff efficiency-security which makes a binary decision between these approaches difficult to make. Table 1 summarizes and compares BF-based BTP schemes and HE-based BTP schemes with respect to the privacy requirements of ISO/IEC 24745 (rows 4, 5 and 6), supported modalities and their nature (rows 7, 8 and 9), biometric recognition protocol (rows 10, 11 and 12), template's characteristics (rows 13 and 14) and performance of the overall BTP (rows 15 and 16). Note that *malleability* means whether the protected template can be inconspicuously altered. A BF-based template can be modified by flipping activated/deactivated bits while HE-based template can be modified by injecting ciphertexts to the encrypted template since HE is malleable by nature. Therefore, a verification mechanism needs to be applied along with BTP schemes to check the validity of the protected template and monitor the correctness of comparison operations.

6 Conclusion

In this paper, we investigated existing BF-based BTPs and HE-based BTPs with regard to the fulfillment of the privacy requirements of ISO/IEC 24745. While both approaches preserve the biometric accuracy, however they present advantages and disadvantages that vary according to the tradeoff efficiency-security. The choice of using one approach over the other depends on the setting where the BTP scheme is intended to be deployed and the level of trustworthiness of the parties involved in processing the protected template. In both approaches, the protected template needs to be treated with cautiousness since according to [Si12] and [AM14] if the parties do not follow the recognition protocol as prescribed, then serious biometric leakage can happen. Unlike BF-based BTPs, HE-based BTPs are more able to deal with this kind of misbehavior since they can be combined with secure and verifiable computation techniques to monitor the flow of the computation and thus satisfy the privacy requirements of ISO/IEC 24745 in a trustless setting.

Acknowledgment

This work was supported by the PriMa project that has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 860315.

We would like to thank Christoph Busch and Marcel Grimmer for their inspiring lecture that motivated the research done in this paper.

References

[Ac18] Acar, Abbas; Aksu, Hidayet; Uluagac, A Selcuk; Conti, Mauro: A survey on homomorphic encryption schemes: Theory and implementation. ACM Computing Surveys (CSUR), 51, 2018.

[AM14] Abidin, Aysajan; Mitrokotsa, Aikaterini: Security aspects of privacy-preserving biometric authentication based on ideal lattices and ring-lwe. In: 2014 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2014.

[Ba10] Barni, Mauro; Bianchi, Tiziano; Catalano, Dario; Di Raimondo, Mario; Labati, Ruggero Donida; Failla, Pierluigi; Fiore, Dario; Lazzeretti, Riccardo; Piuri, Vincenzo; Piva, Alessandro et al.: A privacy-compliant fingerprint recognition system based on homomorphic encryption and fingercode templates. In: 2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS). IEEE, 2010.

[BGN05] Boneh, Dan; Goh, Eu-Jin; Nissim, Kobbi: Evaluating 2-DNF formulas on ciphertexts. In: Theory of cryptography conference. Springer, 2005.

[BGV14] Brakerski, Zvika; Gentry, Craig; Vaikuntanathan, Vinod: (Leveled) fully homomorphic encryption without bootstrapping. ACM Transactions on Computation Theory (TOCT), 6, 2014.

[BM04] Broder, Andrei; Mitzenmacher, Michael: Network applications of Bloom filters: A survey. Internet mathematics, 1, 2004.

[BMR17] Bringer, Julien; Morel, Constance; Rathgeb, Christian: Security analysis and improvement of some biometric protected templates based on Bloom filters. Image and Vision Computing, 58, 2017.

[Br12] Brakerski, Zvika: Fully homomorphic encryption without modulus switching from classical GapSVP. In: Annual Cryptology Conference. Springer, 2012.

[Ch16] Cheon, Jung Hee; Chung, HeeWon; Kim, Myungsun; Lee, Kang-Won: Ghostshell: Secure Biometric Authentication using Integrity-based Homomorphic Evaluations. IACR Cryptology ePrint Archive, 2016.

[Ch17] Cheon, Jung Hee; Kim, Andrey; Kim, Miran; Song, Yongsoo: Homomorphic encryption for arithmetic of approximate numbers. In: International Conference on the Theory and Application of Cryptology and Information Security. Springer, 2017.

[El85] ElGamal, Taher: A public key cryptosystem and a signature scheme based on discrete logarithms. IEEE transactions on information theory, 31, 1985.

[FV12] Fan, Junfeng; Vercauteren, Frederik: Somewhat practical fully homomorphic encryption. IACR Cryptol. ePrint Arch., 2012.

[GA13] Geravand, Shahabeddin; Ahmadi, Mahmood: Bloom filter applications in network security: A state-of-the-art survey. Computer Networks, 57, 2013.

[GBFG16] Gomez-Barrero, Marta; Fierrez, Julian; Galbally, Javier: Variable-length template protection based on homomorphic encryption with application to signature biometrics. In: 2016 4th International Conference on Biometrics and Forensics (IWBF). IEEE, 2016.

[Ge09] Gentry, Craig et al.: A fully homomorphic encryption scheme, volume 20. Stanford university Stanford, 2009.

[Go14] Gomez-Barrero, Marta; Rathgeb, Christian; Galbally, Javier; Fierrez, Julian; Busch, Christoph: Protected facial biometric templates based on local gabor patterns and adaptive Bloom filters. In: 2014 22nd International Conference on Pattern Recognition. IEEE, 2014.

[Go16a] Gomez-Barrero, Marta; Fierrez, Julian; Galbally, Javier; Maiorana, Emanuele; Campisi, Patrizio: Implementation of fixed-length template protection based on homomorphic encryption with application to signature biometrics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2016.

[Go16b] Gomez-Barrero, Marta; Rathgeb, Christian; Galbally, Javier; Busch, Christoph; Fierrez, Julian: Unlinkable and irreversible biometric template protection based on Bloom filters. Information Sciences, 370, 2016.

[Go17a] Gomez-Barrero, Marta; Galbally, Javier; Rathgeb, Christian; Busch, Christoph: General framework to evaluate unlinkability in biometric template protection systems. IEEE Transactions on Information Forensics and Security, 13, 2017.

[Go17b] Gomez-Barrero, Marta; Maiorana, Emanuele; Galbally, Javier; Campisi, Patrizio; Fierrez, Julian: Multi-biometric template protection based on homomorphic encryption. Pattern Recognition, 67, 2017.

[HMP14] Hermans, Jens; Mennink, Bart; Peeters, Roel: When a Bloom filter is a doom filter: security assessment of a novel iris biometric template protection system. In: 2014 international conference of the biometrics special interest group (BIOSIG). IEEE, 2014.

[Hu09] Hurst, Thomas: , Bloom Filter Calculator, 2009.

[IJL20] Im, Jong-Hyuk; Jeon, Seong-Yun; Lee, Mun-Kyu: Practical Privacy-Preserving Face Authentication for Smartphones Secure Against Malicious Clients. IEEE Transactions on Information Forensics and Security, 15, 2020.

[JNN08] Jain, Anil K; Nandakumar, Karthik; Nagar, Abhishek: Biometric template security. EURASIP Journal on advances in signal processing, 2008, 2008.

[Ka15] Karabat, Cagatay; Kiraz, Mehmet Sabir; Erdogan, Hakan; Savas, Erkay: THRIVE: threshold homomorphic encryption based secure and privacy preserving biometric verification system. EURASIP Journal on Advances in Signal Processing, 2015.

[KM08] Kirsch, Adam; Mitzenmacher, Michael: Less hashing, same performance: Building a better Bloom filter. Random Structures & Algorithms, 33, 2008.

[Ko20] Kolberg, Jascha; Drozdowski, Pawel; Gomez-Barrero, Marta; Rathgeb, Christian; Busch, Christoph: Efficiency Analysis of Post-quantum-secure Face Template Protection Schemes based on Homomorphic Encryption. In: 2020 International Conference of the Biometrics Special Interest Group (BIOSIG). IEEE, 2020.

[Ku20] Kumar, Nitin et al.: Cancelable biometrics: A comprehensive survey. Artificial Intelligence Review, 53, 2020.

[Li15] Li, Guoqiang; Yang, Bian; Rathgeb, Christian; Busch, Christoph: Towards generating protected fingerprint templates based on Bloom filters. In: 3rd International workshop on biometrics and forensics (IWBF 2015). IEEE, 2015.

[Ma17] Martiri, Edlira; Gomez-Barrero, Marta; Yang, Bian; Busch, Christoph: Biometric template protection based on Bloom filters and honey templates. IET Biometrics, 6, 2017.

[MSM17] Martins, Paulo; Sousa, Leonel; Mariano, Artur: A survey on fully homomorphic encryption: An engineering perspective. ACM Computing Surveys (CSUR), 50, 2017.

[Pa99] Paillier, Pascal: Public-key cryptosystems based on composite degree residuosity classes. In: International conference on the theory and applications of cryptographic techniques. Springer, 1999.

[PPV17] Peeters, Joep; Peter, Andreas; Veldhuis, Raymond NJ: Fast and Accurate Likelihood Ratio Based Biometric Comparison in the Encrypted Domain. arXiv preprint arXiv:1705.09936, 2017.

[Ra14] Rathgeb, Christian; Breitinger, Frank; Busch, Christoph; Baier, Harald: On application of Bloom filters to iris biometrics. IET Biometrics, 3, 2014.

[Ra15] Rathgeb, Christian; Breitinger, Frank; Baier, Harald; Busch, Christer: Towards Bloom filter-based indexing of iris biometric data. In: 2015 international conference on biometrics (ICB). IEEE, 2015.

[RB14] Rathgeb, Christian; Busch, Christoph: Cancelable multi-biometrics: Mixing iris-codes based on adaptive Bloom filters. Computers & Security, 42, 2014.

[RBB13] Rathgeb, Christian; Breitinger, Frank; Busch, Christoph: Alignment-free cancelable iris biometric templates based on adaptive Bloom filters. In: 2013 international conference on biometrics (ICB). IEEE, 2013.

[Se11] Secretary, ISO Central: Information technology – Security techniques – Biometric information protection. Standard ISO/IEC 24745:2011, International Organization for Standardization, 2011.

[Še14] Šedĕnka, Jaroslav; Govindarajan, Sathya; Gasti, Paolo; Balagani, Kiran S: Secure outsourced biometric authentication with performance evaluation on smartphones. IEEE Transactions on Information Forensics and Security, 10, 2014.

[Si12] Simoens, Koen; Bringer, Julien; Chabanne, Hervé; Seys, Stefaan: A framework for analyzing template security and privacy in biometric authentication systems. IEEE Transactions on Information forensics and security, 7, 2012.

[SP17] Sandhya, Mulagala; Prasad, Munaga VNK: Biometric template protection: A systematic literature review of approaches and modalities. Biometric Security and Privacy, 2017.

[SS18] Sadhya, Debanjan; Singh, Sanjay Kumar: Design of a cancelable biometric template protection scheme for fingerprints based on cryptographic hash functions. Multimedia Tools and Applications, 77, 2018.

[SSNS15] Shahandashti, Siamak F; Safavi-Naini, Reihaneh; Safa, Nashad Ahmed: Reconciling user privacy and implicit authentication for mobile devices. Computers & Security, 53, 2015.

[TKL08] Teoh, Andrew BJ; Kuan, Yip Wai; Lee, Sangyoun: Cancellable biometrics and annotations on biohash. Pattern recognition, 41, 2008.

[Up10] Upmanyu, Maneesh; Namboodiri, Anoop M; Srinathan, Kannan; Jawahar, CV: Blind authentication: a secure crypto-biometric verification protocol. IEEE transactions on information forensics and security, 5, 2010.

[WNK20] Wood, Alexander; Najarian, Kayvan; Kahrobaei, Delaram: Homomorphic encryption for machine learning in medicine and bioinformatics. ACM Computing Surveys (CSUR), 53, 2020.

Toward Practical Adversarial Attacks on Face Verification Systems

Kazuya Kakizaki[1], Taiki Miyagawa[2], Inderjeet Singh[3], Jun Sakuma[4]

Abstract: DNN-based face verification systems are vulnerable to adversarial examples. The previous paper's evaluation protocol (scenario), which we called the probe-dependent attack scenario, was unrealistic. We define a more practical attack scenario, the probe-agnostic attack. We empirically show that these attacks are more challenging than probe-dependent ones. We propose a simple and effective method, PAMTAM, to improve the attack success rate for probe-agnostic attacks. We show that PAMTAM successfully improves the attack success rate in a wide variety of experimental settings.

Keywords: Adversarial example, Face verification, Security.

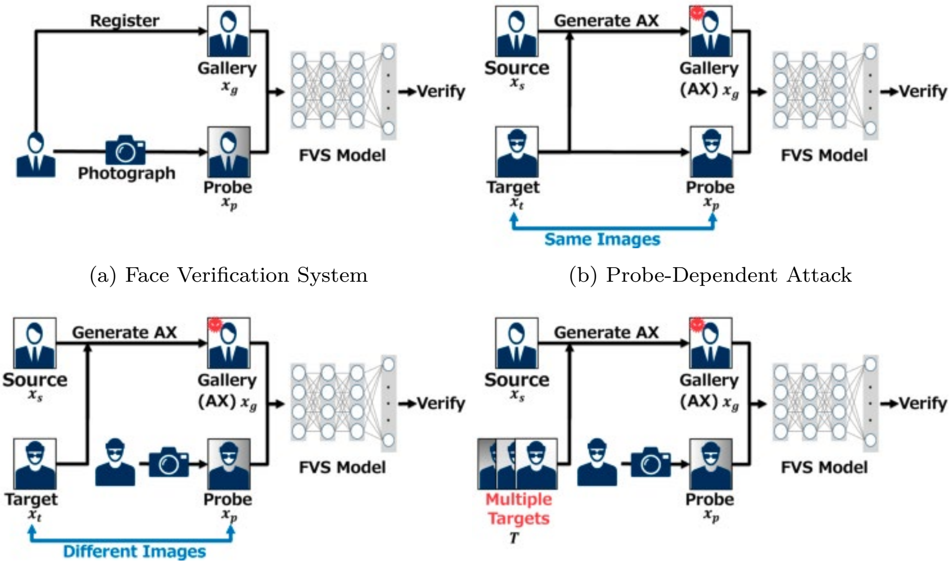
1 Introduction

Face verification systems (FVSs) verify the identity of a person by comparing two face images: *gallery* and *probe* images. The gallery image x_g is registered in the FVS in advance, and the probe image x_p is captured by a camera installed in the FVS at verification time, as shown in Fig. 1a. Recent progress on deep neural networks (DNNs) has significantly improved the performance of FVSs; however, DNNs have been shown to be vulnerable to small, human-imperceptible perturbations to the input data, or *adversarial examples* (AXs) [Sz14], which jeopardize the safety and security of DNN-based FVSs.

There are several studies on adversarial attacks against FVSs [RGB17, ZD20, Do19b]. These studies assume an adversary who generates an AX from images of the victim's and adversary's face (*source image* x_s and *target image* x_t, respectively); the generated AX looks like the victim but is expected to be misidentified as the adversary. Then, they assume an attack scenario in which the adversary can input a generated AX and target image x_t into the DNN in FVSs as a gallery image x_g and probe image x_p, respectively, as shown in Fig. 1b. However, this attack scenario is impractical in real-world settings because the probe images are captured by a camera at verification time [5]. We call this impractical attack scenario ($x_t = x_p$) the *probe-dependent attack*.

[1] NEC Corporation, University of Tsukuba, kazuya1210@nec.com
[2] NEC Corporation, miyagawataik@nec.com
[3] NEC Corporation, inderjeet78@nec.com
[4] University of Tsukuba, RIKEN AIP, jun@cs.tsukuba.ac.jp

[5] If the camera in the FVS is under the control of the adversary, then the adversary can input a generated AX directly into the FVS as the probe image. In our paper, however, we assume



(a) Face Verification System

(b) Probe-Dependent Attack

(c) Probe-Agnostic Attack

(d) Proposed Method (PAMTAM)

Fig. 1: (a) Face verification systems (FVSs) verify the identity of a person by comparing two face images: *gallery* image x_g, which is registered in the FVS in advance, and *probe* image x_p, which is captured by the camera installed in the FVS at the verification time. (b) Probe-Dependent Attack assumes that the adversary can input a generated AX and target image into the DNN in FVSs as a gallery image and probe image ($x_t = x_p$). (c) Probe-Agnostic Attack assumes that the adversary cannot input a target image as a probe image ($x_t \neq x_p$). (d) Our method, PAMTAM, generates AXs using multiple target face images $\boldsymbol{T}$. Note that the aforementioned settings are different from the presentation attack [Hu19], which is outside the scope of the present paper.

In this paper, we consider a more practical but challenging attack scenario, the *probe-agnostic* attack, as shown in Fig. 1c. We do not assume that $x_t = x_p$; thus, there generally exists a domain gap between x_t and x_p depending on when, where, and how the images are captured (*capturing conditions*), e.g., the illumination conditions, head poses, and image resolution. To the best of our knowledge, the probe-agnostic attack is yet to be explored in the literature and is important for assessing the true risk of practical adversarial attacks against FVSs.

The difficulty with the probe-agnostic attack comes from the domain gap between x_t and x_p due to the different capturing conditions. To address this problem, we propose a simple but effective method for increasing the attack success rate for probe-agnostic attacks: the Probe-Agnostic Multiple Target Method (PAMTAM). PAMTAM makes *arbitrary* attack methods robust to variable capturing conditions,

that the adversary cannot hack the FVS, which is the case, e.g., the automated face recognition gates at airports.

irrespective of white- or black-box attacks. PAMTAM generates AXs so that it approaches, on average, *multiple* target images in the feature space to make the features robust to domain gaps, as shown in Fig. 1d. We empirically show that PAMTAM successfully increases the attack success rates of 2 widely used attacks on 3 databases with 8 different model combinations, attaining a maximum relative recovery of 83.3%.

Our contribution is twofold. First, we formulate the probe-agnostic attack, which is yet to be explored in the literature and is important for assessing the true risk of practical adversarial attacks against FVSs. Next, we propose PAMTAM, which makes arbitrary attack methods robust to variable capturing conditions. We empirically show that PAMTAM successfully improves the attack success rate under a wide variety of conditions.

1.1 Related Work

All the following studies focus on the AXs against FVSs but follow the probe-dependent scenario. In the white-box setting, the attacker can access the network structure and parameters of the target DNN. [Sa16] is the first to show that DNN-based feature extractors, not only classifiers, are vulnerable to AXs. [RGB17] proposed LOTS that generates AX that is close to the target face image in the feature space. [ZD19] proposed Iterative Feature Target Gradient Sign Method (IFTGSM), which iteratively updates AX with a gradient sign of the gradient. [SWY18, DZJ19] leveraged Generative Adversarial Networks (GANs) to generate AXs with high perceptual quality. In contrast, the black-box attack assumes that the attacker cannot access the network structure or parameters of the target DNN. [Do19b] proposed a query-based attack method, where the attacker could send queries to FVSs and see the outputs. The query-based attack can relatively high attack success rates but can be easily detected because a number of queries are necessary to generate AXs, causing a suspiciously large amount of accesses to the target FVS. [ZD20] used surrogate models to generate AXs without queries. The authors proposed the dropout face attacking network (DFANet) to enhance transferability. They also showed that [Li17, Xi19, Do18], originally used for classifiers, are effective even for feature extractors.

2 Preliminaries

Face verification systems. Face verification is a task to determine whether two face images are derived from the same identity. Modern FVSs use DNN-based feature extractors [De19, Wa18]. Let $\mathscr{X}$ be a set of images with height $H \in \mathbb{N}$, width $W \in \mathbb{N}$, and the number of channels $C \in \mathbb{N}$, i.e., $\mathscr{X} = \{0,1,...,255\}^{H\times W\times C}$. Let $f : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^d$ be a feature extractor, where $d \in \mathbb{N}$ is the feature dimension. We define a function **Ver**, which represents the internal processes of FVSs, as a mapping from two images ($x_1, x_2 \in \mathscr{X}$) to a binary set ($\{\text{Verified}, \text{NotVerified}\}$):

$$\mathbf{Ver}_{f,\alpha}(x_1, x_2) = \begin{cases} \text{Verified} \quad (\text{if } \mathbf{dist}(f(x_1),\, f(x_2)) \leq \alpha) \\ \text{Not Verified} \quad (\text{otherwise}), \end{cases} \tag{1}$$

where $\alpha \in \mathbb{R}_{\geq 0}$ is a threshold, and $\mathbf{dist} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ is an arbitrary distance function of feature vectors. Typically, the cosine similarity or L^2 norm are used for face verification. We use the latter in the present paper, but the extensions to other distance functions are straightforward.

Our primary focus is the FVSs which involve the following two steps (See also Fig. 1a):

1. **Registration.** A user registers her or his face image (gallery image $x_g \in \mathscr{X}$) with the FVS. The FVS stores $f(x_g)$ in the gallery set.
2. **Verification.** At the verification phase, the FVS takes a photo of the user (probe image $x_p \in \mathscr{X}$) with the internal camera, which is sometimes invisible to the user. Then, the FVS computes $f(x_p)$ and runs Eq. (1) to investigate whether the two identities are the same.

Adversarial attacks against face verification systems. We assume that the gallery image x_g is an AX and the probe image x_p is not, as in Introduction, although there are two other possibilities in principle: (i) x_p is an AX, and (ii) both x_g and x_p are AXs. We can see that both (i) and (ii) are possible but infeasible, because the attacker is required to hack the FVS to input an AX directly to it. A potential solution is the physical adversarial attack [Sh16], but this is out of the scope of the present paper. Therefore, the attacker's goal is formally summarized into the problem of finding adversarial noise $\delta \in \mathbb{R}^{H \times W \times C}$ such that

$$\mathbf{dist}(f(x_g = x_s + \delta), f(x_p)) \leq \alpha \tag{2}$$

$$\|\delta\|_\infty \leq \varepsilon, \tag{3}$$

where $||\cdot||_\infty$ denotes the L^∞ (max) norm. Eq. (3) restricts the size of δ and ensures that the noise is imperceptible to humans. In general, δ is a function of x_s and x_t.

3 PAMTAM

Probe-dependent and probe-agnostic attacks. A common way to generate δ is to define an objective function and minimize it. In the probe-dependent attack, $x_t = x_p$ and $\delta = \delta(x_s, x_p)$ (Fig. 1b); therefore, Eq. (2) can be achieved by minimizing the objective function

$$J(x_s + \delta, x_p, f) = \|f(x_s + \delta) - f(x_p)\|_2^2 \tag{4}$$

with respect to δ. The adversarial noise thus obtained, denoted by $\delta^*(x_s,x_p)$, deceives the target FVS more easily than in the probe-agnostic scenario because $\delta^*(x_s,x_p)$ has the prior knowledge of x_p. In comparison, the probe-agnostic attack assumes $x_t \neq x_p$ and $\delta = \delta(x_s,x_t)$ (Fig. 1c); therefore, the objective function is

$$J(x_s+\delta,x_t,f) = \|f(x_s+\delta)-f(x_t)\|_2^2. \tag{5}$$

The solution $\delta^*(x_s,x_t)$ has no prior knowledge of x_p and is likely to overfit to x_t; therefore, the AX $x_g = x_s + \delta^*(x_s,x_t)$ has no guarantee of being misidentified as x_p if the domain gap between the two (x_t and x_p) is large. In fact, we empirically show in Section 4 the degradation from $\delta^*(x_s,x_p)$ to $\delta^*(x_s,x_t)$; probe-agnostic attacks are more challenging than probe-dependent attacks.

Proposed method. To achieve better attack success rates in the probe-agnostic scenario, we propose *diversifying the target image* x_t, introducing the *target image set* $T = \{x_t^i \in \mathscr{X} \,|\, i=1,...,|T|\}$, and modifying the objective function (5) as

$$J(x_s+\delta,T,f) = \frac{1}{|T|}\sum_{x_t\in T}\|f(x_s+\delta)-f(x_t)\|_2^2. \tag{6}$$

The target image set consists of facial images of the attacker, which should cover the domain gaps between x_t and x_p, such as different head poses, illumination conditions, image resolutions, facial expressions, and makeup. In fact, our experiment shows that a larger T enhances the attack success rate (Section 4). Note that it is easy to increase the sample size of T in practice, compared with source and probe images, because the attacker can take selfies under arbitrary conditions. We set $|T| = 5$ in our experiments unless otherwise noted.

A motivation of Eq. (6) comes from the recent studies showing that the diversity of the input images improves transferability [Do19a, Xi19]; however, no previous work has explored it to attack FVSs especially in the probe-agnostic scenario. Moreover, a crucial difference between our method and [Do19a, Xi19] is that the latter uses automatic, mechanical transformations for the input diversity (random resizing, random padding, and translation). However, such transformations are not sufficient to fill the large, complex domain gaps. In addition, their transormations are applied to the *source image*, while our method diversify the *target image*, to adapt the adversarial example to the probe-agnostic scenario.

The proposed method, *PAMTAM*, is widely applicable to arbitrary objective functions for (2) and arbitrary optimization methods, e.g., [ZD19, RGB17, Sa16, Do19b] . PAMTAM does not even depend on whether the attack is white-box or black-box.

4 Experiment

In this section, we demonstrate that probe-agnostic attacks are more challenging than probe-dependent attacks, as mentioned in Section 3. We then show that PAM-

TAM successfully improves the attack success rate. Our experiments are based on 2 attack methods on 3 datasets under 8 different conditions, as explained below. We focused on three domain gaps, head poses, illumination conditions, and image resolutions, which are likely to occur when we use an actual FVS.

To simulate the most realistic situations, all the experiments assumesd that the attacker cannot access the FVS model. First, we trained a DNN model (*FVS model*) on a training dataset (*FVS dataset*). Second, we trained another DNN model (*surrogate model*) on another dataset (*surrogate dataset*) to perform the surrogate model attack. Third, we sampled (i) source-probe doublets (x_s,x_p) and (ii) source-probe-target triplets (x_s,x_p,x_t) from yet another dataset (*material dataset*), which should have no intersection with the FVS or surrogate datasets. Note that this step distinguishes our experiments from those in preceding papers. (i) and (ii) were used for probe-dependent and probe-agnostic attacks, respectively. x_t in (ii) was replaced with T when PAMTAM was used. Fourth, using (i) and (ii), we generated AXs that deceive the surrogate model. Finally, using the AXs thus generated, we evaluated their attack success rates on the FVS model. The evaluation measure was the attack success rate, i.e., the proportion of the AXs matched to the probe images.

$$\sum_{x_s,x_p,\delta\in D} \frac{\mathbb{1}(\mathbf{Ver}_{f,\alpha}(x_s+\delta,x_p)=\text{Verified})}{|D|}, \tag{7}$$

where the test set D was defined as $\{(x_s^i,x_p^i,\delta(x_s^i,x_p^i))\}_{i=1}^{|D|}$, $\{(x_s^i,x_p^i,x_t^i,\delta(x_s^i,x_t^i))\}_{i=1}^{|D|}$, or $\{(x_s^i,x_p^i,T^i,\delta(x_s^i,T^i))\}_{i=1}^{|D|}$ for the probe-dependent attacks, probe-agnostic attacks, and PAMTAM, respectively ($|D|=200$ for our experiments). The verification threshold α of the FVS model was determined to achieve the best verification accuracy on the LFW dataset [Hu07]. All the FVS models in our experiments achieved a verification accuracy of 98% or higher.

Surrogate and FVS datasets and models. Though not essential, we slightly modified the objective functions (4), (5), and (6) to improve the base attack success rate of all the methods. Following [Li17, Xi19, ZD20], we introduced multiple surrogate models ($F=\{f_i\}_{i=1}^{|F|}$) and stochastic transformations τ of $x_s+\delta$ and took the average over F and τ. F is defined in Tab. 1 ($|F|=5$). τ includes random resizing and padding.

Our experiments used 8 combinations of the surrogate dataset, surrogate models, FVS dataset, and FVS model (Tab. 1). We used seven network architectures: residual network (R50, R100) [He16], inception residual network (IR50, IR100) [Sz17], squeeze-and-excitation inception residual network (SE50, SE100), and MobileFaceNet (MOB) [Ch18], each of which was attached with a state-of-the-art loss function (ArcFace (Arc) [De19] or CosFace (Cos) [Wa18]). We used two datasets: MS1MV2 (MS) [De19] and VGGFace2 (VGG) [Ca18].

Tab. 1: **Surrogate and FVS datasets and models.** We use four conditions (I, II, III, and IV).

	Surrogate			FVS		
	Data	Loss	Architecture	Data	Loss	Architecture
I	MS	Arc	R100,R50,IR100, IR50,SE50	VGG	Cos	MOB
II	MS	Arc	R100,R50,IR100,IR50,SE50	VGG	Cos	SE100
III	MS	Cos	R100,R50,IR100,IR50,SE50	VGG	Arc	MOB
IV	MS	Cos	R100,R50,IR100,IR50,SE50	VGG	Arc	SE100
V	VGG	Arc	R100,R50,IR100, IR50,SE50	MS	Cos	MOB
VI	VGG	Arc	R100,R50,IR100,IR50,SE50	MS	Cos	SE100
VII	VGG	Cos	R100,R50,IR100,IR50,SE50	MS	Arc	MOB
VIII	VGG	Cos	R100,R50,IR100,IR50,SE50	MS	Arc	SE100

Attack methods. We used two standard attack methods: the Sabour's attack (SAB) [Sa16] and the iterative feature target gradient sign method (IFTGSM) [ZD19]. SAB minimized Eq. (2) under the constraint Eq. (3) by using a box-constrained L-BFGS. IFTGSM iteratively updated δ as

$$\delta^{i+1} = C_\varepsilon(\delta_i - \mathrm{sign}(\nabla_{x_s+\delta^i} J(x_s, x_p, \delta, f))), \tag{8}$$

where $C_\varepsilon(\cdot)$ is a clipping function with a max radius ε, and $\mathrm{sign}(\cdot)$ is a sign function. In our experiments, the maximum perturbation ε was 10 in terms of the L^∞ norm; therefore, the perturbation to the pixel range was at most $10/255 \simeq 3.9\%$.

Material datasets. We use Head Pose Image Database [GHC04], Extended Yale Face Database B [LHK05], and VGGFace2 [Ca18]. They allow us to simulate various types of the domain gaps: head poses (Head Pose Image Database); illumination conditions (Extended Yale Face Database B); and combinations of head poses, illumination conditions, and image resolutions (VGGFace2). Head Pose Image Database consists of 2790 color facial images of 15 individuals with variations of vertical and horizontal face angles. These angles are expressed from -90 degrees to 90 degrees. Extended Yale Face Database B contains 16128 grayscale facial images of 28 individuals under 9 poses and 64 illumination conditions. We only use frontal facial images to fix the head pose. VGGFace2 consists of 3.31 million color facial images of 9131 persons, which covers a wide variety of head poses, illumination conditions, and image resolutions. Note that we also use VGGFace2 to train the FVS datasets, but there is no duplication with the material dataset.

4.1 Results

All the results are summarized in Fig. 2. PD and PA are short for probe-dependent and probe-agnostic. PD is the baseline and attained comparable performances with

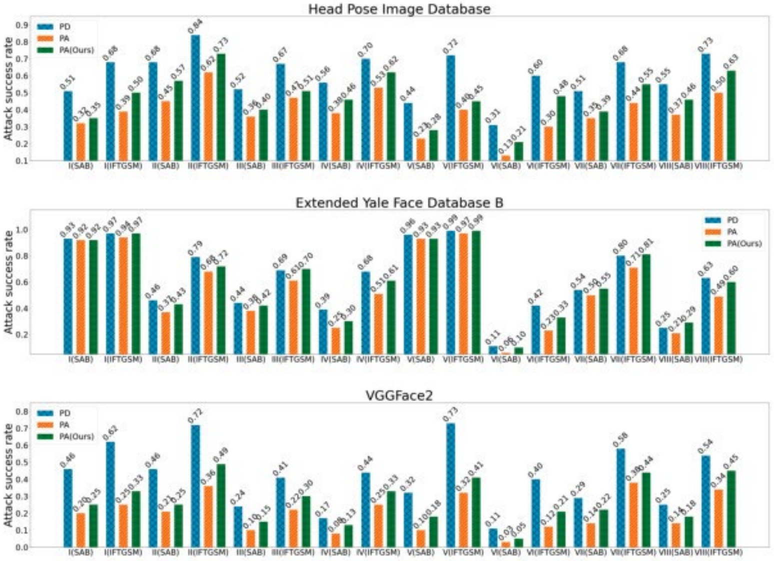


Fig. 2: **Degradation (PD→PA) and recovery (PA→PA(Ours)) of attack success rate.** Three subfigures correspond to three material datasets. PD and PA are short for probe-dependent and probe-agnostic. PA(Ours) is PAMTAM. I, II, III, IV, V, VI, VII, and VIII correspond to those in Tab. 1. Note that the numbers below the third decimal place are omitted.

a recent paper [ZD20][6]. PD achieved the best performance compared with the others PA and PA(Ours), and we found a significant degradation from PD to PA; *PA was more challenging than PD.* The attack success rates decreased by up to 56.4% for the Head Pose Image Database, 45.4% for Extended Yale Face Database B, and 73.9% for VGGFace2. The degradation of VGGFace2 was relatively large because the domain gap between x_t and x_p was larger than the other two. PAMTAM successfully increased the rates under almost all the conditions. The rates increased by up to 61.6% for the Head Pose Image Database, 75.0% for Extended Yale Face Database B, and 83.3% for VGGFace2. Fig. 3 shows PAMTAM's dependence on $|T|$ (evaluated on the VGGFace2 material dataset). We confirmed that large sample sizes help to enhance the attack success rate. The performance gain gradually saturated.

5 Conclusion

This paper considered adversarial attack against FVSs. We defined a more practical attack scenario (probe-agnostic attacks) than that in the previous paper (probe-dependent attacks). We empirically showed that probe-agnostic attacks are more

[6] Note that in face verification, attack success rates fluctuate significantly, depending on the DNN model and dataset (e.g., see [ZD20]).

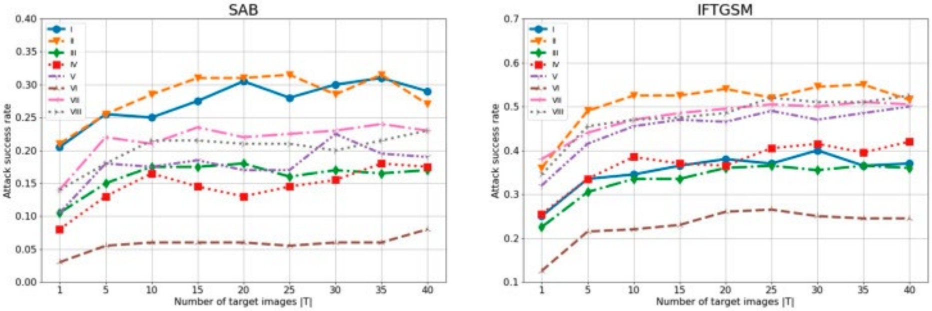


Fig. 3: **Dependence of PAMTAM on $|T|$.** Two subfigures correspond to the two attacks. I, II, III, IV, V, VI, VII, and VIII correspond to those in Tab. 1.

challenging than probe-dependent ones. The results above suggest that previous papers have overestimated the risk of AXs, especially when the domain gaps between x_p and x_t are large. We proposed PAMTAM, which successfully increase the attack success rate of probe-agnostic attacks. We conclude that we should evaluate not only probe-dependent attacks but also probe-agnostic ones under practical domain gaps to correctly capture the threat of AXs to FVSs.

Acknowledgment

We would like to thank to Toshinori Araki and anonymous reviewers for insightful discussions.

References

[Ca18] Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; Zisserman, A.: VGGFace2: A dataset for recognising faces across pose and age. In: International Conference on Automatic Face and Gesture Recognition. 2018.

[Ch18] Chen, Sheng; Liu, Yang; Gao, Xiang; Han, Zhen: Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In: Chinese Conference on Biometric Recognition. Springer, pp. 428–438, 2018.

[De19] Deng, Jiankang; Guo, Jia; Xue, Niannan; Zafeiriou, Stefanos: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4690–4699, 2019.

[Do18] Dong, Yinpeng; Liao, Fangzhou; Pang, Tianyu; Su, Hang; Zhu, Jun; Hu, Xiaolin; Li, Jianguo: Boosting adversarial attacks with momentum. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9185–9193, 2018.

[Do19a] Dong, Yinpeng; Pang, Tianyu; Su, Hang; Zhu, Jun: Evading defenses to transferable adversarial examples by translation-invariant attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4312–4321, 2019.

[Do19b] Dong, Yinpeng; Su, Hang; Wu, Baoyuan; Li, Zhifeng; Liu, Wei; Zhang, Tong; Zhu, Jun: Efficient decision-based black-box adversarial attacks on face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7714–7722, 2019.

[DZJ19] Deb, Debayan; Zhang, Jianbang; Jain, Anil K: Advfaces: Adversarial face synthesis. arXiv preprint arXiv:1908.05008, 2019.

[GHC04] Gourier, Nicolas; Hall, Daniela; Crowley, James L: Estimating face orientation from robust detection of salient facial features. In: ICPR International Workshop on Visual Observation of Deictic Gestures. Citeseer, 2004.

[He16] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778, 2016.

[Hu07] Huang, Gary B.; Ramesh, Manu; Berg, Tamara; Learned-Miller, Erik: Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[Hu19] Husseis, Anas; Liu-Jimenez, Judith; Goicoechea-Telleria, Ines; Sanchez-Reillo, Raul: A Survey in Presentation Attack and Presentation Attack Detection. In: 2019 International Carnahan Conference on Security Technology (ICCST). pp. 1–13, 2019.

[LHK05] Lee, Kuang-Chih; Ho, Jeffrey; Kriegman, David J: Acquiring linear subspaces for face recognition under variable lighting. IEEE Transactions on pattern analysis and machine intelligence, 27(5):684–698, 2005.

[Li17] Liu, Yanpei; Chen, Xinyun; Liu, Chang; Song, Dawn: Delving into transferable adversarial examples and black-box attacks. International Conference on Learning Representations, 2017.

[RGB17] Rozsa, Andras; Günther, Manuel; Boult, Terranee E: LOTS about attacking deep features. In: 2017 IEEE International Joint Conference on Biometrics (IJCB). IEEE, pp. 168–176, 2017.

[Sa16] Sabour, Sara; Cao, Yanshuai; Faghri, Fartash; Fleet, David J: Adversarial manipulation of deep representations. International Conference on Learning Representations, 2016.

[Sh16] Sharif, Mahmood; Bhagavatula, Sruti; Bauer, Lujo; Reiter, Michael K: Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In: Proceedings of the 2016 acm sigsac conference on computer and communications security. pp. 1528–1540, 2016.

[SWY18] Song, Qing; Wu, Yingqi; Yang, Lu: Attacks on state-of-the-art face recognition using attentional adversarial attack generative network. arXiv preprint arXiv:1811.12026, 2018.

[Sz14] Szegedy, Christian; Zaremba, Wojciech; Sutskever, Ilya; Bruna, Joan; Erhan, Dumitru; Goodfellow, Ian; Fergus, Rob: Intriguing properties of neural networks. International Conference on Learning Representations, 2014.

[Sz17] Szegedy, Christian; Ioffe, Sergey; Vanhoucke, Vincent; Alemi, Alexander: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. volume 31, 2017.

[Wa18] Wang, Hao; Wang, Yitong; Zhou, Zheng; Ji, Xing; Gong, Dihong; Zhou, Jingchao; Li, Zhifeng; Liu, Wei: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5265–5274, 2018.

[Xi19] Xie, Cihang; Zhang, Zhishuai; Zhou, Yuyin; Bai, Song; Wang, Jianyu; Ren, Zhou; Yuille, Alan L: Improving transferability of adversarial examples with input diversity. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2730–2739, 2019.

[ZD19] Zhong, Yaoyao; Deng, Weihong: Adversarial learning with margin-based triplet embedding regularization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6549–6558, 2019.

[ZD20] Zhong, Yaoyao; Deng, Weihong: Towards Transferable Adversarial Attack against Deep Face Recognition. arXiv preprint arXiv:2004.05790, 2020.

Transferability Analysis of an Adversarial Attack on Gender Classification to Face Recognition

Zohra Rezgui [1], Amina Bassit[2]

Abstract: Modern biometric systems establish their decision based on the outcome of machine learning (ML) classifiers trained to make accurate predictions. Such classifiers are vulnerable to diverse adversarial attacks, altering the classifiers' predictions by adding a crafted perturbation. According to ML literature, those attacks are transferable among models that perform the same task. However, models performing different tasks, but sharing the same input space and the same model architecture, were never included in transferability scenarios. In this paper, we analyze this phenomenon for the special case of VGG16-based biometric classifiers. Concretely, we study the effect of the white-box FGSM attack, on a gender classifier and compare several defense methods as countermeasures. Then, in a black-box manner, we attack a pre-trained face recognition classifier using adversarial images generated by the FGSM. Our experiments show that this attack is transferable from a gender classifier to a face recognition classifier where both were independently trained.

Keywords: Transferability, adversarial attacks, gender classification, face recognition.

1 Introduction

The cutting edge advances in deep learning (DL) have made computer vision problems more approachable. However, the black box nature of neural networks has made their security questionable. In fact, the majority of DL models are vulnerable to *adversarial attacks* that, based on subtle perturbations applied to the clean samples, mislead the classifier with a high confidence. There has been a number of studies investigating the vulnerabilities of DL-based machine learning systems to different types of adversarial attacks on the input images. The existing attacks can be partitioned into two categories: *white-box attacks*, where an adversary has full access to the attacked model's parameters, and *black-box attacks* where an adversary has no access to such information. Typically, white-box attacks are more powerful than black-box attacks due to their ability to leverage the parameters of the model against its own predictions. In a real-life scenario, a deployed model's parameters would not be accessible leaving the black-box attacks as the only option to disrupt its predictive performance. To benefit from the strength of white-box attacks, [De19, ZD20] show that it is possible to target a model, where its parameters are known, and transfer the resulting effects on an unknown model, as long as the two models are trained for the same task. Particularly in the field of biometrics, the effectiveness of these attacks should not be overlooked, given the variety of biometric applications such as forensics and border control where wrong predictions are not tolerated.

Many biometric applications are inter-connected, specifically those related to the face modality. For instance, there is a plethora of works showing that different face recognition systems can be

[1] University of Twente, DMB Group, Enschede, The Netherlands, z.rezgui@utwente.nl

[2] University of Twente, DMB Group and SCS Group, Enschede, The Netherlands, a.bassit@utwente.nl

enhanced with a soft biometric classifier such as a gender classifier [Go18a]. Similarly, deep face recognition features are known to be discriminative for soft biometric classification [OAE16] via transfer learning. This association incites us to further investigate the transferability potential of adversarial attacks on models sharing the same input space but trained independently to perform different tasks. However, such hypothesis was not included in previous studies on the transferability of adversarial attacks.

In this paper, we investigate the transferability of an adversarial attack against a gender classifier to a face recognition classifier where both classifiers are independently trained and only share the same input space (facial images) and the same model architecture. We start by providing an overview of the hypothesis of transferability between different tasks given the same input space. We then study the impact of an existing gradient-based attack and deep features-based defense on the gender classifier. Subsequently, we use the generated adversarial images along with those resulting from the defense against a pre-trained face recognition model to analyze the tranferability of both the attack and the defense. Our results, illustrated in Figure 1, support the transferability hypothesis of the chosen attack and defense.

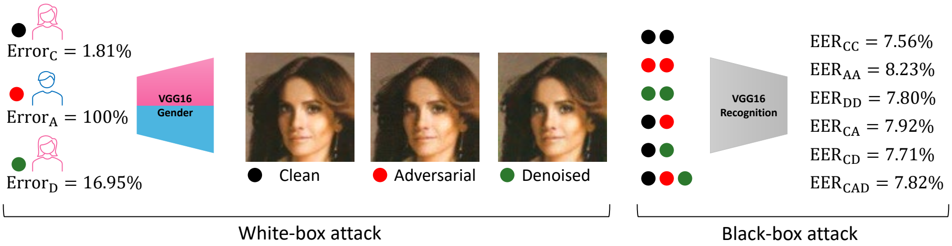

$\text{Error}_\text{C} = 1.81\%$
$\text{Error}_\text{A} = 100\%$
$\text{Error}_\text{D} = 16.95\%$
VGG16
Gender
● Clean
● Adversarial
● Denoised
VGG16
Recognition
$\text{EER}_\text{CC} = 7.56\%$
$\text{EER}_\text{AA} = 8.23\%$
$\text{EER}_\text{DD} = 7.80\%$
$\text{EER}_\text{CA} = 7.92\%$
$\text{EER}_\text{CD} = 7.71\%$
$\text{EER}_\text{CAD} = 7.82\%$
White-box attack
Black-box attack

Fig. 1: Overview of the transferability hypothesis from a white-box attack on a gender classifier to a black-box attack on a face recognition classifier. The black dots refer to the clean images, the red dots refer to the adversarial images generated by the FGSM attack and the green dots refer to the denoised images generated by the defense method. *Error* refers to the binary classification error rate of the gender classifier while *EER* refers to the equal error rate of the recognition classifier.

2 Related Work

Adversarial attacks have become an active area of research as they expose the design vulnerabilities of deep learning-based models. Several white-box attacks are gradient-based such as the Fast Gradient Sign Method (FGSM) [GSS14] its iterative version (IFGSM) [Ku16], and Projected Gradient Descent [KGB16]. Unlike the gradient used in backpropagation to train neural networks, the gradient used in those attacks helps determining the nearest perturbation to the input such that the adversarial image is misclassified. Other methods are based on network architecture information, [MDFF16] finds the minimal perturbation possible to an image that would make it misclassified, via projecting inputs on the closest classification hyperplane. Results in [Go18b] show that deep learning-based face recognition models, such as VGGFace, are vulnerable to such attacks and to image processing methods that perturb the samples in a perceptible manner. Moreover, [Mi18] uses GAN-based image editing to change the direction of the predictions of a binary gender classifier. However, the changes in the resulting images are perceptible to the human eye which contradicts the purpose of adversarial attacks.

To improve the robustness of existing deep learning models, many defense approaches have been proposed to withstand these attacks. [GSS14] and [Hu15] show that incorporating adversarial samples with the training data increases the attacked model's robustness but such an approach can be resource demanding. In practice, the model is trained over a diverse training set where, it learns to correctly classify the clean samples and, at the same time, it rectifies the predictions of the adversarial samples. [AG17] enhances the classifier's predictions by targeting each class and partitioning it into several sub-classes, assuming that only a few of them are sensitive to adversarial attacks. Subsequently, the different predictions of the sub-classes are aggregated via voting. Other approaches are based on input reconstruction such as [GR14] by using a denoising auto-encoder on the adversarial images in order to remove the perturbations. This method has been improved in [Li18] by using a U-Net architecture for the denoiser and defining the reconstruction loss based on the deep features of the classifier.

While white-box attacks are effective on known machine learning models, it was shown in [PMG16] that the resulting adversarial images can be effective against unknown models. The literature refers to such phenomenon as *attack transferability* where the attacked model is called *surrogate model* and the model to which the attack is transferred is called *target model*. [PMG16] shows that adversarial attacks are transferable between the same models and between different models performing the same task, whether these models are differentiable (such as DNNs) or non-differentiable (such as SVM). [De19] analyzes the level of complexity of the surrogate model in an attempt to justify the transferability effectiveness; a surrogate model that has a low variance loss function is more transferable than a model with a high variance loss function. In order to ameliorate transferability across different neural networks performing the same task, [Xi19] modifies the IFGSM attack by randomly resizing the images at each iteration. [DZJ19] proposes a GAN-based approach to generate synthetic adversarial samples with imperceptible perturbations against FaceNet [SKP15] and report effective results across different face recognition models. Similarly, [ZD20] reports transferability of attacks from an open-source surrogate face recognition model to several commercial target face recognition models.

3 Methodology

Let us denote X_F the space of all facial images, X_C the space of clean images, X_{Adv} the space of adversarial images and X_{Den} the space of denoised adversarial images where $X_C \cup X_{Adv} \cup X_{Den} \subseteq X_F$. We denote $Y_G = \{0,1\}$ the space of the gender labels, and $Y_R = \{✓, ✗\}$ the space of recognition labels. We consider $G: X_F \rightarrow Y_G$ a gender classifier and $R: X_F \times X_F \rightarrow Y_R$ a facial recognition classifier.

- **Attack:** An adversarial attack $f_{Adv}: X_C \rightarrow X_{Adv}$ is considered successful if for $x \in X_F$ there is an adversarial sample $f_{Adv}(x) = x_{Adv} \in X_{Adv}$ such that: $G(x) = y_G$ and $G(x_{Adv}) = \overline{y_G}$.

- **Denoising Defense:** Let $f_{Den}: X_{Adv} \rightarrow X_{Den}$ denote a denoising function. Ideally, a denoised image $x_{Den} = f_{Den}(x_{Adv}) \in X_{Den}$ and verifies $G(x_{Den}) = G(x)$ where $x \in X_C$ is the clean image such that $x_{adv} = f_{Adv}(x)$.

- **Gender-Recognition Transferability:** We say that an attack f_{adv} and a defense f_{Den} are transferrable from the gender classifier G to a face recognition model R for $(x_1,x_2) \in X_C \times X_C$ if we have $R(x_1,x_2) \neq R(x_1, f_{Adv}(x_2))$ and $R(x_1,x_2) = R(x_1, f_{Den} \circ f_{Adv}(x_2))$
- **Metrics:** we use the classification accuracy, that is the number of correct predictions divided by the total number of predictions, to measure the performance of the gender classifier and we derive different performance metrics for the face recognition classifier, based on a similarity measure.

Based on the above-mentioned definitions, we adopt the following procedure:

1. Train the gender classifier and measure its classification accuracy.
2. Attack the gender classifier to generate a set of adversarial samples.
3. Train a denoising defense on a subset of adversarial samples and their corresponding clean versions and evaluate it on a separate subset by comparing the classification accuracy of the gender classifier on the adversarial images and their denoised versions.
4. Run a face recognition model on a clean set, its adversarial, its denoised versions and their combinations to assess the transferability of the attack and the defense methods in terms of the sensitivity of the recognition performance across the diverse sets of images.

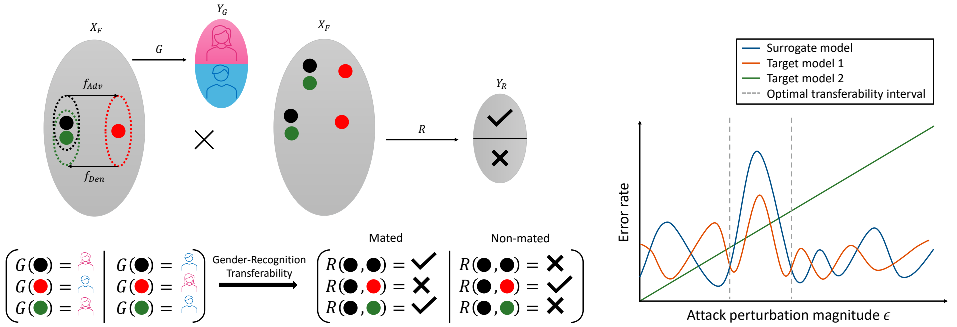


Fig. 2: Methodology overview for analyzing the transferability attack from a gender classifier (surrogate model) to a face recognition classifier (target model). The graph on the right illustrates the expected behaviour of the error variation of surrogate and target models as a function of the applied perturbation. Target model 1 is sensitive to the attack on the surrogate model unlike target model 2 that deteriorates when an image is reduced in quality, considering ε similarly as blurring or random noise.

4 Background

FGSM Attack on Gender Classifier: We use $J(\theta,x,y_G)$ to denote the loss function of the gender classifier G with respect to an input image $x \in X_C$ and its ground truth gender label $y_G \in Y_G$. The FGSM attack maximizes the loss with respect to the input image [GSS14] by adding to the image a step ε in the direction of the loss gradient. An FGSM adversarial attack

$f_{Adv}: X_C \rightarrow X_{Adv}$, with perturbation magnitude $\varepsilon \in \mathbb{R}$, results in adversarial images $x_{Adv} \in X_{Adv}$ such that: $x_{Adv} = x + \varepsilon \cdot sign(\nabla J(\theta, x, y_G))$. Note that FGSM attack does not rely on equalizing the probabilities of the different input classes and thus we cannot expect an equal probability between the classes after the attack. Instead, it relies on changing the prediction by simulating a gradient ascent behaviour on the sample images.

High-level representation and pixel guided denoisers: In this paper, we consider two types of denoisers: pixel-guided denoiser (PGD) and high-level representation guided denoiser (HGD). A PGD learns to reconstruct a clean image x by reducing the loss defined as, $\mathscr{L}_{\text{PGD}} = \|x - x_{Adv}\|_1$, the pixel level difference between a clean image x and its adversarial version x_{Adv}. Whereas, a HGD [Li18] reduces the loss defined as, $\mathscr{L}_{\text{HGD}} = \|f^i_{emb}(x) - f^i_{emb}(x_{Adv})\|_1$, the difference between the deep features of a clean image x and the deep features of its adversarial version x_{Adv} where $f^i_{emb}: X_C \rightarrow \mathbb{R}^n$ denotes the function describing the attacked model until its i^{th} layer that outputs a feature vector of size n.

5 Experiment and Evaluation

Architectures: We used the VGG16 architecture as the gender classification network and restricted its last layer to two classes to suit our classification goal. The same architecture is used for the face recognition model VGGFace pre-trained on the VGGFace dataset [PVZ15]. VGG16 has a straightforward architecture that comprises 13 convolution layers and 3 fully connected layers. For the denoiser, similarly as [Li18], we use a U-Net based Denoising Convolution Neural Network (DnCNN) [3] a denoising model that we will refer to in this work as UDnCNN. The structure of the UDnCNN denoiser has an encoding part sharing skip connections with a decoding part. The skip connections allow the transfer of fine-grained information that could be lost in a regular auto-encoder.

Dataset division: We use the CelebA dataset that comprises 202,599 samples of 10,177 different individuals. We divide this dataset into three sets: A (162,770 samples), B (19,962 samples), and C (19,867 samples) with respect to the train-test-validation partition provided by the authors [Li15] where identities do not overlap. For the FGSM attack against the gender classifier and the defenses experiment, we use sets A and B to train and test the gender classifier and set C to generate FGSM adversarial images against the gender classifier. The resulting adversarial images and their corresponding clean versions are partitioned into four subsets: $C_{AdvTrain}$ and $C_{CleanTrain}$ of equal size (73,779 samples each) as well as $C_{AdvTest}$ and $C_{CleanTest}$ (18,449 samples each). The subsets $C_{CleanTrain}$ and $C_{AdvTrain}$ are used for the training of the denoisers while $C_{CleanTest}$ and $C_{AdvTest}$ are used to evaluate them. For the transferability experiment, we use set B to get the clean images from which we generate the adversarial images and their corresponding denoised images. Since not all the clean images from B are vulnerable to FGSM, we collect for each adversarial image, the clean image it was derived from and its denoised image. As a result, we have a set of clean images, a set of adversarial images, and another set of denoised images of the same size (94,965 samples and 995 identities each). Those three sets are used to analyze the transferability of the FGSM attack on the face recognition classifier.

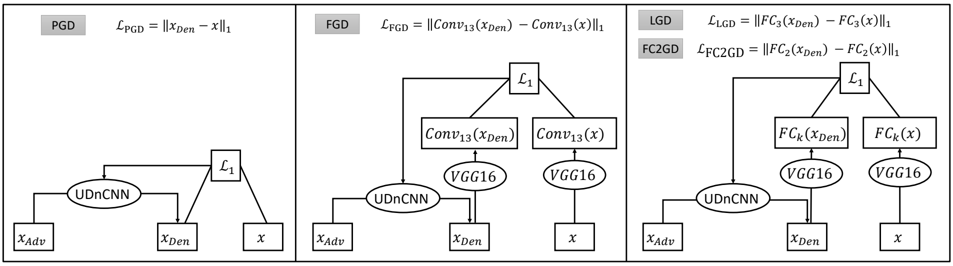
[3] https://github.com/lychengr3x/Image-Denoising-with-Deep-CNNs

Performance metrics: To assess the gender classifier performance, either before the attack and the defense or after, we calculate the classification accuracy. To reason in terms of errors in the two models, we use the classification error rate (1 - accuracy) for the gender classifier in Figure 1. For the face recognition performance, we use cosine similarity to measure the False Non-Match Rate (FNMR) at a fixed False Match Rate (FMR) of 0.1% as well as the Area Under the Detection Error Trade-off Curve (AUC-DET) and finally, the Equal Error Rate (EER).

Training the gender classifier on CelebA: We trained our gender classifier from scratch using batch normalization after convolution layers to speed up the training of the baseline VGG16 achieving a validation accuracy of 98.62 %.

FGSM attack: We run the FGSM attack on the VGG16 gender classifier using various values for the perturbation $\varepsilon \in [0.005, 0.55]$. Figure 4a shows how the classifier behaves for different values of ε. We observe that the accuracy decreases for ε between 0.01 and 0.035 and it starts to increase from 0.04. As our goal is to study the effect of perturbations that are imperceptible to the human, we consider the following range of epsilons $\varepsilon \in \{0.01, 0.015, 0.02, 0.025, 0.03, 0.035\}$ as it is where the classifier is most vulnerable.

Denoising losses: In addition to a PGD, we use three types of HGDs illustrated in Figure 3: FGD based on the last convolutional layer of the gender classifier, FC2GD based on the second fully connected layer and LGD based on the logits layer.

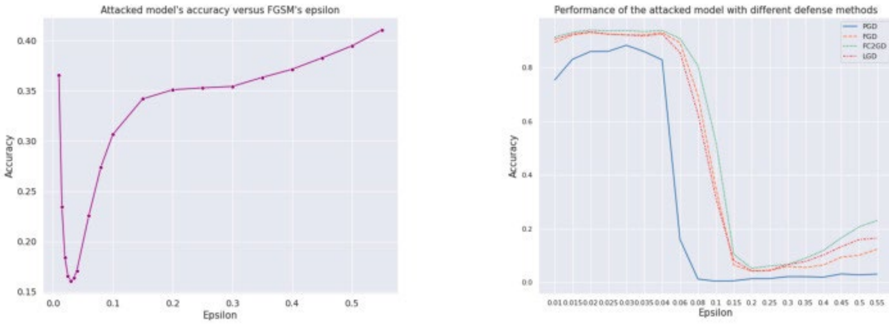


PGD
$\mathcal{L}_{\text{PGD}} = \|x_{Den} - x\|_1$
$\mathcal{L}_1$
UDnCNN
x_{Adv}
x_{Den}
x
FGD
$\mathcal{L}_{\text{FGD}} = \|Conv_{13}(x_{Den}) - Conv_{13}(x)\|_1$
$\mathcal{L}_1$
$Conv_{13}(x_{Den})$
$Conv_{13}(x)$
$VGG16$
$VGG16$
UDnCNN
x_{Adv}
x_{Den}
x
LGD
$\mathcal{L}_{\text{LGD}} = \|FC_3(x_{Den}) - FC_3(x)\|_1$
FC2GD
$\mathcal{L}_{\text{FC2GD}} = \|FC_2(x_{Den}) - FC_2(x)\|_1$
$\mathcal{L}_1$
$FC_k(x_{Den})$
$FC_k(x)$
$VGG16$
$VGG16$
UDnCNN
x_{Adv}
x_{Den}
x

Fig. 3: Training of UDnCNN denoiser when considering the PGD defense, the FGD defense, the LGD defense (k = 3) [Li18] and when considering the FC2GD defense (k = 2).

Figure 4b shows the performance of the defense methods over increasing values of epsilon. FC2GD seems to be the most robust against adversarial examples generated with values of epsilon outside of its training range, followed by LGD and FGD. PGD on the other hand, is the most vulnerable to high epsilons. Nevertheless, we notice that the performance inevitably drops at a certain range for all three HGD methods before slowly increasing again.

Comparison between the defense methods: Table 1 compares the performance of the attacked VGG16 gender classifier when applying the different defense methods (columns 2 to 5) and without (first column), over clean images (row 2) and adversarial images (row 3). For PGD and FC2GD, both considerably help in defending the classifier against adversarial attacks as the accuracy reaches 84.34% on the adversarial test images with PGD denoising and 93.14% with FC2GD denoising. We also observe that there is a deterioration of the performance of the classifier on clean images after they are fed into the denoiser. This effect is particularly noticeable for the FC2GD. The latter seems to infer adversarial noise more effectively than PGD but with the



(a) Sensitivity of the classification accuracy of the VGG16 gender classifier upon the choice of perturbation (epsilon) used in the FGSM attack.

(b) Effect of the defense methods on the classification accuracy of the VGG16 gender classifier over increasing values of ε.

Fig. 4: Classification accuracy of VGG16 gender classifier during FGSM attack and after applying the defense methods over various attack intensities.

expense of reduced discriminative power in clean images. For the HGD methods, we observe the higher the representation (i.e the deeper the target layer) the better the defense method performs on clean images and that LGD seems to be the most convenient method for defense so far.

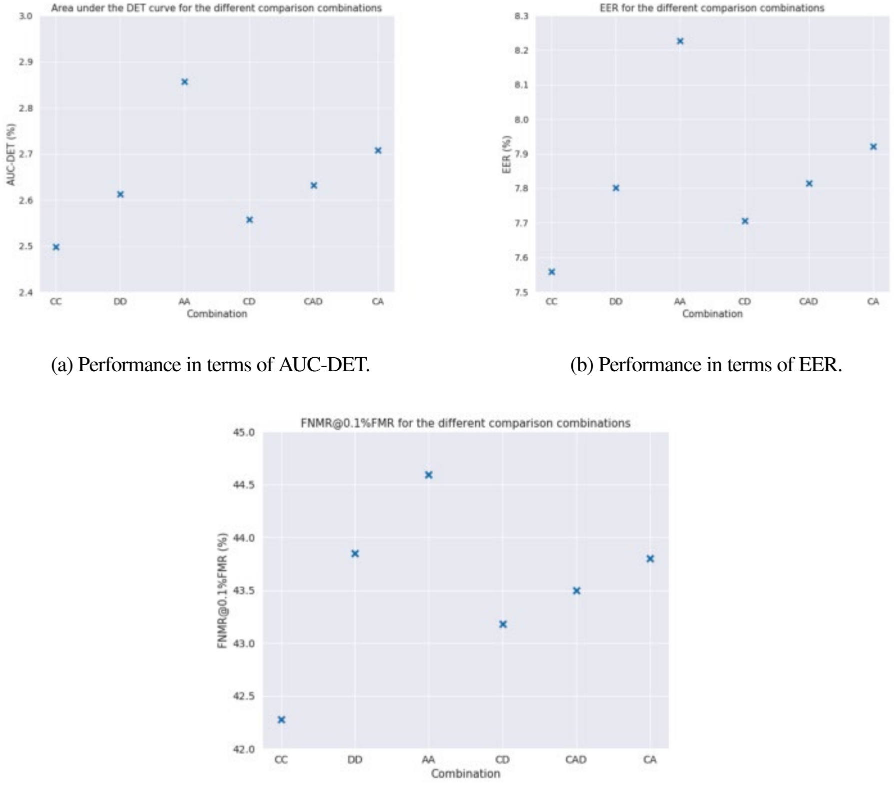
	Without denoising	PGD	FGD	FC2GD	LGD
Clean Test	98.19%	95.61%	57.50%	63.48%	83.05%
Adversarial Test $\varepsilon \in [0.01, 0.035]$	0%	84.34%	91.82%	93.14%	92.02%

Tab. 1: Performance summary of the attacked VGG16 gender classifier in terms of accuracy with and without the defense methods

Transferability of FGSM on Face Recognition Classifier: We study the transferability of the attack from the gender classifier (surrogate) to the face recognition model (target) by performing six comparison combinations of mated and non-mated comparisons depending on the type of the input images, either clean, adversarial or denoised. The totality of these combinations are illustrated in Figure 1. We perform a verification entirely on the clean set (CC) to obtain a base-line performance of VGGFace before running the FGSM attack. We then perform clean/adversarial (CA) and clean/denoised (CD) verifications to evaluate the transferability of both the attack and the defense method. We also report the combinations adversarial/adversarial (AA), denoised/denoised (DD) and a blind verification on the three sets combined (CAD) to further assess the robustness of VGGFace. To realize the comparisons, we select ~ 15 different images per subject where each image should be vulnerable to at least 3 values of ε out of 6. Table 3 summarizes the resulting numbers of mated and non-mated comparisons per epsilon and in total.

We notice in the Figures 5a, 5b and 5c that the presence of non-clean images (denoised and adversarial) regardless of the attack intensity, decreases the recognition performance. The difference between the variation of the performance in the combinations CC, CD and DD, where there is 0% of adversarial samples, and the variation in combinations CAD, CA and AA, where there is 33%, 50% and 100% respectively, shows that VGGFace is prone to degradation as more adversarial images are included in the comparisons. In case of the three comparison combinations CC, CD and CA, we observe that the recognition performance degrades from CC to CA and that the error difference is larger than the error difference between CC and CD. This suggests that the defense partly compensates the performance degradation.

Table 2 (a) and Table 2 (b) show that for each combination involving adversarial or denoised images, the errors are the highest for the smallest perturbation 0.01 then for the subsequent increasing perturbations, the errors decrease until perturbation 0.025 before they start to increase again. This implies a low transferability of the attack in the selected epsilon range. It is possible that a more optimal range of epsilon values exist, that would result in a high transferability of the attack as shown in the illustrative graph in Figure 2.



(a) Performance in terms of AUC-DET.

(b) Performance in terms of EER.

(c) Performance in terms of FNMR@0.1%FMR.

Fig. 5: Performance measures across the different comparison combinations: **C** designates Clean, **A** designates Adversarial and **D** refers to denoised.

ε	0.01	0.015	0.02	0.025	0.03	0.035
CC	**46.96**	44.14	42.05	41.41	41.41	41.39
DD	**47.70**	45.66	43.62	43.08	43.12	43.30
AA	**47.28**	45.16	43.96	43.38	44.78	44.76
CD	**47.39**	45.08	42.95	42.33	42.42	42.52
CAD	**47.28**	44.98	43.20	42.57	43.03	43.08
CA	**47.18**	44.86	43.41	42.78	43.70	43.68

(a) FNMR@0.1%FMR in percentage (%)

ε	0.01	0.015	0.02	0.025	0.03	0.035
CC	**2.98**	2.68	2.51	2.41	2.41	2.41
DD	**3.08**	2.78	2.61	2.50	2.52	2.55
AA	**3.08**	2.87	2.83	2.70	2.86	2.86
CD	**3.04**	2.73	2.57	2.46	2.46	2.48
CAD	**3.04**	2.76	2.63	2.52	2.58	2.58
CA	**3.04**	2.79	2.70	2.58	2.69	2.69

(b) AUC-DET in percentage (%)

Tab. 2: Comparison performance of different combinations per epsilon in terms of FNMR@0.1%FMR in (a) and area under the DET curve (AUC-DET) in (b) where the first rows serve as a reference with only clean images.

ε	0.01	0.015	0.02	0.025	0.03	0.035	All
CC	M = 4.7E3 U = 1.2E6	M = 8.4E3 U = 2.6E3	M = 1.1E4 U = 3.8E6	M = 1.4E4 U = 4.7E6	M = 1.4E4 U = 4.7E6	M = 1.4E4 U = 4.7E6	M = 6.7E4 U = 2.1E7
DD	M = 4.7E3 U = 1.2E6	M = 8.4E3 U = 2.6E6	M = 1.1E4 U = 3.8E6	M = 1.4E4 U = 4.7E6	M = 1.4E4 U = 4.7E6	M = 1.4E4 U = 4.7E6	M = 6.7E4 U = 2.1E7
AA	M = 4.7E3 U = 1.2E6	M = 8.4E3 U = 2.6E6	M = 1.1E4 U = 3.8E6	M = 1.4E4 U = 4.7E6	M = 1.4E4 U = 4.7E6	M = 1.4E4 U = 4.7E6	M = 6.7E4 U = 2.1E7
CD	M = 9.4E3 U = 2.4E6	M = 1.6E4 U = 5.3E6	M = 2.3E4 U = 7.7E6	M = 2.8E4 U = 9.4E6	M = 2.8E4 U = 9.4E6	M = 2.8E4 U = 9.4E6	M = 1.3E5 U = 4.3E7
CAD	M = 1.8E4 U = 4.9E6	M = 3.3E4 U = 1.0E7	M = 4.7E4 U = 1.5E7	M = 5.7E4 U = 1.8E7	M = 5.7E4 U = 1.8E7	M = 5.7E4 U = 1.8E7	M = 2.7E5 U = 8.7E7
CA	M = 9.4E3 U = 2.4E6	M = 1.6E4 U = 5.3E6	M = 2.3E4 U = 7.7E6	M = 2.8E4 U = 9.4E6	M = 2.8E4 U = 9.4E6	M = 2.8E4 U = 9.4E6	M = 1.3E5 U = 4.3E7

Tab. 3: Number of mated (M) and non-mated (U) comparisons

6 Conclusion

In this work, we studied the effect of the FGSM attack on the VGG16 gender classifier over a variety of perturbations. We also applied defense methods from the literature such as a pixel guided denoiser PGD and variants of high-level representation guided denoisers. We studied the transferability of the FGSM attack with a selected range of epsilons and the LGD defense on a pre-trained face recognition model. Our experiments confirmed that the attack and the defense of the gender classifier impact the performance of the face recognition model. This result consolidates the existing literature reporting an association between face recognition and gender classification, except that this time, this association is demonstrated through an adversarial attack and defense. We hope this work opens grounds to think about transferability of adversarial attacks between models built for different tasks while maintaining the same input space domain.

Acknowledgment

This work was supported by the PriMa project that has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 860315.

References

[AG17] Abbasi, Mahdieh; Gagné, Christian: Robustness to adversarial examples through an ensemble of specialists. arXiv preprint arXiv:1702.06856, 2017.

[De19] Demontis, Ambra; Melis, Marco; Pintor, Maura; Jagielski, Matthew; Biggio, Battista; Oprea, Alina; Nita-Rotaru, Cristina; Roli, Fabio: Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In: 28th {USENIX} Security Symposium. 2019.

[DZJ19] Deb, Debayan; Zhang, Jianbang; Jain, Anil K: Advfaces: Adversarial face synthesis. In: 2020 IEEE International Joint Conference on Biometrics (IJCB). 2019.

[Go18a] Gonzalez-Sosa, Ester; Fierrez, Julian; Vera-Rodriguez, Ruben; Alonso-Fernandez, Fernando: Facial soft biometrics for recognition in the wild: Recent works, annotation, and COTS evaluation. IEEE Transactions on Information Forensics and Security, 2018.

[Go18b] Goswami, Gaurav; Ratha, Nalini; Agarwal, Akshay; Singh, Richa; Vatsa, Mayank: Unravelling robustness of deep learning based face recognition against adversarial attacks. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2018.

[GR14] Gu, Shixiang; Rigazio, Luca: Towards deep neural network architectures robust to adversarial examples. arXiv preprint arXiv:1412.5068, 2014.

[GSS14] Goodfellow, Ian J; Shlens, Jonathon; Szegedy, Christian: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

[Hu15] Huang, Ruitong; Xu, Bing; Schuurmans, Dale; Szepesvári, Csaba: Learning with a strong adversary. arXiv preprint arXiv:1511.03034, 2015.

[KGB16] Kurakin, Alexey; Goodfellow, Ian; Bengio, Samy: Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236, 2016.

[Ku16] Kurakin, Alexey; Goodfellow, Ian; Bengio, Samy et al.: , Adversarial examples in the physical world, 2016.

[Li15] Liu, Ziwei; Luo, Ping; Wang, Xiaogang; Tang, Xiaoou: Deep Learning Face Attributes in the Wild. In: Proceedings of International Conference on Computer Vision (ICCV). 2015.

[Li18] Liao, Fangzhou; Liang, Ming; Dong, Yinpeng; Pang, Tianyu; Hu, Xiaolin; Zhu, Jun: Defense against adversarial attacks using high-level representation guided denoiser. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[MDFF16] Moosavi-Dezfooli, Seyed-Mohsen; Fawzi, Alhussein; Frossard, Pascal: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[Mi18] Mirjalili, Vahid; Raschka, Sebastian; Namboodiri, Anoop; Ross, Arun: Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images. In: 2018 International Conference on Biometrics (ICB). IEEE, 2018.

[OAE16] Ozbulak, Gokhan; Aytar, Yusuf; Ekenel, Hazim Kemal: How transferable are CNN-based features for age and gender classification? In: 2016 International Conference of the Biometrics Special Interest Group (BIOSIG). IEEE, 2016.

[PMG16] Papernot, Nicolas; McDaniel, Patrick; Goodfellow, Ian: Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint arXiv:1605.07277, 2016.

[PVZ15] Parkhi, Omkar M.; Vedaldi, Andrea; Zisserman, Andrew: Deep Face Recognition. In: British Machine Vision Conference. 2015.

[SKP15] Schroff, Florian; Kalenichenko, Dmitry; Philbin, James: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2015.

[Xi19] Xie, Cihang; Zhang, Zhishuai; Zhou, Yuyin; Bai, Song; Wang, Jianyu; Ren, Zhou; Yuille, Alan L: Improving transferability of adversarial examples with input diversity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[ZD20] Zhong, Yaoyao; Deng, Weihong: Towards Transferable Adversarial Attack Against Deep Face Recognition. IEEE Transactions on Information Forensics and Security, 2020.

Image Quality Assessment on Identity Documents

Claudio Yáñez[1], Juan Tapia[2]

Abstract: This paper developed a method for performing Image Quality Assessment (IQA) on ID-card images. First, we built the dataset, consisting of 204 images from Chilean ID cards, containing real and tampered images with varying quality levels. Then, we evaluated different features, obtaining the best results using the BRISQUE features and a newly trained SVR, with an R^2 score of 0.5868. This proposed method is called BRISQUE-ID. The IQA on ID cards can be used as a pre-processing stage for discarding lousy quality images and helping the subsequent steps in the processing pipeline.

Keywords: Image quality assessment, IQA, identity documents, ID cards, biometric sample quality.

1 Introduction

Identity Document (ID) cards are used nowadays in a wide variety of remote services, such as digital banking, government services and e-commerce, to verify the identity of customers. Ideally, access to this document would be obtained through Near Field Communication (NFC) —which would guarantee the information is read correctly and increase the difficulty of tampering—, or by scanning the document using dedicated hardware. These options are not always feasible and can be expensive. For instance, in South America, a country such as Brazil has a population of more than 200 million inhabitants, and their national ID card is chipless. However, the widespread availability of smartphones with cameras facilitates remote access to ID cards by photographing them while opening new challenges to ensure proper reading and use of the information.

In order to process an ID card remotely, the first step is to capture a digital image of it using a camera. These images are captured remotely in non-controlled scenarios, with different backgrounds, illumination, distances, and hardware qualities. Additionally, different smartphones have unique camera models. These conditions present many difficulties in the process of getting the information from the ID card. For example, if a blurry image were captured, the Optical Character Recognition (OCR) algorithm could fail reading some important data, like the person's name or the national ID number. Therefore, a method to verify the quality of the capture must be implemented, ensuring the subsequent processes operate on an image with enough quality.

According to the literature [Sc20, ZM20] image quality algorithms have focused in two main branches: Face Image Quality Assessment (FQA) which analyses face images fo-

[1] R+D Center TOC Biometrics, claudio.yanez@tocbiometrics.com

[2] da/sec-Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany, juan.tapia-farias@h-da.de

cused on biometrics applications, and Image Quality Assessment (IQA) oriented to general-purpose images for perceptual quality. This perceptual quality can be objective or subjective.

This paper focuses on performing IQA on ID-card images developing a method based on the BRISQUE features [MMB12], called BRISQUE-ID. Performing IQA on ID-card images early in the pipeline can help save computational resources if low quality images are discarded, and at the same time, improve the results in subsequent stages, such as Presentation Attack Detection systems [GMF14] or OCR.

The objective of this paper is to develop a system for predicting subjective image quality on ID-card images. This is accomplished by studying multiple features for ID image quality assessment and using them to predict subjective image quality scores. An ID-card subjective IQA dataset was generated by surveying 15 subjects on the quality of 204 images, which enabled us to evaluate IQA performance.

The rest of this paper is organized as follows: Section 2 describes previous related work. Section 3 describes the dataset used and the protocol employed for obtaining subjective quality scores. Section 4 describes the experiments performed. Section 5 reports the results obtained. Finally, conclusions and future work are reported in Sections 6 and 7 respectively.

2 Related work

Image ‘quality’ can mean fidelity of an image to its source, utility to perform tasks related to the image —such as facial recognition or OCR—, or perceived subjective quality based on the previous meanings [AFFOG12]. Much of the work done on IQA focuses on Face Quality Assessment (FQA) or general-purpose IQA [Sc20, ZM20]. Some standards describe how biometric sample quality assessments, including FQA, should be performed [IS16]. FQA is used to ensure face pictures in ID documents have been adequately taken, ensuring high sample quality [GNH19].

In [Sc20], over 50 works on FQA were surveyed. Some of the methods shown employ measurements that are specific for evaluating faces, *i.a.* pose, location of facial features and facial expression. Additionally, the scores yielded by FQA algorithms are usually intended to predict facial recognition performance. These two factors may prevent FQA algorithms from being usable for other types of IQA.

Objective blind or No-Reference (NR) IQA refers to automatic quality assessment of an image through an algorithm that only requires the image to be assessed as input information [MMB12]. In contrast, Full-Reference (FR) or Reduced-reference (RR) IQA algorithms require a ‘clean,’ new reference image in the case of FR IQA, or some information about the reference image (such as a watermark or template) in the case of RR IQA [SB12]. In this sense, NR IQA methods have the advantage that they can be used in scenarios where a reference image cannot be obtained beforehand.

Mittal *et al.* developed a NR IQA model called Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [MMB12]. BRISQUE was developed considering certain regular statistical properties present in natural images —images were taken with an optical camera as opposed to synthetic images— that change by the presence of distortions and can be measured. These properties were measured by extracting locally normalized luminance coefficients, as well as pairwise products of these coefficients, and modeling them using natural scene statistics [Ru94]. These properties resulted in 18 parameters, which were extracted at two scales for a total of 36 features. Finally, a Support Vector machine Regressor (SVR) was trained using the BRISQUE features extracted from the LIVE IQA database [SSB06].

The LIVE IQA database was presented by Sheikh *et al.* as one of the most significant subjective image quality studies at the time [SSB06]. The database has 779 distorted images, generated from 29 high-resolution reference images, which included faces, people, animals, nature scenes and artificial objects, among others. Five distortion types were used to generate the distorted images: JPEG2000 compression, JPEG compression, white noise, Gaussian blur, and simulated fast fading Rayleigh channel. These distorted images were evaluated in 7 sessions using a double-stimulus methodology [Se12]. The images were evaluated by an average of 23 subjects per session, on a scale of 0–100, where 0 is the lowest possible quality and 100 the highest. The scale was divided into five equal portions with the adjectives 'Bad', 'Poor,' 'Fair,' 'Good,' and 'Excellent'. In every session, both the original and the distorted version of an image appeared, which enabled the authors to obtain a differential score for every distorted image. These scores were subject-normalized and realigned using the responses of an 8th session, yielding a Difference Mean Opinion Score (DMOS) that indicate the subjective quality of the images.

3 IQA dataset

For this proposal, an IQA dataset was constructed using 204 images taken from the database used in [GVT21]. The dataset is comprised of real Chilean national ID cards ('digital'), printed ID cards ('printed') and ID cards displayed on screens ('screen'), with 68 images per class. The images are of varied quality, ranging from completely out of focus or tampered, to well-focused or real-looking forgeries. The background was removed automatically, and then the images where resized to 320 × 240 pixels. Example images are shown in Figure 1.

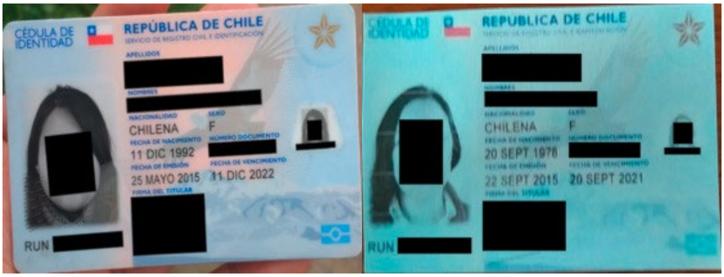


Fig. 1: Left: High quality example. Right: Low quality example. Sensible information was covered with black boxes.

Subjective quality assessments were obtained by surveying 15 subjects. The number of subjects was due to availability, and no subject selection procedure was carried out, besides ensuring the subjects could follow the survey instructions. Out of these subjects, 6 had prior experience in ID card quality assessment or image processing, while the other 9 had no relevant prior experience.

In our attempt to perform ID-card IQA using BRISQUE, we loosely based our survey protocol on [SSB06]. The purpose and the instructions of the survey were explained to the subjects before starting. The survey was separated into three sessions, each with only one class of images to reduce inter-class biases. In each session, the subjects had to evaluate the images one at a time, on a 1-to-5 scale; the five opinion scores (OS) were labeled 'Bad,' 'Poor,' 'Fair,' 'Good,' and 'Excellent'. Subjects were instructed to consider aspects such as focus, lighting and integrity for their evaluation. In most cases, the three sessions were completed in succession, with a short rest between them. Each subject evaluated each of the 204 images, resulting in 3,060 individual quality judgments.

The OS were processed in order to obtain a single quality score for each image. First, a Normalized Opinion Score (NOS) is calculated for each image i and each subject j according to the following equation:

$$NOS_{ij} = \frac{OS_{ij} - \mu_j}{\sigma_j}$$

where OS_{ij} is the OS given by subject j to image i, and μ_j and σ_j are the mean and standard deviation of all scores given by subject j. The resulting NOS are averaged for each image, resulting in a Mean Normalized Opinion Score (MNOS):

$$MNOS_i = \frac{\sum_{j=1}^{N} NOS_{ij}}{N}, \quad N = 15$$

These scores are based on the DMOS scores of [SSB06]. However, since our survey sessions were few and performed over a short time, no realigning step was performed.

The main difference between [SSB06] and our work is in the scope of the work. The dataset generated by [SSB06] was intended for detecting specific distortions, and had images of various kinds. Our dataset contains only images of Chilean ID cards, and includes distortions due to source (photos of real ID cards, photos of printed ID cards and photos of ID cards displayed on screens) and due to image capture conditions (lighting and focus variations). Thus, our dataset is application-specific for training a Chilean ID card IQA model.

4 Experiments

Two experiments on the ID-cards dataset were conducted. First, the BRISQUE score "out-of-the-box" was used as a measure of image quality, obtaining poor results. Then, we trained an SVR using BRISQUE features and compared its performance to other features.

4.1 Experiment 1 – Out-of-the-box BRISQUE tests

Initially, we evaluated BRISQUE "out-of-the-box" (OOB) ability to predict subjective image quality on ID cards. For this, PyBRISQUE[3] implementation was used. This implementation includes the 36 luminance-based BRISQUE features, and the pre-trained SVR. The implementation results closely resemble those obtained by the original BRISQUE paper [MMB12].

The quality scores given by OOB BRISQUE (features and SVR) showed no correlation with the subjective image quality of ID cards. This result was evidenced by a Pearson Correlation coefficient of 0.0985, calculated between the OOB BRISQUE scores and the MNOS subjective scores of our dataset (scores shown in Figure 2). This is further discussed in Section 5.

4.2 Experiment 2 – Feature comparison

The following hand-crafted feature types were selected for comparison: raw pixel intensity, BRISQUE, Local Binary Patterns (LBPs)/quadrant-LBPs (QLBPs) [OPM02], Histograms of Oriented Gradients (HOG) [DT18], and discrete Fourier transform (DFT) [LLJ08]. The number of features per type is shown in Table 1. In all cases, images were transformed to grayscale and scaled to 320×240 pixels using OpenCV before extracting the features. Some of the features have multiple parameters, which are mentioned in the following section. In those cases, only the parameters that yielded the best results are reported.

Feature type	N. of features
Raw image	76,800
BRISQUE	36
LBPs	1,024
QLBPs	4,096
HOG	12,000
DFT	76,800

Tab. 1: Number of features per feature type.

As a baseline, the raw grayscale intensity values of the dataset images were used as features. These images correspond to automatically cropped IDs, which resulted in different sizes. For that reason, all ID cards were resized to 320×240 pixels and flattened to a $1 \times 76,800$ vector. This size kept the image ratio closest to the cropped images.

PyBRISQUE implementation allows for the raw BRISQUE features to be used instead of just obtaining an image quality score. As described previously, BRISQUE extracts 36 features —18 at two different scales— which describe the distributions of locally normalized luminance coefficients. Further on, we call the combination of BRISQUE features with the newly trained SVR BRISQUE-ID, to differentiate it from OOB BRISQUE.

[3] `https://github.com/bukalapak/pybrisque`

LBP features were selected as an attempt to use texture descriptors for predicting image quality. The sklearn Python library was used to extract LBP features. Among all LBP variants, the default and the nonrotation-invariant uniform LBPs yielded the best results. The former was tested using ten neighbors, and the latter (due to size constraints) was tested using 8, 16, and 24 neighbors. In both cases, a radius of one, two, four, and six were used. The best results were obtained using default LBP with ten neighbors and a radius of one among these combinations.

In order to preserve spatial information when using LBPs, the images were divided into quadrants, and LBPs were extracted separately for each quadrant. This is referred to as quadrant-LBPs (QLBPs). The resulting feature vector is obtained by concatenating the four resulting LBP vectors; thus, QLBPs have four times as many features as their corresponding LBPs. The same variants and number of neighbors and radii used in LBPs were explored. The best results were also obtained using default LBP with ten neighbors and a radius of one.

HOG features were selected as an attempt to use shape descriptors for predicting image quality. HOG features were extracted using the sklearn Python library. When using 8 orientations, cells of 8×8, 10×10, 12×12 and 16×16 pixels were used. When using 10 and 12 orientations (due to size constraints), only cells of 8×8 pixels were used. Among these combinations, the best results were obtained using ten orientations and 8x8 cells.

The last feature used was the discrete Fourier transform (DFT) of the ID document image. DFT has successfully been used in the past to detect blur in images [LLJ08]. Lower frequencies of the DFT image are removed by shifting the zero-frequency component to the center of the spectrum, setting the 120×120 pixels area surrounding it to zero, and shifting the zero-frequency back. This feature slightly improved the results obtained when using DFT. Additionally, every feature was scaled to the [0–1] range.

4.3 Model training

An SVR with a Gaussian kernel was used to model and predict MNOS values for ID cards. In order to reduce the chance of biases due to partitioning, accuracy was averaged over ten trials with stratified random 80/20 train/test partitions; stratification keeps the image classes balanced in every trial. Furthermore, the SVR parameters were set using five-fold cross-validation in each trial. The C and γ parameters were selected from the following options: $C: \{10^x, x \in [-3,1]\}$, and $\gamma: \{10^x, x \in [-3,0]\}$.

4.4 Metrics

Mean Squared Error (MSE), Mean Absolute Error (MAE), coefficient of determination (R^2 score), and Pearson correlation index are used to evaluate the regression model performance. The MSE metric penalizes outliers more heavily compared to the MAE metric. The R^2 score is a good indicator of how well new samples are likely to be predicted by the model, with a value of one indicating a perfect model and a value of 0 indicating a model that always outputs the expected value of y. The Pearson correlation index is used to compare OOB BRISQUE results with the features that were evaluated.

5 Results and discussion

The results of the IQA feature comparison are shown in Table 2. BRISQUE-ID yielded the best results using all three metrics. Results using BRISQUE-ID were consistently better across all ten trials, with a minimum R^2 score of 0.4754 and a maximum of 0.7870.

HOG yielded results very close to BRISQUE-ID. However, BRISQUE consists of only 36 features, whereas HOG, with ten orientations and 8×8 cells, consists of 12,000. This makes model training and prediction with HOG much slower when compared to using BRISQUE-ID.

Baseline results using the raw image as a feature yielded the worst results. Results using LBP and QLBP were the least consistent, varying drastically across trials; LBP maximum and minimum R^2 scores were 0.0603 and 0.4985, respectively. QLBP shows improvements when compared to LBPs as a result of preserving more spatial information.

		Metric		
Feature type	R^2 score	MSE	MAE	Pearson c.c.
OOB BRISQUE				0.0985
Raw image	0.1127 ± 0.0510	0.4300 ± 0.0809	0.5363 ± 0.0502	0.3862 ± 0.1018
BRISQUE-ID	$\mathbf{0.5868 \pm 0.0885}$	$\mathbf{0.1972 \pm 0.0511}$	$\mathbf{0.3278 \pm 0.0296}$	$\mathbf{0.7703 \pm 0.0748}$
LBPs	0.1927 ± 0.1228	0.3978 ± 0.1049	0.4728 ± 0.0589	0.4516 ± 0.1336
QLBPs	0.3069 ± 0.1226	0.3412 ± 0.0918	0.4312 ± 0.0610	0.5574 ± 0.1010
HOG	0.5083 ± 0.0534	0.2404 ± 0.0590	0.3891 ± 0.0468	0.7677 ± 0.0305
DFT	0.2385 ± 0.0697	0.3675 ± 0.0649	0.4995 ± 0.0433	0.5545 ± 0.0853

Tab. 2: Comparison of IQA regression results using different features.

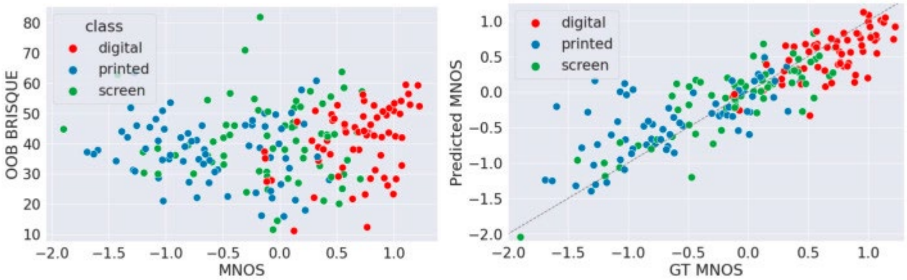


Fig. 2: Left: Comparison between OOB BRISQUE and MNOS scores on the ID-Cards dataset, showing no correlation between them. Right: MNOS, ground truth vs. predicted, using BRISQUE-ID. The dashed line represents perfect prediction. All dataset images are displayed here, although metrics were calculated only on the test partitions.

As mentioned in Section 4.1, the OOB BRISQUE scores show little correlation with the subjective quality MNOS scores. This is reflected both on the low Pearson correlation coefficient (Table 2) and on Figure 2. Note that, while MNOS scores ranged from −2.0 to 1.5 and OOB BRISQUE scores ranged from 10 to 80, correlation does not depend on the scale of the scores.

Because both OOB BRISQUE and BRISQUE-ID use the same features, the difference of results was due to the SVR. As described in Section 3, the dataset used to train the OOB BRISQUE SVR (the LIVE IQA dataset) and our dataset have a different scope. The SVR learns the features that give an image high or low quality index (DMOS or MNOS), but these features may be different in each case —i.e., a JPEG-compressed image of a tree has bad quality for different reasons than a picture of an ID card being displayed on a screen.

This could be extrapolated to more closely-related fields. ID cards of different countries, or even Chilean IDs distorted in a way that was not accounted for, may be evaluated incorrectly by our model. Our model has not yet been tested on images with different content or distortions. In those cases, training a new application-specific model may be required.

6 Conclusions

In this work, we studied image quality assessment on ID-card images. We were able to perform No Reference IQA on Chilean ID cards using BRISQUE-ID. This method performed significantly better than other hand-crafted features, such as LBPs, while using only 36 features. The BRISQUE-ID methodology could be replicated for performing IQA on other types of images, as long as the proper dataset is present. We also showed that the general-purpose OOB BRISQUE was not adequate for Chilean ID–card IQA.

7 Future work

A better and larger dataset could be constructed by surveying more subjects and a wider variety of images. The limited availability of ID-card images makes it harder to generate a large-scale dataset. However, this would improve results so that they reflect more closely real-world application performance. Additionally, using ID-card images taken in a wider range of conditions would let us assess the robustness of our method.

While we were able to perform subjective image quality prediction on ID cards, its impact on the following steps of the pipeline —such as tampering detection or OCR— remains to be studied. These steps should be done by analyzing how bad quality scores correlate with tampering detection or OCR performance.

Acknowledgments

We would like to thank all the volunteers that participated in this research. TOC Biometrics, R+D Center SR-226, supported this work.

References

[AFFOG12] Alonso-Fernandez, Fernando; Fierrez, Julian; Ortega-Garcia, Javier: Quality Measures in Biometric Systems. IEEE Security and Privacy, 10(6):52–62, 2012.

[DT18] Dalal, N.; Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). volume 1. IEEE, pp. 886–893, may 2018.

[GMF14] Galbally, Javier; Marcel, Sebastien; Fierrez, Julian: Image Quality Assessment for Fake Biometric Detection: Application to Iris, Fingerprint, and Face Recognition. IEEE Transactions on Image Processing, 23(2):710–724, feb 2014.

[GNH19] Grother, Patrick; Ngan, Mei; Hanaoka, Kayee: Face Recognition Quality Assessment Concept and Goals. National Institute of Standards and Technology (NIST), 2019.

[GVT21] Gonzalez, S.; Valenzuela, A.; Tapia, J.: Hybrid Two-Stage Architecture for Tampering Detection of Chipless ID Cards. IEEE Transactions on Biometrics, Behavior, and Identity Science, 3(1):89–100, 2021.

[IS16] ISO/IEC JTC1 SC37 Biometrics: ISO/IEC 29794-1:2016 Information technology — Biometric sample quality — Part 1: Framework. Standard, International Organization for Standardization, Geneva, CH, September 2016.

[LLJ08] Liu, Renting; Li, Zhaorong; Jia, Jiaya: Image partial blur detection and classification. 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2008.

[MMB12] Mittal, Anish; Moorthy, Anush Krishna; Bovik, Alan Conrad: No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing, 21(12):4695–4708, 2012.

[OPM02] Ojala, T.; Pietikainen, M.; Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7):971–987, 2002.

[Ru94] Ruderman, Daniel L: The statistics of natural images. Network: computation in neural systems, 5(4):517–548, 1994.

[SB12] Soundararajan, Rajiv; Bovik, Alan C.: RRED Indices: Reduced Reference Entropic Differencing for Image Quality Assessment. IEEE Transactions on Image Processing, 21(2):517–526, 2012.

[Sc20] Schlett, Torsten; Rathgeb, Christian; Henniger, Olaf; Galbally, Javier; Fiérrez, Julian; Busch, Christoph: Face Image Quality Assessment: A Literature Survey. CoRR, abs/2009.01103, 2020.

[Se12] Series, BT: Methodology for the subjective assessment of the quality of television pictures. Recommendation ITU-R BT.500-13, 2012.

[SSB06] Sheikh, Hamid R.; Sabir, Muhammad F.; Bovik, Alan Conrad: A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Transactions on Image Processing, 15(11):3440–3451, 2006.

[ZM20] Zhai, Guangtao; Min, Xiongkuo: Perceptual image quality assessment: a survey. Science China Information Sciences, 63(11):1–52, 2020.

QualFace: Adapting Deep Learning Face Recognition for ID and Travel Documents with Quality Assessment

João Tremoço[1], Iurii Medvedev[2], Nuno Gonçalves[3]

Abstract: Modern face recognition biometrics widely rely on deep neural networks that are usually trained on large collections of wild face images of celebrities. This choice of the data is related with its public availability in a situation when existing ID document compliant face image datasets (usually stored by national institutions) are hardly accessible due to continuously increasing privacy restrictions. However this may lead to a leak in performance in systems developed specifically for ID document compliant images. In this work we proposed a novel face recognition approach for mitigating that problem. To adapt deep face recognition network for document security purposes, we propose to regularise the training process with specific sample mining strategy which penalises the samples by their estimated quality, where the quality metric is proposed by our work and is related to the specific case of face images for ID documents. We perform extensive experiments and demonstrate the efficiency of proposed approach for ID document compliant face images.

Keywords: face recognition, biometric template, document security.

1 Introduction

Security border control applications widely embed biometrics recognition where the face image is one of the most popular biometric source for such applications. The standard approach to the face recognition nowadays implies learning deep face features that are combined into a biometric template. This template is further utilised for distinguishing identities with relatively simple similarity metric and may be stored in a secured database or even embedded to the document itself for performing verification in the match-on-document scenario [Me20].

The features of the template may be learned explicitly by contrastive methods (i.e. by the contrast between match/non-match pairs [Sc15]) or implicitly in the multiclass (identities) classification manner [De19a]. The deep networks, which are used for extracting the biometric template, usually have complex architectures of stacked convolutional layers. These networks are usually trained on big collections of labelled face images of celebrities [Ca18, Gu16].

Face recognition for document security applications possesses specificities. Official identification documents (i.e. biometric passports, national ID cards) adopt only the frontal face

[1] University of Coimbra, Institute of Systems and Robotics - Coimbra, Portugal, joao.tremoco@isr.uc.pt

[2] University of Coimbra, Institute of Systems and Robotics - Coimbra, Portugal, iurii.medvedev@isr.uc.pt

[3] University of Coimbra, Institute of Systems and Robotics - Coimbra, Portugal; Portuguese Mint and Official Printing Office (INCM), Lisbon, Portugal, nunogon@deec.uc.pt

images compliant to ICAO standards [IS18, IS19]. In comparison with unconstrained face recognition systems, which adapts to variations in illumination, pose, occlusion, facial expressions, document security solutions deal with more regular conditions especially in a situation when biometric enrolment tends to become more controlled [Eu18].

At the same time, the collections of ICAO compliant enrolled images, which are usually stored by national institutions, are hardly available for the research and development due to privacy issues. As an example, European GDPR (General Data Protection Regulation) categorise face images as sensitive personal data which results in many constraints for their collecting and distributing [Eu16]. Recently, following this trend, many of the face datasets (even public wild datasets of celebrities) were withdrawn and usually available only in a form of redistribution.

That is why there is a challenge for face recognition in document security when for efficient training of the face recognition algorithms one require large ICAO compliant face image datasets which remain private, and the ones that are public available are of insufficient size. In this situation the most effective approach is to follow training on available wild datasets and then apply some optional measures (like fine-tuning) for achieving better performance in the deploy scenario [SJ18].

In this work we address the problem of this inconsistency between the training and deploy data and introduce a novel approach to mitigate this issue. We propose to emphasise the face features which are more characteristic for ID document compliant images by designing a sophisticated sample mining strategy which regularises the training process. The developed strategy penalises the samples by their quality score (estimated by several metrics). Our approach allows to learn facial biometric template which better suits the document security applications.

2 Related Work

Loss function design have been in a focus of many recent investigations of deep learning face recognition. The general trend of these works was directed onto the increasing the discriminative power of learned features. Most of the current state of the art methods follow the approach of multi-class classification with use of softmax based loss functions. To increase intra-class compactness and inter-class dispersion, several marginal modifications of softmax were proposed. For instance SphereFace, CosFace and ArcFace introduced the margin (in different manner) to the feature logits in the angular domain [Li17, Wa18, De19a]. These methods demonstrated clear geometric interpretation at the same time having relatively simple implementation. Although these loss functions allowed to achieve state of the art performance in several benchmarks, they do not account the hardness and variability of each sample.

Hard sample mining strategies allowed to improve the face recognition performance in several approaches. For instance, MV-Softmax [Wa20] treats miss-classified samples as hard samples increasing their weights in the training process. CurricularFace [Hu20] also uses miss-classification for indicating hard samples and adapts the curricular learning strategy to the face recognition. Hard samples are emphasised increasingly over the training duration with an additional hyper-parameter. NPCFace [Ze20] makes the important distinction between hard positive and hard negative samples and show that for large datasets hard positives will usually be hard negatives for another class as well. The form of the negative logit is defined with use of a binary mask that indicates whether a sample is hard or not. Following the ArcFace approach, the NPCFace also utilises a margin for the positive logits, which is controlled by the hardness of the sample.

These methods try to optimise their performance towards hard samples, however we propose that for the document security applications emphasising higher quality samples during training better suits the target scenario. Unlike the previous works mentioned, MagFace [Me21] includes the quality of the samples in the training process in a way that pulls easy (high quality) samples closer to the class centre and pushes harder (lower quality) samples away. The authors follow a formulation similar to ArcFace where the margin parameter varies for each sample with accordance to its quality. In MagFace, the quality of each sample is defined by magnitude of the feature vector. This approach shares several conceptual similarities to our approach, however we shift our attention to adapting the quality sampling to the document security images scenario.

Document security specific face recognition investigation is reported in several works. DocFace [SJ18] present a method for matching Identification Document (ID) photos to live photos. The authors use a pair of trained sibling networks and fine-tune them on a small private ID-Selfie dataset. The method achieves better performance over general methods, however the dataset used for benchmarking is private. Several improvements on the ID-Selfie dataset and the loss function for fine-tuning were introduced in the DocFace+ [SJ19].

Face Image Quality Assessment (FIQA) inherits aspects from general image QA also considering several other attributes such as pose, illumination, face occlusion or facial expressions. A survey on this topic was done recently by Schlett et. al [Sc20]. Blur is good baseline indicator for the quality of any image. The blur of an image can be extracted by convolving the image with a Laplacian filter and then calculating the variance of the result [Ba16]. BRISQUE [Mi12] is a no-reference generic image quality assessment method. Through the use of scene statistics this method is able to quantify the ”naturalness” and quality of an image. Regarding face specific attributes, several works have been recently developed to extract face specific meta-information from images. The pose of a face in an image can be characterised as a rotation in three dimensions, the yaw pitch and roll. Estimating these angles is helpful to understand a datasets pose distribution. Ruiz et. al [Ru18] use a Convolutional Neural Network (CNN) to estimate these three angles. The quality of facial illumination is also a useful indicator of the quality of a facial image. Zhang et. al [Zh17] use a CNN, which is trained on the FIIQD dataset to score the quality of illumination. FaceQnet [He19] is a face image QA CNN based method. It used a third party framework to calculate ICAO compliance scores used as ground-truth values to train the network. The authors also show high correlation between the resulting scores and face biometric verification performance for a variety of off-the-shelf biometric recognition systems.

Some recently developed methods of face image quality assessment were developed in such a way to remove human perception from the quality estimation process. SER-FIQ [Te20] is a quality estimation method based on the use of dropout during the training of a model. The quality of a sample is defined with respect with the robustness of its embeddings in different sub-networks. The closer the outputs are for different sub-networks, the higher the quality of the sample is. Shi and Jain introduced the concept of Probabilistic Face Embedding (PFE) [Sh19]. This work shows that poor image quality affects the similarity scores of genuine and impostor pairs in such a way that higher degradation of an image leads to higher probability of false reject or false accept of these pairs (named Feature Ambiguity Dilemma). As such, instead of the normal deterministic face embedding, the authors propose to encode the uncertainty in the representation of the face with two different output vectors one representing the Gaussian mean and the other for the Gaussian variance. The authors also introduce a method for matching the PFEs that penalises high levels of uncertainty (variance). SDD-FIQA [Ou21] also bases its quality classification on the recognition performance of the sample in question. This is done by mapping the inter-class and intra-class similarity scores to quality pseudo-labels through the use of a distribution distance metric. Afterwards, these quality values are used to train a network to predict quality scores.

3 Methodology

Deep learning classification approaches usually utilise softmax loss function, which now serves as basis for most of recently developed loss functions in the field of face recognition. It is usually formulated as follows:

$$L_{softmax} = \frac{1}{N}\sum_{i} -\log\left(\frac{e^{f_{y_i}}}{\sum_{j}^{C} e^{f_{y_j}}}\right) \tag{1}$$

where C is the number of classes in the classification problem, y_i is the index of the class of the $i-th$ sample, N is the number of samples in a batch and f_{y_j} is the y_j-th component of the final layer's logits $\mathbf{f}$. If l2 normalisation of the weights $\mathbf{w_j}$ and biometric feature set $\mathbf{x_i}$ is performed, then f_{y_j} can be represented as: $f_{y_j} = w_j^T x_i = \cos(\theta_j)$. The normalised features are constrained on the hyper sphere in $\mathbb{R}^d$ space (where d is the size of $\mathbf{f}$), which leads to the angular similarity metric between samples. By reformulating softmax with this normalisation and adding an angular margin parameter m to the positive logit we obtain the ArcFace loss:

$$L_{arcface} = \frac{1}{N}\sum_{i} -\log\left(\frac{e^{s\cos(\theta_{y_i}+m)}}{e^{s\cos(\theta_{y_i}+m)} + \sum_{j\neq y_i} e^{s\cos\theta_j}}\right) \tag{2}$$

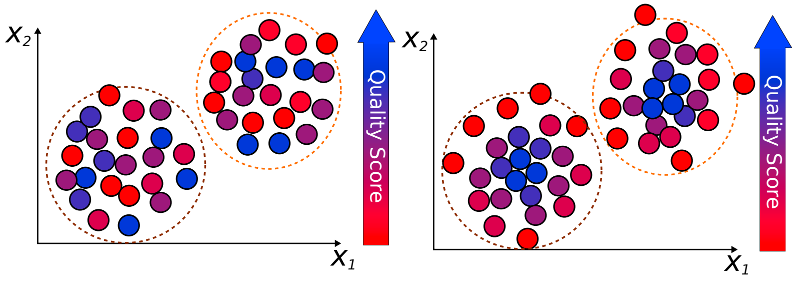
QualFace. Basing on the cooperative margin presented in NPCFace [Ze20], we introduce the concept of adaptive margin with regard to image quality. Our approach, unlike others previously mentioned, implies developing the sample mining strategy, which enhance the impact of higher quality samples instead of harder samples. In this case deep feature distribution is characterised by the concentration of the qualitative samples closer to the class feature centre (see Fig. 1). With this approach, higher impact means higher loss value for samples with better quality. This is done by increasing the margin parameter in the ArcFace loss in an adaptive way, which results in the following formulation:

$$L_q = \frac{1}{N}\sum_i -\log\left(\frac{e^{s\cos(\theta_{y_i}+m_i)}}{e^{s\cos(\theta_{y_i}+m_i)} + \sum_{j\neq y_i} e^{s\cos\theta_j}}\right) \tag{3}$$

where the adaptive margin parameter m_i is defined as a baseline value plus an added constant dependent on the quality of the image:

$$m_i = m_0 + \sum_{j}^{Q} w_j q_{ij} m_1 \tag{4}$$

Here, m_0 and m_1 are hyper-parameters, q_{ij} represents the normalised $j-th$ quality score value for the sample i. Q is the total number of quality attributes and w_j is the weight of each score. For travel document photos, we consider high quality samples as samples that have high ICAO standards compliance [IS18]. For instance, images with frontal poses, clear background, frontal face lighting, no face occlusion, no facial expressions, etc. In our work we use five different indicators of quality that are inspired by ICAO recommendations for portrait photographs: Blur [Ba16], FaceQNet scores [He19], BRISQUE scores [Mi12], Face Illumination quality [Zh17] and a pose score [Ru18]. The pose scores used were calculated as the average of absolute values of the yaw, pitch and roll angles. QualFace strengthens the supervision on higher quality samples through the use of external quality indicators. The following section will show the advantages of QualFace on document security applications.



(a) (b)

Fig. 1: The spatial distribution of two high level features; a) default feature distribution; b) desired distribution in our method.

4 Experiments and Results

We perform extensive training experiments with QualFace and several baseline loss functions and benchmark the result models in a following way.

Training. As a training data we used the subset of public VGGFace2_train dataset [Ca18], selecting classes with more than 400 images per identity. The resulting dataset has a total of 1.34M images and 2842 identities. Face detection and alignment to 299×299 is performed with use of RetinaFace method [De19b]. Each image channel is normalised by subtracting the mean of the training dataset. The scores (FaceQNet, BRISQUE, pose score) were extracted from the aligned images. They are normalised and fed to the model as additional input.

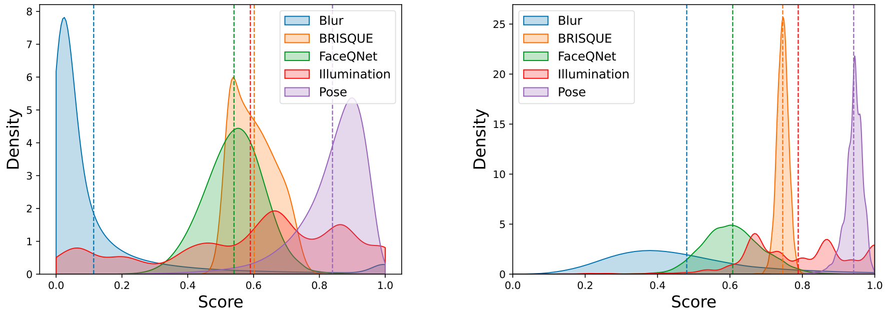
As a backbone CNN architecture we choose the ResNet50V2 [He16], adding the fully connected feature layer with 512 nodes. We initialise all models with the imagenet weights before training. The training was performed on a NVIDIA RTX 3090 GPU. We limit the batch size with 24 images and decay the learning rate with cosine annealing scheduler from $5e-3$ in the beginning to $1e-5$ in the end. The model is trained with SGD optimiser for $6-th$ epochs with a momentum parameter of 0.5 and weight decay of 0.0005.

Benchmarking. In order to demonstrate the effect of our method, and its superiority for ID document compliant images, we designed two different benchmarks datasets. The first one includes ”wild” images, and the second one is comprised of images that are compliant to ICAO standards (we call it ”strict”). The wild benchmark dataset was created basing on a subset of VGGFace2_test part and include identities disjoint from the training set. It contains 31k face images of 147 identities. The strict dataset was created with images from the Face Recognition Grand Challenge V2 (FRGC_V2) dataset [Ph05]. Since its default version includes wild images, we performed its filtering in a semi-automatic way choosing only ICAO compliant images. The final strict dataset contains 11.7k images from 565 identities. For each dataset we generated the protocols for 1-1 for verification by random selecting of comparison pairs. Each protocol contains around 110K pairs for match comparison and 220K pairs for non-match comparison. [3]

To demonstrate the relative difference of distributions across two benchmark datasets we performed min-max normalisation with respect to the minimum and maximum scores values for the VGGFace2_train. One can see that the designed strict benchmark (see Fig. 2b) has better image quality with respect to the five scores presented. The wild benchmark dataset distributions, as expected, turned out to be identical to the train dataset distributions (see Fig. 2a).

Results Discussion. We performed intensive experiments training deep networks with QualFace and observed that the strong applied adaptation usually lead to a problem with the convergence. However, applying regular and careful adaptation, we could attain the superiority of our method. We achieved the best results in two following configurations: $m_0 = 0.4$ with $m_1 = 0.1$ and $m_1 = 0.2$. For each of those we trained five different models using a single score: Blur, BRISQUE, FaceQNet, Illumination and Pose. The Receiver

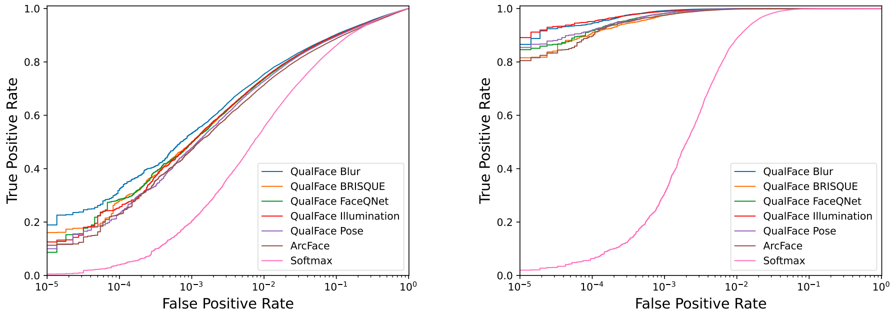
[3] https://github.com/visteam-isr-uc/QualFace



(a) (b)

Fig. 2: Normalised quality scores distributions across the datasets; a) VGGFace2_train dataset (identical to VGGFace2_test); b) FRGC_V2 test strict dataset.

Operating Characteristic (ROC) curves of the trained QualFace models (with $m_0 = 0.4$ and $m_1 = 0.1$) are represented in Fig. 3 as well as ArcFace and Softmax models for comparison.



(a) (b)

Fig. 3: ROC curves; a) Wild Benchmark; b) Strict ICAO compliance Benchmark.

From the ROC curves one can see that most of the QualFace models have better operation curves than ArcFace and Softmax. For the strict benchmark, the illumination score QualFace model exhibits the best results while for the wild benchmark the blur scores is the best performing. We estimate the performance by several metrics: False Non-Match Rate at False Match Rate (FNMR@FMR) and Area Under Curve (AUC) of ROC (see Table 1).

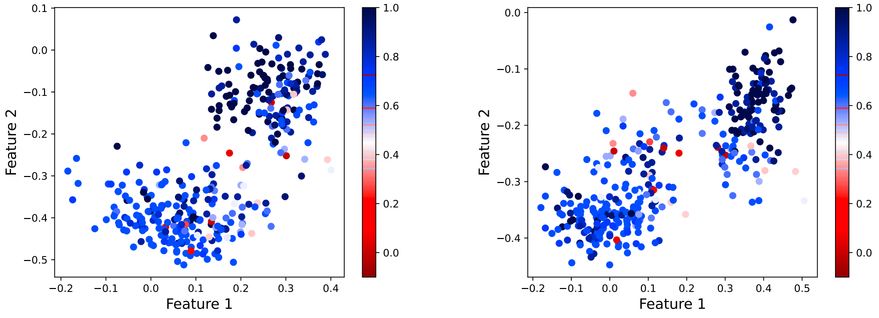
From the results obtained, we conclude that QualFace significantly enhances the biometric verification performance in ICAO compliant face images when compared to a simple margin based loss function like ArcFace. This statement can be verified for most of the models trained in both configurations, however the models with $m_1 = 0.1$ clearly show superior results. Considering wild benchmarks, our approach performs on par with the baseline models. However, most of QualFace experiment results still slightly outperform

Tab. 1: FNMR@FMR thresholds and AUC scores for two benchmarks.

Method		Wild			Strict			
		1e-2	1e-3	AUC	1e-3	1e-4	1e-5	AUC
Softmax		0.44502	0.79633	0.944118	0.69017	0.93655	0.98027	0,995333
ArcFace		0.28680	0.52938	0.951181	0.02486	0.10205	0.19507	0,999871
QualFace (m_0=0.4, m_1=0.1)	Blur	**0.24600**	**0.46806**	**0.959089**	**0.00793**	0.05453	0.13429	**0.999957**
	BRISQUE	0.26185	0.49934	0.954925	0.02556	0.08950	0.18444	0.999878
	FaceQNet	0.26383	0.50290	0.956515	0.01874	0.08284	0.15398	0.999910
	Illumination	0.26037	0.50076	0.956797	0.01066	**0.04835**	**0.10878**	0.999936
	Pose	0.27177	0.52186	0.954926	0.01805	0.08011	0.14550	0.999917
QualFace (m_0=0.4, m_1=0.2)	Blur	0.27460	0.52942	0.956314	0.04183	0.13423	0.19618	0.999813
	BRISQUE	0.28253	0.57363	0.953276	0.04329	0.21139	0.30880	0.999772
	FaceQNet	0.26524	0.54649	0.956252	0.03185	0.05963	0.19958	0.999944
	Illumination	0.29351	0.56006	0.952535	0.04046	0.14946	0.21644	0.999792
	Pose	0.28781	0.51604	0.951155	0.01146	0.11058	0.19958	0.999838

ArcFace. We conclude that our method allows to regularise the training process in deeper manner (not just adapting to qualitative samples) but generally learns better (more qualitative/discriminative) face features. From that point of view, our approach inherently shares conceptual similarities with the curriculum learning strategy.

Feature distribution. To better understand the QualFace impact to the learning process we analysed the real feature distribution for several particular identities in the benchmark datasets. To constrain the analysis in the 2D case we extract two principal components from the 512 dimensional embeddings with PCA (Principal Component Analysis). We represent the resulting feature distributions for two identities from the FRGC_V2 Dataset Fig. 4.



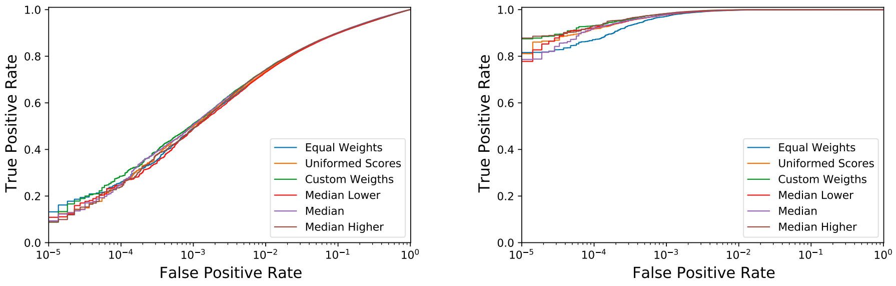
(a) (b)

Fig. 4: Features distribution of 2 different identities (04430 and 02463) from the FRGC_V2 dataset with Illumination scores represented in colour; a) ArcFace Model; b) Illumination Score QualFace Model with $m_0 = 0.4$ and $m_1 = 0.1$.

Basing on our results we make two observations. First, the separation between identities, which is commonly seen in margin based methods can be confirmed both in ArcFace and QualFace cases. Second, while ArcFace does not take into account image quality, Qual-

Face pulls high quality samples towards the class centre and compacts their distribution, while the low quality samples are pushed away as theoretically hypothesised in Fig. 1b.

Combined scores experiments After the experiments with sampling by a single score we intuitively investigated several scores averaging techniques. Namely, we utilised straight forward mean value, weighted mean and several median value implementations. The median implementations used three scores each. The *Median Lower* model averaged the three lower scores, the *Median* model - the three centre scores and the *Median Higher* averaged the three highest scores, for each image. We also made experiments with uniforming scores distributions in the range $[0,1]$ before averaging for equalising their impact. The ROC curves of the combined models are represented on Fig. 5



(a) (b)

Fig. 5: Combined model ROC curves; a) Wild Benchmark; b) Strict ICAO compliance Benchmark.

In our experiments the models with *Custom Weights* and *Median Higher* weights averaging demonstrated the best performance in benchmarks. This can also be confirmed from the AUC and FNMR@FMR metrics, which are represented in table 2.

Tab. 2: FNMR@FMR thresholds and AUC scores for two benchmarks using all five scores QualFace models with m_0=0.4, m_1=0.1.

Models	Wild			Strict			
	1e-2	1e-3	AUC	1e-3	1e-4	1e-5	AUC
Equal Weights	0.26495	0.50912	0.956074	0.02869	0.12879	0.18398	0.999850
Uniformed Scores	0.26393	0.50244	0.957280	0.02171	0.08221	0.18881	0.999897
Custom Weights	**0.25735**	**0.48964**	0.956706	0.01834	**0.06875**	0.12521	0.999907
Median Lower	0.26914	0.51016	0.955681	0.02195	0.07807	0.22204	0.999891
Median	0.25829	0.49400	**0.958679**	0.02087	0.07853	0.21371	0.999905
Median Higher	0.25877	0.51080	0.957604	**0.01629**	0.07027	**0.12184**	**0.999929**

We made several observations regarding the usage of combined scores. Scores uniforming indeed allowed better regularise the training and achieve better performance results. Scores weighing demonstrated its importance and the best performance was achieved when the weights were selected according to the results of single score models (Blur - 0.3, BRISQUE - 0.1, FaceQnet - 0.15, Illumination - 0.3, Pose - 0.15 in our experiments).

In the list of models with median averaging, *Median Higher* case gave the most promising result, which means that the QualFace sampling strategy should be good score biased. In other words, it is better to treat a sample by its best scores rather than consider it as a bad sample even if it has some lower scores.

The use of combined scores did not demonstrate the superiority in any particular benchmark. However, it allowed to achieve more regular results across the two utilised benchmarks (strict and wild) making the face representation more universal in applications with unspecified scenario. This can be verified when comparing the *Custom Weight* and *Median Higher* model with the single score blur and illumination models.

We conclude that sampling of face images with single generic illumination and blur quality metrics allow to learn better face representation when applying the QualFace technique. Particularly, illumination quality is better suitable in application to the document security scenario, while blur score better shifts the performance towards wild face recognition scenario.

5 Conclusions

In this work we proposed a novel approach of adapting deep learning face recognition methods for document security applications. We introduced a sophisticated sample mining strategy that regularises the training process by careful emphasising the impact of samples which are better suitable for document security. The method allows to effectively train face recognition networks on big wild datasets and at the same time reduce the effect of ”wildness” of these datasets. The extensive experiments with the selected marginal loss function (ArcFace) proved the superiority of adapted models against the default ones in tests with ID compliant images. The introduced strategy can also be applied to other loss functions. Our future work will focus on the study of additional image quality metrics more specific to concrete ICAO requirements. Experiments with different loss functions and finding better normalisation for the quality scores are also part of our future work plan.

6 Acknowledgements

The authors would like to thank the Portuguese Mint and Official Printing Office (INCM) and the Institute of Systems and Robotics - University of Coimbra for the support of the project Facing. This work has been supported by Fundação para a Ciência e a Tecnologia (FCT) under the project UIDB/00048/2020.

References

[Ba16] Bansal, Raghav; Raj, Gaurav; Choudhury, Tanupriya: Blur image detection using Laplacian operator and Open-CV. In: 2016 International Conference System Modeling Advancement in Research Trends (SMART). pp. 63–67, 2016.

[Ca18] Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; Zisserman, A.: VGGFace2: A Dataset for Recognising Faces across Pose and Age. In: 2018 13th IEEE International Conference on AFGR. pp. 67–74, 2018.

[De19a] Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S.: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: 2019 IEEE Conference on Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4685–4694, 2019.

[De19b] Deng, J.; Guo, J.; Zhou, Y.; Yu, J.; Kotsia, I.; Zafeiriou, S.: RetinaFace: Single-stage Dense Face Localisation in the Wild. CoRR, abs/1905.00641, 2019.

[Eu16] General Data Protection Regulation. Official Journal of the European Union.

[Eu18] European Enrolment Guide for Biometric ID Documents. CEN/TC 224.

[Gu16] Guo, Y.; Zhang, L.; Hu, Y.; He, X.; Gao, J.: MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In: 2016 ECCV. volume 9907, pp. 87–102, 10 2016.

[He16] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian: Identity Mappings in Deep Residual Networks. CoRR, abs/1603.05027, 2016.

[He19] Hernandez-Ortega, J.; Galbally, J.; Fierrez, J.; Haraksim, R.; Beslay, L.: FaceQnet: Quality assessment for face recognition based on deep learning. arXiv, 2019.

[Hu20] Huang, Y.; Wang, Y.; Tai, Y.; Liu, X.; Shen, P.; Li, S.; Li, J.; Huang, Feiyue: CurricularFace: Adaptive Curriculum Learning Loss for Deep Face Recognition. In: 2020 IEEE Conference on Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5900–5909, 06 2020.

[IS18] Portrait Quality - Reference Facial Images For MRTD. https://www.icao.int/Security/FAL/TRIP/Documents/TR - Portrait Quality v1.0.pdf.

[IS19] Information technology — Extensible biometric data interchange formats — Part 5: Face image data. ISO/IEC 39794-5:2019.

[Li17] Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; Song, L.: SphereFace: Deep Hypersphere Embedding for Face Recognition. In: IEEE Conference on Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6738–6746, 2017.

[Me20] Medvedev, I.; Gonçalves, N.; Cruz, L.: Biometric System for Mobile Validation of ID And Travel Documents. In: 2020 International Conference of the BIOSIG. pp. 1–5, 2020.

[Me21] Meng, Qiang; Zhao, Shichao; Huang, Zhida; Zhou, Feng: MagFace: A Universal Representation for Face Recognition and Quality Assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14225–14234, June 2021.

[Mi12] Mittal, A.; Moorthy, A. K.; Bovik, A. C.: No-Reference Image Quality Assessment in the Spatial Domain. IEEE Transactions on Image Processing, 21(12):4695–4708, 2012.

[Ou21] Ou, Fu-Zhao; Chen, Xingyu; Zhang, Ruixin; Huang, Yuge; Li, Shaoxin; Li, Jilin; Li, Yong; Cao, Liujuan; Wang, Yuan-Gen: SDD-FIQA: Unsupervised Face Image Quality Assessment With Similarity Distribution Distance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7670–7679, June 2021.

[Ph05] Phillips, P.J.; Flynn, P.J.; Scruggs, T.; Bowyer, K.W.; Chang, Jin; Hoffman, K.; Marques, J.; Min, Jaesik; Worek, W.: Overview of the face recognition grand challenge. In: 2005 IEEE Computer Society Conference on Conference on Computer Vision and Pattern Recognition (CVPR). volume 1, pp. 947–954, 2005.

[Ru18] Ruiz, N.; Chong, E.; Rehg, J.: Fine-Grained Head Pose Estimation Without Keypoints. In: The IEEE Conference on Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. 06 2018.

[Sc15] Schroff, F.; Kalenichenko, D.; Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: 2015 IEEE Conference Conference on Computer Vision and Pattern Recognition (CVPR). pp. 815–823, 2015.

[Sc20] Schlett, Torsten; Rathgeb, Christian; Henniger, Olaf; Galbally, Javier; Fierrez, Julian; Busch, Christoph: , Face Image Quality Assessment: A Literature Survey, 09 2020.

[Sh19] Shi, Yichun; Jain, Anil K.; Kalka, N.: Probabilistic Face Embeddings. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6901–6910, 2019.

[SJ18] Shi, Yichun; Jain, Anil K.: DocFace: Matching ID Document Photos to Selfies*. In: 2018 IEEE 9th International Conference on BTAS. pp. 1–8, 2018.

[SJ19] Shi, Yichun; Jain, Anil K.: DocFace+: ID Document to Selfie Matching. IEEE Transactions on Biometrics, Behavior, and Identity Science, 1(1):56–67, 2019.

[Te20] Terhörst, Philipp; Kolf, Jan Niklas; Damer, Naser; Kirchbuchner, Florian; Kuijper, Arjan: SER-FIQ: Unsupervised Estimation of Face Image Quality Based on Stochastic Embedding Robustness. 2020.

[Wa18] Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; Liu, W.: CosFace: Large Margin Cosine Loss for Deep Face Recognition. In: 2018 IEEE/CVF Conference on Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5265–5274, 2018.

[Wa20] Wang, X.; Zhang, S.; Wang, S.; Fu, T.; Shi, H.; Mei, T.: Mis-Classified Vector Guided Softmax Loss for Face Recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 34:12241–12248, 04 2020.

[Ze20] Zeng, D.; Shi, H.; Du, H.; Wang, J.; Lei, Z.; Mei, T.: NPCFace: A Negative-Positive Cooperation Supervision for Training Large-scale Face Recognition. CoRR, abs/2007.10172, 2020.

[Zh17] Zhang, L.; Zhang, L.; Li, L.: Illumination Quality Assessment for Face Images: A Benchmark and a Convolutional Neural Networks Based Model. Lecture Notes in Computer Science, 10636 LNCS:583–593, 2017.

Fingermark Quality Assessment: An Open-Source Toolbox

Tim Oblak[1 2], Rudolf Haraksim[2], Laurent Beslay[2], Peter Peer[1]

Abstract: Fingermark quality assessment is an important step in a forensic fingerprint identification process. Often done in the scope of criminal investigation, it is performed by trained fingerprint examiners whose quality assessment can be rather subjective. The goal of this work is to develop an automated fingermark quality assessment tool, which would assist the fingermark examiners in their work. In this paper, we present a fast, open-source, and well documented fingermark quality assessment toolbox, which contains more than 20 algorithms for feature extraction, segmentation, and enhancement of fingermark images. We demonstrate the utility of the toolbox by assembling a feature vector and training various baseline machine learning models, capable of predicting the quality of fingermark images with high accuracy. The AFQA toolbox source code is publicly available online.

Keywords: fingermark, forensic, biometric, quality, evaluation, open-source.

1 Introduction

Fingerprint comparison is one of the oldest types of biometric identification, used extensively in automated fingerprint identification systems (AFIS) for forensic investigations. A particular challenge in this area present fingermarks (in the USA latent fingerprints), partial friction ridge skin impressions typically from fingertips, left in an unconstrained environment, e.g., a crime scene. Fingermarks are lifted from various surfaces and their quality is influenced by several external factors, which often result in distorted or only partially visible impressions. The scientific method of comparing fingermarks, ACE-V [As99], is well established and followed by trained forensic examiners. An important first step in this process is determining the quality (or value for identification) of a fingermark. This value establishes how the fingermark will be processed and indicates, whether its quality is sufficient for finding a mated fingerprint in an open-set biometric dataset. Due to the involvement of human experts, quality assessment is subjective, prone to bias and can potentially result in evidence mishandling. An Automated Fingermark Quality Assessment (AFQA) toolbox assists forensic fingermark examiners by proposing a probabilistic quality value, helps to reduce their subjectivity and improves their efficiency.

Based on the initial review, a novel automated, reliable, and open-source AFQA implementation would greatly benefit the forensic community. Within the existing work, accessibility and reproducibility are currently the major limiting factors: (i) Commercially available solutions are widespread but require the user to pay for the product. In rare cases,

[0] This work was funded in part by the JRC's GH20 Collaborative Doctoral Partnership programme.

[1] University of Ljubljana, Ljubljana, Slovenia, {tim.oblak,peter.peer}@fri.uni-lj.si

[2] European Commission, Joint Research Centre, Ispra, Italy {rudolf.haraksim,laurent.beslay}@ec.europa.eu

companies offer limited access to their software, but only for research purposes and such arrangements do not include access to the source-code. (ii) Some solutions are designed specifically for local law enforcement or larger intelligence agencies and are not available publicly to minimize intentional tampering or spoofing with the goal of concealing identity. (iii) Existing published work is rarely accompanied with open-source code, data, or data annotations. This is particularly noticeable in novel approaches, which are largely data-driven machine learning (ML) solutions [Yo13, Ch18]. Popular fingermark datasets being recently discontinued [GM00], it makes many methods difficult or in some cases impossible to reproduce. (iv) Some fingerprint-related algorithms, published in the open-source format [TWW04, Ta21], are commonly implemented in low-level programming languages and intended for integration into end products. While this is beneficial in a production environment, research productivity is lowered. Furthermore, the majority of available open-source programs are focused on fingerprints and not on fingermark processing.

To boost research in this field and improve accessibility of related methods, we propose an open-source AFQA toolbox for fingermark analysis and quality assessment. This includes a centralized and ready to use collection of most frequently used computer vision techniques for feature extraction, segmentation, and enhancement of friction ridge images, written in a high-level programming language (*Python*). We use the toolbox to extract features from fingermarks and construct a fixed-length feature vector. Using popular ML algorithms and with the help of existing quality assessment methods to annotate the data, we train predictive models in order to assess the quality of fingermarks. We demonstrate the accuracy of our models on a publicly available fingermark dataset and provide the motivation in favour of using the AFQA toolbox to develop new quality assessment methods.

In Section 2, we describe the related work within friction ridge quality assessment. In Section 3, we present and describe the fingermark toolbox and demonstrate its usage by proposing a quality assessment pipeline in Section 4. Finally, we compare our approaches, discuss the results in Section 5, and conclude with final thoughts in Section 6.

2 Related work

Fingerprint quality. Initially, quality assessment of fingerprints was based on various local image quality indicators, such as local frequency and clarity, deviation of Gabor features, and other pixel intensity or gradient methods [Al07]. The National Institute of Standards and Technology (NIST) fist aimed to standardize fingerprint quality assessment and developed the NIST Fingerprint Image Quality (NFIQ) algorithm [TWW04]. This was the first attempt to define a fingerprint quality measure as being indicative of the probability to find a mated reference fingerprint in a database for a questioned fingerprint. Thanks to the advances in fingerprint recognition technology, NIST initiated work on an upgrade, the NFIQ 2 [Ta21]. The authors implemented 155 feature extraction algorithms, eliminated features with low predictive power and trained an improved random forest classifier to assign quality values. They observed improved predictive capabilities and a faster execution time in comparison with the original NFIQ. The open-source method was gradually improved and culminated with a recent (2021) release of version NFIQ 2.1.

Fingermark quality. After the 2004 Madrid bombings, the FBI misidentified a prime suspect based on a single fingermark. In response, they investigated the decision-making process of their forensic examiners. The findings were used to develop the Universal Latent Workstation (ULW), a toolbox assisting forensic examiners with fingermark analysis, which also contains a fingermark quality metric (LQmetric). The LQmetric is not publicly available and is only partially published [KBH20]. Based on operational feedback from the FBI Laboratory, it performs well on good- and bad-quality fingermarks, but not on borderline cases [HGB19]. To define a quality measure specifically for fingermarks, Yoon *et al.* developed the Latent Fingerprint Image Quality (LFIQ) [Yo13]. The algorithm uses a combination of local clarity indicators and minutiae data to determine quality. The method relies on manual minutiae extraction for best performance. Sankaran *et al.* [SVS13] proposed a heuristic to determine the local clarity and quality of fingermarks, however, they did not consider minutiae data as a qualitative indicator. Chugh *et al.* [Ch18] gathered expert crowd-sourced data and cross referenced it to develop a predictive model for quality assessment. Ezeobiejesi *et al.* [EB18] were the first to utilize deep learning in the context of quality assessment, however, in their approach, the final quality measure is computed trivially only by counting the number of patches of certain quality. In general, published fingermark quality assessment methods use a combination of commonly used algorithms, heuristics, and machine learning practices for processing friction ridge impressions, which we combined into a versatile open-source collection.

3 Automatic Fingermark Quality Assessment Toolbox

In this section we describe the contents of the AFQA Toolbox. The algorithms included are presented graphically in Figure 1. The majority of algorithms (green labels) are implemented in Python and make use of common Python libraries, such as *NumPy*, *SciPy*, *scikit-learn*, *scikit-image* and *OpenCV*. For other methods, which benefit more from their original low-level implementation, we offer a Python wrapper function (red labels), which enables seamless integration of complex methods into the toolbox. The AFQA toolbox is provided ”as-is” under the MIT open-source licence and can be accessed online: `https://github.com/timoblak/OpenAFQA`.

Preprocessing. Determining the region of interest is the first step when processing a friction ridge image. The foreground, containing friction ridge information, is separated from the often noisy background. This task is particularly challenging in cases, in which fingermarks are captured directly on the surface, on which they were deposited. The friction ridge is typically described using the features distributed at different levels. The toolbox includes a heuristic algorithm, which determines the fingermark foreground based on the analysis of pixel values in a local area. The friction ridge impression can often include distorted regions, which are recoverable with the use of enhancement algorithms. Such algorithms exploit the deterministic structure of friction ridge to correct local damage and improve ridge clarity. The initial set of tools includes a Difference of Gaussian-based filter [MH80], which enhances local contrast, and Hong’s method [HWJ98], which uses oriented Gabor filters to enhance the structure.

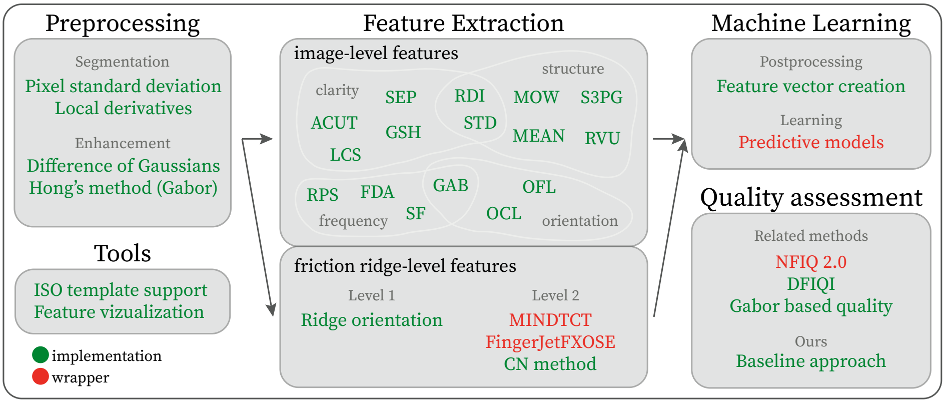


Fig. 1: **AFQA toolbox.** The toolbox (Python) contains pre-processing and feature extraction modules and includes other useful tools for friction ridge impression analysis.

Feature extraction. Friction ridge skin has inherent features, categorized into 3 levels of detail: (i) Level 1 features represent the flow of the friction ridge and its class, which is based on abrupt changes in friction ridge orientation. These features are usually detectable even in low quality or low resolution images. (ii) Level 2 details describe the salient points (endings and bifurcations) of individual ridges, called minutiae points. Many automatic fingerprint identification systems (AFIS) are heavily dependent on level 2 details. Similarly, they are used extensively in mark-to-print comparisons by forensic experts. (iii) Level 3 details are most visible on high resolution images and describe friction ridge at the highest level, e.g., skin pores, shapes of ridges, etc. Such features are highly deterministic, but can be hard to detect and compare. Due to the widespread use of level 2 features, we include in our toolbox several methods for minutiae extraction. The first is MINDTCT, an algorithm included in the NIST Biometric Image Software (NBIS)[3] distribution and used by NFIQ as well as in FBI's ULW. The second is a robust open-source minutiae extractor called FingerJetFXOSE[4], also used in the NFIQ 2. For these methods, we provide a wrapper function, which calls either a compiled binary or a library, implemented in a low-level language. Additionally, we include a simple and customizable Python implementation of the Crossing Number algorithm [Ka08].

Another class of features originates from the more general field of image quality. Through time, these were adopted specifically for analysing friction ridge impressions and are used commonly within the related literature [LJY02, CJY04, OŠB16, Sw21, Ta21]. These features can be used to capture the following friction ridge properties: (i) *Frequency* of ridges on human fingers has a known value of around 2.1 and 2.4 ridges/mm for males and females, respectively. A deviation from this value indicates the absence of friction ridge or the presence of local deformations. To calculate frequency, we use Gabor filtering of 2D Fourier transform. (ii) *Clarity* describes the separability between pixel values of nearby

[3] https://www.nist.gov/services-resources/software/nist-biometric-image-software-nbis
[4] http://github.com/FingerJetFXOSE/FingerJetFXOSE

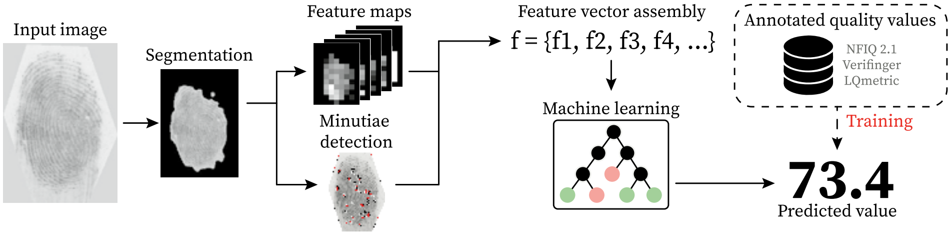


Fig. 2: **Predictive pipeline.** Predictive pipeline of AFQA toolbox, consisting of pre-processing, feature extraction, and machine learning algorithms.

ridges and valleys of the impression. Higher clarity ridges in close proximity to detected minutiae points indicate higher probability of their existence. We use image derivatives or other pixel intensity methods to calculate local clarity. (iii) *Orientation* of ridges should not change drastically in a local neighborhood of image blocks. A large difference in orientation could indicate a presence of local distortions or singular points. We calculate orientation with image gradients or by analyzing the frequency domain. (iv) *Structure* is another important factor. Ideally, the width of ridges should be consistent and comparable to the width of valleys. The ridge structure is extracted by using various pixel intensity methods.

Feature vector assembly. Due to the unconstrained nature of fingermark imposition, their image can vary drastically. To represent all fingermarks with a unified description, our toolbox enables automatic construction of a fixed-length feature vector from minutiae data and feature maps of different sizes.

4 Baseline quality assessment

In this section, we propose a baseline quality assessment method, which is derived from the AFQA toolbox feature vectors. The pipeline is visualized in Fig. 2.

For an input fingermark impression, we use image equalization and a heuristic analysis of local pixel values to determine the friction ridge area. Then, in a block-wise manner, the 15 feature extraction algorithms calculate local features, which result in 15 feature maps. FingerJetFXOSE algorithm is used to detect minutiae points within the segmented region. Each of the 15 feature maps are then compressed into a vector of 12 values. The first two values represent the mean and standard deviation of the entire feature map. The remaining 10 values represent a histogram with 10 bins, where each bin amounts to the number of values within a specific range. The edges of histogram bins are computed from an average feature map of the entire training dataset. Minutiae data are described in a similar fashion with a vector of length 12. First value represents the mean minutia quality, the second value is the number of all detected minutiae, and the remaining 10 values again represent a histogram of qualities of detected minutiae. The minutiae quality is calculated for each

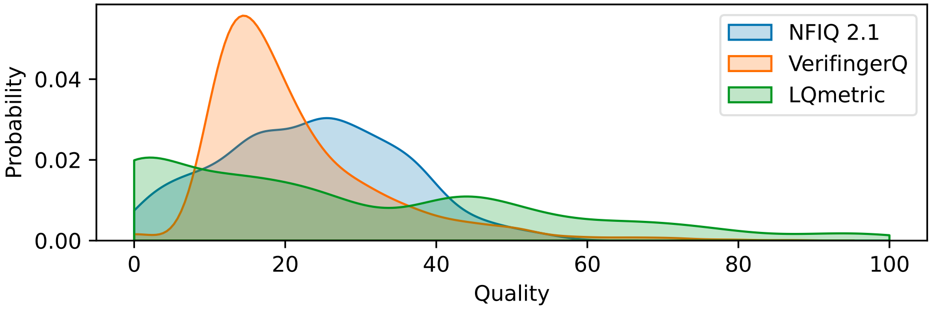


Fig. 3: **Distribution of quality scores.** Shown here are the annotated labels of the test set (NIST SD301 dataset) by three different quality assessment methods. The majority of scores is clustered towards the low quality side of the spectrum.

detected minutiae using FingerJetFXOSE. The aggregated fingermark feature vector is 192 features long.

Different ML techniques are used to train three predictive regression models for fingermark quality assessment, all of which are implemented in the *scikit-klearn* Python library. The first is a fully connected neural network, which has a single hidden layer with 100 neurons using ReLu activation and a single linear output neuron. The second is a random forest regressor, which uses an ensemble of 750 decision trees with a maximum depth of 110. The third method is a support vector regressor (SVR), for which default parameters are used. For all methods, the hyper-parameters were determined by using random search across a wide range of available values and validated on a reserved part of the training data. The task for each quality estimator is then to minimize the square difference between predicted quality values and the annotated quality values. Since dataset annotation with trained forensic examiners was not available at this point of time, we employ existing quality assessment algorithms to provide the necessary baseline quality labels for fingermark images. We use the following methods to annotate the public datasets:

- **NFIQ 2.1** [Ta21] is an open-source software, originally trained and intended for predicting quality values of flat fingerprints.
- The fingerprint quality assessment method included in the Verifinger SDK [Ne98], sold by a commercial vendor Neurotechnology. We refer to it as the **VerifingerQ** metric in this paper.
- **LQmetric** [KBH20] is fingermark quality measure, used within the FBI's Universal Latent Workstation software.

All three quality assessment methods are used to generate three sets of ground-truth labels, which are then used to train three different models for each of the ML approaches.

Tab. 1: **Evaluation results on the test set.** Random forest achieved best results in terms of performance metrics. Our feature vector is able to capture NFIQ 2.1 and VerifingerQ properties but struggles with LQmetric scores.

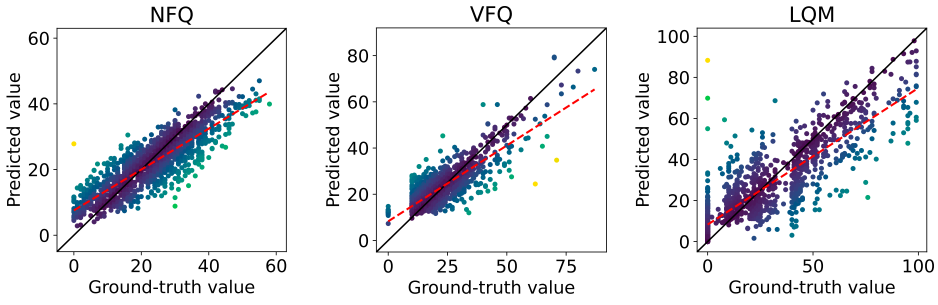
	Neural network			**Random Forest**			**SVR**		
Model	MSE	MAE	r^2	MSE	MAE	r^2	MSE	MAE	r^2
NFQ	71.63	6.09	0.50	41.62	5.16	0.71	49.84	5.51	0.65
VFQ	47.51	5.07	0.64	41.50	4.84	0.68	56.81	5.49	0.57
LQM	183.00	9.35	0.72	181.50	9.66	0.72	281.15	12.65	0.56

5 Evaluation

Experimental setup. We use NIST SD 302 [Fi18a] and SD 301 [Fi18b] fingermark image datasets. Both datasets contain fingermarks, lifted from various surfaces by trained forensic experts in a simulated environment. To better capture the properties of the whole spectrum of friction ridge quality, we use fingermark as well as fingerprint images to train the models. We split the data into a training set of 10,000 fingermarks (SD 302) and 2,000 fingerprints (SD 302), and a test dataset of 1,200 fingermarks (SD 301). The distribution of scores attributed to fingermarks by the three quality values is shown in Fig. 3. We evaluate the performance of the three trained ML models with common regression metrics. We monitor the Mean Squared Error (MSE), as well as the Mean Absolute Error (MAE) to reduce the effect of the predicted outliers. To assess the correlation between the predicted and ground-truth quality values, we also calculate the coefficient of determination r^2.

Results and comparison. By using the annotations from existing quality assessment methods NFIQ 2.1, VerifingerQ, and LQmetric, we produce three models, which we label NFQ, VFQ, and LQM, respectively. The results are shown in Table 1. Based on the metrics alone, the random forest regressor was able to approximate fingermark quality the closest for all annotated sets of scores. Neural network and SVR performed slightly worse on average. The neural network model was able to estimate VerifingerQ and LQmetric scores better while, in contrast, the SVR achieved better results on NFIQ 2.1 scores.

There are notable differences in MSE and MAE metrics between models trained on different sets of scores, particularly LQM stands out out of the three. This, however, is due to the different distribution spread of the annotated scores. A more comparable metric here is r^2, which shows the amount of variance, that can be explained by the learned model. Despite higher MAE and MSE metrics, the r^2 achieved with LQmetric scores is the highest, which means the correlation of predicted scores with original scores is higher. The high r^2 score of around 0.7 suggests that the assembled feature vector is able to capture the properties of all quality assessment algorithms, which were used to annotate the data. Overall, the VFQ model, trained using random forest, achieves the smallest MSE and MAE values. While the same metrics are slightly higher for NFQ and LQM random forest models, the coefficient of determination r^2 is higher, which indicates predicted NFIQ 2.1 and LQmetric scores are better correlated with the ground-truth values in comparison with VerifingerQ scores.

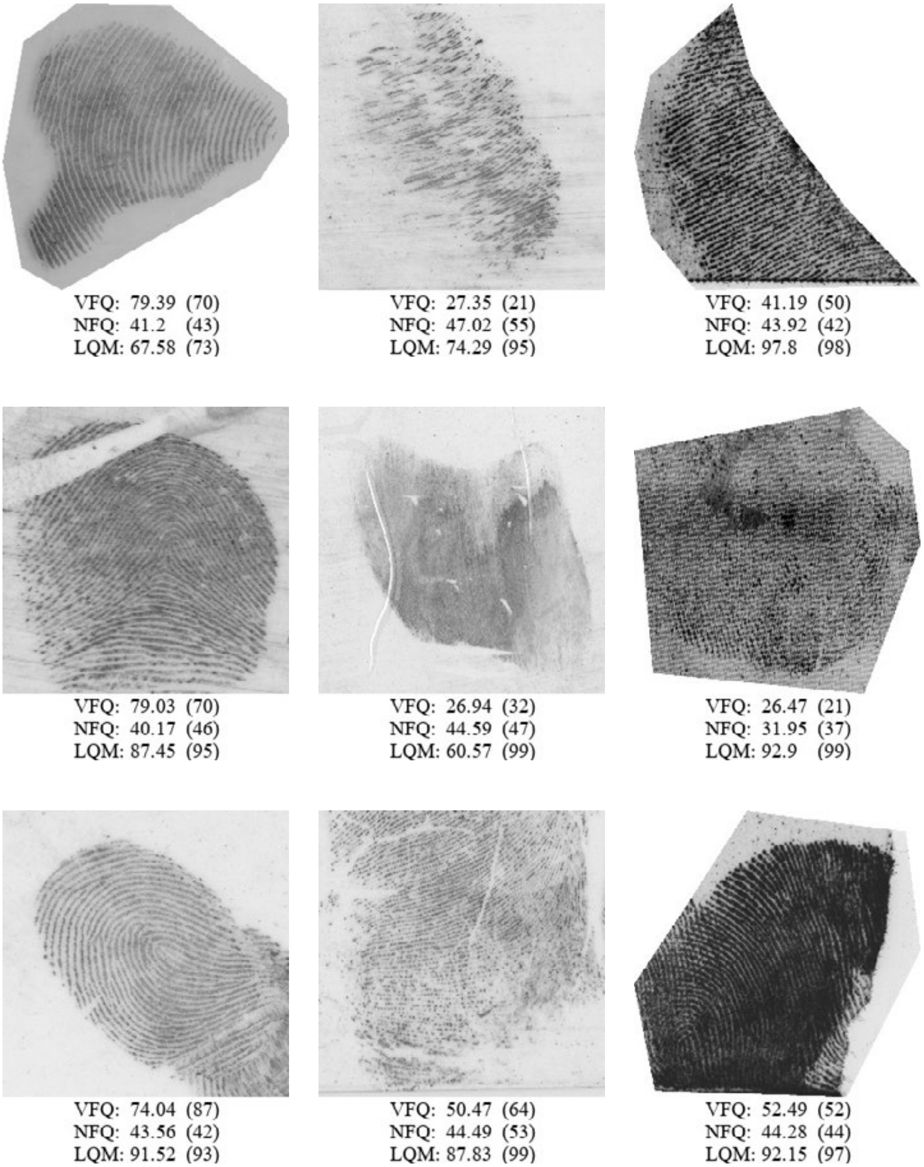


(a) (b) (c)

Fig. 4: **Random forest predictions.** The models are trained using (a) NFIQ 2.1, (b) VerifingerQ, and (c) LQmetric quality values. All models appear to slightly overestimate the lower quality fingermark values and underestimate higher quality fingermarks, as indicated by the red regression line. Lighter color indicates larger error between the prediction and ground truth values.

In Fig. 4, we show scatter plots of predicted and ground-truth quality scores for the best performing random forest regressor. The NFQ model forms the most uniform distribution with only a few outliers and shows a clear trend, following the ideal diagonal line. The VFQ model shows a similar picture with small differences. First, there is a gap with no values in the bottom left corner of the graph, which exists because 10 was the lowest score that VerifingerQ attributed to fingermark images. The exception are a few examples where the method failed and a score of 0 was assigned instead. In contrast with the NFQ, the VFQ attributes a wider range of quality values. Finally, the predictions of the LQM model are most scattered across the spectrum but still follow a clear trend. The reason again is due to the larger variance in initial annotations. The red regression line is displayed for all models and shows that all models tend to slightly overestimate bad quality fingermarks and underestimate good quality fingermarks. Around the area, where the regression line crosses the ideal diagonal line, the models are most accurate in their predictions. The exact location of crossing is at a quality value of 20.0 for the NFQ model, 24.1 for VFQ model and 24.7 for the LQM model. These values are consistent with the distributions annotated quality values, where the majority of fingermark examples are considered to be of lower quality, as shown in Fig. 3.

These experiments show how the feature vector, constructed using the AFQA toolbox, can capture the properties of NFIQ 2.1, VerifingerQ and LQmetric quality assessment methods. Since our toolbox implements the majority of NFIQ 2.1 features, the compatibility between NFIQ 2.1 and our trained model was expected. We do not know how Verifinger calculates their quality values, but the AFQA toolbox features are able to represent the properties of their qualtiy assessment method well. Finally, the LQmetric was designed specifically for assessing the quality of fingermarks and like the remaining 2 methods, our feature vector can capture its properties. This means that the AFQA toolbox features are sufficient for representation of fingerprints as well as fingermarks. These models will serve as as a baseline for the future development of a new, independent, and open-source fingermark quality assessment methods.



(a) VFQ top 3 (b) NFQ top 3 (c) LQM top 3

Fig. 5: **Qualitative comparison between random forest models.** We demonstrate model capabilities based on actual examples from the SD 301 dataset. Ordered from top to bottom, we display fingermark images, which were given highest quality scores by (a) the VFQ model, (b) NFQ model, and (c) LQM model. For each image, we provide scores for all three models in format "MODEL: predicted_value (true_value)". We can observe that the most consistent is the VFQ model, which attributes high value only when clear ridge structure is present.

Qualitative evaluation. For the best performing random forest model, we visualize some of the examples from the test dataset together with their respective predicted and ground-truth quality values. This is shown in Fig. 5. For each model, we show the top three fingermarks based on their predicted quality value.

We begin with the (a) column examples, which the VFQ model considers to be of best quality. All of the examples contain a clear ridge structure, which could be easily recovered with various enhancement methods. The width of ridges and valleys is uniform and the area of the impression is relatively large. In the middle column (b) are top examples based on the NFQ model. Here we see examples with a large amount of high frequency ridge-like formations, which are in most cases not actually friction ridges, but rather specks or smudges. The NFIQ 2.1 method is in essence not intended to be used with fingermarks, which might explain why the trained model cannot differentiate between real ridges and impression distortions. In the right column are top quality examples based on the VFQ model. Here we observe a stronger bias toward fingermarks with a darker average color. It appears that the middle fingermarks contains no recoverable friction ridge, but the LQ-metric falsely detects the high frequency background patterns as friction ridge and consequently assigns to it a high quality value. Another comparison can be made between VFQ and LQM. As discussed, VFQ model assigns good quality to fingermarks with high clarity in the left column (a), but LQM scores for the same marks are more proportional to the area of the visible friction ridge, giving the smaller fingermark a smaller estimate, despite good quality of ridges.

The predicted values for these examples are a relatively close approximation of the ground-truth scores. However, our intent here was not to evaluate the suitability of individual methods for the task of fingermark quality assessment. As apparent from the qualitative results, each method assigns quality based on different friction ridge features, which can result in high differences between scores for a singe fingermark. To leverage the collective power of multiple quality assessment methods, a fusion of predicted scores could improve the overall consistency of the quality assessment process and produce even more objective quality values for fingermarks.

6 Conclusion

In this paper we proposed the AFQA toolbox for fingermark analysis, which contains a large collection of established algorithms, intended for friction ridge feature extraction, as well as various pre-processing for segmentation and enhancement. By making the toolbox open-source, we want to improve the accessibility of existing methods and the reproducibility of future work for the biometric and forensic communities.

We demonstrated the usability of the toolbox by extracting friction ridge features and creating a compact feature vector to represent individual fingermarks efficiently. We then trained three baseline fingermark quality assessment models, based on annotations from existing methods, and evaluated them on a public dataset. The results indicate a high compatibility between the proposed feature vector and the inner workings of existing friction ridge quality assessment methods.

In the future work we plan to expand the toolbox with additional algorithms and better define, what friction ridge properties influence quality the most. We also plan to make use of more contemporary ML methods, such as deep learning, with the objective to further improve the fingermark image quality assessment.

References

[Al07] Alonso-Fernandez, Fernando; Fierrez, Julian; Ortega-Garcia, Javier; Gonzalez-Rodriguez, Joaquin; Fronthaler, Hartwig; Kollreider, Klaus; Bigun, Josef: A comparative study of fingerprint image-quality estimation methods. IEEE Transactions on Information Forensics and Security, 2(4):734–743, 2007.

[As99] Ashbaugh, David R.: Quantitative-qualitative friction ridge analysis : an introduction to basic and advanced ridgeology. CRC press, 1999.

[Ch18] Chugh, Tarang; Cao, Kai; Zhou, Jiayu; Tabassi, Elham; Jain, Anil K.: Latent Fingerprint Value Prediction: Crowd-Based Learning. IEEE Transactions on Information Forensics and Security, 13(1):20–34, 2018.

[CJY04] Chen, Tai Pang; Jiang, Xudong; Yau, Wei-Yun: Fingerprint image quality analysis. In: International Conference on Image Processing. volume 2, pp. 1253–1256, 2004.

[EB18] Ezeobiejesi, Jude; Bhanu, Bir: Latent fingerprint image quality assessment using deep learning. In: Conference on Computer Vision and Pattern Recognition Workshops. pp. 508–516, 2018.

[Fi18a] Fiumara, Gregory; Flanagan, Patricia; Grantham, John; Ko, Kenneth; Marshall, Karen; Schwarz, Matthew; Tabassi, Elham; Woodgate, Brian; Boehnen, Christopher: National Institute of Standards and Technology Special Database 302: Nail to Nail Fingerprint Challenge. Technical Note 2007, National Institute of Standards and Technology, 2018.

[Fi18b] Fiumara, Gregory; Flanagan, Patricia; Schwarz, Matthew; Tabassi, Elham; Boehnen, Christopher: National Institute of Standards and Technology Special Database 301: Nail to Nail Fingerprint Challenge Dry Run. Technical Note 2002, National Institute of Standards and Technology, 2018.

[GM00] Garris, Michael D; McCabe, R. Michael: NIST special database 27: Fingerprint minutiae from latent and matching tenprint images. US Department of Commerce, NIST, 2000.

[HGB19] Haraksim, R; Galbally, J; Beslay, L: Study on Fingermark and Palmmark Identification Technologies for their Implementation in the Schengen Information System. EUR 29755 EN, Publications Office of the European Union, 2019.

[HWJ98] Hong, Lin; Wan, Yifei; Jain, A.: Fingerprint image enhancement: algorithm and performance evaluation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20:777–789, 1998.

[Ka08] Kaur, Manvjeet; Singh, Mukhwinder; Girdhar, Akshay; Sandhu, Parvinder S: Fingerprint verification system using minutiae extraction technique. World Academy of Science, Engineering and Technology, 46:497–502, 2008.

[KBH20] Kalka, Nathan D; Beachler, Michael; Hicklin, R Austin: LQMetric: A Latent Fingerprint Quality Metric for Predicting AFIS Performance and Assessing the Value of Latent Fingerprints. Journal of Forensic Identification, 70:443–463, 2020.

[LJY02] Lim, Eyung; Jiang, Xudong; Yau, Weiyun: Fingerprint quality and validity analysis. In: International Conference on Image Processing. volume 1, pp. 469–472, 2002.

[MH80] Marr, David; Hildreth, Ellen: Theory of edge detection. Royal Society of London. Series B. Biological Sciences, 207(1167):187–217, 1980.

[Ne98] Neurotechnology: , VeriFinger, 1998. Available online: https://www.neurotechnology.com/verifinger.html. [Accessed 1.6.2021].

[OŠB16] Olsen, Martin Aastrup; Šmida, Vladimír; Busch, Christoph: Finger image quality assessment features – definitions and evaluation. IET Biometrics, 5(2):47–64, 2016.

[SVS13] Sankaran, Anush; Vatsa, Mayank; Singh, Richa: Automated clarity and quality assessment for latent fingerprints. International Conference on Biometrics: Theory, Applications and Systems, pp. 1–6, 2013.

[Sw21] Swofford, H.; Champod, C.; Koertner, A.; Eldridge, H.; Salyards, M.: A method for measuring the quality of friction skin impression evidence: Method development and validation. Forensic Science International, 320:1–13, 2021.

[Ta21] Tabassi, Elham; Olsen, Martin; Bausinger, Oliver; Busch, Christoph; Figlarz, Andrew; Fiumara, Gregory; Henniger, Olaf; Merkle, Johannes; Ruhland, Timo; Schiel, Christopher; Schwaiger, Michael: NIST Fingerprint Image Quality 2, NISTIR 8382. NIST, 2021.

[TWW04] Tabassi, Elham; Wilson, Charles; Watson, Craig I: Fingerprint Image Quality, NISTIR 7151. NIST, 2004.

[Yo13] Yoon, Soweon; Cao, Kai; Liu, Eryun; Jain, Anil K.: LFIQ: Latent fingerprint image quality. In: International Conference on Biometrics: Theory, Applications and Systems. IEEE, pp. 1–8, 2013.

BIOSIG 2021

Further Conference Contributions

The effect of face morphing on face image quality

Biying Fu [1], Noémie Spiller [1], Cong Chen [1], Naser Damer [1,2]

Abstract: Face morphing poses high security risk, which motivates the work on detection algorithms, as well as on anticipating novel morphing approaches. Using the statistical and perceptual image quality of morphed images in previous works has shown no clear correlation between the image quality and the realistic appearance. This motivated our study on the effect of face morphing on image quality and utility, we, therefore, applied eight general image quality metrics and four face-specific image utility metrics. We showed that MagFace (face utility metric) shows a clear difference between the bona fide and the morph images, regardless if they were digital or re-digitized. While most quality and utility metrics do not capture the artifacts introduced by the morphing process. Acknowledged that morphing artifacts are more apparent in certain areas of the face, we further investigated only these areas, for instance, tightly cropped face, nose, eyes, and mouth regions. We found that especially close to the eyes and the nose regions, using general image quality metrics as MEON and dipIQ can capture the image quality deterioration introduced by the morphing process.

Keywords: Face components, Image Quality, Face Morphing Attacks, Face Image Quality.

1 Introduction

Face morphing is used to create an image, out of two or more faces from different individuals, so that this attack face can successfully be verified to multiple identities [FFM14]. The perceptual quality of morphing attacks is important for their success, and therefore, different works presented new morphing methodologies that focus on the image appearance. Human operators can commonly only identify visible image artifacts (perceptual quality), which is important in the document issuing and identity check processes that do not include automatic attack detection.

Measuring the statistical and perceptual image quality of the used morphs in previous works have shown no clear correlation between the image quality and the realistic appearance when dealing with Morphing attacks (typically of ICAO standard [ICA18] with not large quality variation) [Zh20]. Full-reference image quality metrics like PSNR and SSIM performed on the same morphing pairs used in [Zh20] showed an insignificant difference between the different types of morph attacks. The little difference was not consistent with the visual perception score. The work in [Da19a] showed that the clearly visibly unrealistic images of MorGAN [Da18] achieve even higher statistical quality metrics (6 different metrics). Debiasi et al. further showed in a clear study that even though MorGAN attacks have clearly low realistic appearance, they show closer perceptual quality distributions to bona fide (BF) images than attacks of the more realistic appearance.

[1] Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany, biying.fu@igd.fraunhofer.de
[2] Interactive Graphics Systems Group, TU Darmstadt, Germany

To enable an informed quantitative measurement of morphed image quality, a wider study on multiple quality/utility metrics is required. This will allow researchers to make clear statements on the quality of the created morphing attacks by knowing which metrics does actually correlate with such artifacts, regardless of processes like re-digitization. To address this research gap, we investigated the effect of face morphing on image quality and utility by using 12 different metrics. To ensure that the captured quality differences are those related to morphing, we performed our investigation on a digital and re-digitized version of a newly created morphing attack database. Our conclusions indicated that the artifacts introduced by the morphing process were not clearly detected by most quality-/utility metrics, unlike those introduced by the re-digitization process. Only the MagFace [Me21] did show clear differentiation between morphed and bona fide samples in both, digital and re-digitized samples. We further look into the quality changes in artifact-prone areas of the face (e.g. nose, mouth, eyes) and found a higher correlation between some quality metrics and morphing when closely looking at such areas.

2 Quality and utility metrics

To investigate the perceptual image quality of morphs, we evaluated in total 12 methods separated into two families of quality metrics: 1) the general image quality assessment (IQA) methods, and 2) the face image quality (utility) assessment (FIQA) methods.

Four FIQA methods (FaceQnet [He19], rankIQ [Ch15], MagFace [Me21], and SER-FIQ [Te20]) are selected, either supervised or unsupervised. They are fine-tuned for FR solutions and operate on face images. **FaceQnet** [He19] is trained on VGGFace2 dataset using ICAO compliance software to setup the target labels for supervised training. A regression layer is trained on top of the ResNet-50 backbone. However, to quantify the face utility in an absolute manner is difficult, we selected unsupervised methods to avoid the groundtruth labeling.**MagFace** [Me21] is a recently developed method to access both the face embeddings for FR and the quality score of the face image. While training the magnitude of the feature vector (i.e. the face embedding) is made proportional to the cosine distance to its class center. **rankIQ** [Ch15] is trained to assess FIQA based on images with ranked qualities. The method is trained using three different databases with varying quality face images. **SER-FIQ** [Te20] is an unsupervised DL-based FIQA method. This method fully mitigates the need for any training or human labeling. It uses a stochastic method to relate the robustness of a face embedding towards the face utility by applying dropout to the face embeddings. Here, we used the SER-FIQ (on ArcFace) method to evaluate the results.

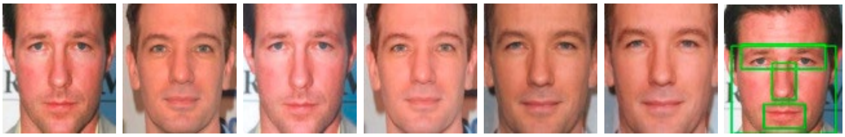
Eight IQA methods are chosen for evaluation. They are categorized into the following three categories: (1) based on natural image statistics, (2) convolutional neural network-based, and (3) ranking-based methods. These methods are more generalized approaches by considering image distortions and artifacts and its measure related to quality metrics.

BRISQUE [MMB12], **NIQE** [MSB13], and **PIQE** [N.15] belong to the first category and are all based on studying the deviation from the general statistics of natural images. These statistics are based on the finding by Rudermann [Ru94] that natural scene images have a distribution similar to a normal Gaussian distribution. The degree of deviation from the normal Gaussian relates to the degradation in image quality. The second category uses convolution layers to automatically extracting features without the need for a priori special design. CNNIQA [Ka14] , DeepIQA [Bo18], and MEON [Ma18] are counted to this category. While **CNNIQA** only have one single convolutional layer, **DeepIQA** increased the number of the base convolutional layers to increase the ability to deal with more complex images and also colored images. **MEON** is a multitask network combining the advantages of two complementary tasks of distortion type classification (Subtask I) and quality score estimation (Subtask II). The third category uses ranking-based image pairs to avoid the need for annotating absolute quality score for training images. Both methods (dipIQ[Ma19] and RankIQA [LvdWB17]) are categorized to this class. One benefit of these methods is that it is easier to generate image distortions and synthesize ranked image pairs. The network structures of these methods are build using two parallel streams with shared weights. Here, only one trained stream is used to assess the quality of the input.

Definition of Face areas: As the morphing process introduces blending artifacts especially apparent in certain areas of the face, we looked at the image quality of morphed and BF images grouped in these areas. For this task, the MTCNN framework [XZ17] is used to detect the face. Face images are aligned and standardized to 260 × 260 pixels, such that the eyes, nose, and mouth are on the same relative position within the aligned image. We crop the face regions into separate face components. The considered regions are: (1) **Eyes:** The eyes region includes both the eyes and eyebrow region. (2) **Nose:** The image patch includes only the nose region. (3) **Mouth:** The mouth region includes only the lips limited by the left and right corners of the mouth. (4) **Tightly cropped face:** Tightly cropped face region covers the area from chin to the eyebrows excluding the forehead. The information derived from the forehead, hairstyles, or hair colors is neglected under this setting. (5) **Aligned Face:** by using MTCNN are the images required for FIQA algorithms. The images and patches are all adequately resized matching the input size of the examined methods. A sample image is showed in Fig. 1(g).

3 Experimental setup

Database We manually chose images from the VGGFace2 database [Ca18] so that they are frontal, with a neutral expression, have no glasses, and have an eye-distance of over 90 pixels and frontal as described in [Da21] (54010 total selected images). From these, morphing pairs were selected by choosing the most similar to each other (using pre-trained OpenFace model). The morphing was performed by the approach and parameters presented in [Ra17]. An additional bona fide image is selected for each morphed identity, when available, such that it follows the mentioned manual quality check. In total, the used morphing database contains 364 digital Bona fide (D-BF) images and 276 digital morphing attacks (D-M). These images were printed on 11,5cmx9cm glossy photo paper in a professional studio and scanned with 600dpi scanner. They resulted in the same number of re-digitized bona fide (PS-BF) and attacks (PS-M). No split between training and testing is required as we do not train any solution, but rather analyze the quality of the images. The vulnerability (to the morphing attacks) of a pre-trained ResNet-100 ArcFace [De19] face recognition model was measured, resulting in Mated Morph Presentation Match Rate (MMPMR) of 80.07 % for the D-M, and 77.17% for the PS-M, both at false match rate (FMR) of 0.1%. Samples of the images included in the database are shown in Fig. 1. The Database is described in details in [Da21].



(a) D-ID1 (b) D-ID2 (c) PS-ID1 (d) PS-ID2 (e) D-M (f) PS-M (g) Segment

Fig. 1: Samples of the used database, including the digital image (D-ID1, D-ID2), the re-digitized images (PS-ID1, PS-ID2), the digital MA (D-M), and the re-digitized attack (PS-D). Note the different nature of effect between the re-digitization process and the morphing process. (g) displays the selective face areas considered.

Evaluation metrics We present distribution plots showing the quality score distribution of (1) D-BF (blue), (2) D-M (orange), (3) PS-BF (green), and (4) PS-M (red) for the 12 methods individually. The curves displays the probability density function (PDF) scaled by the number of observations across the seen quality score, such that the AUC sums to one across the seen quality scores. We further report the Fisher Discriminant Ratio (FDR) [DON14] [LdLFdC10] to measure the separability of classes, i.e., between D-BF and D-M and between PS-BF and PS-M. The higher the separability, the more different the two distributions are, and thus the more is the morphing effect captured by the specific metric.

Experiments The experiments are performed individually for (1) aligned faces, (2) cropped faces, (3) eye regions, (4) nose regions, and (5) mouth regions. In Tab. 1, to enable better understanding, the ordering of the quality score metric is provided for each method, where a descending order means high quality in case it is True and vice versa. Because FIQA methods rely on the entire face to determine the face image utility, we only report the quality metrics from the more generalized IQA methods for face areas.

FIQA methods				IQA methods							
rankIQ	FaceQnet	MagFace	SER-FIQ	BRISQUE	PIQE	NIQE	CNNIQA	DeepIQA	MEON	rankIQA	dipIQ
False	False	False	False	True	True	True	True	True	True	False	False

Tab. 1: True indicates small value for high-quality images. Hence, if the score distribution is on the left side, the sample has a higher quality and vice versa.

4 Results

In Fig. 2, we show the score distribution for all investigated methods on aligned face images. MagFace and SER-FIQ are the only FIQA methods that show differences between BF and MA, while the MagFace performance shows that more clearly and without differentiating between PS and D, which points out that the difference is correlated to the morphing process. Most solutions did differentiate between D and PS images, regardless of being BF or MA.

Fig. 3 shows the score distribution for all investigated methods on tightly cropped face areas. For most methods, the score distributions largely overlap. However, MEON shows slightly deviating peak locations and displays a better quality distribution for BF compared to MA. The differences here are mostly between PS and D, regardless of the morphing.

In Fig. 4(a), score distribution for the eye region is shown. Similar results for tightly cropped face area can be observed here. Although, the separability between BF and MA increased for MEON as indicated by the peak positions. Fig. 4(b) presents the score distribution

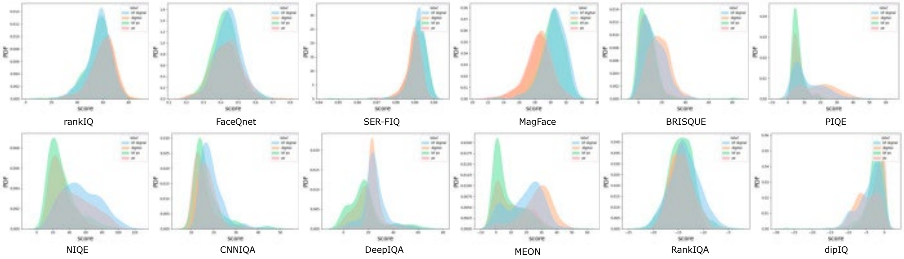


Fig. 2: PDF of the quality score distribution for the settings: (1) D-BF (blue), (2) D-M (orange), (3) PS-BF (green), and (4) PS-M (red) on aligned face images.

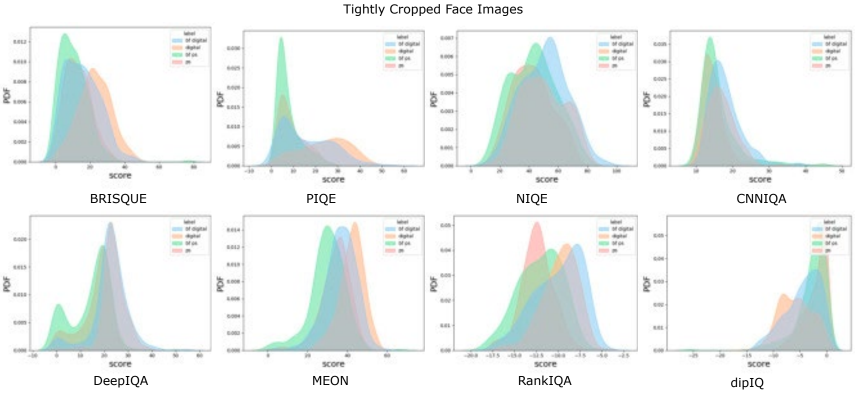


Fig. 3: PDF of the quality score distribution for the settings: (1) D-BF (blue), (2) D-M (orange), (3) PS-BF (green), and (4) PS-M (red). The first two rows are on the tightly cropped face images.

for the nose region. Looking at the nose region only, further increased the performance for MEON, as the shift of BF to the left side is more apparent compared to the previous face areas. This suggested a better quality distribution for BF images and also less effect of the re-digitization process. In Fig. 4 (c), we show the score distribution for the mouth region. In this area, most methods show a strong overlap between the quality distributions. Only dipIQ revealed a strong separability for the BF images and the morphs. The shift to the right indicates that the quality of the BF images is higher compared to the morphs.

In Tab. 2, we present the FDR for (1) d-BF vs D-M, and (2) Ps-BF vs PS-M. For aligned faces, all metrics are provided, while for face parts only IQA methods are available. The highest separability can be observed by using MagFace to distinguish the distributions between the BF images and the morphs.

The following findings can be drawn from the conducted experimental results: (1) Most quality and utility metrics do not capture the artifacts introduced by the morphing process, for instance, PIQE and NIQE. This makes them of less importance to measure the deterioration of the visual quality introduced by morphing. These metrics examined in this paper, however, clearly capture the artifacts introduced by the re-digitization process. Only the MagFace utility did show clear separability between the MA and BF. Moreover, this separability was consistent between both the PS and the D versions of the images. (2) Knowing

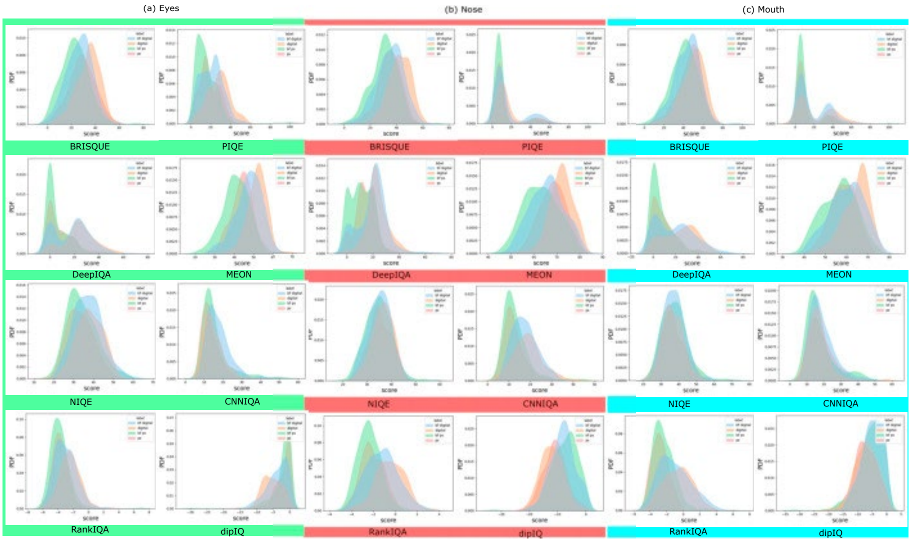


Fig. 4: PDF of the quality score distribution for the settings: (1) D-BF (blue), (2) D-M (orange), (3) PS-BF (green), and (4) PS-M (red) on the eyes, nose, and mouth regions.

	FIQA methods				IQA methods							
	rankIQ	FaceQnet	**MagFace**	SER-FIQ	BRISQUE	PIQE	NIQE	CNNIQA	DeepIQA	**MEON**	rankIQA	dipIQ
	Aligned Faces											
FDR D-BF vs D-M	0.06	0.00	**0.69**	0.13	0.12	0.20	0.02	0.09	0.01	0.24	0.02	0.03
FDR PS-BF vs PS-M	0.04	0.00	**0.63**	0.13	0.00	0.00	0.00	0.06	0.00	0.04	0.01	0.06
	Tightly Cropped Faces											
FDR D-BF vs D-M	-	-	-	-	0.27	0.25	0.00	0.03	0.03	**0.33**	0.00	0.09
FDR PS-BF vs PS-M					0.04	0.04	0.00	0.06	0.03	**0.13**	0.01	0.04
	Eyes Region											
FDR D-BF vs D-M	-	-	-	-	0.27	0.21	0.01	0.08	0.11	**0.27**	0.09	0.11
FDR PS-BF vs PS-M	-	-	-	-	0.15	0.08	0.01	0.08	0.01	**0.09**	0.00	0.00
	Nose Region											
FDR D-BF vs D-M	-	-	-	-	0.30	0.00	0.01	0.02	0.09	**0.50**	0.14	0.42
FDR PS-BF vs PS-M	-	-	-	-	0.10	0.00	0.04	0.01	0.07	**0.29**	0.03	0.34
	Mouth Region											
FDR D-BF vs D-M	-	-	-	-	0.12	0.03	0.01	0.04	0.05	**0.13**	0.02	0.13
FDR PS-BF vs PS-M	-	-	-	-	0.09	0.00	0.01	0.00	0.00	**0.08**	0.05	0.08

Tab. 2: The FDR value for (1) D-BF vs D-M and (2) PS-BF vs PS-M values are provided for all 12 methods investigated. FIQA metrics can only be applied to aligned images. For aligned face images, the **MagFace** show the highest separability between the MA and BF, while for face components, the multitask method **MEON** outperforms the other IQA methods.

that the blending artifacts introduced by the morphing process are commonly more apparent in certain areas of the face, we additionally looked at the image quality of MA and BF images, for only these areas, the nose, eyes, mouth, and finally tightly cropped face area. In general, MEON shows differences between the BF and the MA for certain face areas, like for instance tightly cropped face, eyes, and nose, whereas dipIQ also shows a similar trend for mouth region. This has shown clearer separability between MA and BF images. Especially the areas eyes and nose together with the metric by MEON and dipIQ showed to capture the image quality deterioration introduced by the morphing process. This was consistent between the digital samples and the printed and scanned samples.

5 Conclusion

In this paper, we investigated the effect of face morphing on image utility. Performing experiments on a proposed morphing database, including digital and re-digitized images, by exploring 12 different metrics both IQA and FIQA. Only MagFace shows a clear separability between MA and BF, both in digital and re-digitized images. Additionally, after investigating several facial parts, MEON shows clear separability on certain areas of the face. It can be used to discriminate between the BF and the MA for areas, such as the eyes and nose. Future work will consider the effect of synthetically generated morphs [Da18, Zh20], morph pair selection [Da19b], and image compression, on image quality.

Acknowledgements: This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[Bo18] Bosse, Sebastian; Maniry, Dominique; Müller, Klaus-Robert; Wiegand, Thomas; Samek, Wojciech: Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment. IEEE Trans. Image Process., 27(1):206–219, 2018. 3

[Ca18] Cao, Qiong; Shen, Li; Xie, Weidi; Parkhi, Omkar M.; Zisserman, Andrew: VGG-Face2: A Dataset for Recognising Faces across Pose and Age. In: 13th IEEE, FG 2018, Xi'an, China, May 15-19, 2018. IEEE Computer Society, S. 67–74, 2018. 3

[Ch15] Chen, Jiansheng; Deng, Yu; Bai, Gaocheng; Su, Guangda: Face Image Quality Assessment Based on Learning to Rank. IEEE Signal Proc. Lett., 22:90–94, 2015. 2

[Da18] Damer, Naser; Saladie, Alexandra Mosegui; Braun, Andreas; Kuijper, Arjan: MorGAN: Recognition Vulnerability and Attack Detectability of Face Morphing Attacks Created by Generative Adversarial Network. In: BTAS. IEEE, S. 1–10, 2018. 1, 7

[Da19a] Damer, Naser; Boutros, Fadi; Saladie, Alexandra Mosegui; Kirchbuchner, Florian; Kuijper, Arjan: Realistic Dreams: Cascaded Enhancement of GAN-generated Images with an Example in Face Morphing Attacks. In: BTAS. IEEE, S. 1–10, 2019. 1

[Da19b] Damer, Naser; Saladie, Alexandra Mosegui; Zienert, Steffen; Wainakh, Yaza; Terhörst, Philipp; Kirchbuchner, Florian; Kuijper, Arjan: To Detect or not to Detect: The Right Faces to Morph. In: ICB. IEEE, S. 1–8, 2019. 7

[Da21] Damer, Naser; Spiller, Noemie; Fang, Meiling; Boutros, Fadi; Kirchbuchner, Florian; Kuijper, Arjan: PW-MAD: Pixel-wise Supervision for Generalized Face Morphing Attack Detection. CoRR, 2021. 3

[De19] Deng, Jiankang; Guo, Jia; Xue, Niannan; Zafeiriou, Stefanos: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: CVPR. Computer Vision Foundation / IEEE, S. 4690–4699, 2019. 3

[DON14] Damer, Naser; Opel, Alexander; Nouak, Alexander: Biometric source weighting in multi-biometric fusion: Towards a generalized and robust solution. In: EUSIPCO. IEEE, S. 1382–1386, 2014. 4

[FFM14] Ferrara, Matteo; Franco, Annalisa; Maltoni, Davide: The magic passport. In: IJCB. IEEE, S. 1–7, 2014. 1

[He19] Hernandez-Ortega, Javier; Galbally, Javier; Fiérrez, Julian; Haraksim, Rudolf; Beslay, Laurent: FaceQnet: Quality Assessment for Face Recognition based on Deep Learning. In: ICB. IEEE, S. 1–8, 2019. 2

[ICA18] Portrait quality (reference facial images for MRTD). ICAO Technical Report, 2018. 1

[Ka14] Kang, Le; Ye, Peng; Li, Yi; Doermann, David S.: Convolutional Neural Networks for No-Reference Image Quality Assessment. In: CVPR. IEEE Computer Society, S. 1733–1740, 2014. 3

[LdLFdC10] Lorena, Ana Carolina; de Leon Ferreira de Carvalho, André Carlos Ponce: Building binary-tree-based multiclass classifiers using separability measures. Neurocomputing, 73(16-18):2837–2845, 2010. 4

[LvdWB17] Liu, Xialei; van de Weijer, Joost; Bagdanov, Andrew D.: RankIQA: Learning from Rankings for No-Reference Image Quality Assessment. In: ICCV. IEEE Computer Society, S. 1040–1049, 2017. 3

[Ma18] Ma, Kede; Liu, Wentao; Zhang, Kai; Duanmu, Zhengfang; Wang, Zhou; Zuo, Wangmeng: End-to-End Blind Image Quality Assessment Using Deep Neural Networks. IEEE Trans. Image Process., 27(3):1202–1213, 2018. 3

[Ma19] Ma, Kede; Liu, Wentao; Liu, Tongliang; Wang, Zhou; Tao, Dacheng: dipIQ: Blind Image Quality Assessment by Learning-to-Rank Discriminable Image Pairs. CoRR, abs/1904.06505, 2019. 3

[Me21] Meng, Qiang; Zhao, Shichao; Huang, Zhida; Zhou, Feng: MagFace: A Universal Representation for Face Recognition and Quality Assessment. CoRR, abs/2103.06627, 2021. 2

[MMB12] Mittal, Anish; Moorthy, Anush Krishna; Bovik, Alan Conrad: No-Reference Image Quality Assessment in the Spatial Domain. IEEE Trans. Image Process., 21(12):4695–4708, 2012. 2

[MSB13] Mittal, Anish; Soundararajan, Rajiv; Bovik, Alan C.: Making a ”Completely BlindÏmage Quality Analyzer. IEEE Signal Process. Lett., 20(3):209–212, 2013. 2

[N.15] N., Venkatanath; D., Praneeth; Bh., Maruthi Chandrasekhar; Channappayya, Sumohana S.; Medasani, Swarup S.: Blind image quality evaluation using perception based features. In: NCC. IEEE, S. 1–6, 2015. 2

[Ra17] Raghavendra, Ramachandra; Raja, Kiran B.; Venkatesh, Sushma; Busch, Christoph: Face morphing versus face averaging: Vulnerability and detection. In: IJCB. IEEE, S. 555–563, 2017. 3

[Ru94] Ruderman, Daniel L: The statistics of natural images. Network: Computation in Neural Systems, 5(4):517–548, 1994. 2

[Te20] Terhörst, Philipp; Kolf, Jan Niklas; Damer, Naser; Kirchbuchner, Florian; Kuijper, Arjan: SER-FIQ: Unsupervised Estimation of Face Image Quality Based on Stochastic Embedding Robustness. In: CVPR. IEEE, S. 5650–5659, 2020. 2

[XZ17] Xiang, Jia; Zhu, Gengming: Joint face detection and facial expression recognition with MTCNN. In: 2017 4th International Conference on Information Science and Control Engineering (ICISCE). IEEE, S. 424–427, 2017. 3

[Zh20] Zhang, Haoyu; Venkatesh, Sushma et al.: MIPGAN - Generating Robust and High Quality Morph Attacks Using Identity Prior Driven GAN. CoRR, abs/2009.01729, 2020. 1, 7

Will new definitions of emotion recognition and biometric data hamper the objectives of the proposed AI Act?

Jan Czarnocki[1]

Abstract: The paper explains how the definition of biometric data copied from the GDPR may hamper the regulation of emotion recognition—as defined in the proposed AI Act. A replicated definition of biometric data is suitable for biometric systems, but not emotion recognition technologies. It is because, under the proposed AI Act, an emotion recognition system is understood as such if it processes biometric data—as defined in the GDPR. But the definition from the GDPR does not encompass all biometric data, which are technically biometric data and are processed in the emotion recognition systems. Also, in the proposed AI Act the definition of emotion recognition does not recognize emotion recognition systems not relying on biometric data processing. That is why the obligation in the proposed AI Act for users to inform natural persons about their exposure to the emotion recognition system is unapplicable in the majority of cases. The flawed definition may also put at risk the proposed AI Act-based assessment of whether AI systems should be prohibited. Therefore, a new definition of emotion recognition and biometric data is needed.

Keywords: biometric data, emotion recognition, AI Act, AI, data protection, privacy, biometrics.

1 Introduction

The proposed Artificial Intelligence Act (AI Act) is the first legislative attempt to comprehensively regulate AI systems [Eu21]. Presented recently by the European Commission[2] it is an important part of building a new EU digital constitutionalism [Gr21]—a broad effort to address the challenges of digitalization. The goal of the proposed AI Act is i.a. to safeguard the trustworthiness of AI systems by ensuring their deployment is aligned with the values enshrined in the Charter of the Fundamental Rights of the EU [Eu20]. The proposal is the effect of an EU-wide effort and was preceded by the work of High Group on AI, which culminated in the White Paper on AI [Eu20]. Based on experience with enacting i.e. GDPR [Eu16] it will be some time until the implementation of the new regulation (it took GDPR 6 years from the first draft to come into force). Still, once enacted the AI Act will impact the fate of AI systems development in the EU for the next decades.

[1] Doctoral Researcher, KU Leuven Faculty of Law, Center for IT & IP Law, jan.czarnocki@kuleuven.be. This article has been made possible by received funding from the European Union's Horizon 2020 research and innovation programme in the context of the Privacy Matters Innovative Training Network (prima-itn.eu). I would like to thank prof. Els Kindt for her support and critical review.

[2] It was presented by the European Commission in April 2021. My paper relies on the first draft of the proposal, as presented therein.

The scope of the proposed AI Act encompasses AI applications such as machine learning, logical, statistical, and knowledge-based approaches [Eu21b]. It categorizes AI applications according to their purpose [Eu21c] and according to the risks posed to fundamental rights. The proposed AI Act prohibits or limits AI practices considered too risky, such as remote biometric identification or harmful manipulation of a natural person's behaviour [Eu21d]. The proposed AI Act introduces numerous compliance and due diligence requirements for high-risk AI systems [Eu21e].

One of the new requirements is the transparency obligation for emotion recognition systems using biometric data. Unless it is obvious from the context, users of AI emotion recognition systems are obliged to notify natural persons that they are interacting or are exposed to the workings of such a system [Eu21f]. This obligation can prove ineffective as it may be bypassed if left in its current form. It is due to a faulty definition of emotion recognition, which relies on the definition of biometric data from GDPR for AI applications to be classified as emotion recognition. Definition of biometric data replicated to the proposed AI Act is not suitable for AI systems—especially for emotion recognition, because it will leave most of the systems recognizing emotions out of the scope of the definition of emotion recognition.

2 Why regulate emotion recognition

Emotion recognition is an interdisciplinary research field, encompassing i.e psychology, cognitive science, and computer science. Its goal is to enable computers to understand human emotions and affects, to act accordingly [Pi03]. Emotion recognition is divided into fields of affective computing—mainly related to speech, video, and image processing and real-time analysis, and sentiment analysis—mainly related to longer-term opinions forming analysis, through natural language processing and describing what content is emotional [Pi03]. They are crucial fields to AI development [Mi06][3].

An example emotion recognition system could be embedded in an automated facial recognition system (AFRS), detecting face, extracting needed features, and classifying a natural person according to his or her gender, age, and six basic facial emotions: anger, happiness, fear, surprise, disgust, and sadness [LTL16]. However, emotion recognition can also range to uses of non-obvious types of data, such as stemming from i.e. measuring galvanic skin response (measurement of skin sweat to infer emotional arousal), electrocardiograms (cardiac cycle measurement), electroencephalograms (brain waves activity measurement), electromyograms (electrical measurement of skeletal muscles activity), or measurement of respiration, and skin temperature [UDRS17].

[3] These approaches assume that a key to embedding the machine with human intelligence is the capacity of a machine to understand human emotions and affects. Therefore, this approach teaches machines how to recognize emotions and then adjust actions accordingly. Through analysis of face scans, speech samples, or written excerpts engineers teach AI systems to recognize in what emotional state the natural person is, and what physical traits and behavior are related to what emotion.

Exposure to emotion recognition systems poses numerous risks to privacy and data protection, by revealing what an individual may not want to disclose. Emotions can be added to profiles, putting privacy and data protection at risk. The ability to recognize emotions can give data processors and controllers precise and accurate information about the state of mind of a person, disclosing sensitive knowledge about him or her [Ko21]. It is through emotion recognition systems that numerous predatory practices online are possible, such as profiling [GH08] and nudging [ST09], including using dark patterns to monetize emotions [Cl19]. These practices are the backbone of data power, surveillance capitalism, and attention economy [Zu19]—the dark sides of the digital world.

Emotion recognition systems can be an important part of AI systems, which systems use is prohibited or limited in the new regulation. Judging the impact of an AI system on fundamental rights might be determined by whether the system is capable of recognizing emotion. For example one of the prohibited AI practices in Article 5 is a subliminal, harmful, material distortion of a natural person's behavior [Eu21h]. The capability to recognize emotions and affect is crucial for some AI systems to capacitate such harm. Therefore, a proper legal definition determining what counts as an emotion recognition system is of crucial importance for protecting natural persons. Also, the effectiveness of the obligation to inform the natural person about their exposure to an emotion recognition system depends on how the emotion recognition system is defined.

3 Inheriting the wrong definition of biometric data

According to the proposed AI Act an emotion recognition system is "*AI system for the purpose of identifying or inferring emotions or intentions of natural persons on the basis of their biometric data*" [Eu21i]. Using conjunction the definition determines that only systems that recognize emotions based on biometric data are emotion recognition systems. However, what is understood from a technical point of view as biometric data, is not always what is understood as biometric data from the legal point of view [Ja15]. The definition of biometric data in the proposed AI Act is duplicated from GDPR [Eu16a][4]. It is a functional definition—affordances of the technology used to process the data define the legal nature of the data [Ki18]. It assumes that personal data becomes biometric data once processed through the system, which allows or confirms unique identification. Otherwise, personal data is not biometric data—it does not enjoy a higher level of protection reserved for the category of sensitive data [Ki18]. The threshold for falling under the definition of biometric data is whether biometric data results from specific, technical processing which allows identifying a natural person. Therefore, unless personal data is processed through technical means that enable to recognize the identity of an individual it is not construed as biometric data. For example, unless scanned through a facial recognition system, or other biometric system for the purpose of "unique

[4] where it is defined as "*personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data*"

identification", a photo or video with a natural person's image is not considered biometric data [Eu16b]. As pointed out by Kindt and Jasserand there is a misalignment between technological affordances and legal definition [Eu16b, Ja15]. The current definition of biometric data fails to properly delineate the scope of what must be considered biometric data. It limits the understanding of what is biometric data from a legal perspective [Ki18]. Therefore, this definition leaves numerous data, which technically is biometric data, outside of the scope of higher protection, granted to biometric data by GDPR.

Therefore, since the proposed AI Act replicates the same definition of biometric data, unless an emotion recognition system processes such narrowly understood biometric data and unless it is a biometric system capable to recognize the identity, it is not considered an emotion recognition system. Thus, unless it uses biometric data specifically for purpose of "unique identification" it is not an emotion recognition system. This leaves most of the AI systems capable of recognizing emotion outside of the scope of the definition. Most of them neither process biometric data—as narrowly understood under the definition in the proposed AI Act, or are biometric systems—capable of recognizing the identity or verifying it [KRBDJ18].

The proposed definition of emotion recognition misses its intended object because according to the GDPR and the proposed AI Act's definition [Eu16a] biometric data only becomes biometric data if it is processed through technical means allowing for uniquely identifying a natural person—biometric systems. They are systems for *"automated recognition of individuals based on their biological and behavioural characteristics"* [IS17]. Biometric recognition can form a part of the emotion recognition system but is not a necessary component—it is not essential to identify or verify a natural person's identity, to recognize his or her emotions. The old definition of biometric data is well suited for biometric systems [IS17], but not for other affective computing and sentiment analysis technologies.

For instance, most AFRSs detect emotion, attention, and different states using face modeling, extracting only the features needed, and then compressing them. A particular psychological state or sociodemographic characteristic is recognized through comparison with a dataset of such modeled images using an artificial neural network [LTL16]. Therefore, biometric data as legally defined are not processed, because identification of a natural person is not possible in this case. In one example, the emotions of gamers were measured and analyzed in an experiment, through EEG (electroencephalography – measurement of electrical activity on the scalp, representing the macroscopic activity of the surface layer of the brain) and using machine learning. The system was able to recognize the valence of gamer's emotions like anger, anticipation, joy, trust, fear, surprise, sadness, and disgust [BGT20]. However, the system was not able to tell anything about the gamer's identity, because that was not its purpose. According to the current regulation it would not be branded as an emotion recognition system, because personal data processed through it do not fall under the definition of biometric data. Similarly, mobile applications for emotion recognition can extract facial representations from images or videos. Then face detection module is applied to extract frames of face regions and then compare them to the model based on training data set containing other faces [HM17]. In this way, emotion can be detected without the need for an AI system to be a biometric system, able to "uniquely identify" a natural person [LTL16].

Emotion recognition system can extract particular features from either video, image (e.g. like muscle movement, wrinkles, or other key facial regions [FM13]), or speech record (e.g. particular qualities of voice), which allows it to recognize and classify emotion, without the need to extract an entire biometric template [VChK15]. According to the current definition only when a biometric template is extracted, the system is processing biometric data. But other methods of emotion recognition can be used, such as through e.g. measurement of galvanic skin response, electrocardiogram, electroencephalogram (brain waves), electromyogram, or respiration, and skin temperature—without processing of legally understood biometric data [UDRS17]. In consequence emotion recognition systems do not have to be able to "uniquely identify" to recognize or infer emotions. Thus, under the GDPR definition, they will not be processing biometric data. However, technically speaking it will be biometric data. Therefore, due to its dependence on narrowly understood biometric data the definition of an emotion recognition system in the proposed AI Act will be ineffective and will consider most emotion recognition systems as outside its scope. This may also hamper the due diligence and compliance process for assessing the impact of an AI system on fundamental rights and whether it can enter the Single Market.

4 The need for new definitions

For the definition in the proposed AI Act to effectively encompass all biometric data, it would require recognition of not only the personal data processed through a biometric system but of data directly relating to biometric characteristics of a person, which is a view proposed by Kindt [Ki13][5]. In her definition condition for data to be biometric data is its relation to certain traits of natural persons, not the fact of being processed through specific technical means—the condition present in deficient definition in the proposed AI Act, repeated after the GDPR.

Perhaps the intention behind what is classified as biometric data in the current definition of emotion recognition is what would have been if Kindt's definition were applied. But the proposed AI Act takes the definition of biometric data from the GDPR. And this understanding of biometric data effectively erases the potential effectiveness of the definition of "emotion recognition system" in the proposed AI Act. Hence, due to deficiencies in the definition from the GDPR, the original sin of the bad definition of biometric data was replicated to the proposed AI Act. As a consequence, if the definition

[5] Under definition proposed by her biometric data are *"all personal data which (a) relate directly or indirectly to unique or distinctive biological or behavioral characteristics of human beings and (b) are used or are fit to be used by automated means (c) for purposes of identification, identity verification or verification of a claim of living natural persons."*.

of the emotion recognition system in the proposed AI Act is deficient—dependent on the definition of biometric data, then the majority of the AI systems recognizing emotions may fall outside of the scope of transparency and notification obligations therein. Thus, natural persons will not have to be notified if they are exposed to most emotion recognition systems. Moreover, if certain AI systems are not recognized as emotion recognition technologies under the current definition, then there is a risk of numerous harmful AI systems squeezing through the AI Regulation regime and entering the Single Market. It is because during assessment whether they pose risk to fundamental rights their capacity or property of being categorized as emotion recognition systems may be taken into consideration. Under the definition of emotion recognition systems from the proposed AI Act, many of them may fall outside of its scope—even though they are capable of recognizing emotion.

Therefore, we either need a new definition of emotion recognition or a new definition of biometric data. The first option is still feasible because work on the AI Act has just started. It will require a small amendment during the legislative process. However, in the long term, there is a need to also amend the definition of biometric data—both in the future AI Act and the GDPR. It will be more complicated because even if the definition of biometric data is changed in the AI Act—for example in a way suggested by Kindt, it will leave a faulty definition of biometric data still present in the GDPR. This would carry the risk of jeopardizing the legal system and sowing confusion. Hence, from a pragmatic point of view objectives of the proposed AI Act concerning emotion recognition can be saved by upgrading the definition of the emotion recognition system. The definition should not encompass biometric data and should not be dependent on it. It should be neutral as to the methods used for recognizing or inferring emotions. It should refer to the capacity to recognize or infer emotion or affects, without specifying how it is done. Thus, the definition should focus on the purpose and effect of the AI system. Each AI system capable of inferring or recognizing emotions should be treated as such. Focusing on the effects of the AI system will render the definition more robust and resilient to technological changes.

That is why, instead of defining an emotion recognition systems as an "*AI system for the purpose of identifying or inferring emotions or intentions of natural persons on the basis of their biometric data*" it suffices to define them as AI systems to identify or infer emotions, affects or intentions of natural persons.

References

[BGT20] Burak Alakus T.; Gonen M.; Turkoglu I.: Database for an emotion recognition system based on EEG signals and various computer games – GAMEEMO, Biomedical Signal Processing and Control, Volume 60, 2020.

[Cl19] Clifford, D.: The Legal Limits to the Monetisation of Online Emotions, KU Leuven, Faculteit Rechtsgelerdheid, Leuven, 2019.

[Eu16] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation – GDPR).

[Eu16a] Article 4(14) GDPR.

[Eu16b] Recital 51 GDPR, Article 4(14).

[Eu20] European Commission, White Paper On Artificial Intelligence - A European approach to excellence and trust, Brussels, 19.2.2020

[Eu21] European Commission, Proposal for a Regulation Of The European Parliament And Of The Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts, Brussels, 21.4.2021 COM(2021) 206 final.

[Eu21b] Annex I to the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts, Brussels, 21.4.2021 COM(2021) 206 final.

[Eu21c] Annex III to the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts, Brussels, 21.4.2021 COM(2021) 206 final.

[Eu21d] Article 5, Artificial Intelligence Act.

[Eu21e] Article 52.1, Artificial Intelligence Act.

[Eu21f] Article 52.2, Artificial Intelligence Act.

[Eu21g] Article 5.1(a) Artificial Intelligence Act.

[Eu21h] Recital 7 “The notion of biometric data used in this Regulation is in line with and should be interpreted consistently with the notion of biometric data as defined in Article 4(14) of Regulation (EU) 2016/679 of the European Parliament and of the Council”.

[Eu21i] Article 3(34) Artificial Intelligence Act.

[FM13] Filko D.; Martinović G.: Emotion Recognition System by a Neural Network Based Facial Expression Analysis, Automatika, 54:2, 263-272, 2013.

[GH08] Gutwirth, S.; Hildebrandt, M.: Profiling the European Citizen: Cross-disciplinary Perspectives. 1ed, Springer, New York, 2008.

[Gr21] De Gregorio, G.: The Rise of Digital Constitutionalism in the European Union., International Journal of Constitutional Law 19.1, 41-70, 2021.

[HM17] Hossain M. S.; and Muhammad G.: An Emotion Recognition System for Mobile Applications, in IEEE Access, vol. 5, pp. 2281-2287, 2017.

[IS17] ISO/IEC 2382-37, Information technology, Vocabulary, Part 37:Biometrics, 2017.

[Ja15] Jasserand, C. A.: Avoiding Terminological Confusion between the Notions of ‘biometrics’ and ‘biometric Data’: An Investigation into the Meanings of the Terms from a European Data Protection and a Scientific Perspective, International Data Privacy Law 6.1, 2015.

[Ki13] Kindt, E.: Privacy and Data Protection Issues of Biometric Applications - A Comparative Legal Analysis, Springer, 2013.

[Ki18] Kindt, E.: Having yes, using no? About the new legal regime for biometric data, Computer Law & Security Review, Volume 34, Issue 3, Pages 523-538, 2018.

[Ko21] Kosinski, M.: Facial recognition technology can expose political orientation from naturalistic facial images, Sci Rep 11, 100, 2021.

[KRBD18] Kartali, A.; Roglic, M.; Barjaktarovic, M.; Duric-Jovicic, M.; and Jankovic, M. M.: Real-time Algorithms for Facial Emotion Recognition: A Comparison of Different Approaches, 14th Symposium on Neural Networks and Applications (NEUREL): 1-4, 2018.

[LTL16] Lewinski, P.; Trzaskowski, J.; and Luzak, J.: Face and Emotion Recognition on Commercial Property under EU Data Protection Law, Psychology & Marketing 33.9: 729-46. 2016.

[Mi06] Minsky, M.: The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind, Simon & Schuster, New York, 2006.

[Pi03] Picard, R. W.: Affective Computing: Challenges. International Journal of Human-computer Studies 59.1: 55-64, 2003.

[ST09] Sunstein, C. R.; Thaler, R. H.; Nudge: Improving Decisions on Health, Wealth, and Happiness, Penguin, London, 2009.

[UDRS17] Udovičić, G.; Đerek, J.; Russo, M.; and Sikora M.: Wearable Emotion Recognition System based on GSR and PPG Signals, In Proceedings of the 2nd International Workshop on Multimedia for Personal Health and Health Care (MMHealth '17), Association for Computing Machinery, New York, NY, USA, 53–59, 2017.

[VChK15] Varghese A. A.; Cherian J. P.; and Kizhakkethottam J.J.: Overview on emotion recognition system, 2015 International Conference on Soft-Computing and Networks Security (ICSNS), pp. 1-5, 2015.

[Zu19] Zuboff, S.: The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power., First ed., PublicAffairs, New York, 2019.

On Recognizing Occluded Faces in the Wild

Mustafa Ekrem Erakın[1*], Uğur Demir[1*], Hazım Kemal Ekenel[1]

Abstract: Facial appearance variations due to occlusion has been one of the main challenges for face recognition systems. To facilitate further research in this area, it is necessary and important to have occluded face datasets collected from real-world, as synthetically generated occluded faces cannot represent the nature of the problem. In this paper, we present the Real World Occluded Faces (ROF) dataset, that contains faces with both upper face occlusion, due to sunglasses, and lower face occlusion, due to masks. We propose two evaluation protocols for this dataset. Benchmark experiments on the dataset have shown that no matter how powerful the deep face representation models are, their performance degrades significantly when they are tested on real-world occluded faces. It is observed that the performance drop is far less when the models are tested on synthetically generated occluded faces. The ROF dataset and the associated evaluation protocols are publicly available at the following link https://github.com/ekremerakin/RealWorldOccludedFaces.

Keywords: Face recognition, face occlusion, deep learning, real-world occluded faces.

1 Introduction

With the recent advancements in deep learning and its application to computer vision problems, state-of-the-art face recognition systems have achieved excellent results on various datasets, such as LFW [Hu08], AgeDB-30 [Mo17], and MegaFace [Ke16]. As the performance on these well-known datasets converges, researchers started to divert their attention towards more challenging problems. One of these challenges is recognizing occluded faces in the wild [ZVS20]. To catalyze further research on this topic, in this paper, we present the Real World Occluded Faces (ROF) dataset, that contains faces with both upper face and lower face occlusions. To test the authenticity of the dataset, we participated in a masked face recognition challenge [Bo21]. Our model, fine-tuned on real life masked images, outperformed models trained on larger, synthetically generated masked face training sets, leading to the best performance among 16 other academic submissions [Bo21].

There have been several works that studied the effects of many different appearance variations on face recognition performance [GE16], [Ka16], [Gr18]. In this paper, we will be addressing specifically the occlusion problem using a real-world occluded face dataset. [GE16] used AR face dataset [MB98] that contains occluded face images collected in a constrained environment, while [Ka16] and [Gr18] used synthetic occlusions built on top of LFW [Hu08]. Our experiments show that real world occlusions are more challenging than their synthetic counterparts.

[1] Faculty of Computer and Informatics, Dept. of Computer Engineering, 34469 Maslak, Istanbul, TURKEY, erakin20@itu.edu.tr, demirug16@itu.edu.tr, ekenel@itu.edu.tr

[*] Equal contribution

Previous studies show that deep CNN based face recognition models trained on VGGFace [PVZ15], faces major performance drops when confronted by sunglasses and scarves. [GE16] reports the performance with sunglasses occlusion in the range of 30-35% on the AR face dataset [MB98], which is a 110 identity face image dataset that is collected in a controlled environment with cooperating subjects, a rather easy benchmark for a modern face recognition model. Another study [Ka16] uses synthetic occlusions to test the face recognition performance again using a VGGFace pre-trained model [PVZ15]. Occluded face images are generated by applying black boxes on samples from the LFW dataset [Hu08]. Different occlusion types are simulated by applying these black boxes in different locations. The study reports 25.94% face recognition accuracy against sunglasses effect.

The main contributions of this paper can be summarized as follows: (i) We introduced an in-the-wild occlusion dataset for face recognition, (ii) we proposed two evaluation protocols and analyzed the impact of upper face and lower face occlusion on face recognition performance, (iii) we show that real-world face occlusion poses a more challenging problem for face recognition systems. We also visualized the results and discuss the outcomes in detail.

2 The Dataset

Real World Occluded Faces (ROF) dataset contains face images with real-life upper-face and lower-face occlusions, due to sunglasses and face masks, respectively. The dataset consists of 6421 neutral face images, 4627 face images with sunglasses, and 678 face images with masks. There are 47 subjects with neutral, masked, and sunglasses images, 114 subjects with neutral and sunglasses images, while 20 subjects have only neutral and masked images. The identities are collected from a list of celebrities and politicians. All of the images are from real-life scenarios and contain large variations in terms of pose and illumination. The images were downloaded from Google Image Search using the pipeline described in VGGFace2 study [Ca18]. On average there are 50 neutral images, 30 sunglasses images and 15 masked images per identity.

Dataset Collection

Dataset collection is done using a modified version of the pipeline described in VGGFace2 [Ca18]. A name list consisting of public figures, i.e., politicians, celebrities, sports players, etc., were collected. For every name in the list 100 images were downloaded for each type of face image we are after. A reference image was extracted from the collected neutral images for every name using the image size and face count within the image to try to get the best possible reference. Duplicates were removed using perceptual hashing and faces were detected and cropped from the remaining images using a combination of RetinaFace [De19b] and MTCNN [XZ17].

Then using the reference images and a ResNet50 [He16] trained on VGGFace2 [Ca18], face embeddings were extracted for every remaining image and compared with the subject's respective reference image's embedding, using cosine distance as the similarity metric. For neutral images, candidate images with a similarity above 0.5 were selected while for occluded images, the threshold was set to 0.2. Finally, filtered face images were manually verified and stored. Overall, manual work was limited and the bottleneck was finding the appropriate identities that would have both sunglasses and face mask images, which proved to be a niche category. Figure 1 shows sample images from the ROF dataset.

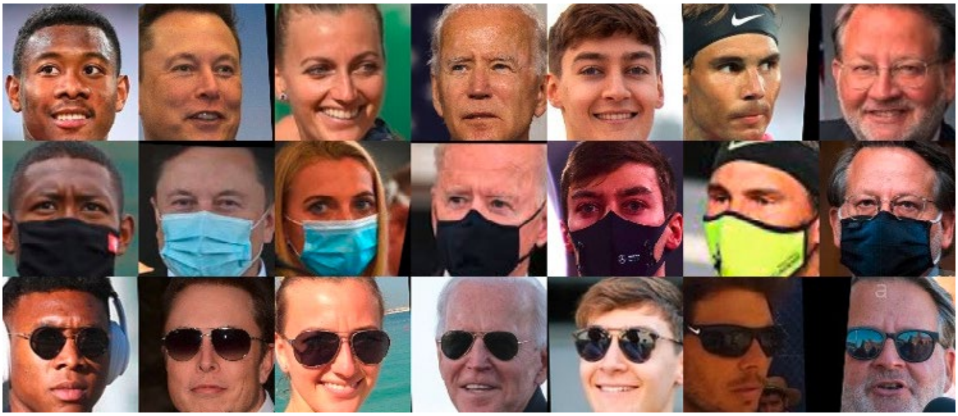


Fig. 1: Samples of neutral, masked, and sunglasses images for the same subjects from the ROF dataset

3 Deep Face Models

We utilized three different deep learning architectures to examine the performance degradation when encountered with facial occlusions, namely ArcFace [De19a], VGGFace2 [Ca18], and MobilFaceNet [Ch18].

ArcFace [De19a] is a state-of-the-art face recognition model that achieved excellent performance on various face recognition datasets, such as LFW [Hu08], AgeDB-30 [Mo17], and MegaFace [Ke16]. ArcFace is trained on MS1MV2 [De19a], which is a revised version of the MS-Celeb-1M dataset [Gu16]. MS1MV2 contains 85,000 identities and 5.8 million images. In this work, we used three different ArcFace architectures to represent various model complexities. Different Arcface architectures are denoted as Arcface-N which corresponds to a ResNet-N model pre-trained on MS1MV2 dataset.

VGGFace2 is a large-scale dataset containing 9131 identities and 3.3 million samples. Researchers used the dataset to train deep learning models, and it was one of the state-of-the-arts. We used their ResNet-50 pre-trained model throughout our experiments [Ca18].

Since ArcFace and VGGFace2 mainly use ResNet as the backbone architecture, their model complexities are not suitable for mobile devices. Therefore, we tested the model also on MobilFaceNet [Ch18] to analyze the performance degradation of a smaller model. MobilFaceNet used in this work is trained on the MS1MV2 dataset [De19a].

Throughout our experiments, we employed 512-dimensional feature embeddings. For distance metrics, Euclidean distance is utilized for ArcFace and MobilFaceNet, and cosine similarity is used for VGGFace2. ArcFace and MobilFaceNet pre-trained models are adopted from the Insightface repository[2]. VGGFace2 pre-trained model is adopted from the verified VGGFace2 repository.

4 Experimental Setup

In this section, we present the experimental setups to evaluate the occlusion robustness of the deep face recognition methods. We also analyze and compare the differences between the effects of synthetically crafted and real-world occluded face images. We present two experiment protocols. The first protocol investigates the effects of upper face occlusions, while the second one assesses the performance against lower face occlusions. Both protocols also probe with synthetic occlusions and compare the results with the ones obtained on the real world occluded samples. The image and identity counts across protocols are given in Table 1.

Protocol	Identities	Gallery	Synthetic Probe	Sunglasses	Masked
1	161	483	5322	4627	-
2	67	199	1800	-	464

Tab. 1: Total number of identities and images for each protocol

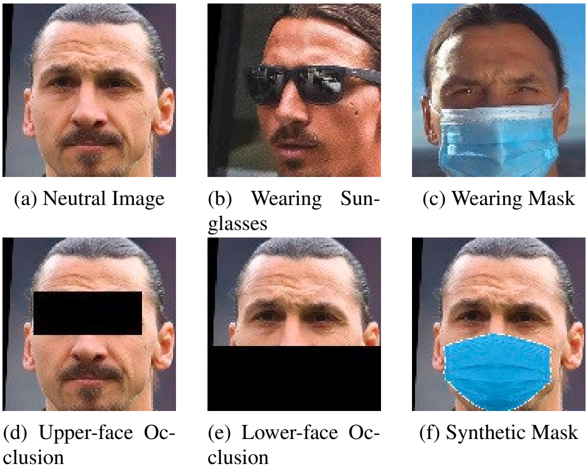

(a) Neutral Image (b) Wearing Sunglasses (c) Wearing Mask
(d) Upper-face Occlusion (e) Lower-face Occlusion (f) Synthetic Mask

Fig. 2: a) Non-occluded face image, b) Upper face occlusion due to wearing sunglasses, c) Lower face occlusion due to wearing a mask, c) Synthetic upper face occlusion, d) Synthetic lower face occlusion, e) Synthetic mask generation for lower face occlusion

For data preprocessing, CosFace [Wa18] and SphereFace [Li17] papers are followed. First, the bounding box and five facial landmarks, namely, eyes, nose, and mouth corners, are obtained using MTCNN [Zh16]. Afterwards, similarity transform is applied to images for face normalization. Then, images are cropped and resized to 112×112.

For testing against upper-face occlusions, we used the ROF sunglasses dataset in the first protocol. We also generated synthetic fixed upper-face occlusions that cover the eye region. Real sunglasses and synthetic occlusions can be seen in Figure 2b and 2d, respectively.

[2] https://github.com/deepinsight/insightface

In the second protocol, for lower-face occlusions we used the ROF mask dataset. Synthetic lower face occlusions are generated by fixing the nose and mouth area and using the mask generator [AR20] published in a recent study. Samples from real and synthetic lower face occlusions can be seen in Figure 2c and 2f, respectively.

5 Experimental Results and Discussion

In this section we present the experimental results. We performed both face identification and verification experiments and assessed the effect of occlusion in both scenarios.

5.1 Impact of Occlusions on Face Identification

Protocol 1: The experimental results are presented in Table 2 for all used deep face models. Each row corresponds to the obtained correct classification rates on a specific probe set. Samples from probe sets are shown in Figure 2. All models are found to be very successful when classifying face images that do not contain occlusion, as can be seen from the first row. Arcface is found to be more robust compared to MobileFaceNet and VGGFace2, when a part of the face is occluded synthetically, either by painting the corresponding region with black or generating an artificial mask. However, even Arcface's performance deteriorates when it is tested on real world occluded faces that contain sunglasses.

	Arcface-100	**Arcface-50**	**Arcface-34**	**MobilFaceNet**	**VGGFace2**
Occlusion Type	**Top 1**	**Top 1**	**Top 1**	**Top 1**	**Top 1**
No occlusion	99.57%	99.34%	99.17%	98.89%	98.12%
Wearing sunglasses	86.60%	84.18%	83.51%	77.16%	76.83%
Upper occlusion	98.25%	95.92%	95.43%	83.13%	75.65%
Lower occlusion	98.21%	96.81%	96.64%	86.98%	88.56%
Synthetic masked	98.53%	97.16%	96.56%	89.57%	89.59%

Tab. 2: Face identification results using protocol 1 (Arcface-N denotes the ResNet architecture with N layers, for VGGFace2 ResNet50 was used)

Protocol 2: In Table 3, we present the experimental results using protocol 2. The outcomes are similar to the ones obtained using protocol 1. The deep face models are found to be very successful when there is no occlusion in the probe images. Arcface is found to be superior to MobileFaceNet and VGGFace2, when a part of the face is occluded synthetically, either by painting the corresponding region with black or generating an artificial mask. However, again, even Arcface's performance deteriorates when it is tested on real world occluded faces that contain masks.

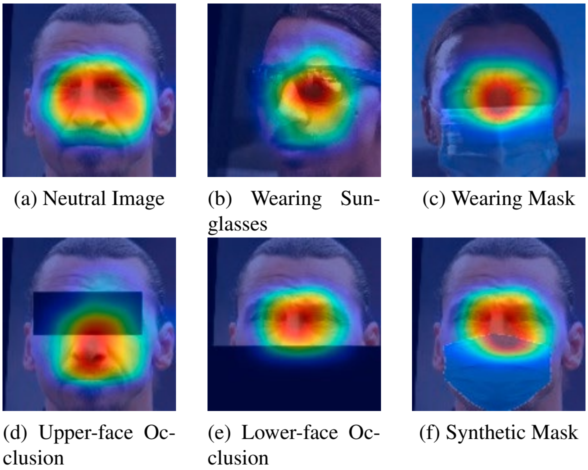
These results show that synthetically generated occlusions do not reflect the nature of the real-world occlusions. One reason could be due to the fact that the synthetic occlusions contain the same texture and covers the same regions across different faces. However, real world occlusions contain different textures and cover different parts of the faces depending on the style of the sunglasses or the type of the mask and the way the person wears it.

To analyze the results further, we also visualized the regions that the deep face model focuses using Grad-CAM method [Se19]. The obtained results are illustrated in Figure 3.

As can be seen the model mainly focuses on the inner face region, where eye and nose are contained. This is expected, since, especially, eye region is known to have a high discrimination power. However, as the models learn from the data the highly discriminative parts and focuses on these, when they are occluded they suffer from a performance loss. Therefore, while developing an occlusion-robust deep face recognition system, this fact has to be taken into account.

	Arcface-100	**Arcface-50**	**Arcface-34**	**MobilFaceNet**	**VGGFace2**
Occlusion Type	**Top 1**	**Top 1**	**Top 1**	**Top 1**	**Top 1**
No occlusion	99.61%	99.33%	99.39%	99.06%	99.28%
Wearing mask	85.34%	76.08%	73.71%	70.04%	79.31%
Upper occlusion	98.39%	96.89%	96.67%	89.78%	89.28%
Lower occlusion	98.83%	97.67%	97.11%	92.06%	93.94%
Synthetic masked	99.00%	97.78%	97.78%	93.22%	94.83%

Tab. 3: Face identification results using protocol 2


(a) Neutral Image

(b) Wearing Sunglasses

(c) Wearing Mask

(d) Upper-face Occlusion

(e) Lower-face Occlusion

(f) Synthetic Mask

Fig. 3: The regions that VGGFace2 targets the most during embedding extraction [Se19].

5.2 Impact of Occlusions on Face Verification

For the sake of completeness, we also run face verification experiments using the proposed ROF dataset. The results of experiments using protocol 1 and 2 are presented in Tables 4 and 5, respectively. Similar observations can be derived from these experiments: ArcFace is found to be more robust to synthetic occlusions. The EER increases significantly when the models are tested on real world occluded faces.

6 Conclusion

In this study, we present a real-world occluded face dataset and explore the effects of occlusion on the state-of-the-art face recognition methods' performance. We have shown that

	Arcface-100	**Arcface-50**	**Arcface-34**	**MobilFaceNet**	**VGGFace2**
Occlusion Type	**EER**	**EER**	**EER**	**EER**	**EER**
No occlusion	0.011	0.014	0.014	0.021	0.017
Wearing sunglasses	0.088	0.091	0.095	0.106	0.096
Upper occlusion	0.024	0.036	0.041	0.073	0.076
Lower occlusion	0.021	0.028	0.033	0.054	0.053
Synthetic masked	0.025	0.033	0.036	0.068	0.059

Tab. 4: Face verification results using protocol 1

	Arcface-100	**Arcface-50**	**Arcface-34**	**MobilFaceNet**	**VGGFace2**
Occlusion Type	**EER**	**EER**	**EER**	**EER**	**EER**
No occlusion	0.013	0.019	0.021	0.024	0.017
Wearing mask	0.083	0.119	0.119	0.119	0.082
Upper occlusion	0.031	0.035	0.043	0.067	0.058
Lower occlusion	0.019	0.036	0.038	0.054	0.045
Synthetic masked	0.028	0.036	0.043	0.074	0.049

Tab. 5: Face verification results using protocol 2

synthetically generated occlusions do not reflect the nature of the real-world occlusions. We have observed significant performance drops when deep face models are tested on real world occluded faces that contain masks or sunglasses. Visualization of the results indicate that the deep face models mainly focus on the inner face region. Therefore, the models experience a performance loss, when this region is occluded. For our future work, we aim to expand the collected dataset and develop an occlusion-robust deep face recognition system by benefiting from the findings of this work.

7 Acknowledgement

This study is supported by the Istanbul Technical University Research Fund, ITU BAP, project no. 42547 and by the Scientific and Technological Research Council of Turkey (TUBITAK) project no. 120N011.

References

[AR20] Anwar, A.; Raychowdhury, A.: Masked Face Recognition for Secure Authentication. In: https://arxiv.org/abs/2008.11104. 2020.

[Bo21] Boutros, F. et al.: MFR 2021: Masked Face Recognition Competition. In: IJCB 2021. pp. 1–10, 2021.

[Ca18] Cao, Q.; Shen, L.; Xie, W.; Parkhi, O.M.; Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: FG 2018. IEEE, pp. 67–74, 2018.

[Ch18] Chen, S.; Liu, Y.; Gao, X.; Han, Z.: Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In: CCBR. Springer, pp. 428–438, 2018.

[De19a] Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR. pp. 4690–4699, 2019.

[De19b] Deng, J.; Guo, J.; Zhou, Y.; Yu, J.; Kotsia, I.; Zafeiriou, S.: Retinaface: Single-stage dense face localisation in the wild. arXiv preprint arXiv:1905.00641, 2019.

[GE16] Ghazi, M.M.; Ekenel, H.K.: A comprehensive analysis of deep learning based representation for face recognition. In: CVPR workshops. pp. 34–41, 2016.

[Gr18] Grm, K.; Štruc, V.; Artiges, A.; Caron, M.; Ekenel, H.K.: Strengths and weaknesses of deep learning models for face recognition against image degradations. IET Biometrics, 7(1):81–89, 2018.

[Gu16] Guo, Y.; Zhang, L.; Hu, Y.; He, X.; Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: ECCV. Springer, pp. 87–102, 2016.

[He16] He, K.; Zhang, X.; Ren, S.; Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778, 2016.

[Hu08] Huang, G. B; Mattar, M.; Berg, T.; Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In: Workshop on faces in'Real-Life'Images: detection, alignment, and recognition. 2008.

[Ka16] Karahan, S.; Yildirum, M.K.; Kirtac, K.; Rende, F.S.; Butun, G.; Ekenel, H.K.: How image degradations affect deep CNN-based face recognition? In: BIOSIG. IEEE, pp. 1–5, 2016.

[Ke16] Kemelmacher-Shlizerman, I.; Seitz, S.M.; Miller, D.; Brossard, E.: The megaface benchmark: 1 million faces for recognition at scale. In: CVPR. pp. 4873–4882, 2016.

[Li17] Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: CVPR. pp. 212–220, 2017.

[MB98] Martinez, A.; Benavente, R.: The AR Face Database: CVC Technical Report, 24. 1998.

[Mo17] Moschoglou, S.; Papaioannou, A.; Sagonas, C.; Deng, J.; Kotsia, I.; Zafeiriou, S.: Agedb: the first manually collected, in-the-wild age database. In: CVPR Workshop. volume 2, p. 5, 2017.

[PVZ15] Parkhi, O.M.; Vedaldi, A.; Zisserman, A.: Deep Face Recognition. In: British Machine Vision Conference. 2015.

[Se19] Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. IJCB, 128(2):336–359, Oct 2019.

[Wa18] Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: CVPR. pp. 5265–5274, 2018.

[XZ17] Xiang, J.; Zhu, G.: Joint face detection and facial expression recognition with MTCNN. In: 2017 4th ICISCE. IEEE, pp. 424–427, 2017.

[Zh16] Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y.: Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. IEEE Signal Processing Letters, 23(10):1499–1503, Oct 2016.

[ZVS20] Zeng, D.; Veldhuis, R.; Spreeuwers, L.: A survey of face recognition techniques under occlusion. In: https://arxiv.org/abs/2006.11366. 2020.

On Brightness Agnostic Adversarial Examples Against Face Recognition Systems

Inderjeet Singh[1], Satoru Momiyama[2], Kazuya Kakizaki[3], Toshinori Araki[4]

Abstract: This paper introduces a novel adversarial example generation method against face recognition systems (FRSs). An adversarial example (AX) is an image with deliberately crafted noise to cause incorrect predictions by a target system. The AXs generated from our method remain robust under real-world brightness changes. Our method performs non-linear brightness transformations while leveraging the concept of curriculum learning during the attack generation procedure. We demonstrate that our method outperforms conventional techniques from comprehensive experimental investigations in the digital and physical world. Furthermore, this method enables practical risk assessment of FRSs against brightness agnostic AXs.

Keywords: Adversarial examples, Face recognition, Brightness variations, Curriculum learning.

1 Introduction

The recent advancement in Adversarial Machine Learning (AML) has discovered that state-of-the-art (SOTA) Deep Learning (DL) models are vulnerable to well-designed input samples called *Adversarial Examples* (AXs) [GSS14]. The vulnerability to AXs becomes a significant risk for applying deep neural networks in safety-critical applications like Face Recognition Systems (FRSs). Face Recognition is a process of validating a claimed identity based on the image of a face. An adversary can conveniently attack practical FRSs from the digital and the physical world, e.g., in ID photo-matching systems [Fo16].

In digital attacks, the digital adversarial noise is directly added to the target digital image. In the physical attacks, digital AX is transferred to the physical world (by printing, etc.) and then used to attack a target system. The generated AXs can be white-box, gray-box, or black-box depending on whether they are generated leveraging complete, partial, or no access, respectively, to the target system's information. Various digital and physical perturbations affect these AXs because the AXs are typical images with a few highly correlated adversarial features with the target ML model's predictions. The perturbations can be in color corrections,

[1] NEC Corporation, inderjeet78@nec.com
[2] NEC Corporation, satoru-momiyama@nec.com
[3] NEC Corporation, University of Tsukuba, kazuya1210@nec.com
[4] NEC Corporation, toshinori_araki@nec.com

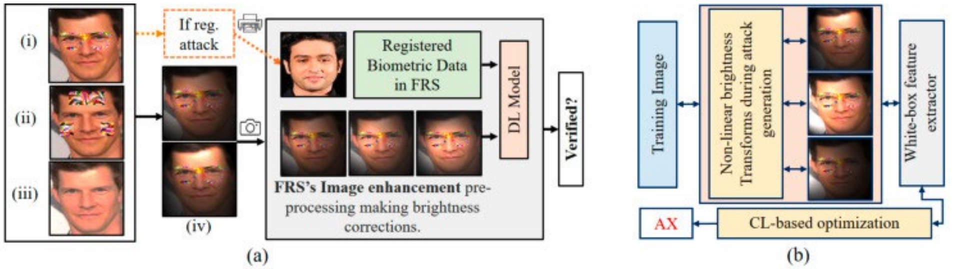

(i)
(ii)
(iii)
If reg. attack
(iv)
Registered Biometric Data in FRS
DL Model
Verified?
FRS's Image enhancement pre-processing making brightness corrections.
(a)
Training Image
Non-linear brightness Transforms during attack generation
White-box feature extractor
AX
CL-based optimization
(b)

Fig. 1: (a) illustrates an example of practical FRS and brightness corrections in input face images [Ca18] during FRS's pre-processing. (i), (ii), and (iii) illustrate eyeglass, sticker, and imperceptible noise AXs. (iv) represents face images taken under variable brightness. (b) demonstrates an outline of our proposed method.

contrast change, hue shift, and brightness changes. The brightness change is one of the critical parameters, causing a significant change in AX's performance.

The practical risk assessment of the FRSs scans the possible vulnerabilities of the ML model used in the FRS from different kinds of AXs. However, the brightness changes weaken the AX, making it non-suitable for the practical risk assessment of the target system. Thus powerful AXs robust to the brightness changes must be adopted. When an AX succeeds even in altering brightness environments, it is called *brightness agnostic AX*. In practical scenarios, brightness changes *non-linearly*. In the digital world, *non-linear* brightness changes occur due to the use of image enhancement techniques [Yi17][Re15] by FRSs for improved performance, which can be seen in `Fig.1a`. In the physical world, four primary factors cause brightness changes: printer specification, printing surface properties, environmental illumination conditions, and camera specifications.

Yang et al. [Ya21] proposed an adversarial example generation method based on random transformations of image brightness. They reduced the overfitting, thereby improving black-box transferability of generated attacks, by applying linear brightness transformations on the training[5] image optimized for an adversarial objective. However, [Ya21] did not evaluate the robustness of the generated attacks in changing brightness conditions. Also, they assumed only linear brightness changes. Additionally, the FRSs were not considered in their evaluation. Therefore, in this work, in addition to our proposed method, we also evaluate (1) the robustness of the attacks generated from their method in the changing brightness conditions, (2) the improvement in the black-box transferability, and (3) performance for FRSs.

Our main contributions are: We propose a *novel Curriculum Learning (CL)-based method* for generating AXs robust to *real-world brightness changes*. To our best knowledge, this is the very first attempt for generating brightness agnostic adversarial attacks. We conduct extensive experiments on four SOTA face verification models under a well-known PGD (Projected Gradient Descent) attack [Ma17]

[5] In AML, we call the input image optimized for adversarial objective as the training image.

Algorithm 1: CL-BA-PGD Algorithm for Adversarial Patch Attacks

```
Input: Source image $X^s$ of identity $s$; target image $X^t$ of identity $t$; face-matcher $f$;
       adversarial loss function $J_{adv}$; random noise $\delta$; patch mask $M_p$; brightness
       mask $M_b$; stopping criteria $T$; step functions $g_1$ & $g_2$; batch constant $N$;
       similarity constant $K$; number of brightness ensembles $N_b$; learning rate $\alpha$.
Output: Brightness agnostic patch adversarial example $X^{adv} = X^{adv}_{T-1}$
1  $X_0^{adv} \leftarrow X^s \cdot M_p^{'} + \delta \cdot M_p$, $l_0 \leftarrow 1$, $h_0 \leftarrow 1$, $p \leftarrow 0$, $loss_0^{cum} \leftarrow 0$
2  for $i$=0 to T-1 do
3      for $j$=0 to $N_b - 1$ do
4          $X_{i,j} \longleftarrow CNBT_j\left(X_i^{adv};p;M_p;M_p^{'};M_b;M_b^{'};l_i;h_i\right) =$
           $\left(Y_j \cdot \left(BT\left(X_i^{adv} \cdot M_p\right) + X_i^{adv} \cdot M_p^{'}\right)\right) \cdot \left(M_{b,j} \cdot X_u + M_{b,j}^{'}\right)$
5      end
6      $X_{i+1}^{adv} \longleftarrow clip_{0-1}\left(X_i^{adv} - \alpha \cdot sign\left(\sum_{j=0}^{N_b-1} \nabla J_{adv}(f(X_{i,j}), f(X^t))\right)\right)$
7      $loss_{i+1}^{cum} = loss_i^{cum} + \frac{\sum_{j=0}^{N_b-1} J_{adv}(f(X_{i,j}), f(X^t))}{N_b}$
8      $l_{i+1} \longleftarrow g_1(l_i)$; $h_{i+1} \longleftarrow g_2(h_i)$
9      if $i \neq 0$ and $i\%N = 0$ then
10         $p = max\left(0, \left(K - \frac{loss_i^{cum}}{N}\right)\right)$
11         $loss_{i+1}^{cum} = \frac{\sum_{j=0}^{N_b-1} J_{adv}(f(X_{i,j}), f(X^t))}{N_b}$
12     end
13 end
```

setting. We evaluated the *white-box* and *black-box* attack performance in the *digital* as well as the *physical world*. We also evaluate our method against the FRSs deployed with adversarial defenses in the pre-processing pipeline.

2 Our method for generating brightness agnostic AXs

The proposed method `[Alg.1]` yields *non-linear* brightness changes during the attack generation process, as it can be seen in `Fig.1b`. The non-linear change in the brightness during attack generation makes the generated AXs robust to them during inference. To better optimize the challenging non-linear brightness changes, our method uses the concept of ***CL*** for generating ***B****rightness* ***A****gnostic AXs* in the **PGD** attack setting; thus, we call our method a CL-BA-PGD attack. CL is an approach proposed by [El93] in which training difficulty is gradually increased while training DL models for better performance.

To generate attacks using our algorithm `[Alg.1]`, non-linear brightness transformations $CNBT_j()$ are applied to the training[5] image X_i^{adv} after the initialization `[Alg.1; 1]`. The transformations `[Alg.1; 4]` are applied while regulating the optimization difficulty based on the loss J_{adv}. The loss J_{adv} in gradient descent setting calculates the inverse of cosine similarity between the predictions of f for X^s and X^t for the impersonation attacks and simply similarity for dodging attacks. For impersonation attacks, the adversary with identity s, tries to mimic the deep features of target identity t. For dodging attacks, s and t are same because adversary tries to minimize the similarity from its clean image's deep features. The predefined step-functions g_1 and g_2 change lower l_i and upper h_i limits for a uniform random variable $X_u \sim U(l_i, h_i)$, thus controlling the non-linear brightness changes.

The function BT changes the brightness of an image tensor X as $BT(X) = X_u \cdot X$ with probability p. The 0-1 mask $M_{b,j}$ with the same dimensions as the X_i^{adv}, randomly chooses a rectangular area $\mathcal{R}_b$ inside X_i^{adv} in each j^{th} iteration to scale brightness by X_u. Thus, $(M_{b,j})_{(m,n)} = 1$ if $(m,n) \in \mathcal{R}_b$ and $(M_{b,j})_{(m,n)} = 0$ if $(m,n) \notin \mathcal{R}_b$. The patch masks M_p is used to separate the predefined patch area inside X_i^{adv}. Also, $M_p^{'} = I_1 - M_p$ and $M_{b,j}^{'} = I_1 - M_{b,j}$ where I_1 is all one matrix. The random variable $Y_j \sim N(\mu_j, \sigma_j)$ follows Gaussian distribution.

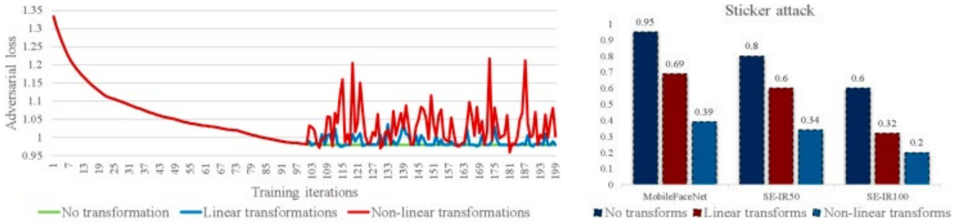
The PGD updates are then performed on X_i^{adv} following [Alg.1; 6]. Note that it is assumed that images are normalized in the [0,1] range. The parameters responsible for the CL are updated in the subsequent steps [Alg.1; 7,8,9,11] following the idea of gradually changing the amount of brightness changes depending upon J_{adv}. The parameter K is loss function specific and serves to provide a margin for the minimum values of the p parameter.

2.1 Sorting optimization difficulty

We define the optimization difficulty in the i^{th} iteration of the attack generation process as directly proportional to $\Delta\mathcal{L}^k_{V_B^{\cdot}}$, which is the change in the adversarial loss caused by variation in the brightness $V_B^{\cdot}$ of input image due to application of a $(\cdot)$-type transformation. The large change in the adversarial loss causes significant variations in gradient-based methods' descent direction, making the optimization process challenging. Also, $\Delta\mathcal{L}^k_{V_B^{\cdot}}$ is calculated for k training[5] images for a DL model f trained for $t \leq T$ iterations.

We applied linear and non-linear brightness transformations to a face image with adversarial eyeglass patch noise to assess the increased training[5] difficulty. We saw from Fig.2a that maximum variations in the adversarial loss (hence the optimization difficulty) was caused by non-linear brightness transformations followed by linear and no brightness transformations, i.e. $\Delta\mathcal{L}^k_{V_B^{nl}} > \Delta\mathcal{L}^k_{V_B^{l}} > \Delta\mathcal{L}^k_{V_B^{0}} = 0$. Thus our hypothesis is that increased optimization difficulty reduces the chances of convergence to optimal solutions thereby reducing attack success probability.

To investigate the effect of the brightness changes and adversarial loss variations on the adversarial potential of successful AXs, we evaluated the reduction in attack success rate (ASR) due to linear [Ya21] and non-linear brightness transformations. ASR is the fraction of AXs which successfully fooled the DL model during inference.



(a) (b)

Fig. 2: (a) depicts the variation of the adversarial loss for the trained (till 100 iterations) MobileFaceNet [Ch18] when subjected to no, linear, and non-linear brightness transformations. (b) demonstrates the performance (ASR) reduction due to the linear and non-linear brightness transformations for sticker attacks.

After evaluating eyeglass, sticker, and imperceptible noise attacks, the results for the ASR confirmed our hypothesis that non-linear brightness transformations cause a significant reduction in the ASR compared to the linear transformations.

3 Experimental Setup

For an adequate assessment, we considered *four* SOTA feature extractors: Residual Network (ResNet50) [He16], MobileFaceNet [Ch18], and Squeeze-and-Excitation Inception Residual Networks (SE-IR50, SE-IR100) [HSS18]; trained on the Vggface2 data [Ca18] using the arcface loss. The test accuracy on Vggface2 data of trained ResNet50, MobileFaceNet, SE-IR50, and SE-IR100 was found as 99.03%, 99.17%, 99.01%, and 99.02%, respectively. For each feature extractor, we implemented a simple PGD attack [Ma17], the existing method [Ya21], our method without CL, and our method with CL `[Alg.1]`.

We implement our algorithm for the patch and the imperceptible noise attacks while evaluating in digital and physical worlds. The patch attacks were generated using `[Alg.1]`. For the imperceptible noise attacks, the adversarial noise δ with size constraints ($\delta \leq (4/255)^{th}$ of input image's pixel range) was distributed over the entire area of the face image. In this case, we implemented `[Alg.1]` by changing `[Alg.1; 4]` with $CNBT_j\left(\cdot\right) = Y_j \cdot \left(BT\left(X_i^{adv} \cdot M_{b,j}\right) + X_i^{adv} \cdot M_{b,j}^{'}\right)$.

To evaluate the white-box and black-box performance of the generated attacks, the white-box attacks were generated and tested directly on the target FRS `(Tab.1)`. In contrast, the black-box attacks were generated using a surrogate FRS and tested on the target FRS `(Tab.2)` to evaluate the transferability of generated attacks.

For adversarial patch attacks, we considered an eyeglass patch (`Fig.1a(i)`) and a sticker patch (`Fig.1a(ii)`). For the practical evaluation of the generated attacks (100 AXs for each case) in the *digital domain*, the mean ASR for each AX was calculated after applying `[Alg.1; 4]` transformations 100 times to simulate the practical brightness variations. For the evaluation in the *physical world*, the fol-

Attack Type	White-box model	Mean IASR				Mean DASR			
		A1	A2	A3	A4	A1	A2	A3	A4
Eyeglass Attack	ResNet50	0.39	0.42	0.56	**0.58**	0.63	0.66	0.74	**0.78**
	MobileFaceNet	0.44	0.51	0.53	**0.55**	0.52	0.60	0.63	**0.69**
	SE-IR50	0.41	0.48	0.51	**0.60**	0.56	0.65	0.62	**0.73**
	SE-IR100	0.29	0.37	0.40	**0.45**	0.42	0.49	0.48	**0.53**
Sticker Attack	ResNet50	0.54	0.69	0.72	**0.86**	0.52	0.69	0.85	**0.95**
	MobileFaceNet	0.54	0.66	0.68	**0.75**	0.48	0.62	**0.68**	0.66
	SE-IR50	0.49	0.59	0.64	**0.69**	0.42	0.48	0.52	**0.60**
	SE-IR100	0.43	0.55	**0.65**	0.60	0.41	0.50	0.63	**0.70**
Imperceptible Noise Attack	ResNet50	0.34	0.43	**0.51**	0.50	0.44	0.55	0.44	**0.62**
	MobileFaceNet	0.30	0.33	0.36	**0.45**	0.38	0.44	0.52	**0.63**
	SE-IR50	0.21	0.25	0.38	**0.52**	0.47	0.57	0.60	**0.63**
	SE-IR100	0.24	0.26	0.28	**0.34**	0.16	0.23	0.32	**0.52**

Tab. 1: The mean ASR of the ***white-box*** attacks in the digital domain. DASR & IASR are Dodging & Impersonation ASRs. A_1, A_2, A_3, and A_4 stands for naive, existing, our method w/o CL, and our CL-based methods, respectively.

lowing steps were followed: (1) Generate digital AX.(2) Transfer it to the physical world by printing at 9 different brightness levels. (3) Capture the printed AXs from various angles. (4) Clean the captured data. We got approximately 20 images for each captured image. (5) Feed the data to the MTCNN face detection and alignment [XZ17]. (6) Feed the preprocessed data to the target feature extractor and check the predictions.

4 Results

`Tab.1` and `Tab.2` shows the results for white-box and black-box ASRs, respectively, in the digital domain. The mean ASR is calculated for 100 AXs after applying the transformations mentioned in section 3 on each AX. Our method with CL results in a significantly higher ASR as compared to the existing method [Ya21] and the naive method. Also, the existing method achieves better results than the naive method. The effect of better optimization due to CL can also be seen from the increased ASR from `A3` to `A4` columns of `Tab.1`. Our method also achieves better ASR for the digital black-box attacks `(Tab.2)`. However, in this case, the performance difference was not as significant as in the white-box setting. Additionally, sticker attacks were found to be having the highest ASRs (`Tab.1` and `Tab.2`) due to the larger area for the adversarial noise region and absence of imperceptible size constraints.

The evaluation of the generated attacks in the *physical domain* also exhibited similar patterns as *digital white-box ASR* `(Tab.1)`. Our method achieves 24.67% and 39.96% better mean ASR than the existing [Ya21] and the naive PGD attack meth-

Attack Type	Surrogate model	Black-box models	Mean IASR				Mean DASR			
			A1	A2	A3	A4	A1	A2	A3	A4
Eyeglass Attack	M_1	M_2, M_3	0.07	**0.08**	0.05	**0.08**	0.1	0.11	0.11	**0.13**
	M_2	M_3, M_4	0.13	**0.14**	**0.14**	**0.14**	0.16	0.20	0.20	**0.23**
	M_3	M_2, M_4	0.16	0.16	0.15	**0.17**	0.34	0.40	0.34	**0.45**
	M_4	M_2, M_3	0.14	0.19	0.18	**0.21**	0.25	0.35	**0.40**	0.37
Sticker Attack	M_1	M_2, M_3	0.04	0.05	**0.06**	0.04	0.18	0.17	0.19	**0.20**
	M_2	M_3, M_4	0.11	0.12	0.12	**0.13**	0.24	**0.40**	0.27	0.35
	M_3	M_2, M_4	**0.15**	0.14	**0.15**	0.13	0.38	0.40	**0.53**	0.50
	M_4	M_2, M_3	0.19	0.24	0.21	**0.25**	**0.58**	0.46	0.52	**0.58**
Imperceptible Noise Attack	M_1	M_2, M_3	0.08	0.10	0.11	**0.17**	0.09	0.12	0.11	**0.18**
	M_2	M_3, M_4	0.22	0.29	0.23	**0.32**	0.19	0.18	**0.26**	0.18
	M_3	M_2, M_4	0.19	0.20	0.22	**0.34**	0.34	0.36	0.32	**0.38**
	M_4	M_2, M_3	0.15	0.14	0.27	**0.24**	0.25	0.39	0.37	**0.43**

Tab. 2: The ASR for the ***black-box*** attacks in the digital domain. DASR & IASR are Dodging & Impersonation ASRs. M_1, M_2, M_3, and M_4 represents ResNet50, MobileFaceNet, SE-IR-50, and SE-IR-100 face feature extractors. A_1, A_2, A_3, and A_4 stands for naive, existing, our method w/o CL, and our CL-based methods, respectively.

ods, respectively, for the eyeglass patch attack. Additionally, we evaluate the robustness of the brightness agnostic AXs against the model with JPEG compression [DGR16], bit squeezing, and median blur defenses [XEQ17] in the pre-processing pipeline. These defenses do not directly cause brightness changes in the input images. After evaluation, we did not find sufficient evidence to validate the better ASR of the brightness agnostic AXs generated using our method against them.

5 Conclusions

This paper contributed a novel CL-based method for generating AXs robust to the practical brightness changes. While considering attacks from digital and physical worlds, we found that our approach significantly exceeds the conventional techniques in white-box and black-box settings from our detailed analysis of the dodging and impersonation attacks. However, we did not find sufficient evidence for the superiority of our method against adversarial defenses that do not cause a direct change in the brightness of input images. A possible weakness of our approach is that it requires careful manual initialization of a few hyper-parameters responsible for CL that can directly affect attack performance. The generated attacks by our method enable practical risk assessment of the FRSs against such attacks. In the future, we would like to consider utilizing color space transformations, and assessing provided robustness improvements through adversarial training by our method.

References

[Ca18] Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; Zisserman, A.: VGGFace2: A dataset for recognising faces across pose and age. In: International Conference on Automatic Face and Gesture Recognition. 2018.

[Ch18] Chen, Sheng; Liu, Yang; Gao, Xiang; Han, Zhen: Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In: Chinese Conference on Biometric Recognition. Springer, pp. 428–438, 2018.

[DGR16] Dziugaite, Gintare Karolina; Ghahramani, Zoubin; Roy, Daniel M: A study of the effect of jpg compression on adversarial images. arXiv preprint arXiv:1608.00853, 2016.

[El93] Elman, Jeffrey L: Learning and development in neural networks: The importance of starting small. Cognition, 48(1):71–99, 1993.

[Fo16] Folego, Guilherme; Angeloni, Marcus A; Stuchi, José Augusto; Godoy, Alan; Rocha, Anderson: Cross-domain face verification: matching ID document and self-portrait photographs. arXiv preprint arXiv:1611.05755, 2016.

[GSS14] Goodfellow, Ian J; Shlens, Jonathon; Szegedy, Christian: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

[He16] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778, 2016.

[HSS18] Hu, Jie; Shen, Li; Sun, Gang: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141, 2018.

[Ma17] Madry, Aleksander; Makelov, Aleksandar; Schmidt, Ludwig; Tsipras, Dimitris; Vladu, Adrian: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017.

[Re15] Ren, Dafeng; Ma, Hui; Sun, Laijun; Yan, Tingchun: A novel approach of low-light image used for face recognition. In: 2015 4th International Conference on Computer Science and Network Technology (ICCSNT). volume 1. IEEE, pp. 790–793, 2015.

[XEQ17] Xu, Weilin; Evans, David; Qi, Yanjun: Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv preprint arXiv:1704.01155, 2017.

[XZ17] Xiang, Jia; Zhu, Gengming: Joint face detection and facial expression recognition with MTCNN. In: 2017 4th International Conference on Information Science and Control Engineering (ICISCE). IEEE, pp. 424–427, 2017.

[Ya21] Yang, Bo; Xu, Kaiyong; Wang, Hengjun; Zhang, Hengwei: Random Transformation of Image Brightness for Adversarial Attack. arXiv preprint arXiv:2101.04321, 2021.

[Yi17] Ying, Zhenqiang; Li, Ge; Ren, Yurui; Wang, Ronggang; Wang, Wenmin: A new image contrast enhancement algorithm using exposure fusion framework. In: International Conference on Computer Analysis of Images and Patterns. Springer, pp. 36–46, 2017.

Learning by Environment Clusters for Face Presentation Attack Detection

Tomoaki Matsunami[1], Hidetsugu Uchida[2], Narishige Abe[3], Shigefumi Yamada[4]

Abstract: Face recognition has been used widely for personal authentication. However, there is a problem that it is vulnerable to a presentation attack in which a counterfeit such as a photo is presented to a camera to impersonate another person. Although various presentation attack detection methods have been proposed, these methods have not been able to sufficiently cope with the diversity of the heterogeneous environments including presentation attack instruments (PAIs) and lighting conditions. In this paper, we propose Learning by Environment Clusters (LEC) which divides training data into some clusters of similar photographic environments and trains bona-fide and attack classification models for each cluster. Experimental results using Replay-Attack, OULU-NPU, and CelebA-Spoof show the EER of the conventional method which trains one classification model from all data was 20.0%, but LEC can achieve 13.8% EER when using binarized statistical image features (BSIFs) and support vector machine used as the classification method.

Keywords: face anti-spoofing, presentation attack detection, face image clustering.

1 Introduction

Face recognition has been used widely such as access control of personal use devices and the border controls because of its high accuracy and convenience. However, it is vulnerable to Attack Presentations (APs) that try to impersonate someone by presenting a photo, a video, or other item. In recent years, since target face images such as photos to impersonate can be easily obtained through social networking service, attackers can easily try to authenticate using the obtained photos as PA. That is why the presentation attack detection (PAD) could be mandatory to realize the secure face recognition.

The PAD approaches include hardware-based or software-based, and the software-based approach can be divided into motion-based method and methods using image features [RB17]. The hardware-based approach uses dedicated equipment to obtain the information for classification real and fake, Raghavendra et al. used light field cameras [RRB15] and Kose et al. used depth cameras [KD13]. As a motion-based method, Kollreider et al. proposed a method using facial expression changes [KFF07] and

[1] Fujitsu Limitd, 4-1-1 Kamikodanaka Nakahara-ku Kawasaki Kanagawa Japan, t.matsunami@fujitsu.com
[2] Fujitsu Limitd, 4-1-1 Kamikodanaka Nakahara-ku Kawasaki Kanagawa Japan, u.hidetsugu@fujitsu.com
[3] Fujitsu Limitd, 4-1-1 Kamikodanaka Nakahara-ku Kawasaki Kanagawa Japan, abe.narishige@fujitsu.com
[4] Fujitsu Limitd, 4-1-1 Kamikodanaka Nakahara-ku Kawasaki Kanagawa Japan, yamada.shige@fujitsu.com

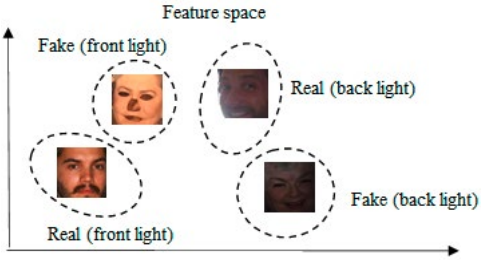


Fig. 1. Illustration of face feature distribution acquired with different light sources

Frischholz et al. proposed a method using head-pose [FW03]. As a method for realizing PAD against a RGB camera image commonly used in face recognition, there are many methods using image features [Wa13]. At first, there are many methods to classify images into two classes of real and fake based on texture features such as local binary feature (LBP) by Pereira et al. [Pe 13]. Haralick feature (HF) by Agarwal et al. [ASV 16], and binarized statistical image features (BSIF) by Soler et al. [SBB 20] and Boulkenafet et al. evaluated combinations of multiple texture features [BKH 18]. Moreover, many techniques using deep learning have also been proposed, such as convolutional neural network (CNN) based [YLL14, PHJ16] and long short-term memory (LSTM) [SSL18]. On the other hand, since various PAIs are used, there is an issue that fake images have a large diversity, and the distribution of fake features is complicated. To solve this issue, 1-class classification method which trains only 1 class of real has been proposed. In the 1-class classification, by training fake as an outlier, real can be identified with high accuracy even if fake feature distribution becomes complicated. Agarwal et al. proposed 1-class SVM [AKW 17], Nikisins et al. proposed 1-class GMM [Ni 18], Arashloo proposed 1-class FV [AR 20], and Bawja et al. proposed 1-class CNN method [Ba 20].

However, in the actual face recognition, there are factors that increase the distribution of features for both real and fake. One of the most prominent examples is the change in the lighting environment, such as the position and intensity of the light source, and this paper focuses on the change of light. For example, as illustrated in Fig. 1, the distance due to the change in the lighting environment can be larger than the distance between real and fake in the feature space. In the existing methods using image features in such a case, the fake is included in the real distribution and the real is included in the fake distribution, and the boundary between the real and the fake cannot be accurately estimated. Therefore, the purpose of our study is to focus on PAD using image features from RGB images, and to train a model that can accurately classify between real and fake, even for training data that includes variations of PAIs and lighting environment. Hereinafter, in this paper, PAI and lighting environment are referred to as photographing environment. The main contributions of our work are as follows:

- We propose Learning by Environment Clusters (LEC) which divides training data into some clusters of similar photographic environments and trains real and fake classification models for each cluster.

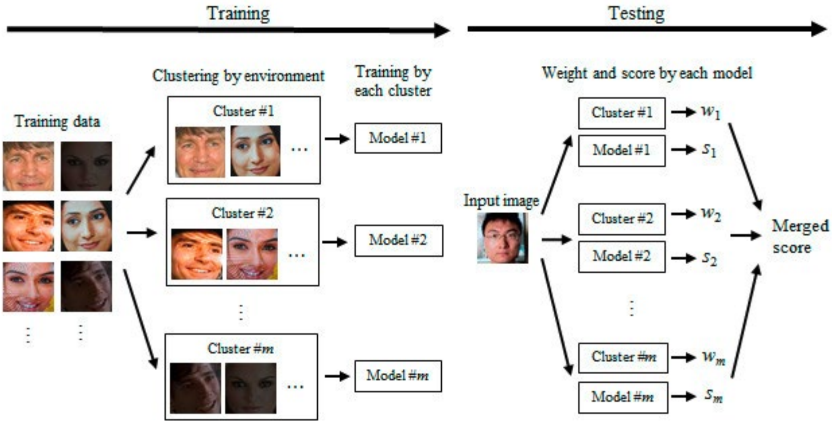


Fig. 2. Process flow of the proposed method

- Using hierarchical cluster analysis, LEC can classify training data into environment clusters even when the photographing environment of them is not known.

- We confirmed the effectiveness of the proposed method on HF+SVM, BSIF+SVM, and BSIF+CNN as a classification method using three PAD datasets, Replay-Attack [CAM12], OULU-NPU [Bo17], and CelebA-Spoof [Yu20].

The structure of this paper is as follows. Chapter 2 shows the protocol of the proposed method. Chapter 3 describes the evaluation results of the proposed method. Finally, Chapter 4 summarizes the results and discusses future prospects.

2 Learning by Environment Clusters

Fig. 2 shows the process flow of the proposed method. In the training phase, the training data is divided into clusters for each similar photographing environment, and a classification model is trained for each cluster. It is assumed that the training data includes images acquired under various photographing environments and that the photographing environment of each image is not given. When the PAI or the lighting environment changes, the way of reflection changes and appears as a local change. LBP can represent the local change in the image, that’s why we adopt the LBP as a descriptor. Since the distance of the LBP histogram becomes small between images acquired in the similar environment, clustering is enabled even when the photographing environment of the training data is not given by using the LBP histogram. In the LBP histogram, the $LBP_{8,1}$ proposed in [OPM02] is calculated from the HSV and YCbCr channels of the face image, and a 354-dimensional feature vector is obtained. Next, the training images are divided into clusters using LBP histograms by hierarchical cluster analysis (HCA) [Wa63]. In HCA, combining the most similar clusters is repeated with N clusters each including one LBP histogram as a start. The Ward‘s method was used to calculate the

```
Algorithm  Algorithm for hierarchical cluster analysis
Input: $N$ LBP histograms
Output: Clusters by environment
Initialize: Generate $N$ clusters, each of which contains 1 LBP histogram and set $nc$ to the number of clusters ($nc = N$)
Definition: cluster distance $cd(C_i, C_j) = L(C_i \cup C_j) - L(C_i) - L(C_j)$, where $L(C_i)$ is sum of squares of Euclid distance between each LBP histogram contained in $C_i$ and LBP histogram centroid of $C_i$
Calculate $mcd$ as minimum of $cd$ for all combinations of 2 clusters
while $mcd <$ threshold of $cd$
    Conbine $C_i$ and $Cj$ where $cd(C_i, C_j)$ is equal to $mcd$ into $C_k$, where $k$ is new ID
    Update $L(C_k) \leftarrow L(C_i \cup C_j)$ and $nc \leftarrow nc - 1$
    Update $mcd$ in new $nc$ clusters
end while
```

similarity between two clusters. The algorithm for HCA is shown in Algorithm 1. HCA can divide training data into the arbitrary number of clusters by reaching at the end condition. In this paper, the termination condition was determined as the time when the minimum value of *cd* exceeded the threshold value from the pre-experimental results. By using HCA, even when it is not known how many face images of what kind of photographing environment are included in the training data, a cluster for each similar photographing environment can be generated. In the following, the number of clusters in the generated photographing environment is represented by m, and the *i*-th cluster is represented by C_i. It then learns a model that classifies real and fake for each of the m clusters. There is no particular limitation on the learning method of the model. In this paper, we verified SVM learning using hand-crafted features and CNN learning. As a result of the learning phase of the proposed method, a model specific to the photographing environment indicated by each cluster is generated, and the model learned by the cluster C_i is represented by M_i. In the test phase, a score S_i, which is an output of the model M_i for an input image, and a weight w_i of the score are calculated in an i ($1 \leq i \leq m$) th photographing environment clustered during learning. w_i is calculated as $w_i = 1/Ed_i$, where Ed_i is Euclidean distance between the LBP histogram of the input image and the center of gravity of C_i. By making the w_i larger when the distance between the input image and C_i is small, even if it is unclear in what kind of shooting environment the input image was acquired, real and fake can be determined for various photographic environments by calculating a merged score where the weight of the result of a model in a similar photographic environment is increased from the training data.

3 Experiments

The experimental protocol aims to address that our proposed method (LEC) is effective regardless of datasets or classification models. The experimental evaluation

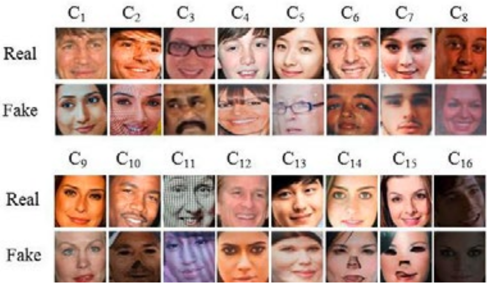


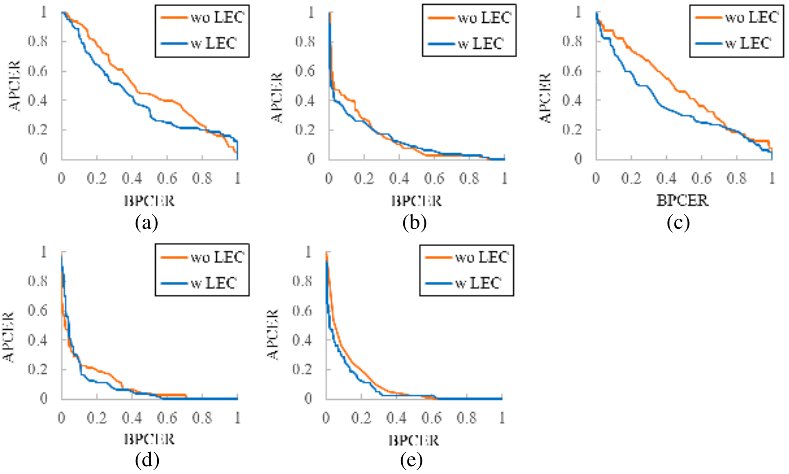
Fig. 3. Image examples in the each cluster

was conducted over three PAD datasets, Replay-Attack [CAM12], OULU-NPU [Bo17], and CelebA-Spoof [Yu20]. Replay-Attack includes photos and videos (replay), OULU-NPU includes prints and videos, and CelebA-Spoof includes prints, videos, and paper cuts. For each data set, MTCNN [Zh16] is used as a face detector, and each detected face is aligned to a size of 112 x 112. Replay-Attack and OULU-NPU are video data sets, so we extract one frame of face image per a video. Either OULU-NPU or CelebA-Spoof is used for training, and CelebA-Spoof is used for testing. When the minimum value of *cd* is more than 200 in HCA, the clustering is terminated to generate four clusters for Oulu-NPU and 16 clusters for CelebA-Spoof. Fig. 3 shows examples of face images included in each cluster C_i ($1 \leq i \leq 16$) of the CelebA-Spoof. The photographic environment is divided into clusters such that C_1 is less affected by the lighting, C_{11} is shaded on the surface, and C_{16} is backlit. In terms of the classification model, we use three models generated by HF + SVM, BSIF + SVM, and BSIF + CNN. As described in [ASV16] for HF, RDWT is applied to each RGB channel of a face image to obtain four sub-bands, and the original image and the four sub-bands are divided into 3x4 patches respectively, and 13 features are calculated from each patch to obtain feature vectors with 2,340 (= 3 x 5 x (3x4) x 13) dimensions. For BSIF, there are 60 pre-learned filters from natural images [KR 12]. We use one of the filters with a filter size of 7 x 7 and a filter number of 8 and obtain 768-dimensional feature vectors by extracting features from 6 channels of HSV and YCbCr. The SVM determined the hyperparameters by 5-fold cross validation using a linear kernel. CNN is a one-dimensional CNN having three convolution layers with 768-dimensional feature vectors of BSIF as input, and a batch normalization is inserted immediately after a second or third convolution layer. The activation function is Leaky ReLU (Negative slope factor is 0.2) and the dropout rate was 0.25. First, a model for each cluster is obtained by transfer learning of all coupling layers for the data of each cluster using a model trained with all data.

We use the evaluation protocol based on the international standard ISO/IEC 30107-3 [ISO17] using attack presentation classification error rate (APCER) which indicates the rate at which the attack (fake) is erroneously determined as Bona-Fide (real), the bona fide presentation classification error rate (BPCER), which indicates the rate at which the bona fide is erroneously determined as fake, the equal error rate (EER), which is the error rate at APECR = BPCER, and the half total error rate (HTER), which is the average of APECR and BPCER.

Tab.1 Evaluation results on LEC using public datasets

Train	Test	Classification method	EER		HTER	
			woLEC	wLEC	woLEC	wLEC
OULU-NPU	Replay-Attack	HF + SVM	45.0%	41.5%	44.6%	39.6%
CelebA-Spoof	Replay-Attack	HF + SVM	23.8%	22.5%	22.4%	20.6%
OULU-NPU	Replay-Attack	BSIF + SVM	46.3%	37.8%	46.0%	36.6%
CelebA-Spoof	Replay-Attack	BSIF + SVM	20.0%	13.8%	16.8%	13.8%
CelebA-Spoof	Replay-Attack	BSIF + CNN	20.0%	16.3%	19.1%	15.9%



(a) (b) (c) (d) (e)

Fig. 4. DET curves of (a) HF + SVM trained with Oulu-NPU, (b) HF + SVM trained with CelebA-Spoof, (c) BSIF + SVM trained with Oulu-NPU, (d) BSIF + SVM trained with CelebA-Spoof, and (e) BSIF + CNN trained with CelebA-Spoof.

The evaluation results are shown in Tab. 1. by comparing proposed method (wLEC) with conventional method when one classification model is generated using all the training data (woLEC). Fig. 4 shows the respective detection error tradeoff (DET) curves. LEC improves both EER and HTER for both datasets and feature extraction methods. From these results, it can be said that the classification model generated from each cluster by LEC is specialized for each photographic environment, and that by increasing the weight of the result obtained from the model of the photographic environment similar to the input image, a robust classification method for the change of the photographic environment can be realized. Further, in the proposed technique, the optimal number of divisions of the training data depends on the diversity of the photographic environment of the training data, and when the number of divisions is small with respect to the diversity of the training data, the model is learned from a data set having a large diversity, and a problem similar to the case of no division occurs. On the contrary, when the number of divisions is large, a plurality of models of similar environments are learned, and the effect of division is reduced. Therefore, it is considered optimal to dynamically determine the number of partitions according to the distance between clusters as we proposed. LEC can be applied regardless of the feature extraction method or discriminator, therefore we used typical hand-crafted features, SVM, and shallow CNN for evaluation in this paper. Experimental results show the effectiveness of LEC in all combinations, hence we consider that LEC is effective even when the state-of-the-art method is used for feature extraction and discriminator.

4 Conclusion

In this paper, in order to deal with the various Presentation Attack Instruments and various photographing environment by lighting environment at the time of face image acquisition in PAD, we proposed LEC which divides training data into clusters depending on the photographing environment and trains the classification model of real and fake for each cluster. Experimental results using Replay-Attack, OULU-NPU and CelebA-Spoof shows the effectiveness of the proposed method, for example, EER of the conventional method, which trains one classification model from all data, was 20.0%, while the accuracy was improved to 13.8% EER using LEC in the case of using BSIF + SVM. As a future work, the proposed method is evaluated by focusing on photographs and videos using the 2-class classification method in this paper, however, it is necessary to confirm whether the proposed method can be used in a more general way by verifying it using the 1-class classification method and evaluating it using various datasets of different types of presentation attack such as 3D face masks.

References

[AKW17] Arashloo, S. R.; Kittler, J.; Christmas, W.: An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol. IEEE Access, 5: pp. 13868 -13882, 2017.

[Ar20] Arashloo, S. R.: Unseen face presentation attack detection using sparse multiple kernel fisher null-space, IEEE Trans. on Circuits and Systems for Video Technology, 2020.

[ASV16] Agarwal, A.; Singh, R.; Vatsa, M.: Face Anti-Spoofing using Haralick Features, IEEE 8th Int. Conf. on biometrics theory, applications and systems, 2016.

[Ba20] Bawja, Y. et. al.: Anomaly Detection-Based Unknown Face Presentation Attack Detection, International Joint Conference on Biometrics 2020.

[BKH18] Boulkenafet, Z.; Komulainen, J.; Hadid, A: On the generalization of color texture-based face anti-spoofing. Image and Vision Computing, 2018

[Bo17] Boulkenafet, Z. et. al.: OULU _ NPU: A mobile face presentation attack database with real-world variations, in Proc. FG, pp. 612 – 618, 2017.

[CAM12] Chingovska, I.; Anjos, A.; and Marcel, S.: On the efficiency of local binary patterns in face anti-spoofing, International Conference of the Biometrics Special Interest Group 2012.

[FW03] Frischholz, R. W.; Werner, A.: Avoiding replay-attacks in a face recognition system using head-pose estimation, IEEE International Workshop on Analysis and Modeling of Faces and Gestures, 2003.

[ISO17] ISO/IEC JTC 1 SC 37 Biometrics. ISO/IEC FDIS 30107 -3. Information Technology - Biometric presentation attack detection - Part 3: Testing and Reporting. The International Organization for Standardization, 2017.

[KD13] Kose, N.; Dugelay, J.-L.: Countermeasure for the protection of face recognition systems against mask attacks, IEEE Intl. Conf. on Automatic Face and Gesture Recognition, 2013.

[KFF07] Kollreider, K: et. al.: Real-time face detection and motion analysis with application in liveness assessment, IEEE Trans. Inf. Forensics Security. 2, 2007.

[KR12] Kannala, K.; Rahtu, E.: BSIF: Binarized statistical image features, in Proc. IEEE International Conference on Pattern Recognition, Nov. 2012, pp. 1363 – 1366.

[Ni18] Nikisins, O. et. al.: On efficiency of anomaly detection approaches against unseen presentation attacks in face anti-spoofing, International Conference on Biometrics, 2018

[OPM02] Ojala, T.; Pietikäinen, M.; Mäenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Transactions on Pattern Analysis and Machine Intelligence. 24 (2002) pp. 971 – 987.

[Pe13] Pereira, T. de Freitas, et. al.: Can face anti-spoofing countermeasures work in a real world scenario?, International Conference on Biometrics, 2013.

[PHJ16] Patel, K.; Han, H.; Jain, A. K.; Cross-database face anti spoofing with robust feature representation, Chinese Conference on Biometric Recognition, 2016.

[RB17] Raghavendra, R.; Busch, C.: Presentation Attack Detection Methods for Face Recognition Systems: A Comprehensive Survey. ACM Computing Surveys, 2017.

[RRB15] Raghavendra, R.; Raja, K. B.; Busch, C.: Presentation attack detection for face recognition using light field camera, IEEE Trans. Image Process. 24, 2015.

[SBB20] Soler, L. J. G.-; Barrero, M. G.-; Busch, C.: Fisher Vector Encoding of Dense-BSIF Features for Unknown Face Presentation Attack Detection, International Conference of the Biometrics Special Interest Group, 2020.

[SSL18] Sun, Z.; Sun, L.; and Li, Q.: Investigation in Spatial-Temporal Domain for Face Spoof Detection, International Conference on Acoustics, Speech and Signal Processing 2018.

[Wa63] Ward, J. H.: Hierarchical Grouping to Optimize an Objective Function, Journal of the American Statistical Association. 58 (301), pp. 236–244, 1963.

[Wa13] Waris, M.A. et. al.: Analysis of textual features for face biometric anti-spoofing, the 21st European Signal Processing Conference, 2013.

[YLL14] Yang, J.; Lei, Z.; and Li.S.Z.: Learn convolutional neural network for face anti-spoofing. arXiv preprint arXiv: 1408.5601, 2014.

[Yu20] Yuanhan, Z. et. al.: CelebA-Spoof: Large-Scale Face Anti-Spoofing Dataset with Rich Annotations. European Conference on Computer Vision 2020.

[Zh16] Zhang K. et. al.: Joint face detection and alignment using multitask cascaded convolutional networks. Signal Processing Letters (SPL), pp. 1499 – 1503, 2016.

The relative contributions of facial parts qualities to the face image utility

Biying Fu [1], Cong Chen [1], Olaf Henniger [1], Naser Damer [1, 2]

Abstract: Face image quality assessment predicts the utility of a face image for automated face recognition. A high-quality face image can achieve good performance for the identification or verification task. Some recent face image quality assessment algorithms are established on deep-learning-based approaches, which rely on face embeddings of aligned face images. Such face embeddings fuse complex information into a single feature vector and are, therefore, challenging to disentangle. The semantic context however can provide better interpretable insights into neural-network decisions. We investigate the effects of face subregions (semantic contexts) and link the general image quality of face subregions with face image utility. The evaluation is performed on two difficult large-scale datasets (LFW and VGGFace2) with three face recognition solutions (FaceNet, SphereFace, and ArcFace). In total, we applied four face image quality assessment methods and one general image quality assessment method on four face subregions (eyes, mouth, nose, and tightly cropped face region) and the aligned faces. In addition, the effect of fusion of different face subregions was investigated to increase the robustness of the outcomes.

Keywords: Face subregions, image quality assessment, face image utility, face image quality.

1 Introduction

Automatic face recognition (FR) can facilitate and accelerate the process of FR, by making contact-less verification on the fly possible. The quality score of a biometric sample should express the utility of the sample, i.e. its usefulness for telling mated and non-mated samples apart [ISO16]. Therefore, face image quality assessment (FIQA) algorithms are devised to assess the quality/utility of a face image towards the FR solution. A review of publications [Sc21] shows that many of these algorithms work with learned embeddings that contain complex information of an aligned face image in a single discriminant vector.

Towards explainable AI, it is important to understand the neural network's decision in detail. However, it is often hard to disentangle complex feature vectors into separate explainable segments. Therefore, in this paper, we intend to gain insights into the contributions of face subregions to the face image quality (FIQ) by using a general image quality (IQ) measure on face subregions. It is interesting to understand which face subregion contributes the most to the neural network results. The goal is to relate semantic context to better interpretable FIQ metrics.

[1] Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany, biying.fu@igd.fraunhofer.de

[2] Technical University of Darmstadt, Department of Computer Science, Darmstadt, Germany

2 Related work

To enable explainability, researchers try to visualize the output of neural networks by using gradients of the activation layers [DG17] or back-propagating the classification score-based attention areas [Zh16] in the input of the network. Not much visualization for the output of FIQA methods are found in the current literature. The decisions of FIQA methods, such as RankIQ[Ch15] and SER-FIQ[Te20a] cannot be easily visualized due to their complexity in network structures. FIQA methods apply only to aligned images and their vector embeddings are learned from different loss functions, whereas general image quality assessment (IQA) is a more generalized approach that need not be applied to aligned face images.

Lienhard et al. [LLC15] investigated the aesthetic quality of face images by using 15 features that depict technical aspects of images such as contrast, sharpness, or colourfulness computed on different subregions (face, eyes, mouth) and showed the usefulness of subregions with respect to the aesthetic quality of face images. In this paper, to study the effect of face subregions on the face image utility, we evaluated the following setups: 1) using general IQ scores on face subregions, 2) using general IQ scores on aligned face images, 3) fusing the general IQ scores of face subregions, and 4) to put the achieved results in perspective, using FIQA methods on aligned face images. We introduce the methodology of our experiments in Section 3. The evaluation and a thorough discussion of the results are in Section 4. We conclude with the main message from this research work in Section 5.

3 Methods

3.1 Image quality assessment on face subregions

We use the MTCNN (multitask cascaded convolutional neural network) framework [XZ17] to detect the face in the input image. The face images are aligned and standardized to 260 × 260 pixels, such that the eyes, nose, and mouth are on the same relative position within the aligned images. Based on these standardized outputs, we cut the face regions into separate face subregions: eyes, nose, and mouth regions. Several examples of aligned face images are seen in Fig. 1 illustrating the regions of interest used in this paper.

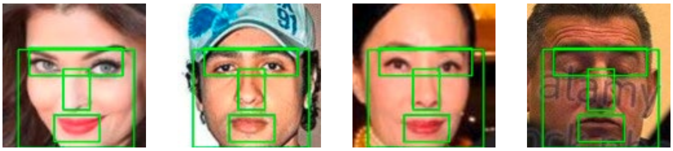


Fig. 1: Examples of aligned face images from VGGFace2 with the face subregions indicated by rectangles. The relative positions of the regions remain the same in the standardized face images.

(1) **Eyes:** The eyes region includes both the eyes and eyebrow region to catch more details and variations. Due to genetic predisposition, some people have thick or thin eyebrows. People could wear eyeglasses or sunglasses as an accessory. We chose to view the

eye region plus the eyebrow region as one face subregion under investigation. (2) **Nose:** The rectangular image patch includes only the nose region. The nose region is the most homogeneous part of the face area. Only the shape and the skin colour of the nose may vary across the dataset. (3) **Mouth:** The mouth region includes only the lips limited by the left and right corners of the mouth. The variations in an image can be multi-fold due to different shapes (e.g., opened/closed) and people applying lipsticks, wearing beards or not. This face subregion is supposed to provide more contextual information than the nose region. (4) **Fusion:** The fusion combines the eyes, nose, and mouth regions using equal weights. It is expected to provide an improved quality score superior to each single face subregion. (5) **Tightly cropped face:** This face region covers the area from the chin to the eyebrows excluding the forehead. The tightly cropped face is limited by the facial landmarks found in the face region. The intuition is to focus on the main face excluding the potentially noisy backgrounds (such as the forehead, hairstyles, or hair colours). The richness of information derived from the forehead, hairstyles, or hair colours is neglected in this setting. (6) **Aligned face:** FR solutions form the face embeddings from aligned face images, and aligned face images are used for FIQA. The image pre-processing is described above. Such aligned face images may contain noisy background as illustrated.

CNNIQA [Ka14] is a DL-based IQA algorithm based on convolutional neural network (CNN) as backbone. The network architecture consists of only one convolutional layer with max and min pooling, completed by two fully connected layers and one output node. It is a patch-based approach to receive more fine-grained quality distribution across the image. We have chosen this method because of its strength in automatically extracting low-level features from images and its light computation. In this paper, we link this quality measure from face subregions to the face utility. The results will be further compared with quality scores from DL-based FIQA methods proposed in Section 3.2.

3.2 Face image quality assessment algorithms

We selected four FIQA algorithms, which used various training strategies: rankIQ [Ch15], FaceQnet [He19], MagFace [Me21], and SER-FIQ [Te20a]. We apply these methods on the aligned face images to determine FIQ scores.

(1) **rankIQ** [Ch15] by Chen et al. is a FIQA method based on learning to rank. The method is trained using three face image datasets of different quality ranks. (2) **FaceQnet** [He19] by Hernandez-Ortega et al. is trained to predict the normalized similarity score between an input face image and ICAO-compliant high-quality mated reference images. A pretrained ResNet-50 is used to associate automatically learned face feature vectors of input images to targeted FIQ scores. (3) **MagFace** [Me21] is a method developed by Meng et al. to provide face embeddings for both FR and FIQA. During training, the magnitude of the feature vector (i.e. the calculated face embedding) is made proportional to the cosine distance to its class center. Hence, the norm is directly related to FIQ. (4) **SER-FIQ** [Te20a] by Terhörst et al. is an unsupervised machine-learning-based FIQA method. It assesses the quality of a face image by determining the robustness of the feature vector of a specific DL-based FR system against random dropout patterns. This method fully eliminates the

need for any labeling. Here, we used the SER-FIQ method trained on ArcFace without retraining for the other FR models (SphereFace and FaceNet). Such a method was also used to investigate the quality of morphing attacks [Fu21] or to gain deeper understanding of demographic fairness in face recognition [Te20b].

4 Evaluation and discussion

In this section, we first introduce the datasets used for the evaluation followed by the three FR systems. Secondly, we present the outcomes of the conducted experiments and discuss the findings. The results are evaluated on two datasets with complex and in-the-wild capture conditions.

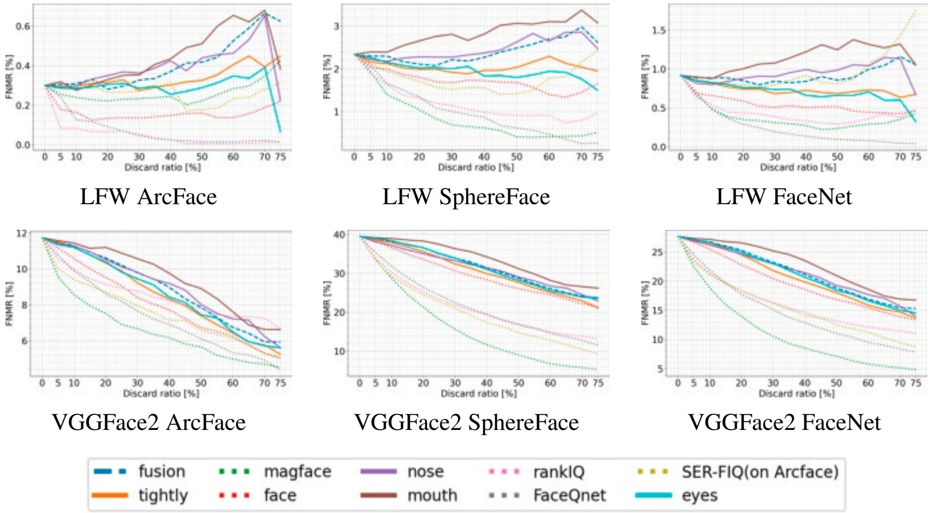
Datasets We chose (1) the Labeled Faces in the Wild (**LFW**) dataset [Hu07] as it is one often cited face image dataset and is a widely used standard benchmark for automatic face verification [Te20a, Me21]. However, this dataset is strongly imbalanced regarding the number of images for each subject. Therefore, we used the test protocol as reported in [HLM14]. (2) **VGGFace2** [Ca18] test setup contains 500 subjects. We chose VGGFace2 as it contains mostly uncontrolled and unconstrained complex acquisition conditions. The dataset has a large variety in quality data distribution. To manage the heavy computation due to the number of images, we randomly selected 30 out of 300 images from each subject to perform the verification task.

Face recognition systems Three FR solutions were used to deduce the face embeddings according to different metrics and evaluate the result: FaceNet [SKP15], SphereFace [Li17], and ArcFace [De19].

FaceNet [SKP15] by Schroff et al. is a FR system trained on deep CNN structures and triplet loss. We chose this system because the reported accuracy on the LFW dataset is 99.63%±0.09 and on the YouTube Faces dataset (YTF) [WHM11] 95.12%±0.39. **SphereFace** [Li17] by Liu et al. uses a modified softmax loss with multiplicative margin for the training to improve the recognition performance. We chose this FR model as it achieved a competitive state-of-the-art benchmark verification accuracy on LFW dataset of 99.42% and on YTF dataset of 95.0%. **ArcFace** [De19] by Deng et al. is trained using the ResNet-100 [HKK17] structure on the MS1M dataset [Gu16]. The loss function further uses additive angular margin to improve the discriminative power of the FR model. This model is chosen because it is one of the top-performing FR solutions used in most of the recent FIQA work [Te20a, Me21]. It further improved the accuracy on LFW to 99.83% and YTF dataset to 99.02%.

Performance metric The evaluation metric used here is the false non-match error vs. reject characteristic (ERC) [GT07]. An ERC shows the dependence of the false non-match rate (FNMR) at a fixed decision threshold on the percentage of discarded images with a low quality score. The ERCs vary for different decision threshold values. The decision threshold was fixed to give an initial FMR value of 0.1%, named FMR1000. The FNMR is expected to go down as the ratio of discarded images increases.

Results and discussion Fig. 2 depicts the ERC at the fixed FMR of 0.1%. CNNIQA evaluated on face subregions are drawn in solid lines, while the FIQA methods using aligned faces are drawn in dashed lines. The FIQA methods on the aligned faces achieved the lowest error rate across all different settings. This can be due to the face-specific training procedures of the used algorithms. However, according to the trend observed in the ERCs for VGGFace2, the general IQ scores of face subregions are correlated with the FIQ scores of aligned face images. Looking at the ERC for VGGFace2 using ArcFace embeddings, the FNMR reduced for both the general IQ metric for face subregions as well as for the specific FIQ scores MagFace, FaceQnet, and SER-FIQ. The tightly cropped face region and the eye region show the strongest correlation with the aligned face. Looking at the ERC with the same setting for a discard ratio greater than 40%, the ERC of the tightly cropped face region and the aligned face images using CNNIQA come close to the ERC of the SER-FIQ method. This indicates the usefulness of linking the general IQ score of face subregions to the FIQ score. The nose and mouth regions perform worse than the eyes region. However, fusing the three regions (mouth, nose, and eyes) using equal weights improved the performance towards the performance of the tightly cropped region. The ERC for the fusion (dashed dark blue curve) shows less oscillation from the discard ratios of 30% to 60% compared to the ERC for the eyes region (in solid orange curve).



LFW ArcFace LFW SphereFace LFW FaceNet

VGGFace2 ArcFace VGGFace2 SphereFace VGGFace2 FaceNet

Fig. 2: False non-match error vs. reject characteristics at FMR1000. The rows reveal the results for different face embeddings on LFW and VGGFace2. ERCs using the IQ scores of face subregions are drawn in solid lines, while ERCs using FIQ scores are drawn in dashed lines. Out of the considered face subregions, IQ scores of the eye region and of the tightly cropped face region provide the strongest correlation with the FIQ scores.

Tab. 1 shows the ERC using general IQ scores on face subregions and FIQ scores on aligned face images. The FNMR at FMR1000 is shown for three different FR systems and two datasets at two discard ratio values of 20% and 40%. The top-3 methods across different settings are drawn in bold. As anticipated, the FIQ metrics outperform the other methods owing to including entire face images. However, a strong correlation is observed between the regional IQ scores and the FIQ metrics. Considering the VGGFace2 at FMR1000 using ArcFace embeddings, the FNMR was reduced by 19.4% and 18.8% for the tightly cropped face region and the eyes region, respectively, from the discard ratio values of 20% to 40%, while a similar reduction of FNMR by 18.95% can be observed for aligned faces. The mouth and nose regions perform inferior compared to the eyes region and the tightly cropped face region. A smaller reduction in terms of FNMR can be observed for the same setup. Only a drop of 12.7% and 12.8% was seen for the mouth and nose regions, respectively. Reviewing the images of the mouth region in Fig. 1 revealed a possible reason for this finding: There is more noise in the mouth region due to the shape variation (opened/-closed, beard/no beard, etc.) compared to the nose region, which makes FR more difficult. Surprisingly, the eyes region performs similarly to the tightly cropped face region. Most of the time, eyes performed best among the different face subregions. One possible cause for the strong correlation of these two metrics may be the overlap of regions. Looking at the example images in Fig. 1, we noticed that the eyes region contains the most significant information within the whole face area.

LFW at FMR1000							
		ArcFace		Sphereface		FaceNet	
		20%	40%	20%	40%	20%	40%
DL based FIQA	rankIQ	**0.066%**	**0.026%**	**1.322%**	0.983%	0.436%	0.333%
	FaceQnet	**0.092%**	**0.023%**	1.337%	**0.871%**	**0.288%**	**0.138%**
	SER-FIQ(on ArcFace)	0.264%	0.313%	1.708%	1.559%	0.790%	0.918%
	MagFace	0.222%	0.244%	**1.045%**	**0.620%**	**0.352%**	**0.272%**
IQA method CNNIQA	Aligned face	0.134%	0.148%	1.803%	1.723%	0.687%	0.507%
	Tightly cropped face	0.315%	0.302%	1.986%	1.961%	0.741%	0.727%
	Eyes	**0.305%**	**0.255%**	**2.018%**	**1.834%**	**0.759%**	**0.660%**
	Nose	0.349%	0.424%	2.278%	2.368%	0.880%	0.992%
	Mouth	0.330%	0.434%	2.652%	2.810%	1.004%	1.221%
	Fusion	0.281%	0.370%	2.111%	2.266%	0.848%	0.841%
VGGFace2 at FMR1000							
		ArcFace		Sphereface		FaceNet	
		20%	40%	20%	40%	20%	40%
DL based FIQA	rankIQ	9.131%	8.239%	25.282%	19.288%	18.381%	14.677%
	FaceQnet	**8.612%**	**6.908%**	26.368%	19.502%	**17.772%**	**12.874%**
	SER-FIQ(on ArcFace)	8.703%	7.137%	**24.466%**	**17.279%**	18.421%	14.167%
	MagFace	**7.520%**	**6.171%**	**21.329%**	**11.650%**	**13.993%**	**8.540%**
IQA method CNNIQA	Aligned face	9.527%	7.721%	33.564%	28.385%	22.814%	18.625%
	Tightly cropped face	10.327%	8.325%	35.559%	29.904%	24.547%	19.877%
	Eyes	**10.389%**	**8.431%**	36.570%	**31.045%**	25.019%	**20.680%**
	Nose	10.527%	9.193%	**35.051%**	31.481%	**24.658%**	21.472%
	Mouth	11.201%	9.761%	38.305%	34.231%	26.590%	23.554%
	Fusion	10.612%	8.982%	36.576%	31.433%	25.337%	21.156%

Tab. 1: False non-match error vs. reject characteristics using general IQ scores on face subregions and FIQ scores on aligned face images at two discard ratios (20% and 40%) at FMR1000 for three FR models (ArcFace, SphereFace, FaceNet). The three best-performing results are marked in bold across settings. Additionally, the best performing region out of eyes, nose, and mouth is in bold. Note that (1) FIQA methods on the aligned face images outperform the face subregions, while (2) the eyes region and tightly cropped face region are close to the performance of the rankIQ from the FIQA category. (3) Most of the time, the eyes region performed best among the face subregions.

5 Conclusion

This paper investigated the correlation of semantic contexts, represented by the face subregions, with FIQ scores. Using a general IQ method on face subregions, we relate this measure to the quality scores extracted using the specifically trained FIQA methods rankIQ [Ch15], FaceQnet [He19], SER-FIQ [Te20a], and MagFace [Me21].

The following take-home messages can be extracted from our experimental results: (1) In general, the FIQA methods outperform the general IQ metric on aligned faces or on the individual face subregions. (2) However, the general IQ scores of face subregions are strongly correlated with the specific FIQ scores. The eyes region works best most of the time across different settings under all considered face subregions. This could indicate that the eyes region includes significant information for the FR task. (3) Finally, the fusion of face subregions (eyes, nose, and mouth) improved the performance of the nose and mouth regions. In addition, it further smoothed out the oscillation in terms of ERC compared to the eyes region alone. It indicates that fusion of face subregions can further increase the robustness and provide a better generalization of this quality measure.

Acknowledgements: This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[Ca18] Cao, Qiong; Shen, Li; Xie, Weidi; Parkhi, Omkar M.; Zisserman, Andrew: VGGFace2: A Dataset for Recognising Faces across Pose and Age. In: FG. IEEE Computer Society, pp. 67–74, 2018. 4

[Ch15] Chen, Jiansheng; Deng, Yu; Bai, Gaocheng; Su, Guangda: Face Image Quality Assessment Based on Learning to Rank. IEEE Signal Process. Lett., 22(1):90–94, 2015. 2, 3, 7

[De19] Deng, Jiankang; Guo, Jia; Xue, Niannan; Zafeiriou, Stefanos: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: CVPR. Computer Vision Foundation / IEEE, pp. 4690–4699, 2019. 4

[DG17] Dabkowski, Piotr; Gal, Yarin: Real Time Image Saliency for Black Box Classifiers. In: NIPS. pp. 6967–6976, 2017. 2

[Fu21] Fu, Biying; Spiller, Noémie; Chen, Cong; Damer, Naser: The effect of face morphing on face image quality. In: BIOSIG. LNI. Gesellschaft für Informatik e.V., 2021. 4

[GT07] Grother, Patrick; Tabassi, Elham: Performance of Biometric Quality Measures. IEEE Trans. Pattern Anal. Mach. Intell., 29(4):531–543, 2007. 4

[Gu16] Guo, Yandong; Zhang, Lei; Hu, Yuxiao; He, Xiaodong; Gao, Jianfeng: MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In: ECCV (3). volume 9907 of Lecture Notes in Computer Science. Springer, pp. 87–102, 2016. 4

[He19] Hernandez-Ortega, Javier; Galbally, Javier; Fiérrez, Julian; Haraksim, Rudolf; Beslay, Laurent: FaceQnet: Quality Assessment for Face Recognition based on Deep Learning. CoRR, abs/1904.01740, 2019. 3, 7

[HKK17] Han, Dongyoon; Kim, Jiwhan; Kim, Junmo: Deep Pyramidal Residual Networks. In: CVPR. IEEE Computer Society, pp. 6307–6315, 2017. 4

[HLM14] Huang, G.B.; Learned-Miller, E.: Labeled faces in the wild: Updates and new reporting procedures. Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep, 14(003), 2014. 4

[Hu07] Huang, G.B.; Ramesh, M.; Berg, T.; Learned-Miller, E.: Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical Report 07-49, University of Massachusetts, Amherst, Oct. 2007. 4

[ISO16] Information technology – Biometric sample quality – Part 1: Framework. International Standard ISO/IEC 29794-1, 2016. 1

[Ka14] Kang, Le; Ye, Peng; Li, Yi; Doermann, David S.: Convolutional Neural Networks for No-Reference Image Quality Assessment. In: CVPR. IEEE Computer Society, pp. 1733–1740, 2014. 3

[Li17] Liu, Weiyang; Wen, Yandong; Yu, Zhiding; Li, Ming; Raj, Bhiksha; Song, Le: SphereFace: Deep Hypersphere Embedding for Face Recognition. In: CVPR. IEEE Computer Society, pp. 6738–6746, 2017. 4

[LLC15] Lienhard, Arnaud; Ladret, Patricia; Caplier, Alice: Low Level Features for Quality Assessment of Facial Images. In: VISAPP (1). SciTePress, pp. 545–552, 2015. 2

[Me21] Meng, Qiang; Zhao, Shichao; Huang, Zhida; Zhou, Feng: MagFace: A Universal Representation for Face Recognition and Quality Assessment. CoRR, abs/2103.06627, 2021. 3, 4, 7

[Sc21] Schlett, T.; Rathgeb, C.; Henniger, O.; Galbally, J.; Fierrez, J.; Busch, C.: , Face image quality assessment: A literature survey. arXiv preprint arXiv:2009.01103, 2021. 1

[SKP15] Schroff, F.; Kalenichenko, D.; Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: Proc. of the IEEE Conf. on CVPR. pp. 815–823, 2015. 4

[Te20a] Terhörst, P.; Kolf, J.N.; Damer, N.; Kirchbuchner, F.; Kuijper, A.: SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness. In: Proc. of the IEEE/CVF CVPR. pp. 5651–5660, 2020. 2, 3, 4, 7

[Te20b] Terhörst, Philipp; Kolf, Jan Niklas; Damer, Naser; Kirchbuchner, Florian; Kuijper, Arjan: Face Quality Estimation and Its Correlation to Demographic and Non-Demographic Bias in Face Recognition. In: IJCB. IEEE, pp. 1–11, 2020. 4

[WHM11] Wolf, L.; Hassner, T.; Maoz, I.: Face recognition in unconstrained videos with matched background similarity. In: Proc. of the Conf. on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 529–534, 2011. 4

[XZ17] Xiang, J.; Zhu, G.: Joint face detection and facial expression recognition with MTCNN. In: 4th Int. Conf. on Information Science and Control Engineering (ICISCE). IEEE, pp. 424–427, 2017. 2

[Zh16] Zhou, Bolei; Khosla, Aditya; Lapedriza, Àgata; Oliva, Aude; Torralba, Antonio: Learning Deep Features for Discriminative Localization. In: CVPR. IEEE Computer Society, pp. 2921–2929, 2016. 2

Biometric Recognition in a Multi-sample Multi-Subject Facial Image Database: The 1:M:N System Model

DeWayne Halfen[1], Srinivasan Rajaraman[2], James L. Wayman[3]

Abstract: Over the last 50 years, biometric recognition has advanced from localized "identity verification" applications [GU77][RY74] to include large-scale systems in which "a determination is made as to the identity of an individual independently of any information supplied by the individual" [GU77]. Models for estimating and expressing system error rates (both false matches and false non-matches) have been largely limited to so-called "1-to-1" and "1-to-N" systems in which each identity is represented by only one enrolled reference [Gr21]. In this paper, we create a highly simplified simulation model for a common current situation in which each known identity record has multiple stored references. We call this the "1:M:N" model and show that both DET and CMC performance depend upon the number of identities and images per identity, not simply the total number of references images, as usually assumed. Although trialed here on very simple decision policies, this model will be extended in future work to more complex decision criteria.

Keywords: Biometric System Modeling, Large-scale systems, Detection Error Trade-off curve, Cumulative Match Characteristics.

1 Introduction

As early as the 1970s, it was recognized that systems for "Automated Personal Identification" (now widely called "biometrics") could operate in two use cases: "verification" of identity and "absolute" identification [GU77][RY74]. The former used bodily characteristics to verify a claim to association with a single, previously enrolled record. The latter sought the enrolled record of a person in a large database indexed by bodily characteristics without any claim of association to a specific enrolled record, necessitating the search of an entire database. Early work focused only on the "verification" use case because applications of that era were limited to comparison of a submitted fingerprint, voice, signature, or hand shape sample to establish linkage to an enrolled record as claimed by the data subject. By the late 1980s, there was expanding interest in "absolute" identification as a way of restricting persons to one identity record in a system [TR88] with the assumption that each identity record would have but one associated biometric reference. By the end of the 20th century, the terms "one-to-one" and "one-to-many" [Ne95] (or alternatively "one-on-many" or "one-to-N") were in common usage to differentiate between the "verification" and "absolute identification" use cases based on the number of biometric comparisons required. However, the feasibility of any

[1] Booz Allen Hamilton Arlington, VA Halfen_DeWayne@bah.com
[2] Booz Allen Hamilton, McLean, VA Rajaraman_Srinivasan@bah.com
[3] U.S.Department of Homeland Security, OBIM, James.Wayman@obim.dhs.gov

system involving a large number of "one-to-N" comparisons was hotly debated owing to the exponential growth of errors with increasing N in the simplest Bernoulli model [Ho99][Wa98]. The weaknesses of this over-simplified system model became apparent with the development of large-scale fingerprint recognition systems based on the recording of multiple, but fixed number of references, one for each finger enrolled, for each person in the database. These new systems required mathematical characterization considerably different from the contested "one-to-N" concept [Wa99]. In recent decades, large-scale biometric applications have advanced beyond fingerprinting, with works on the performance of facial recognition systems [OB14] [Ke16] [BJ17] [Gr21] discussing the existence of multiple facial images, M, stored for each of N data subjects. Clearly, the simplistic "one-to-N" Bernoulli model does not apply to these systems, thus requiring new models.

In the context of facial recognition systems, [Gr21] discusses three types of one-to-many systems employing different strategies for use of multiple references stored for each data subject or "identity":

- Recent: Only the most recent reference image is used in the comparison.
- Lifetime-consolidated: All but the most recent image is used in "producing a single proprietary undocumented 'black-box' template from the images. This affords the algorithm an ability to generate a model of the individual, rather than to simply extract features from each image on a sequential basis" [Gr21].
- Lifetime-unconsolidated: All but the most recent image are compared separately, "with different identifiers, such that the algorithm is not aware that the images are from the same face" [Gr21].

Many operational facial recognition systems we have studied correspond to a different model than any of the three given above. Multiple facial images are "consolidated" by data subject into an identity record, but considered individually to develop comparison scores against a probe image. No multi-image model is created, thus avoiding the complexity of updating the model every time new data is received for a data subject.

2 Contributions of this Project

In this paper, we simulate a large-scale facial recognition system for which each data subject, i, i=1… N, has only one type of biometric characteristic (e.g., a full-frontal face image) but may have several such images M_i stored in an identity record within the database. In an "absolute identification" search, a single probe image of a single data subject is examined against the database of $M=\Sigma M_i$. images in total but, because these M images are consolidated into N identity records, various search strategies can be used.

In our study, the fundamental input data used for the mated and non-mated score distributions and for the distribution of the number of enrolled images M_i, i=1,..N for each data subject is intended to be "realistic", meaning similar to what is seen in extant operational systems, but not representing any particular system. Our data is taken from a variety of actual systems, meaning that the resulting parameter data sets are "chimeric", in the sense of blindly combining, as though independent, correlated data. Further, simulated comparison scores used within each of the N identities are drawn independently from the mated or non-mated distributions, thus ignoring any real-world effects of score correlations based on probe quality or persistent data subject characteristics.

Using this artificial data set, we create a very simplified simulation model to explore the dependencies of the Detection Error Trade-off (DET) and Cumulative Match Characteristic (CMC) curves on: 1) decision policies; 2) the total number of identities, N; and 3) the distribution of the number of images, Mi, over the N identities enrolled in the database. We call this the "1:M:N" model. The ultimate goal of our efforts is to expand the simulation model to consider a variety of decision policies for which analytic modeling would be either intractable or impossible.

3 Chimeric Data Sets

Figure 1 shows the mated (green on right) and non-mated (red on left) score distributions chosen for our simulation model. Figure 2 shows the distribution of number of facial images for the various identities.

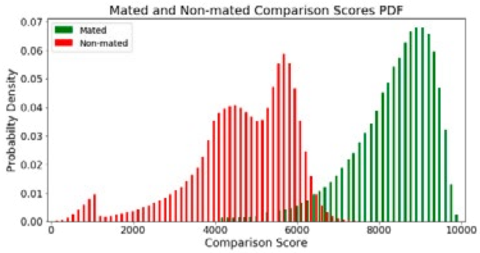


Fig. 1: Mated and Non-Mated Score Distributions

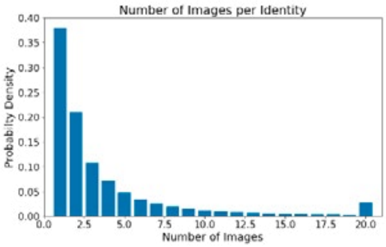


Fig. 2: Number of Images for Each Identity

The data in Figure 2 was further reduced in 3 additional ways: limiting all N enrolled data subjects to no more than 5, 10 or 15 stored images, Mi, with identities having images exceeding the upper bound given the maximum number of images. The 4 distributions used in this study, $M_i \leq \{5,10,15,20\}$ are all dominated by identities with only one image available, $M_i = 1$.

4 Model Validation

The first task is to validate the simulation model against analytic results for a very simple decision policy. We start with the case where each identity has the same number of images,

M_i = constant ≡ M_0, such that the distribution of Figure 2 is not used. This case can be considered analytically as N sets of M_0 simple Bernoulli trials, each trial given a comparison score drawn from one of the distributions of Figure 1 depending upon whether the trial is to be considered "mated" or "non-mated". Under the simple decision rule that a "match" is declared for the identity if any of the M_0 scores exceeds a threshold and a "non-match" declared if all of the M_0 scores fall short of the threshold,- simple Bernoulli models for false non-match and false match rates are

$$\text{Identity record false non-match rate} = \text{FNMR}(\tau)^{M_0} \quad (1)$$

$$\text{Identity record false match rate} = 1 - (1 - \text{FMR}(\tau))^{M_0} \quad (2)$$

where FNMR is the area under the mated distribution in Figure 1 with score less than τ and FMR is the area under the non-mated distribution with score greater than τ.

Figure 3 shows the analytic DET and Figure 4 shows the simulated DET, with $M_0 = \{1, 2, 5, 10, 20\}$ for each identity. DET performance improves as M_0 increases, but with only minimal gains above $M_0 = 10$.

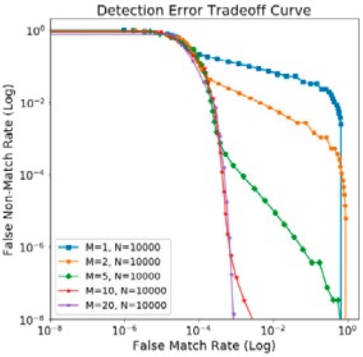


Fig. 3: Analytic DET using Bernoulli equations for $M_0 = \{1,2,5,10,20\}$ with 10^4 identities.

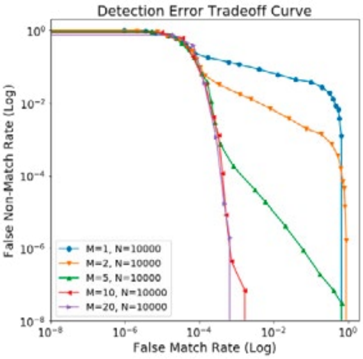


Fig. 4: Simulation DET using $M_0 = \{1,2,5,10,20\}$ with 10^4 identities

The results shown below in Figures 5 (analytic) and 6 (simulation) use the same simple decision policy as above (maximum of the M_i scores for each identity exceeding threshold τ) for the cases where the number of images, $M_{i,}$ is distributed based on Figure 2, and includes cases of $M_i \leq \{5,10,15,20\}$. For these cases, equations (1) and (2) can be rewritten as

$$\text{Identity false non-match rate} = \Sigma_i\, P(M_i)\, \text{FNMR}(\tau)^{M_i} \quad (3)$$

$$\text{Identity false match rate} = 1 - \Sigma_i\, P(M_i)\, (1-\text{FMR}(\tau))^{M_i} \quad (4)$$

where $P(M_i)$ is the percentage of identities with M_i reference images, as from Figure 2. The close correspondence of Figures 3 and 4 and Figures 5 and 6 validates the simulation model for these simple cases.

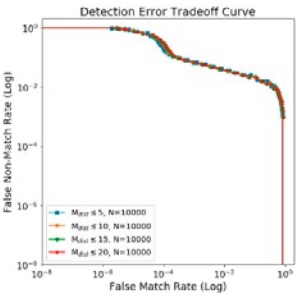


Fig. 5: Analytic DET Curve with Varying M_i and 10^4 identities

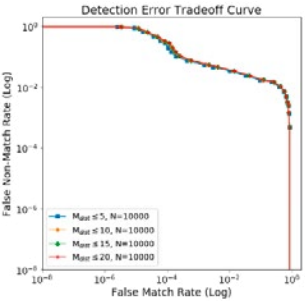


Fig. 6: Simulated DET Curve with Varying M_i and 10^4 identities

5 CMC Curves Using the Analytically Validated Simulation Model

5.1 CMC with M_i constant for all identities with "maximum" decision policy

We now use the validated simulation model to compute the Cumulative Match Characteristic (CMC), exploring the effect of multiple images for each identity on the rank ordering of the database against a randomly chosen probe image. The rank of each identity is determined by the maximum of the M_i comparison scores. Figure 7 shows the case where there are 10^4 identities, each having fixed $M_0 = \{2, 5, 10, 20\}$ number of images. The total number of images, $\Sigma_i M_i = N \cdot M_0$, searched increases as M_0 increases and the system with largest total images is generally seen to perform best, but increasing beyond 5 images per identity yields decreasing gains. We note the intriguing anomaly that this relationship is inverted for rank 1 results, with the lower values of M_0 performing better than the higher values.

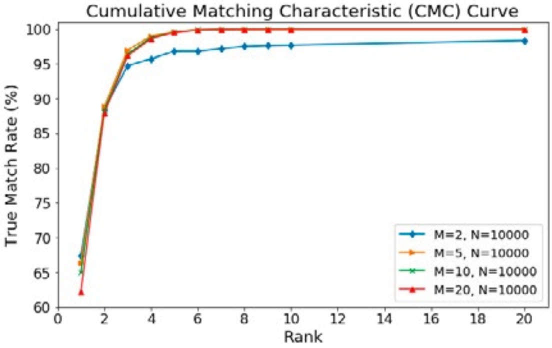


Fig. 7: CMC curve with "maximum" decision policy, 10^4 identities, varying number of images

5.2 CMC for fixed total number of images but varying numbers of identities

Figure 8 shows the simulation CMC for the condition under which the total number of images is fixed at 10^4, but those images are distributed over varying numbers of identities, N = {500, 1000, 2000, 5000, 10000}. The CMC performance improves as number of images per identity increase and number of identities decrease, even at rank 1.

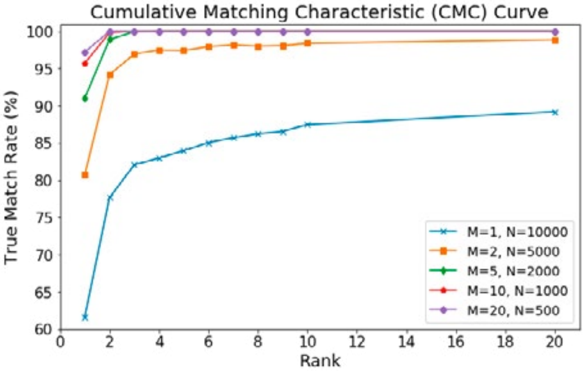


Fig. 8: CMC curves for “maximum” decision policy, 10^4 images, varying number of identities

Under the “maximum” decision policy, there is a close relationship between the cases of “unconsolidated” identities [Gr21] and “consolidated” identities when the distribution of images per identity is known. The parameter most relevant to CMC performance is the total number of identities, not total number of images.

6 Results for Alternate Decision Policies

Our ultimate goal is to create a simulation model for more complex conditions. We compared decision policies based on the maximum, mean, and minimum of each of the M_i comparison scores for each identity.

6.1 DET using alternate decision policies

Figure 9 shows the great improvement of DETs over the same conditions as Figure 4 by taking the mean of the M_0 = {1,2,5,10} comparison scores. The DET improves with increasing M. Figure 10 shows the different DETs resulting from using the maximum, mean and minimum values of the non-uniform $M_i \leq 5$ comparison scores, distributed as suggested in Figure 2. Surprisingly, the min decision policy dominates the other two approaches except at extreme thresholds.

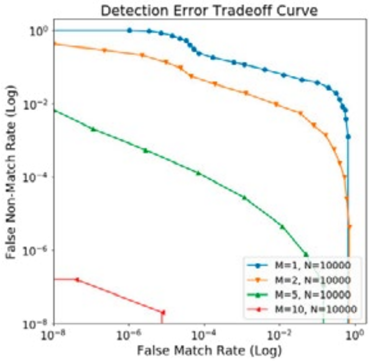


Fig. 9: Mean decision policy with M_0={1,2,5,10} and 10^4 identities

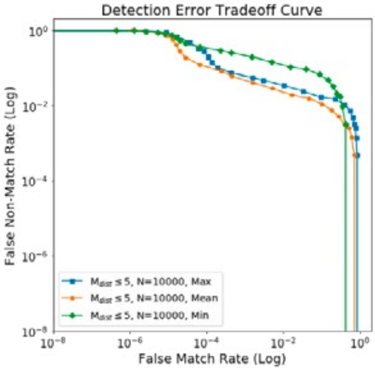


Fig. 10: DET using different decision policies with $M_i \leq 5$ and 10^4 identities

6.2 CMC using alternate decision policies

Figure 11 shows CMC curves under a variety of conditions: fixed $M_0 = 5$ and distributed $M_i \leq 5$; maximum, mean and minimum score decision policy. The decision policy based on the mean of the M_i scores for each identity dominates and having 5 images for each identity improves performance over having fewer. We see that the "minimum" is better than the "maximum" decision policy at rank 1.

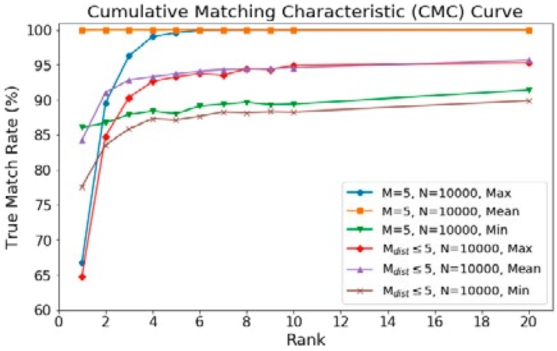


Figure 11: CMC curve with max, mean and min decision policies for $M_0 = 5$ and $M_i \leq 5$

7 Conclusions

We have created a validated simulation model that demonstrates several important properties of 1:M:N systems under various simple decision policies. Our primary findings, caveated by use of the score distributions in Figure 1, are:

1. When each identity has the same fixed number of images and each comparison has scores drawn from the distributions of Figure 1, there is only minor improvements to DET performance by using more than 10 images per identity. This finding is supported by both analytic and simulation models.

2. The policy of using the mean of the comparison scores for each identity produces the best DET except at extreme thresholds. The DET curve improves for the mean policy as the number of images for each identity increases

3. When the maximum score is used for each of a fixed number of identities, the CMC improves steadily as the number of images per identity increases except at rank 1. Using the mean value of the scores gives the best CMC at all ranks. Using the maximum value is better than the minimum value except at rank 1.

4. CMC performance depends on the number of N identities and number of images per identity, not the total number of images, $\Sigma_i M_i$.

References

[BJ17] Best-Rowden, Lacey; Jain, Anil K.: Longitudinal study of automatic face recognition. IEEE transactions on pattern analysis and machine intelligence 40.1 (2017): 148-162.

[Gr21] Grother, Patrick; Ngan,Mei; Hanaoka, Kayee: Face Recognition Vendor Test (FRVT) Part 2: Identification, NISTIR 8271 Draft Supplement, 16 April, 2021

[GU77] Guideline on the Evaluation of Techniques for Automated Personal Identification, U.S. National Bureau of Standards, FIPS 48, 1 April, 1977

[Ho99] Hopkins, Richard: An introduction to biometrics and large scale civilian identification, International Review of Law, Computers & Technology 13.3 (1999): 337-363.

[Ke16] Kemelmacher-Shlizerman, Ira; Seitz, Steven; Miller, Daniel; Brossard, Evan: The megaface benchmark: 1 million faces for recognition at scale. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4873-4882. 2016.

[Ne95] Newham, Emma (ed). Biometric Checks for Singapore's Commuters in Biometric Technology Today 3.1, September 1995

[OB14] Ortiz, Enrique G.; Becker, Brian: Face recognition for web-scale datasets. Computer Vision and Image Understanding 118 (2014): 153-170.

[RY74] Raphael, David; Young, James: Automated Personal Identification, SRI International, Palo Alto, CA, 1974.

[TR88] Truck and Bus Safety and Regulatory Reform Act, U.S. Public Law 100-690, Section 9105, 1988

[Wa98] Wayman, James: Continuing Controversy Over the Technical Feasibility of Large-Scale Systems, Biometrics In Human Services User Group Newsletter #11, State of Connecticut, 1998

[Wa99] Wayman, James: Error-Rate Equations for the General Biometric System," IEEE Robotics & Automation Mag. 6.1, pp. 35-48, 1999

Evaluation on Biometric Accuracy Estimation Using Generalized Pareto (GP) Distribution

Shigefumi Yamada[1], Tomoaki Matsunami[2]

Abstract: The accuracy of biometric authentication technology is becoming more sophisticated with its progress. For this reason, a huge number of biometric samples are required for accuracy evaluation, and the increased collection cost is an issue for biometric vendors. This work establishes a biometric accuracy estimation method using an extreme value theory to reduce the collection cost. It also explains the estimation procedure of false match rate using the generalized Pareto distribution and shows results applied to the face, gait, and voice comparison score data with an estimation effect of about 5–10 times. We investigate the criteria for the applicability of extremum statistics through application cases.

Keywords: accuracy evaluation, false match rate, rule of three, extreme value theory, generalized Pareto distribution.

1 Introduction

As the utilization of Information Technology permeates various corporate activities and consumer life, high-precision biometric authentication (face, fingerprint, vein, etc.) is required to effectively provide and use more secure, safe, and detailed services. Iris, voice, signature, and walk are being used for personal authentication.

The method for evaluating the accuracy of biometric authentication is specified and recommended in the ISO / IEC19795 (Biometric performance testing and reporting) [I21] series and has been applied to many biometric authentication devices and software. On the other hand, with the progress and higher accuracy of biometric authentication technology, a huge number of biometric samples (data obtained from fingerprints, faces, veins, etc.) are required for the evaluation. There is false match rate (FMR) and false non match rate (FNMR) as biometric recognition accuracy. Traditionally, Rule of 3 have been used to estimate the number of data required to determine FMR with a 95% confidence interval. The Rule of 3 requires several comparisons that is three times the reciprocal of the error rate you want to evaluate. Each time a biometric vendor develops a product, the accuracy must be verified with many biometric samples. The accuracy of

[1] Japan Automatic Identification Systems Association, Biometrics Research Group, Performance testing SIG., FK Bldg.7F, 1-9-5, Iwamoto-cho, Chiyoda-Ku, Tokyo, 101-0032, Japan, yamada.shige@fujitsu.com
[2] Japan Automatic Identification Systems Association, Biometrics Research Group, Performance testing SIG., FK Bldg.7F, 1-9-5, Iwamoto-cho, Chiyoda-Ku, Tokyo, 101-0032, Japan, t.matsunami@fujitsu.com

current biometric products is extremely high. For example, the number of biometric data required to evaluate FMR is shown in Tab. 1. Therefore, the cost of collecting biometric samples is a burden for biometric vendors.

FMR	Required number of non-mated trials in the case of zero false matches	Required number of test subjects in the case of full cross-comparison
0.001%	Over 300,000	Over 775
0.0001%	Over 3 million	Over 2450
0.00001%	Over 30 million	Over 7746

Tab. 1: Number of biometric samples required to evaluate FMR

If the tail of the non-mated comparison score distribution that is unstable due to the small number of scores could be modeled appropriately, the score distribution could be extrapolated and a more accurate FMR could be estimated with the small numbers. Therefore, this work adopted the idea of extreme value theory (EVT), which is a statistical inference about the probability of occurrence of rare events, and used the EVT [C01] as a distribution model of the region where false match occur. An attempt to solve this problem is made by adopting the generalized Pareto (GP) distribution.

There are a few cases in which the EVT has been applied to the accuracy evaluation of biometrics. For the design of large-scale identification systems, it has been proposed to approximate the tail of the comparison score distribution for others with the GP distribution [HJ05]. At the tail of the distribution, which varies widely due to the lack of data, the GP distribution provided a more reliable estimate of the FMR than the measured values. When determining the decision threshold for the 1:1 comparison, it is important to appropriately approximate the tail of the comparison score distribution between the mated and non-mated pairs. As a method of estimating the tail of the distribution of comparison scores, the application of GP, which is one of the extreme value statistics, has been proposed [ZFJV08]. Extreme statistics are used for score normalization in the score level fusion of a multimodal biometric system [WART10]. The generalized extreme value (GEV) distribution, especially the Weibull distribution, of the extreme value statistics is applied to estimate the tail of the comparison score distribution of the non-mated and mated pairs. Robustly estimating the tail of the distribution, which was unstable due to the small comparison score, improves the accuracy of the score fusion. Similarly, the GEV distribution has been applied for score normalization in the score level fusion, and better discrimination results have been reported compared to those of conventional Z-score normalization [RSP15].

With these prior arts, the FMR can be estimated with high reliability by approximating the tail of the comparison score distribution with extremum statistics. Furthermore, the FMR that could not be measured with actual data can be estimated by extrapolating the tail of the comparison score distribution. Using this property, it can be expected that the desired FMR can be estimated with the smaller number of samples than using the rule of three. On the other hand, the applicability of the EVT for biometric authentication has not been discussed much in the prior arts. In applying the EVT, the comparison scores should be independent and identically distributed (i.i.d.). As a unique problem of biometric technologies, the distribution of comparison scores for non-mated pairs is known to be biased. For example, there are genetically identical/similar samples such as twins. The authenticated user who achieves a high score for many registered users is called "Wolf," and the registered user who gives a high score to a large number of authenticated users is called "Lamb." Such users can easily cause many false acceptance errors and affect the i.i.d. condition due to the high comparison scores. To spread the application of the EVT in the accuracy evaluation of biometric technology, it is important to accumulate applicable and difficult-to-apply cases through various application cases and clarify the criteria for applicability.

The main contributions of this work are as follows:

- The GP distribution is applied to various comparison score data of face, gait, and voice, and the estimation effect of the FMR is evaluated through an experiment that estimates the FMR required for evaluation from a small amount of data.
- The criteria for the applicability of extremum statistics are investigated through application cases.

This work is based on the research results of the project conducted by the Ministry of Economy, Trade and Industry in Japan from FY2019 to FY2021 [ME20].

2 Extreme Value Theory (EVT)

EVT [C01] typically targets natural phenomena that would cause very large disasters, such as heavy rains, large earthquakes, high waves, typhoons, and droughts, and their potential. It has been used for scale prediction and evaluation, i.e., in EVT, statistics are focused on the extreme value data at the tail, not the main part of the population distribution, which is noted in general statistical applications. It is characterized by making estimates based on target extrapolation.

In the method described in this work, extreme value data is classified into three types: "maximum data in a large block," "upper *r* data in that block," and "all data exceeding sufficiently large values in the observed data." Distributions that apply to these data include the GEV distribution, the simultaneous asymptotic distribution of the top r ordinal statistics (rGEV), and the GP distribution. These distributions assume that the original data are i.i.d. and nondegenerate.

Because the GP distribution is used in this work, its definition is described. The threshold excess data $\{x_1, x_2, ..., x_n\}$ are measured values of random variables that follow the generalized Pareto distribution $GP(\sigma, \xi)$ independently and identically. The GP distribution [C01] has a cumulative distribution function:

$$F(x) = \begin{cases} 1 - \left(1 + \xi \dfrac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}}, & \xi \neq 0 \\ 1 - exp\left(-\dfrac{x-\mu}{\sigma}\right), & \xi = 0 \end{cases} . \quad (1)$$

where σ is the scale parameter for the GP distribution, ξ is the shape parameter, and μ is the location parameter (threshold).

The advantage of the EVT is that most continuous distributions, such as the normal and exponential distributions, belong to the suction region of one of the extreme value distributions. Therefore, the tail region can be approximated by the GEV or GP distributions in many continuous distributions.

3 Biometric Accuracy Estimation Method Using GP distribution

If the tail of the non-mated similarity score distribution that was unstable due to the small number of scores can be accurately approximated using the GP distribution, then the FMR that could not be measured due to the lack of scores can be estimated using extrapolating the score distribution. This method's procedure is described below.

1. Extraction of extreme value data

By selecting the threshold value μ, the data exceeding μ is extracted as extreme value data. When selecting μ, the parameter estimation of the scale parameter σ and the shape parameter ξ is performed while changing μ. μ is plotted on the x-axis, while the estimated value is plotted on the y-axis. If the estimated values can be constant to the right of a certain value, their minimum value can be determined as a threshold.

2. Estimation of the GP distribution parameters

Assuming that the distribution of the extreme value data follows the GP distribution, the scale and shape parameters σ and ξ, respectively, of the GP distribution function are obtained using the maximum likelihood estimation method.

3. Diagnosis of the estimation results

Create a quantile–quantile plot (Q–Q plot) with the percentile values of both the GP distributions obtained via the parameter estimation and actual measurement values used for the estimation on the x- and y-axes, respectively. The Q–Q plot is a method for graphically comparing two probability distributions. If the two distributions to be compared are similar, then the points in the Q–Q plot are located near the straight line y = x. In the observation of the Q–Q plot, the suitability of the model shall be confirmed by the degree of deviation from the straight line y = x. Here there is no method for quantitatively determining the degree of deviation, and the suitability of EVT is conventionally determined via human eyes [C01].

4. Estimation of the FMR

Based on the probability density function of the obtained estimation model, the FMR when the threshold is set can be obtained. Fig. 1 shows an example of a graph plotting the FMR estimated by the GP distribution. The y-axis is plotted on the common logarithm axis to improve visibility. The solid blue line shows the measured value, and the dashed-red line shows the estimated FMR by the GP distribution. The green vertical line indicates the maximum value of the measured value and the score value larger than that indicates the extrapolated value. The FMR at any threshold can be obtained from this graph. If the threshold is selected as 50, the GP estimate is $10^{-7.6224} = 2.3858 \times 10^{-8}$.

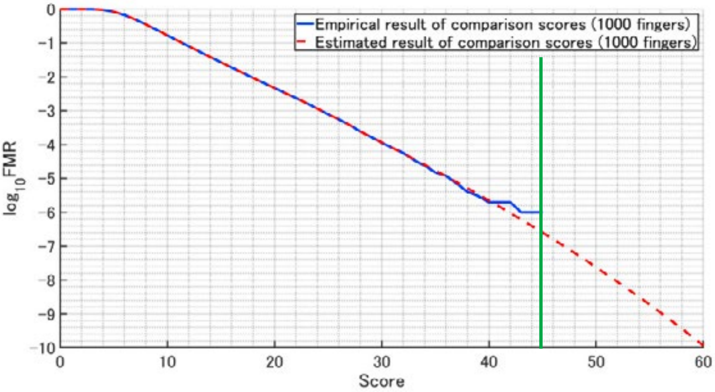


Fig. 1: Estimation result of FMR using GP distribution

4 Experimental Results and Discussions

4.1 Database and Implementation Details

To investigate the effectiveness of the GP distribution in biometric accuracy evaluation, we tried to apply the GP distribution to non-mated similarity score data in face, gait, and voice recognition. Tab.2 shows the details of the similarity score data used in these experiments. As a protocol, the test data was created by randomly extracting a small number of data (10% or 20%) from all score data. GP was applied to these test data, and the validity of the estimation result was verified by the Q–Q plot. If there were no problems in the verification results, the FMR was calculated. The calculated FMR was compared to the FMR measured from all score data. If the results were close, we could, in a sense, estimate the FMR that would require 5 to 10 times more scores. When the approximation using the GP distribution was not appropriate for 10% of the test data, the evaluation was made by increasing the number of scores to 20%.

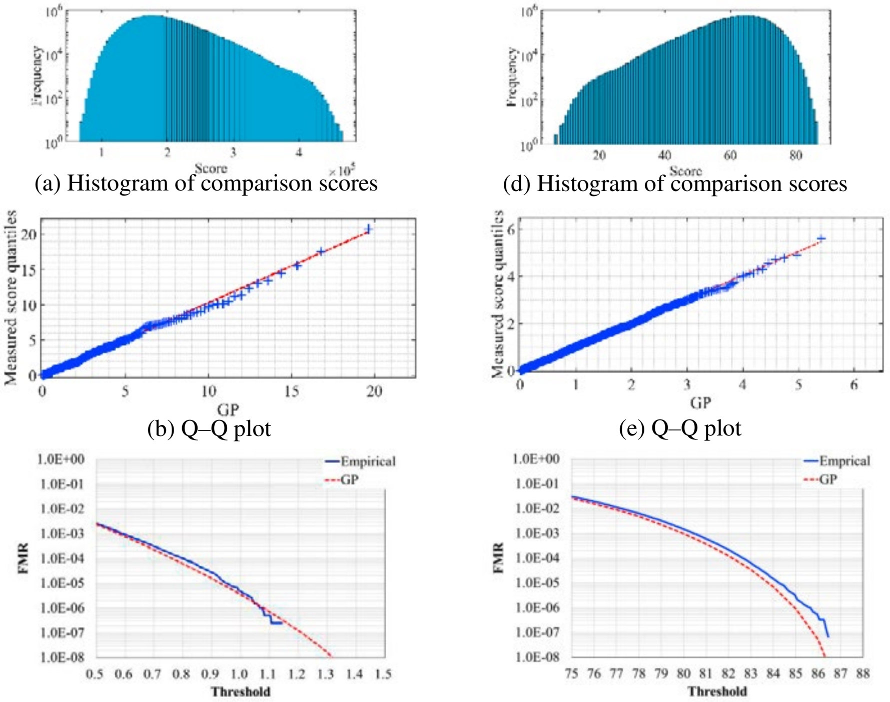
Biometrics	Corpus of biometric samples	Comparison score dataset	Number of comparisons
Face	FRGC	Idiap BIOSCOTE 2014, Face Recognition Grand Challenge v2.0 [L14]	4 million
Gait	GEI	Osaka University, Gait Energy Image (GEI) [Ha12], [Ng12]	13.73 million
Voice	SRE2012	Idiap BIOSCOTE 2014, NIST Speaker Recognition Evaluation 2012 [L14]	13.49 million

Tab. 2: Description of the non-mated similarity score data used in the experiments

4.2 Results

The face recognition grand challenge (FRGC) and gait energy image (GEI) showed good results for the GP distribution. Fig. 2 (a) shows a histogram of the similarity score of the FRGC. Twenty percent of the scores (approximately 800,000 scores) were randomly extracted from all score data and applied to the GP distribution as test data. Fig. 2 (b) shows the Q–Q plot when approximated by the GP distribution (threshold value μ = 0.485), while Fig. 2 (c) shows the comparison between the estimated and measured FMRs. The Q–Q plot is located near $y = x$, and the right tail of the similarity score distribution can be well approximated by the GP distribution. With the FRGC score, the FMR for all scores can be estimated from 20% of the scores, so it can be said that the FMR requiring about 5 times more data can be estimated by the GP distribution. Finally, Figs. 2 (d), (e), and (f) show the experimental results of the GEI. Ten percent of comparison scores (about 1.37 million scores) were randomly extracted. The threshold μ was 79.5. The Q–Q plot is located near $y = x$. With the GEI scores, the FMR that requires about 10 times more data with the GP distribution can be estimated. In Fig. 2 (f), the red line is slightly below the blue line because the subset randomly extracted for testing is below the entire set.

The result of SRE2012 in Fig. 3 is shown as an example of the difficult application of the GP distribution. Twenty percent of the similarity scores (about 2.68 million scores) were randomly extracted. The threshold μ was 2.8. In Fig. 3 (a), the score values are in a very high region and the right tail of the similarity score distribution is discretely distributed. In Fig. 3 (b), the upper right of the Q–Q plot deviates from the line $y = x$, indicating that the GP distribution cannot adequately approximate the right tail of the similarity score distribution. Moreover, the estimated FMR was different from the measured FMR in Fig. 3 (c). It can be seen that the application of the GP distribution is difficult when its right tail is discrete.



(a) Histogram of comparison scores

(d) Histogram of comparison scores

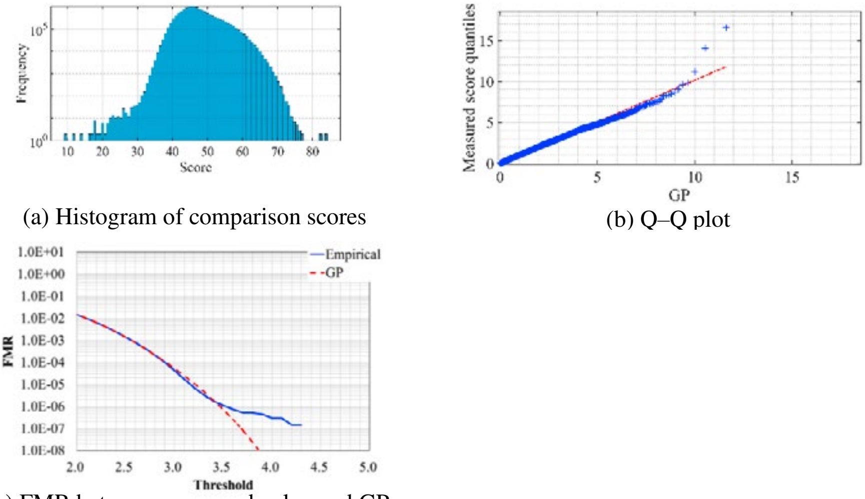
(b) Q–Q plot

(e) Q–Q plot

(c) FMR between measured value and GP

(f) FMR between measured value and GP

Fig. 2: Results of FRGC ((a), (b), (c)) and GEI ((d), (e), (f)) using GP distribution



(a) Histogram of comparison scores

(b) Q–Q plot

(c) FMR between measured value and GP

Fig. 3: Results of SRE2012 ((a), (b), (c)) using GP distribution

5 Conclusions and Future Works

This work shows the procedure for applying EVT, especially the GP distribution, to the accuracy evaluation for biometrics. This method is applied to the non-mated similarity score data of face, gait, and voice recognition, and the estimation effect is confirmed, which is 5 times for FRGC and 10 times for GEI. It was confirmed that it is difficult to apply the GP distribution in SRE2012 because there are discrete scores at the tail of the similarity score distribution. In the future, we plan to apply this method to the various cases of biometric technology and try conditions such as data size to clarify the applicable conditions of this method. In addition, we would like to make it easier to use this method by a creating criteria for determining the suitability of the GP distribution by human eyes on the Q–Q plot.

References

[I21] ISO/IEC JTC 1/SC 37 Biometrics, ISO/IEC 19795-1:2021, 2021.

[C01] Coles, S.: An Introduction to Statistical Modeling of Extreme Values, Springer, 2001.

[HJ05] Herve J.; Jean-Christophe F.: Large-Scale Identification System Design: Chapter 9 of Biometric Systems, Springer, 2005.

[ZFJV08] Zhixin, S.; Frederick, K.; John, S.; Venu, G.: Modeling Biometric Systems Using the General Pareto: In proceedings of SPIE, March 2008.

[WART10] Walter, S.; Anderson, R.; Ross, M.; Terrance, B.: Robust Fusion: Extreme Value Theory for Recognition Score Normalization: ECCV 2010, pp 481-495, 2010.

[RSP15] Renu, S.; Sukhendu, D.; Padmaja, J.: Score Normalization in Multimodal Systems using Generalized Extreme Value Distribution. Conference: Proceedings of the British Machine Vision Conference (BMVC) 2014.

[ME20] Ministry of Economy, Trade and Industry, https://www.meti.go.jp/english/press/2020/0923_003.html, 9.23.2020

[L14] Laurent, S.: Scalable Probabilistic Models for Face and Speaker Recognition, PhD thesis, 2014. http://publications.idiap.ch/index.php/publications/show/2830

[Ha12] Haruyuki, I.; Mayu, O.; Yasushi, M.; Yasushi, Y.: The OU-ISIR Gait Database Comprising the Large Population Dataset and Performance Evaluation of Gait Recognition: IEEE Trans. on Information Forensics and Security, Vol. 7, No. 5, pp. 1511-1521, Oct., 2012. (Data Set 1, 2, and 4).

[Ng12] Ngo, T.; Yasushi, M.; Hajime, N.; Yasuhiro, M.; Yasushi, Y.: Performance Evaluation of Gait Recognition using the Largest Inertial Sensor-based Gait Database: Proc. of the 5th IAPR Int. Conf. on Biometrics, Paper ID 182, pp. 1-7, New Delhi, India, Mar., 2012.(Date Set 3).

Influence of Test Protocols on Biometric Recognition Performance Estimation

Teodors Eglitis[1], Emanuele Maiorana[1], Patrizio Campisi[1]

Abstract: The performance of a biometric system is commonly evaluated by the obtained recognition rates and comparing the results against the ones reported in the literature on the same database. An aspect that has not received the deserved attention in the literature concerns the influence, on the achieved rates, of the test protocol employed to select the enrol and probe data. We provide a detailed analysis of the impact of the experimental choices on the estimated performance, considering the recommendations provided by ISO/IEC 19795 standard. We use the UTFVP finger vein database, reproducing results presented in the literature using multiple protocols. Our experiments highlight the possibility of obtaining equal error rates reduced by half simply by changing the test protocol.

Keywords: Database protocols, reproducible research, vascular biometrics.

1 Introduction

The recognition capabilities of a biometric system are evaluated by running tests on datasets of biometric samples captured from a set of subjects. To this aim, in-house databases are often collected, especially for innovative modalities at an early stage of development. The availability of public datasets enables researchers to perform in-depth research and reproduce others' work on more established modalities according to a test protocol, which determines how the considered data are used. However, the details regarding such employed protocols are often not provided with due care, making it hard to compare the new results against those previously achieved in literature, even when performing tests on the same data.

In this paper, we conduct an extensive analysis of the impact of the used experimental protocols on estimating a biometric recognition system performance. We want to highlight the importance of adequately describing the procedure followed when conducting tests on a given database by investigating the extent to which recognition rates may vary depending solely on how the considered data is exploited. We consider the recommendations of the ISO/IEC 19795 standard "Biometric performance testing and reporting" [In06] when defining the testing procedures. Vascular biometrics is used as the reference scenario. This modality has recently attracted considerable attention from industry and academia due to its advantages over more traditional biometric traits, with an ever-increasing number of papers published recently. At least eleven public databases containing finger-vein samples,

[1] Department of Industrial, Electronic, and Mechanical Engineering, Roma Tre University, Via Vito Volterra 62, 00146 Rome, Italy
{teodors.eglitis, emanuele.maiorana, patrizio.campisi}@uniroma3.it

and eight databases with palm-vein images, are currently publicly available [Uh20]. The University of Twente Finger Vascular Pattern (UTFVP) database [TV13], one of the first publicly available finger-vein datasets and among the most cited ones, is employed in this paper to assess the influence of the used test protocol on the recognition rates.

The remainder of this paper is structured as follows: Section 2 summarizes the testing recommendations provided by the ISO/IEC 19795 standard. Section 3 describes the database, recognition approach, the test protocols, and the performance metrics used. Finally, a discussion on the obtained results and conclusions are given in Section 4.

2 Test Protocol Recommendations

Guidelines for designing and testing biometric recognition systems are provided by the ISO/IEC joint technical committee (JTC) 1/SC 37. Standards for test protocols are defined in the ISO/IEC 19795, “Information technology – Biometric performance testing and reporting” documents, currently consisting of ten parts. Those relevant for our study are “Part 1: Principles and framework” [In06], first published in 2005, and “Part 2: Testing methodologies for technology and scenario evaluation" [In07], initially released in 2006, which describe the recommended scientific practices for technical performance testing. Recommendations from ISO/IEC 19795-1 [In06] for the definition of test protocols can be summarized as follows:

- the test phase should be conducted on data unavailable during algorithm development [In06, § 5.5.3.a];
- collection of enrolment and probe data should be separated at least by days [In06, § 6.5.5];
- when reporting error rates, the “rule of 3” and “rule of 30” [In06, § 6.6.1], which relate the number of probes with the achievable error confidence intervals, should be taken into account. It is remarked that handling ten probes for ten subjects is not equivalent to having a hundred subjects each with only a single probe, although, for certain protocols, this produces an equal number of comparisons;
- data from the same subject and the same modality, yet different instances (e.g., distinct eyes, fingerprints, finger-veins) can be used to represent distinct users [In06, § 6.6.3.b];
- collected samples should be excluded from the database only if a predetermined criterion is violated [In06, § 7.1.6];
- each test subject should be enrolled only once [In06, § 7.3.1.1];
- impostor comparisons involving data captured from the same subject (e.g., vascular data from different fingers of the same person, representing different *virtual* users) should not be performed because intra-individual data are likely to contain more similarities than data from different individuals [In06, § 7.6.1.3];
- zero-effort impostors can be selected by randomly choosing biometric templates or by doing a full cross-comparison [In06, § 7.6.3.1.1];
- enrolment templates can be used as impostor data in case different feature extractors are applied to enrolment and probe samples [In06, § 7.6.3.3.b].

Several of the ISO/IEC 19795 recommendations mentioned above, e.g., enrolment and probe data being captured at different days, or computing a minimum number of comparisons to validate error rates, are often not respected in the employed test protocols, thus affecting the reliability of the reported performance.

Additional suggestions on the test protocols to be used have been proposed in the literature. For instance, when evaluating biometric systems performing verification, in [JKR15] it has been suggested to use training, validation, and testing sets derived from different subjects, to avoid positive bias in the estimation of performance such as false match rate (FMR), false non-match rate (FNMR), and equal error rate (EER). [Ma15] (published in 2015) recommends using a Receiver Operating Characteristic (ROC) and Detection Error Tradeoff (DET) curves; providing False Acceptance Rate (FAR) in the range $\{10^{-4}, 10^{-2}\}$; compare algorithms using verification (1:1) experimental setup instead of 1:N, arguing that 1:1 comparisons more clearly indicates the algorithm effectiveness (if the specific research does not concern identification), and disclose detailed information about software, database, algorithms, and computational efficiency. Paper [MZB16], published in 2016, analyze different methods of data division as enrol and probe data, namely hold-out (selecting percentage of the data as probe samples), cross-validation (using *n* folds and repeating calculations *n* times, every time using the different fold as probe data) and leave-one-out methodology (cross-validation where the number of folds equals the number of samples in the dataset). Authors summarize published results and offer their own, using different data division scenarios on three face databases.

Our investigation is similar to the [MZB16], the main differences are that we focus on more exotic data division in protocols often used in vascular biometrics experiments, we summarize and follow the recommendations provided by the ISO standards, hoping that our research will be beneficial to the novices in the field.

3 Method

The UTFVP database, upon which the performed tests are conducted, comprises data recorded from 60 subjects. Two samples from three fingers (index, ring, and middle finger) of both hands have been captured during two sessions separated by 15 days for each of the involved individuals. For each finger, images 1-2 are obtained in the first recording session, and images 3-4 in the second one, for a total of 4 biometric samples. The database, therefore, consists of 360 different finger-vein classes for a total of 1440 vascular pattern images. Samples from UTFVP are processed using the maximum curvature (MC) feature extractor [MNM07]. The similarity score between two templates is evaluated using the Miura match (MM) algorithm [MNM07]. Such comparison is not symmetrical and can generate different scores if the two templates are switched in places.

Biometric systems working in verification modality have been considered, with the EER used to characterize their recognition performance. A summary of EERs reported in papers using MC for feature extraction and MM for template comparison is given in Table 1. These results already show the significant variability exhibited in literature for tests conducted on the same database with the same processing pipeline, yet resorting to different test protocols. Nonetheless, since a different number of classes has been considered in the referenced papers, it is impossible to properly evaluate the influence of the employed test protocols on the obtained performance analyzing these data. Conversely, our analysis investigates the test protocols used in the papers mentioned in Table 1, whose details are given in the following, while keeping as unaltered as possible any other aspect in the performed comparisons. Furthermore, additional testing strategies highlight the variability of the achieved performance depending on the employed protocol.

Tab. 1: Reported recognition results, UTFVP database, MC extractor, MM matcher.

Paper	# classes	# gen. comp.	# imp. comp.	EER (%)
[TV13]	325	1950	842 400	0.4
[Va14]	325	3900	1 684 800	0.49
[Va14]	108	216	46 224	1.39
[Id21]	360	5760	2 067 840	0.6 [1]
[Id21]	325	3900	1 684 800	0.7 [1]
[Id21]	192	768	146 688	1.1 [1]
[KRU14]	35	210	9 520	6.443
[KU20]	360	3600	64 620	0.37
[KU15]	360	2160	10 620	0.6 [2]

3.1 Considered Test Protocols

A test protocol is defined specifying which similarity scores are computed to estimate the achievable recognition capabilities. The possible score types are depicted in Figure 1, showing a confusion matrix obtained comparing samples belonging to two classes from the UTFVP database. The following groups of scores, identified with the symbols reported hereafter:

⟍ : scores generated by genuine comparisons, with the enrolment image compared against itself, resulting in a perfect similarity. Such scores correspond to the diagonal line of the confusion matrix, with 4 scores for each class in UTFVP;

◹ : genuine scores obtained comparing a probe sample with an enrolment sample having a lower index number (e.g., image 1 of finger 1 serves as enrolment template and image 2 of finger 1 as probe). Such group comprises genuine scores located above the diagonal line (⟍) of the confusion matrix;

◺ : genuine scores located below the diagonal line (⟍) of the confusion matrix. For every class, there are 6 ◹ and 6 ◺ scores;

◥ : impostor scores, located above the diagonal line ⟍. Such scores are calculated comparing probe samples with enrolment samples having a lower class index (e.g., enrolment image 2 of finger 1 compared with the probe image 1 of finger 2);

◣ : impostor scores, located below the diagonal line ⟍.

If scores from only one side of the diagonal (◹ or ◺; ◥ or ◣) are used, then all the samples in a database are compared only once. For consistency, if genuine scores from only *one side* (e.g., ◹) are considered, the same *side* for impostor scores (e.g., ◥) is also taken into account in our analysis. We evaluate testing strategies implemented in two of the most commonly used open-source, reproducible research frameworks available to test vein-based biometric recognition systems: PLUS OpenVein Toolkit (PLUS) [KU20], written using Matlab, and BOB [An17], a comprehensive signal processing framework, writ-

[1] Reproduced in this work

[2] Employs histogram equalization in preprocessing

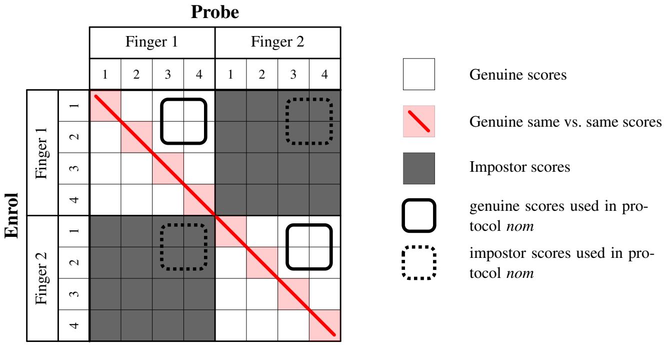

Probe
Finger 1
Finger 2
1
2
3
4
1
2
3
4
Enrol
Finger 1
Finger 2
1
2
3
4
1
2
3
4
Genuine scores
Genuine same vs. same scores
Impostor scores
genuine scores used in protocol *nom*
impostor scores used in protocol *nom*

Fig. 1: Graphical representation of the considered comparison scores.

ten in Python, designed for biometric experiments. The specific library dedicated to vein recognition is *bob.bio.vein*. In more detail, the test protocols executable in the aforementioned open-source frameworks, and employed in the performed tests on UTFVP, are:

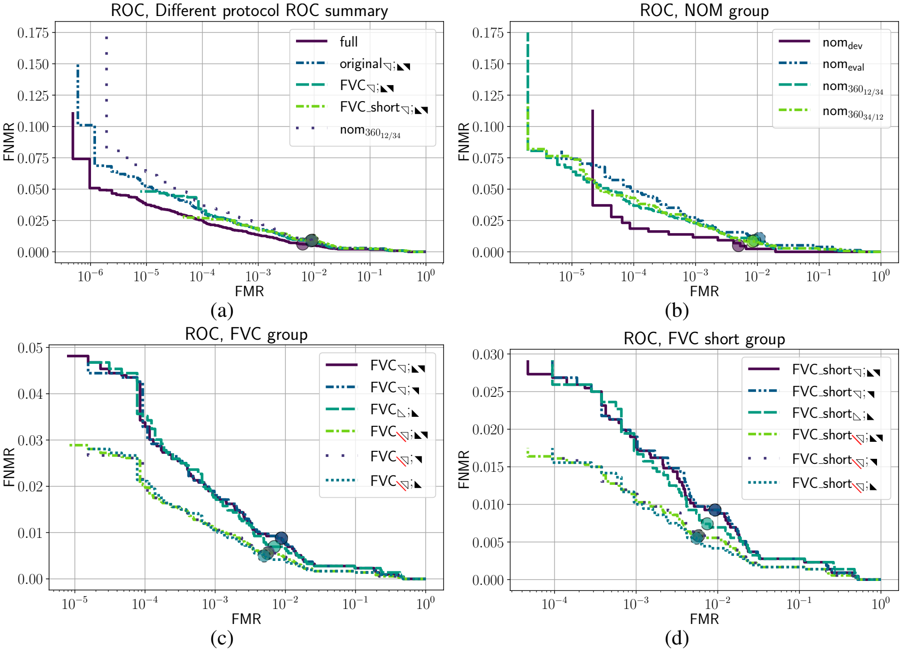
- **original** [TV13]: considering the UTFVP database, it reserves 35 class data for parameter tuning, and uses the remaining 325 classes for performance evaluation. Unsymmetrical genuine comparisons (◹ or ◺) are employed to estimate recognition capabilities, resulting in $325 \cdot 6 = 1\ 950$ genuine scores;
- **FVC** (PLUS): derived from the FVC2004 fingerprint verification contest, uses all 360 classes for testing. Since there are discrepancies between the description and formulas in [KU20] and the source code [Ka21], it is unclear whether the protocol uses ╲ genuine scores. Thus, different genuine score combinations are here used;
- **FVC_short** (PLUS): unlike FVC, a reduced number of impostor scores is considered, with only the first image from the same fingers as the enrolled sample used as impostor probe (e.g., left hand, middle fingers).
- **full** (BOB): considers all possible comparisons from 360 classes, including those in ╲, therefore consisting of $(360 \cdot 4)^2$ computations, with $360 \cdot 4 \cdot 4 = 5\ 760$ genuine and $360 \cdot 4\ \cdot 359 \cdot 4 = 2\ 067\ 840$ impostor scores;
- **1vsall** (BOB): analogous to *full*, but using only 325 class data, excluding those used for parameter estimation in the *original* protocol;
- **nom** (BOB): designed according to ISO/IEC 19795-1 suggestions, with Session-1 data used as enrolment templates, and Session-2 data as probes [Id21]. Furthermore, as recommended in [JKR15], the 60 available subjects are split into three disjoint subsets:
  - the train subset comprises samples from 10 subjects (60 fingers), used for setting the feature extractor parameters;
  - the development subset comprises samples from 18 subjects (108 fingers), used for parameter determination, including system threshold;
  - the evaluation subset comprises data from 32 subject (192 fingers), employed to estimate the achievable performance.

  We also report results derived from comparisons carried out on all the available 360 finger-vein classes, denoting with $\text{nom}_{360_{S1vsS2}}$ the use of Session-1 data for enrolment and Session-2 data for probes, and vice versa for $\text{nom}_{360_{S2vsS1}}$

Since it often happens that not enough information about impostor scores is provided, for protocols *original*, *FVC* and *FVC_short* we explore { ◥; ◣; ◣◥ }. All the scores needed in the considered test protocols are computed exploiting the BOB framework with *full* protocol. We then select the specific scores needed for each protocol and compute the associated results[3]. Such an experiment ensures that the only aspect varying between different tests is the used protocol, with no other implementation detail impacting the results reported in the next section.

4 Discussion and Conclusions

The obtained EERs are summarized in Table 2, while Figure 2 depicts the associated ROC curves in terms of FNMR vs FMR. For protocols involving all the 360 available subjects, the most notable difference is between FVC⟍◹;◣ and FVC_short◹;◣◥, with the latter EER being 86% worse than the former.



(a)

(b)

(c)

(d)

Fig. 2: ROC curves for the performed tests (FMR and FNMR reported in absolute values, not as percentage). Circles indicate EER values.

It has to be noted that ⟍ scores should not be considered to estimate recognition rates since they can severely distort the obtained results, as shown in Table 2 and by the ROC curves in Figures 2c and 2d. A considerable impact on recognition performance is produced from choices regarding the employed impostor scores, using a single probe for each impostor resulting in a misleading worsening of the performance obtained in our tests. Moreover, Figures 2c and 2d show that, when adopting a non-symmetrical comparison approach such as the MM, selecting only ◥ or ◣ impostor scores may have a significant influence on the obtained results, and therefore both groups should be considered to represent an average behaviour.

[3] Code available at `https://gitlab.com/biomedia4n6-public/biosig2021-influence-of-test-protocols/`

The ISO standard suggestion of using data from different acquisition sessions as enrolment and probe samples should be followed whenever possible. The results obtained using the full (*All vs All*) protocol are notably better than those referred to a $\mathrm{nom}360_{S1vsS2}$ approach. Nonetheless, only this latter resembles how a biometric system works in real-life applications. Such observation is also in line with the ISO standard suggestions arguing that generating more scores from fewer subjects is not equivalent to having more subjects with the same number of comparison scores. It is also to be remarked that the obtained results could depend on which session is employed to provide enrolment data, as observed in Figure 2a, where protocol $\mathrm{nom}360_{S1vsS2}$, using Session-1 data for enrolment and Session-2 as probes, turns out to be more challenging than $\mathrm{nom}360_{S2vsS1}$. The observed behaviour confirms the need for collecting multi-session databases to test biometric systems properly.

Tab. 2: Recognition results obtained exploiting scores generated according to the *full* protocol. Protocols with less than 360 classes are shaded. Corresponding ROC curves are shown in Figure 2.

Protocol	# classes	# gen. comp.	# imp. comp.	EER (%)
full ◺╲◹;◣◥	360	5 760	2 067 840	0.6238
1vsall ◺╲◹;◣◥	325	5 200	1 684 800	0.6731
parameter tuning	35	210	9 520	0.1103
original◹;◣◥	325	1 950	1 684 800	0.9149
original◹;◥	325	1 950	842 400	0.9231
original◺ ;◣	325	1 950	842 400	0.7692
FVC◹;◣◥	360	2 160	129 240	0.8797
FVC◹;◥	360	2 160	64 620	0.8793
FVC◺ ;◣	360	2 160	64 620	0.6946
FVC╲◹;◣◥	360	3 600	129 240	0.5556
FVC╲◹;◥	360	3 600	64 620	0.5780
FVC╲◹;◣	360	3 600	64 620	0.4968
FVC_short◹;◣◥	360	2 160	21 240	0.9267
FVC_short◹;◥	360	2 160	10 620	0.9244
FVC_short◺ ;◣	360	2 160	10 620	0.7423
FVC_short╲◹;◣◥	360	3 600	21 240	0.5836
FVC_short╲◹;◥	360	3 600	10 620	0.5836
FVC_short╲◹;◣	360	3 600	10 620	0.5508
$\mathrm{nom}_{\mathrm{dev}}$	108	432	46 224	0.4954
$\mathrm{nom}_{\mathrm{eval@dev}}$	192	768	146 688	1.0781
$\mathrm{nom}360_{S1vsS2}$	360	1 440	516 960	0.9032
$\mathrm{nom}360_{S2vsS1}$	360	1 440	516 960	0.8333

It can be observed that the "rule of 3" and "rule of 30" mentioned in the ISO specifications, although often overlooked, should be instead considered when reporting low error rates, e.g., in the order of 10^{-5}. Even if the EERs reported in Table 2 are not so low as to require special care, the FNMR and FMR rates reported in Figure 2 should be carefully evaluated under this perspective. As a general recommendation, if a rough performance estimate is needed, as for grid-type parameter search, it could be reasonable to not take the rules mentioned above into account, whereas they should be considered when reporting the outcomes of the performed research.

In conclusion, the performed tests and the obtained results demonstrate the need to accurately describe comprehensive test protocols when evaluating the recognition performance on a given biometric database.

References

[An17] Anjos, A.; Günther, M.; de Freitas Pereira, T.; Korshunov, P.; Mohammadi, A.; Marcel, S.: Continuously Reproducing Toolchains in Pattern Recognition and Machine Learning Experiments. In: International Conference on Machine Learning (ICML). August 2017.

[Id21] UTFVP Fingervein Database User's Guide, `https://www.idiap.ch/software/bob/docs/bob/bob.db.utfvp/v3.0.5/guide.html` (accessed February 17, 2021).

[In06] International Organization for Standardization: Information technology – Biometric performance testing and reporting – Part 1: Principles and framework. ISO/IEC, pp. 19795–1, 2006.

[In07] International Organization for Standardization: Information technology – Biometric performance testing and reporting – Part 2: Testing methodologies for technology and scenario evaluation. Standard, ISO/IEC, Geneva, CH, 2007.

[JKR15] Jain, A.; Klare, B.; Ross, A.: Guidelines for best practices in biometrics research. In: 2015 International Conference on Biometrics (ICB). pp. 541–545, 2015.

[Ka21] OpenVein-toolkit, Matcher.m, `https://gitlab.cosy.sbg.ac.at/ckauba/openvein-toolkit/-/blob/master/Matcher.m#L1008` (accessed February 17, 2021).

[KRU14] Kauba, C.; Reissig, J.; Uhl, A.: Pre-processing cascades and fusion in finger vein recognition. Lecture Notes in Informatics (LNI), Proceedings - Series of the Gesellschaft fur Informatik (GI), pp. 99–110, 01 2014.

[KU15] Kauba, C.; Uhl, A.: Sensor ageing impact on finger-vein recognition. In: 2015 International Conference on Biometrics (ICB). pp. 113–120, 2015.

[KU20] Kauba, C.; Uhl, A.: An Available Open-Source Vein Recognition Framework. In (Uhl, Andreas; Busch, Christoph; Marcel, Sébastien; Veldhuis, Raymond, eds): Handbook of Vascular Biometrics. Springer International Publishing, Cham, pp. 113–142, 2020.

[Ma15] Matey, J. R.; Quinn, G. W.; Grother, P.; Tabassi, E.; Watson, C.; Wayman, J. L.: Modest proposals for improving biometric recognition papers. In: 2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS). pp. 1–7, 2015.

[MNM07] Miura, N.; Nagasaka, A.; Miyatake, T.: Extraction of Finger-Vein Patterns Using Maximum Curvature Points in Image Profiles. IEICE - Trans. Inf. Syst., E90-D(8):1185–1194, August 2007.

[MZB16] Mery, D.; Zhao, Y.; Bowyer, K.: On accuracy estimation and comparison of results in biometric research. In: 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS). pp. 1–8, 2016.

[TV13] Ton, B. T.; Veldhuis, R. N. J.: A high quality finger vascular pattern dataset collected using a custom designed capturing device. In: Proceedings of the 2013 International Conference on Biometrics (ICB), Madrid, Spain. pp. 1–5, 2013.

[Uh20] Uhl, A.: State of the Art in Vascular Biometrics. In (Uhl, Andreas; Busch, Christoph; Marcel, Sébastien; Veldhuis, Raymond, eds): Handbook of Vascular Biometrics. Springer International Publishing, Cham, pp. 3–61, 2020.

[Va14] Vanoni, M.; Tome, P.; El Shafey, L.; Marcel, S.: Cross-database evaluation using an open finger vein sensor. In: 2014 IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS) Proceedings. pp. 30–35, 2014.

Detecting Sexual Predatory Chats by Perturbed Data and Balanced Ensembles

Parisa Rezaee Borj[1], Kiran Raja[2], Patrick Bours[3]

Abstract: Securing the safety of the children on online platforms is critical to avoid the mishaps of them being abused for sexual favors, which usually happens through predatory conversations. A number of approaches have been proposed to analyze the content of the messages to identify predatory conversations. However, due to the non-availability of large-scale predatory data, the state-of-the-art works employ a standard dataset that has less than 10% predatory conversations. Dealing with such heavy class imbalance is a challenge to devise reliable predatory detection approaches. We present a new approach for dealing with class imbalance using a hybrid sampling and class re-distribution to obtain an augmented dataset. To further improve the diversity of classifiers and features in the ensembles, we also propose to perturb the data along with augmentation in an iterative manner. Through a set of experiments, we demonstrate an improvement of 3% over the best state-of-the-art approach and results in an F_1-score of 0.99 and an F_β of 0.94 from the proposed approach.

Keywords: Imbalanced dataset, Online Conversation, Predatory Detection, Sampling.

1 Introduction

Children are vulnerable due to the new sexual norms caused by advanced technology and increased time spending on online communities where chats with unknown persons are fully possible. The children can thus be targeted by sexual predators by convincing text messages [Be16]. Detecting and identifying the predatory chats has been a major problem for parents and law enforcement agencies. However, predatory conversation detection is a complex problem as the offenders apply many techniques to avoid disclosure. The predators may not necessarily discuss about sex in the conversations, but apply different strategies and variations in time, type, and intensity to keep the victim interested and eventually exploit them. The process of gaining the trust of victim is usually called grooming [CBG06]. A common challenge in detecting online sexual predators is collecting the data as the chat providers do not make it publicly available, and accessing them requires legal permission. Of the few available datasets like PAN 2012 competition [IC12], one can observe the common problem encountered in most machine learning problems [IC12]. The datasets are heavily imbalanced due to normal conversations representing higher proportions than the predatory conversations. In reality, the percentage of sexual predatory data is 0.25% of the total online data that causes many problems for designing an automated machine learning driven detection models [LMF11]. Such composition of dataset makes the predatory detection a challenging problem as handling the imbalanced dataset for the sexual predatory detection is critical.

[1] Norwegian University of Science and Technology (NTNU), Norway, parisa.rezaee@ntnu.no

[2] Norwegian University of Science and Technology (NTNU), Norway, kiran.raja@ntnu.no

[3] Norwegian University of Science and Technology (NTNU), Norway, patrick.bours@ntnu.no

In this work, we present a new approach for detecting predatory chat detection by providing a new strategy in handling the imbalance to provide a new approach. Specifically, we present an approach which first creates a balanced class distribution by increasing the minor class with a set of augmented and perturbed data. The balanced class distribution is increased until a 50% balance is obtained by simply augmenting and perturbing the data. With the refined class distribution, we create an ensemble of HistogramBoostedGradient classifiers which directly benefit from the augmented and perturbed data in selecting different set of features for creating ensembles. With the set of experimental validation, we evaluate the proposed approach on PAN 2012 [IC12] dataset where the proposed approach outperforms the existing approaches. The proposed approach results in a precision of 99%, a recall of 99% and a $F_{0.5}$ score of 94% with a gain of 3% over the recent work which reported a recall of 96%. In the rest of this paper, we first present briefly detail the dataset employed and discuss the imbalanced nature of the dataset in Section 2. We then list out few related works which have tried to address the imbalanced nature of sexual predatory data for the convenience of the reader in Section 3. We present the proposed approach in Section 4 followed by the discussion on results in Section 5. To the end, we make concluding remarks and list out few potential future works.

2 Database for Sexual Predatory Detection

A chat conversation typically is one of three types of conversations such as (a) a conversation without sexual topics, (b) a conversation between adults on sexual topics, or (c) a conversation between a predator and a minor victim which is considered as a predatory conversation. The PAN 2012 [IC12] competition dataset deals with the third category and the data contains the conversations between police officers who pretended to be minors and convicted predators extracted from the PJ website (`http://www.perverted-justice.com/`). The data also contains the ordinary chat without any sexual content extracted from `http://www.irclog.org`, and sexual conversation between consenting adults from Omegle (`www.omegle.com`). In addition to the conversation data, the data also consists of a unique conversation ID to distinguish between the conversations. Each message in a conversation further includes an author ID, a timestamp, and the text of the message [BRB20]. In training data, there are 951 predatory conversations and 8477 non-predatory conversations. The test data contains 1697 predatory samples and 19922 non-predatory conversations. More detail about the applied data and the pre-processing method can be found in [IC12] and [BRB20].

2.1 Constraints of Dataset

As with any other type of data investigation, predatory detection requires pertinent data. The amount of predatory data is much lower than the normal chatlogs, making it challenging to find appropriate subset of data. Further, analyzing the data, we note the heavily imbalance in the data where predatory data is less than 0.25% of the total data [LMF11]. Such imbalance leads to sub-optimal classifiers favoring one class over the other resulting either in underfitting or overfitting. When the training data is highly imbalanced, it becomes more critical as the class with fewer samples is severely under-sampled and causes to not capturing the complete information of the given data. If one does not consider the class imbalance problem, the learning techniques can be overwhelmed by the majority class, and the minority class will be easily ignored. An imbalanced classification problem is a problem where the datasets have skewed distributions. It has several characteristics, including class overlapping, small sample size, and small disjuncts [Li17]. A predatory dataset can suffer from all these characteristics as there are many overlaps and disjuncts between a predatory conversation and a non-predatory one. In addition, the number of predatory conversation samples is much lower than the non-predatory ones. Also, a chat conversation might contain some sentences or topics that are common in both predatory and non-predatory talks.

3 Related Works

Earlier research works mainly have used conventional methods for sexual predatory detection disregarding the imbalanced dataset [EHS11, BRS12, Vi12, IC12, BRS14, Eb16, BRB20]. However, we restrict our focus to few sample works and focus on works that deal with data imbalance problem within predator detection. Cardei et al. [CR17] tested several techniques for coping with the imbalanced data in sexual predatory detection, such as cost-sensitive technique and sampling techniques including BalanceCascade [LWZ08], and CBO - a clustering-based method using k-means [CJK04]. The authors found that the cost-sensitive model where a cost matrix gave a penalty for misclassifying gained the best performance experimenting on PAN 2012 dataset [IC12]. Their proposed model had two stages where it investigated behavioural features that cover the users' behaviour on the online platform. It considered the ratio of questions, underage expressions, slang words, and the bag of word feature vectors and obtained an $F_{0.5}$ score of 0.95 [CR17]. Zuo et al. [Zu19] presented an adaptive fuzzy method for artificial neural networks (ANNs) to address the imbalanced data in sexual predatory detection. They used conventional fuzzy inference based on dense rule and fuzzy rule interpolation to handle the imbalanced dataset in the sexual predatory detection problem. Their method was a combination of an adaptive fuzzy inference-based activation function with the artificial neural networks (ANNs) that extracted BoW and TFIDF as feature sets, classified the data sets, and gained an accuracy of 0.766.

4 Proposed Approach

The overall pipeline of the proposed approach is illustrated in the Figure 1. The proposed approach starts with the preprocessing of the data, followed by feature extraction using Word2Vec [Ro14] and the proposed strategy of learning the ensemble classifiers as detailed below in this section. As the data is heavily noisy, we first preprocess the data to eliminate the irrelevant entries from PAN 2012 dataset. Based on the common properties of the predator victim contacts, we assert that these kinds of conversations have only two authors. We therefore eliminate all the other conversations that involve multi-parties or have only one author. Further, we discard all the conversations with less than seven messages as such conversations contain too little information to be classified as either predatory or non-predatory. Further, as another refinement, we analyze chat messages to eliminate non-English words that did not provide any special meaning or do not follow standard grammar in a remote manner, have many slang and emoticons. To keep the information as much as it is possible, no stemming or lemmatization in the preprocessing of the data was performed. We then extract the features the from the chat logs that cover the word relationships in different contexts with a low dimensional feature vector. In order to fully exploit the word analogies, we extract features using the Word2Vec embedding model with pre-trained networks with a 300 dimensional vectors. Word2Vec provides distributed representation of the text data which we further use to design the classifier.

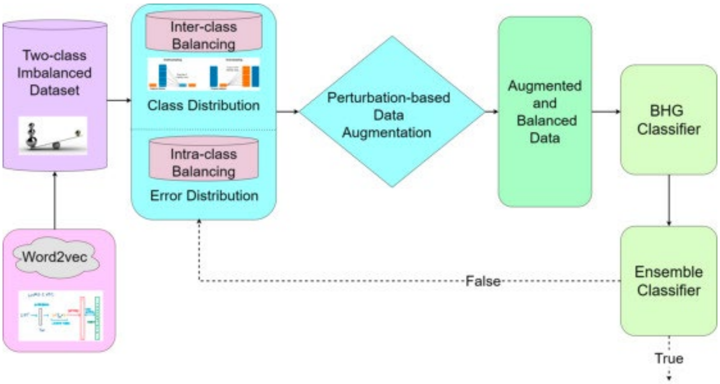


Fig. 1: Proposed approach for predatory chat detection

4.1 Balanced and Augmented Dataset

Given the dataset $\mathscr{D}$ with n classes and m features, $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^{n}$, each class can consist of k samples. When all the classes have equal number of samples, i.e., $k \approx k_{ave}$ one can effectively learn a classifier. However, when the number of samples k_i for a chosen class i is significantly lower than average number of samples from all other classes, the classifier is challenged with skewed data distribution. As it happens, the number of predatory samples is much lower than the number of non-predatory conversation with a sample distribution ratio $1.00 : 8.91$ for predatory to non-predatory samples in our case. Thus, irrespective of the sampling approaches to be used, the minor class will contribute to imbalance for learning a classifier. Thus, we first propose to create an augmented dataset $\mathscr{D}'$ for T number of iterations. For each class C_i in the n classes, we employ two kind of sampling such that class with higher samples is under-sampled and the class with lower samples is over-sampled. In order to achieve this, we simply resort to progressively balanced hybrid sampling using the class distribution. The balanced classes for each iteration is then used to compute the error distribution for the true class distribution and inverse error distribution. Further, as the number of samples in one class can be much higher than the other class, for instance, in our case non-predatory conversations are much higher than the predatory samples, we augment the features in both classes such that the minor class is represented equally with a set of perturbation. The perturbation factor p therefore leads to new samples of the minor class which can be represented as $x' \rightarrow x + \alpha . x^p$ where α is a linear scaling factor. Thus, the new augmented samples lead to creation of $\mathscr{D}'$. For each of these samples obtained, we obtain new class distribution C' for a given iteration t in total number of iterations T. Using the newly augmented dataset $\mathscr{D}'$ with new class distribution C' with balanced, augmented and perturbed data, we learn a classifier Histogram Gradient

Boosted Decision Trees as detailed in the next section. In every iteration t, the class distribution and inverse class distribution is used to balance the samples chosen to learn the classifier.

Algorithm 1: Pseudocode for Proposed Approach

```
initialization : T iterations, Number of base estimators, Number of bins for
  HistogramBoostedGradient;
procedure CLASS DISTRIBUTION BALANCE;
t ←∈ $Titerations$
while $K$ do
    Compute class distributions;
    Compute balanced hybrid sampling;
    Compute the expected class distribution (number of samples from each class);
    Compute the intra-class balanced sampling weights by inversing prediction error
    distribution;
    Undersample or oversample the features;
end

procedure AUGMENT DATA;
Perturb and augment data;
    procedure CREATE ENSEMBLE;
For each set of augmented data, create classifier - HistogramBoostedGradient;
Fit HistogramBoostedGradient estimator;
Choose features if the loss is less than iteration t-1;
```

The Algorithm 1 represents the pseudocode of the approach.

4.2 Histogram Gradient Boosted Decision Trees

Given the augmented, balanced and perturbed dataset $\mathscr{D}'$ with n samples and m features, $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^m$ and $y_i \in \mathbb{R}$, the predictions from the boosted decision tree model, $\hat{y}_i$, is defined as a tree-based additive ensemble model, $\phi(x_i)$, comprising of K additive functions, f_k, defined as:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i),\ f_k \in \mathscr{F}$$

where $\mathscr{F} = \{f(x) = w_{q(x)}\}$ is a collection of Classification and Regression Trees, such that $q(x)$ maps each input feature x to one of T leaves in the tree by a weight vector, $w \in \mathbb{R}^T$. Given the function defined above, the Gradient Boosted algorithm minimizes the following regularized objective function:

$$\tilde{\mathscr{L}} = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k)$$

where $l(y_i, \hat{y}_i)$ is the loss function of the ith sample between the prediction $\hat{y}_i$ and the target value y_i, and $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2$ is the regularization component to penalize k^{th} tree in growing additional leaves by λ - a regularization parameter and a weight vector w. We approximate the loss function using a second-order Taylor expansion [CG16], and we omit the details for the brevity of the paper considering the page limit.

4.3 Ensemble Construction

Based on the augmented features selected in each iteration, a classifier is chosen if the loss $l(y_i, \hat{y}_i)$ is the loss function of the ith sample between the prediction $\hat{y}_i$ and the target value y_i, and $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2$ is the regularization component to penalize k^{th} tree in growing additional leaves by λ - a regularization parameter and a weight vector w.

Ref	**Accuracy**	F_1	F_β
Bogdanova et al. [BRS12]	0.97	-	-
Villatoro et al. [Vi12]	0.92	0.87	0.93
Borj & Bours [BB19]	0.98	0.86	-
Fauzi & Bours [FB20]	0.95	0.90	0.93
Bours & Kulsrud [BK19]	-	0.94	0.97
Borj et al. [BRB20]	0.99	0.96	-
Ebrahimi et al. [ESO16]	-	0.80	-
Ebrahimi et al. [Eb16]	0.99	0.77	-
Imbalance based approaches			
Cardei et al. [CR17]	-	-	0.95
Zuo et al. [Zu19]	0.76	-	-
Proposed Model	**0.99**	**0.99**	**0.94**

Tab. 1: Performance of various approaches against proposed approach. The blocks in gray color indicate the approaches that handle data imbalance and can be directly compared to our proposed approach.

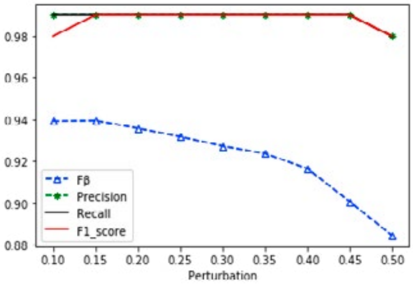


Fig. 2: Performance variation to perturbation factor in data.

5 Experimental Results

For detection of predatory conversation, all the messages of a single conversation were merged into a single text block. Then, we extracted the Word2Vec feature vector for each of the merged texts. The main focus of this analysis is to handle the imbalanced nature of the dataset applying the proposed method. Thus, we select two state-of-the-art approaches which are close to our work to provide a comparison. Specifically, we compare our results against Cardei et al. [CR17] and Zuo et al. [Zu19] who propose strategies to handle the imbalance in the predatory data. Further, we also compare our results against other state-of-the-art approaches to give a broader comparison. Predatory detection techniques have been evaluated using different metrics such as accuracy, precision, recall, and F_1-score. Further, to avoid many false-positive detection F_β is also recommended as another primary metric for analyzing the performance [IC12] with $\beta = 0.5$. Table 1 demonstrates the obtained results and compares them with the baseline of various works. The proposed approach obtains a gain of 3% over the best benchmark, while it gains more than 23% more accuracy compared to the earlier approach [Zu19] in a similar category of using balancing strategies. Further, we also analyze the effect of perturbation factor in augmenting the dataset, and the obtained accuracy is presented in Figure 2. As noted from Figure 2, the performance changes slightly when the perturbation factor is increased to more than 20%. Despite the slight drop in performance, one can note the superiority of the proposed approach as compared to the accuracy reported in Table 1. Thus, we deduce that the perturbation factor should not be more than 50% to obtain a reliable classification accuracy.

6 Conclusion

Predatory conversation detection based on text messages is a crucial problem to avoid exploiting under-aged or minors for sexual favors. Owing to the limited real datasets available, current works employ a standard dataset with less than 10% predatory data leading to a heavy imbalance in the dataset resulting in a classifier that may be sub-optimal. This work has proposed a new approach for handling the imbalanced nature of predatory data by hybrid sampling and class re-distribution to obtain an augmented dataset. Further, to improve the diversity of classifiers and features in the ensembles, this work also proposes to perturb the data along with augmentation in an iterative manner. With the set of experiments on the state-of-the-art dataset, we demonstrate that the proposed approach obtains an improvement over the best state-of-the-art approach by 3% and results in a F_1-score of 0.99 and a F_β of 0.94. Unlike this work, in future works, we also intend to explore different feature extraction approaches to validate the scalability of the proposed approach for predatory detection. Further, this work can also be extended by generating the predatory data through advanced approaches, including Generative Adversarial Networks.

References

[BB19] Borj, Parisa Rezaee; Bours, Patrick: Predatory conversation detection. In: 2019 International Conference on Cyber Security for Emerging Technologies (CSET). IEEE, pp. 1–6, 2019.

[Be16] Bentley, Holly; O'Hagan, Orla; Raff, Annie; Bhatti, Iram: How safe are our children. The most comprehensive overview of child protection in the UK, 2016.

[BK19] Bours, Patrick; Kulsrud, Halvor: Detection of cyber grooming in online conversation. In: 2019 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, pp. 1–6, 2019.

[BRB20] Borj, Parisa Rezaee; Raja, Kiran; Bours, Patrick: On Preprocessing the Data for Improving Sexual Predator Detection: Anonymous for review. In: 2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMA. IEEE, pp. 1–6, 2020.

[BRS12] Bogdanova, Dasha; Rosso, Paolo; Solorio, Thamar: On the impact of sentiment and emotion based features in detecting online sexual predators. In: Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis. pp. 110–118, 2012.

[BRS14] Bogdanova, Dasha; Rosso, Paolo; Solorio, Thamar: Exploring high-level features for detecting cyberpedophilia. Computer speech & language, 28(1):108–120, 2014.

[CBG06] Craven, Samantha; Brown, Sarah; Gilchrist, Elizabeth: Sexual grooming of children: Review of literature and theoretical considerations. Journal of sexual aggression, 12(3):287–299, 2006.

[CG16] Chen, Tianqi; Guestrin, Carlos: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794, 2016.

[CJK04] Chawla, Nitesh V; Japkowicz, Nathalie; Kotcz, Aleksander: Special issue on learning from imbalanced data sets. ACM SIGKDD explorations newsletter, 6(1):1–6, 2004.

[CR17] Cardei, Claudia; Rebedea, Traian: Detecting sexual predators in chats using behavioral features and imbalanced learning. Nat. Lang. Eng., 23(4):589–616, 2017.

[Eb16] Ebrahimi, Mohammadreza; Suen, Ching Y; Ormandjieva, Olga; Krzyzak, Adam: Recognizing predatory chat documents using semi-supervised anomaly detection. Electronic Imaging, 2016(17):1–9, 2016.

[EHS11] Egan, Vincent; Hoskinson, James; Shewan, David: Perverted justice: A content analysis of the language used by offenders detected attempting to solicit children for sex. Antisocial behavior: Causes, correlations and treatments, 20(3):273297, 2011.

[ESO16] Ebrahimi, Mohammadreza; Suen, Ching Y; Ormandjieva, Olga: Detecting predatory conversations in social media by deep convolutional neural networks. Digital Investigation, 18:33–49, 2016.

[FB20] Fauzi, Muhammad Ali; Bours, Patrick: Ensemble Method for Sexual Predators Identification in Online Chats. In: 2020 8th International Workshop on Biometrics and Forensics (IWBF). IEEE, pp. 1–6, 2020.

[IC12] Inches, Giacomo; Crestani, Fabio: Overview of the International Sexual Predator Identification Competition at PAN-2012. In: CLEF (Online working notes/labs/workshop). volume 30, 2012.

[Li17] Lin, Wei-Chao; Tsai, Chih-Fong; Hu, Ya-Han; Jhang, Jing-Shang: Clustering-based undersampling in class-imbalanced data. Information Sciences, 409:17–26, 2017.

[LMF11] Latapy, Matthieu; Magnien, Clémence; Fournier, Raphaël: Quantifying paedophile queries in a large p2p system. In: 2011 Proceedings IEEE INFOCOM. IEEE, pp. 401–405, 2011.

[LWZ08] Liu, Xu-Ying; Wu, Jianxin; Zhou, Zhi-Hua: Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39(2):539–550, 2008.

[Ro14] Rong, Xin: word2vec parameter learning explained. arXiv preprint arXiv:1411.2738, 2014.

[Vi12] Villatoro-Tello, Esaú; Juárez-González, Antonio; Escalante, Hugo Jair; Montes-y Gómez, Manuel; Pineda, Luis Villasenor: A Two-step Approach for Effective Detection of Misbehaving Users in Chats. In: CLEF (Online Working Notes/Labs/Workshop). volume 1178, 2012.

[Zu19] Zuo, Zheming; Li, Jie; Wei, Bo; Yang, Longzhi; Chao, Fei; Naik, Nitin: Adaptive activation function generation for artificial neural networks through fuzzy inference with application in grooming text categorisation. In: 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). IEEE, pp. 1–6, 2019.

On the Relevance of Minutiae Count and Distribution for Finger Vein Recognition Accuracy

Michael Linortner[1], Andreas Uhl[1]

Abstract: Vein recognition usually uses binary features, but besides deep learning-based approaches key-point and minutiae-based ones started to become popular as well. Statistical measures for vein minutiae points, like spatial point distribution, have not been investigated in literature so far. In this work the number of vein minutiae points and their spatial distribution is analyzed in relation to recognition accuracy. The goal is to initiate a discussion on statistical behavior of vein minutiae points and deriving possible quality measures for vein minutiae point sets.

Keywords: Vein recognition, vein minutiae distribution, spatial point distribution.

1 Introduction

Finger or hand vein recognition has become an established and accepted technology in biometrics. One approach is to use branches of the blood vessels as minutiae points analogously to minutiae points in fingerprint recognition. Due to low image quality of the raw sample images, a crucial step is the parameter selection for the feature extraction process, consisting of preprocessing, image enhancement and vein segmentation. Varying these parameters alters the segmented vein output from which minutiae points are extracted. As a consequence, the minutia sets may vary. In literature there is no generally accepted approach to derive optimal parameters for vein minutiae extraction and there is no accepted and standardized quality measure for vein images with respect to minutiae-based vein recognition. It is of interest to gather knowledge on how variations in the minutiae sets influence the recognition accuracy. The first aspect worth investigating is the impact of the number of minutia points on the recognition performance. Further it is of interest to learn about the minutiae's spatial point distribution, whether they follow spatial randomness, tend to disperse or to cluster and hence, the influence on the recognition accuracy. Knowledge about minutia point distribution is a key aspect in biometric individuality studies. For fingerprint minutiae points there exists corresponding literature, for example [Bo04, Sc79, PPJ01, JY06].
For vein minutiae no statistical investigations regarding number of minutiae and minutiae point distribution have been done so far. The goal of this work is to initiate a discussion on statistical behavior of vein minutiae points and deriving possible quality measures for vein minutiae extraction. Therefore, the relation between the number of vein minutiae points and recognition accuracy is analyzed. To determine spatial distribution of minutiae points, two measures are employed, which characterize the spatial distribution of minutiae points

[1] Department of Computer Sciences, University of Salzburg, Salzburg, Austria, {mlinortner,uhl}@cs.sbg.ac.at

in a single number. These numbers are set in context to the recognition performance and the number of minutiae points to reveal potential correlations.
In a previous work [LU21] we showed that utilizing finger vein minutiae points in combination with standard minutiae-based fingerprint recognition software can compete with and even outperform classic vein recognition techniques in terms of recognition accuracy and comparison time, which motivates this work to analyze statistical behavior of finger vein minutiae points.

2 Methods

Two distance-based point pattern measures are utilized to describe the spatial point distribution of vein minutia points. Let M denote a (minutiae) point set containing n points p, U_{ij} a set containing the Euclidean distances $d(p_i, p_j)$ of each point $p_i, p_j \in M, i \neq j$ and the mean nearest neighbor distance $\bar{d}_{\text{min}}$ with $\bar{d}_{\text{min}} = \frac{1}{n}\sum_{i=1}^{n} \min\{U_{ij}\}$. The overall density λ of a point pattern can be estimated with $\hat{\lambda} = n/A$, where $\hat{\lambda}$ is the estimated intensity, n the number of points in M within a region of area A. Under assumption of complete spatial randomness (CSR) the expected value for $\bar{d}_{\text{min}}$ is $E(d) = \frac{1}{2\sqrt{\lambda}}$. Thus, a ratio R can be defined [OU10]:

$$R = \frac{\bar{d}_{\text{min}}}{0.5\sqrt{A/n}} \tag{1}$$

describing a pattern's point distribution relative to CSR. An R value < 1 indicates a tendency towards clustering and > 1 towards dispersing, respectively.
The second measure utilizes the K-function $K(t)$, which incorporates all distances between a point and its neighbors within a radial distance t. The estimator for $K(t)$ is defined as follows [Di14]:

$$\hat{K}(t) = \frac{A}{n(n-1)} \sum_{i=1}^{n} \sum_{i \neq j} \frac{1}{w_{ij}} I(U_{ij} < t) \tag{2}$$

with $I(\cdot)$ as indicator function and w_{ij} computed as in equations (4.16) and (4.17) in [Di14]. The variance $v_{LS}(t)$ of $\hat{K}(t)$ is computed as in equation (4.19) in [Di14]. The expected value of $K(t)$ under CSR is πt^2 and the difference is given with $D(t) = \hat{K}(t) - \pi t^2$. $D(t) > +2\sqrt{v_{LS}(t)}$ indicates clustering, $D(t) < -2\sqrt{v_{LS}(t)}$ indicates dispersion whereas in between the CSR assumption holds [Di14]. $D(t)$ has been used in [JY06] to investigate the distribution of fingerprint minutiae points. In this work we use

$$Q = F\left(D(t), +2\sqrt{v_{LS}(t)}), t_{\text{min}}, t_{\text{max}}\right) \tag{3}$$

as our second measure to describe the minutiae point distribution in a single number. $F(\cdot)$ computes the area between $D(t), +2\sqrt{v_{LS}(t)})$ in the interval $t \in [t_{\text{min}}, t_{\text{max}}]$. Thus, Q gives a measure for the tendency to cluster.

3 Experiments

The experiments are conducted on four publicly available finger vein data sets: the UTFVP data set [TV13], containing 1440 images from 360 fingers, the HKPU-FV data set (1. session) [KZ12], containing 1872 images from 312 fingers and two data sets of the PLUS-3FV database [KPU18] each consisting of 1880 images from 360 fingers captured under near-infrared laser illumination. One data set shows the palmar view of the finger and the other the dorsal view. They are denoted a PLUS-Las-P and PLUS-Las-D, respectively.
Four standard minutiae-based fingerprint recognition tools are utilized to compute the recognition accuracy for each setting, expressed as equal error rate (EER): two publicly available tools, the Bozorth3 as part of the NIST Biometric Image Software (NBIS) Release 5.0.0[2] and the minutiae cylinder code (MCC) SDK [CFM10, CFM11, FMC12, FMC14], as well as two state of the art commercial products, the IDKit SDK Version 9.0[3] and the VeriFinger 11.2 Extended SDK[4]. As these tools are design for fingerprint recognition they are not suitable to retrieve minutiae points from finger vein images, but their minutiae-based comparison algorithms are utilized. Therefore, the vein minutiae points are extracted as proposed in [LU21] and stored in a standardized format, in order to be usable for the comparison algorithms in the above mentioned fingerprint recognition tools. Briefly summarized, the minutiae points in [LU21] are extracted by firstly applying image enhancement on the a vein sample. Then the veins are segmented, subsequently thinned and from the resulting skeleton bifurcation points are retrieved which serve as minutiae points. On each data set 308 different parameter settings are applied to extract a variation of minutiae sets. The parameters for image enhancement, vein segmentation and spur removal in the thinning process are modified and the combinations of all selected parameter values generate 308 different settings, where each individual setting produces a single set of minutiae points on which in the following the statistical analysis is performed. We use the same parameter settings as in [LU21] (ergo the same minutiae points), extended by additional parameter settings to retrieve additional minutiae sets with a lower number of minutiae points in average to extend the variability. As in [LU21] the results in terms of recognition accuracy show it is suitable to utilize finger vein minutiae points in combination with minutiae-based fingerprint comparison software as a biometric recognition technique. Therefore, it motivates to investigate the extracted minutiae points regarding correlations between number or distribution and recognition performance employing the proposed measures. For more detailed information on utilizing finger vein minutiae points in combination with standard fingerprint recognition tools and the performance in recognition accuracy and template comparison time, also related to standard vein recognition techniques, the interested reader is referred to [LU21].

On each minutiae set three indicators are investigated: the number of minutiae points, mean nearest neighbor ratio R and Q, all in relation to the EER. To compute a single value for each setting representing the whole data set, for each indicator the average is computed over all vein samples in a data set. Thus, the mean number of minutiae points per sample is computed with $1/N \sum_i |M_i|$, with N as the number of samples in a data set and $|M_i|$ as the

[2] https://www.nist.gov/services-resources/software/nist-biometric-image-software-nbis
[3] https://www.innovatrics.com
[4] https://www.neurotechnology.com/verifinger.html

amount of minutiae points in a sample i. The mean nearest neighbor ratio is computed with $\bar{R} = 1/N\sum_i R_i$, applying equation (1) on M_i to obtain R_i and $\bar{Q} = 1/N\sum_i Q_i$ by applying equation (3) on M_i to obtain Q_i. The parameter t is varied between 0 and $\min(a,b)/4$ where a and b describe the width and height of the finger vein region of interest (ROI) from which the minutiae points have been extracted.

4 Results

Figure 1 shows a minutiae point density map for each data set averaged over all containing image samples exemplary using the parameter configuration leading to extracted minutiae points on which VeriFinger performs best. It visualizes how densely minutiae points populate certain areas of the used finger vein ROI. Therefore, the ROI region is tiled into bins of size 4×4 pixels and a density histogram is computed. The visual impression suggests that the minutiae points are somehow randomly distributed, but especially on the PLUS-Las and HKPU-FV data sets a tendency for clustering in certain areas can be observed. There are noticeable differences between the minutiae point density maps of the different data sets.

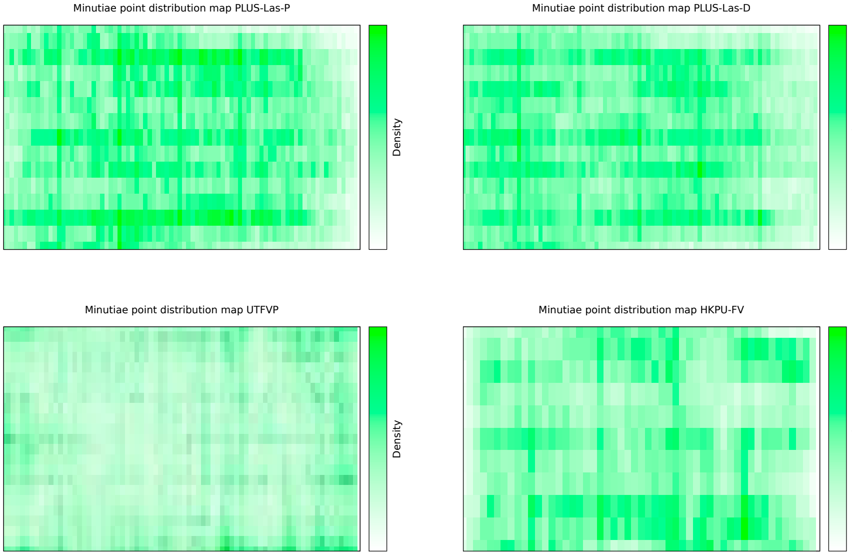


Fig. 1: Density map of vein minutiae points within the utilized finger ROI.

Figure 2 shows the relation between recognition performance and mean amount of minutiae points per finger vein sample. On the PLUS data set the behavior is as expected: there is a range, between 40 to 60 minutiae points per sample, where the recognition methods perform best, while with a increasing or decreasing number of points the recognition performance decreases. Having too few minutiae points means that important information is lost. On the other end, if there are too many false and noisy minutiae points included it

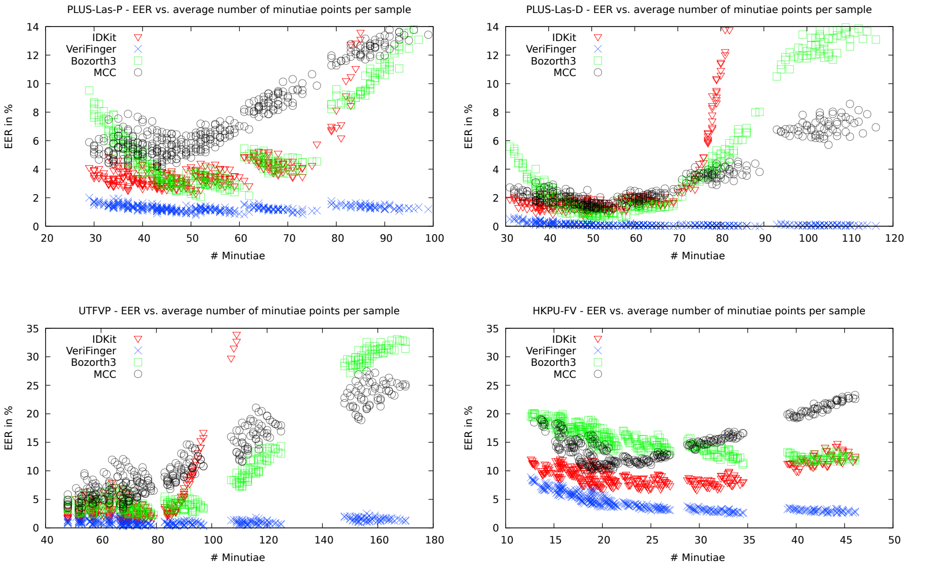


Fig. 2: EER vs. average number of minutiae points per vein sample.

causes a drop in recognition accuracy. Interestingly, VeriFinger is able to maintain high recognition accuracy with an increasing number of minutiae. On the UTFVP data set it was not possible to extract minutiae sets which contain less than 40 minutia points per sample on average, so we can only see the trend of decreasing EER with an increasing number of points. The HKPU-FV is a challenging data set for minutiae-based methods. Compared to the other data sets the samples' image quality is lower, often parts of a finger are overexposed and without visible vein structure. VeriFinger and Bozorth3 show a clear trend of increasing recognition accuracy towards 40 and more minutiae per sample, while for IDKit and MCC the optimum is around 20 minutiae points for the HKPU-FV data set.

In figure 3 the relation of the mean nearest neighbor ratio $\bar{R}$ to the EER is plotted. First we can observe that for the PLUS data sets there a is general tendency towards clustering ($\bar{R} < 1$). The EER decreases with increasing $\bar{R}$ which could indicate that randomly distributed minutiae points are better for the recognition accuracy than clustered points. For the UTFVP data set it can be stated that the points in all extracted minutiae sets are randomly distributed because $\bar{R}$ varies within a narrow range around 1. This suggests, that the trend visible in the plot is most likely caused by the amount of minutiae point rather than by the distribution. On the HKPU-FV data set the minutiae points slightly tend to disperse.

The relation between the third measure $\bar{Q}$ and EER is visualized in figure 4. It can bee seen that for each data set every single setting produces minutiae sets which, on average, show some clustering. Note that it is no contradiction to the results produced by measuring $\bar{R}$, where it is indicated that there are settings which generate on average randomly distributed minutia sets ($\bar{R} = 1$). $\bar{R}$ is a global measure on a point set, while $D(t)$ operates on different scales (by varying t) and can detect local clusters even when on global scale the distribution

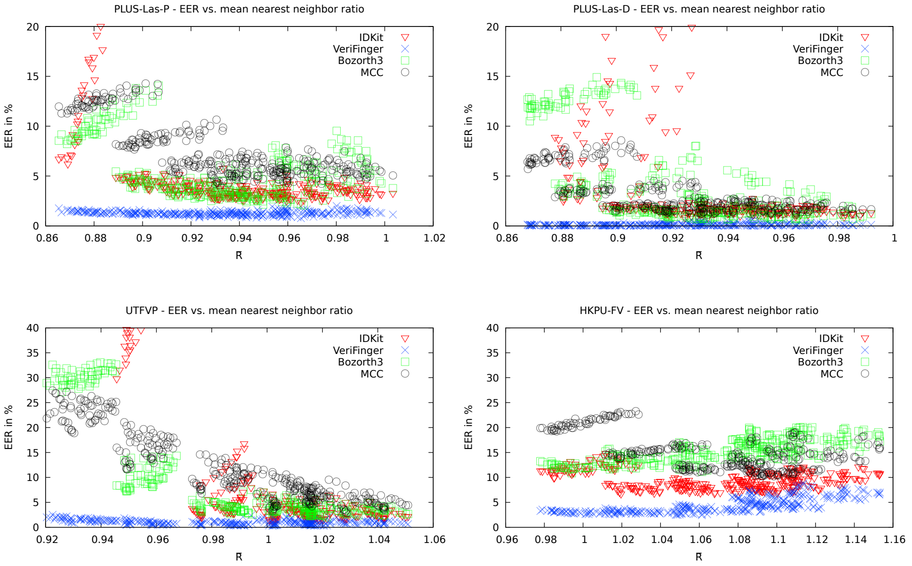


Fig. 3: EER vs. $\bar{R}$.

is closer to random. Clustering at any scale is accumulated in $\bar{Q}$. The standard deviation of Q is around 2 times $\bar{Q}$ which indicates that in a data set there are minutiae sets which do not cluster and others may cluster considerably. VeriFinger's recognition performance shows a clear trend of dropping with increasing $\bar{Q}$ (more clustering). On the PLUS data set there are more outliers for the other recognition methods but those settings which produce the best recognition accuracy follow the same trend as VeriFinger. On the UTFVP data set the $\bar{Q}$ values of the MCC methods are more scattered, but they indicate that for a low EER the $\bar{Q}$ value needs to be small.

To investigate whether there is a correlation between the average minutiae count and the distribution measures $\bar{R}$ and $\bar{Q}$, respectively, they are plotted against each other. As all data sets show the same behavior only the plots for the PLUS-Las-P data sets are depicted exemplary in figure 5. For $\bar{R}$, shown in figure 5(a), a trend is noticeable: with an increasing number of minutiae points $\bar{R}$ decreases, meaning the points start to cluster. Figure 5(b) shows that there is no obvious correlation between $\bar{Q}$ and the amount of minutiae. Combining the insights gained by the coherences in figure 2 and 4 we know that the recognition performance tends to be high when an optimum number of minutiae points is available and $\bar{Q}$ is low. Using this information in combination with the results in plot 5(b) may help to identify settings which produce “high quality” minutiae set on which a high recognition accuracy can be achieved.

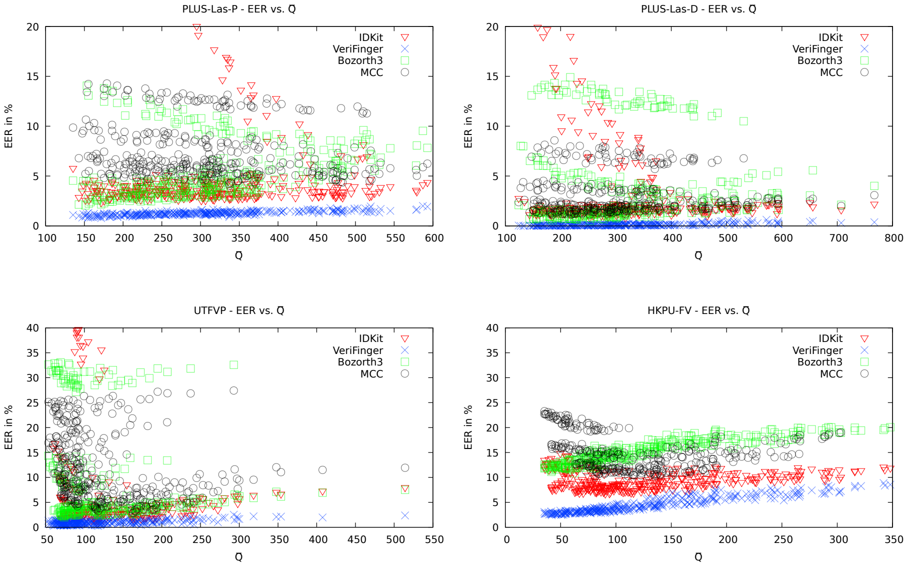
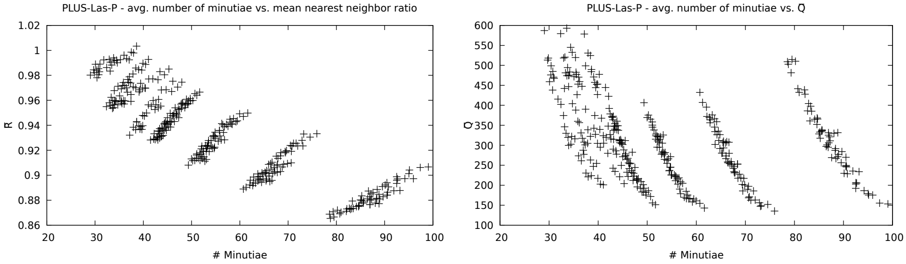


Fig. 4: EER vs. $\bar{Q}$.



(a) (b)

Fig. 5: Average number of minutiae points per sample vs. $\bar{R}$ (a) and $\bar{Q}$ (b).

5 Conclusion

In this work the influence of the number of finger vein minutiae points and their spatial distribution on the recognition accuracy was investigated. The results showed that there are correlations between recognition accuracy, minutiae number and minutiae distribution. Based on the discussion of the results for a possible usage of the proposed methods to estimate the quality of vein minutiae sets, the outcome of this work motivates further investigation on the proposed measures or a combination of them to derive a suitable measure for estimating the quality of vein samples and/or vein minutiae sets regarding minutia-based recognition techniques.

6 Acknowledgment

This project was partly funded by the Austrian Science Fund FWF project “Advanced Methods and Applications for Fingervein Recognition” under grant No. P 32201-NBL.

References

[Bo04] Bolle, R. M.; Connell, J.; Pankanti, S.; Ratha, N. K.; Senior, A. W.: Guide to Biometrics. Springer New York, 1 edition, 2004.

[CFM10] Cappelli, R.; Ferrara, M.; Maltoni, D.: Minutia Cylinder-Code: A New Representation and Matching Technique for Fingerprint Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(12):2128–2141, Dec 2010.

[CFM11] Cappelli, R.; Ferrara, M.; Maltoni, D.: Fingerprint Indexing Based on Minutia Cylinder-Code. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(5):1051–1057, 2011.

[Di14] Diggle, Peter J: Statistical analysis of spatial and spatio-temporal point patterns. CRC press, 3rd edition, 2014.

[FMC12] Ferrara, M.; Maltoni, D.; Cappelli, R.: Noninvertible Minutia Cylinder-Code Representation. IEEE Transactions on Information Forensics and Security, 7(6):1727–1737, 2012.

[FMC14] Ferrara, M.; Maltoni, D.; Cappelli, R.: A two-factor protection scheme for MCC fingerprint templates. In: 2014 International Conference of the Biometrics Special Interest Group (BIOSIG). pp. 1–8, 2014.

[JY06] Jiansheng Chen; Yiu-Sang Moon: A Statistical Study on the Fingerprint Minutiae Distribution. In: 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings. volume 2, pp. II–II, 2006.

[KPU18] Kauba, Christof; Prommegger, Bernhard; Uhl, Andreas: The Two Sides of the Finger - An Evaluation on the Recognition Performance of Dorsal vs. Palmar Finger-Veins. In: Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG’18). Darmstadt, Germany, pp. 1–8, 2018.

[KZ12] Kumar, A.; Zhou, Y.: Human Identification Using Finger Images. IEEE Transactions on Image Processing, 21(4):2228–2244, April 2012.

[LU21] Linortner, Michael; Uhl, Andreas: Towards Match-on-Card Finger Vein Recognition. In: Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security. IH&amp;MMSec ’21, Association for Computing Machinery, New York, NY, USA, pp. 87–92, 2021.

[OU10] O’Sullivan, David; Unwin, David: Geographic information analysis. John Wiley & Sons, 2nd edition, 2010.

[PPJ01] Pankanti, S.; Prabhakar, S.; Jain, A.K.: On the individuality fingerprints. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. volume 1, pp. I–I, 2001.

[Sc79] Sclove, Stanley L.: The Occurrence of Fingerprint Characteristics as a Two-Dimensional Process. Journal of the American Statistical Association, 74(367):588–595, 1979.

[TV13] Ton, B. T.; Veldhuis, R. N. J.: A high quality finger vascular pattern dataset collected using a custom designed capturing device. In: 2013 International Conference on Biometrics (ICB). pp. 1–5, June 2013.

Vein Enhancement with Deep Auto-Encoders to improve Finger Vein Recognition

Victor Bros[1], Ketan Kotwal[1], Sébastien Marcel[1]

Abstract: The field of Vascular Biometric Recognition has drawn a lot of attention recently with the emergence of new computer vision techniques. The different methods using Deep Learning involve a new understanding of deeper features from the vascular network. The specific architecture of the veins needs complex model capable of comprehending the vascular pattern. In this paper, we present an image enhancement method using Deep Convolutional Neural Network. For this task, a residual convolutional auto-encoder architecture has been trained in a supervised way to enhance the vein patterns in near-infrared images. The method has been evaluated on several databases with promising results on the UTFVP database as a main result. In including the model as a preprocessing in the biometric pipelines of recognition for finger vein patterns, the error rate has been reduced from 2.1% to 1.0%.

Keywords: Finger Vein Recognition, Deep Residual Convolutional Auto-Encoder, Vein Enhancement.

1 Introduction

Automatic biometric recognition has become a reliable technology to perform identification and verification of an individual. This domain has a wide range of applications in everyday life to assess the identity of an individual or to attribute one to them. Since the vascular patterns are believed to be unique from one person to another, they are well-suited for the task of verification. For the convenience of regular use, the veins of the most accessible parts of the body, such as the hands (palm, finger, wrist, etc.), are preferred as a recognition modality.

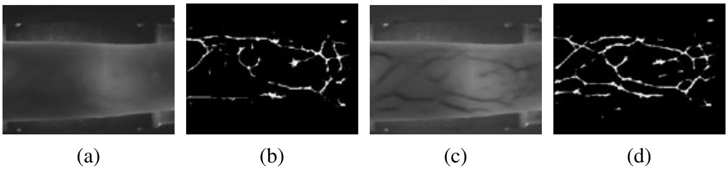


Fig. 1: Samples of finger vein images from the SDUMLA dataset. From left to right : (a) shows the original vein image, and (b) depicts its maximum curvature (MC) without enhancement. The resulting image after vein enhancement is shown in (c); while (d) depicts the MC obtained from it.

Usually biometric verification experiments consist of the following stages. The sensor acquires the biometric characteristic. The feature extractor generates the feature descriptor from the input presentation. The feature descriptor is compared against the precomputed templates to obtain the matching score. Despite several finger vein (FV) recognition

[1] Biometrics Security and Privacy Group, Idiap Research Institute, Martigny, Switzerland,
{vbros, ketan.kotwal, sebastien.marcel}@idiap.ch

methods being developed [Hu10, KZ11], it may be noted that their performance is often strongly correlated to the quality of the input presentations captured in the NIR spectra. Due to the nature of the FV structure (under the skin with many flesh artefacts), the quality of the captured presentations may often be poor, especially with lack of contrast. For instance, Figure 1 (a) and (b) show the input FV presentation and its corresponding FV pattern extracted using Maximum Curvature. Since the input presentation exhibits low contrast, it is indeed a challenging task to identify relatively darker vein patterns. As we may observe from Figure 1 (b), the FV extraction algorithm may miss several smaller veins, which could significantly enhance the discriminative capabilities of the FV recognition system. An efficient mechanism for preprocessing of FV presentations may result in the extraction of subtle vein patterns, and thus, obtaining a robust feature for the subsequent recognition pipeline.

In this work, we propose a deep learning-based preprocessing method that, in particular, enhances the vein patterns acquired in the NIR spectra. The objective of our method is to improve the separation between the background and the vascular networks in FV images. We use a Deep Convolutional Neural Network (DCNN) Auto-Encoder, more particularly a residual convolutional auto-encoder (RCAE), that can function as a preprocessor in the recognition pipeline before the feature extractor stage. Our model has been trained to reconstruct the enhanced versions of the input FV presentations with darker vein patterns. Using these enhanced presentations, the feature extractor has been able to identify even subtle FV patterns, that the standalone Maximum Curvature was unable to extract from unprocessed presentations. Figure 1 (c)-(d) depict the output of the proposed method and the extracted FV pattern, respectively. When compared against Figures 1 (a-b), that represent the equivalent results without any preprocessing; a simple visual inspection can demonstrate the improvement in the quality of presentations in terms of higher contrast and better separation between vein patterns from the rest of the content. In the context of FV, an objective evaluation of quality of the presentation is an open problem; and there is no universally accepted measure for the same. Therefore, in this work, we measure efficacy of the proposed FV enhancement method in an indirect manner, *i.e.,* through the gain observed in overall recognition accuracy after incorporating the proposed preprocessing mechanism.

Although the present work has been conducted in the context of finger vein recognition, it is possible to imagine other applications of such enhancement across a variety of domains. Some examples could be improvement in the contouring of a vascular network, a correction of low quality captured images which are quite expensive tasks in medical imaging.

The remainder of the paper is structured as follows: In the Section 2, we briefly describe the relevant work and the related work on enhancement of finger vein patterns. The deep convolutional auto-encoder model that reconstructs the image with enhanced vein patterns is described in Section 3. We discuss the details of the experiments based on the proposed method in Section 4. Section 5 summarizes the conclusions.

2 Related Work

To the best of our knowledge, the proposed work is the first attempt to use a DCNN Auto-Encoder to enhance FV patterns aimed towards improving the subsequent feature extraction. Therefore, in this section, we present a brief overview of commonly used building blocks of the FV recognition pipeline. We also discuss a few similar image enhancement methodologies- that have been developed for different end applications. Inspired by the idea of preprocessing images for their specific use, the method aims to enhance veins in NIR images of fingers in the context of biometric recognition. This method has been incorporated into verification pipelines, whose algorithms have been previously chosen for their reliability and performance. These are canonical algorithms that will be the baseline for the comparison with and without enhancement. The extraction of vein patterns is done by Maximum Curvature [MNM07]. The principle is to compute a binary map of the vascular patterns by calculating the centers of the veins by their intensity profiles and connecting them via a filter operation. Then the comparator, designed by the same team, is the Miura Matching [MNM04]. This method calculates a similarity score between two binary patterns by retrieving their maximum superposition score with possible displacements represented by a sliding window.

Within the pipelines, other preprocessing have been tested to improve the image quality. The first idea was to center the image on the finger, either by performing a crop of the region of interest (ROI) [YS12], or by including as much as possible the minutiae of the veins [LLP09]. Also since the patterns have a third dimension component, the second preprocessing reduces the distortion of the capture in adding a normalization for the position [Hu10]. Hence for the baseline, the whole sequence has been considered [Pe13].

With the idea of learning vein patterns by deep neural networks, other works have shown the use of convolutional networks for FV experiments. A study was able to use a convolution neural network as a comparator between two patterns with great success [Li17]. It was then questioned if an auto-encoder can compress the patterns in a reduced bottleneck embedding which is then compared with a Support Vector Machine classifier for the verification task [HY20]. From these experiments, it seems possible that network models have the ability to retrieve latent information from vascular patterns, in particular those mentioned above. The combination of an auto-encoder with convolutional layers has been tested for vascular patterns in the fundus of the eye. In [Li20], Li *et al* have proposed a neural network, NuI-Go, that aims to reduce the non-uniform illumination of the images of the eye. Here, they generated a dataset of retina vessel images with a synthetic degradation of the illumination. Their method is built on a deep residual convolutional auto-encoder to perform the reconstruction from the degraded image towards the original high quality image of the fundus.

3 Proposed Method

For the enhancement of FV images, we design an RCAE consisting of the encoder and decoder blocks that are linked through a residual connection. The encoder accepts an input FV image, and the decoder attempts to reconstruct its enhanced version. In the next subsection, we first outline the architecture of our RCAE network, and then provide details of the training process.

3.1 RCAE Architecture

For FV enhancement, we need to design a network structure that selectively identifies specific patterns in the input, and enhances their representation, in terms of good contrast and sharpness, during the reconstruction phase. Recently, Li *et al* have demonstrated the use of convolutional auto-encoder with a non-local unit for enhancement of the fundus of the eye image [Li20]. Although their use-case is quite different from ours, the quality of their results indeed suggests that a convolutional AE can be a highly effective method for enhancement of finer structures in the image.

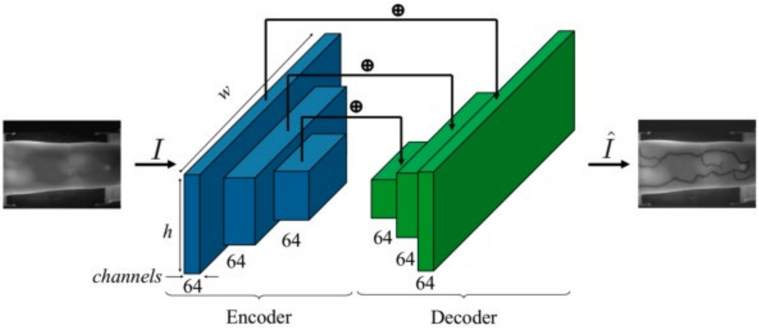


Fig. 2: The proposed architecture, a Residual Convolutional Auto-Encoder (RCAE), for the vein enhancement.

Following an extensive study, we design a RCAE network for the enhancement of FV aimed at improved detection of vein patterns. Figure 2 shows the schematic of the proposed RCAE. The encoder of our RCAE consists of 3 blocks: each of which includes a convolution, activation, and normalization. These blocks are connected to each other in sequence through pooling operation- thereby reducing the spatial dimensions of the effective input at every stage. On the decoder side, we have a succession of 3 decoding blocks: each consisting of a convolution and transposed convolution filters, along with activation and normalization layers. The network includes a residual transmission across every pair of blocks in encoder-decoder. This results into a differential component that learns the difference between the input and the target across layers. In the proposed RCAE, we have performed max-pooling, across the encoder, over the window of 2×2, and retained the stride of 2 for transposed convolutions in the decoder. The ReLU (Rectified Linear Unit) has been chosen as the activation operator; and each block is interspersed with a batch normalization layer to reduce the dependence on the training dataset and help the generalization [IS15].

The size of convolutional kernel is an important factor that determines the (effective) receptive field of the input and later layers of the deep network. Since the width or thickness of the FV patterns varies within a given presentation, the optimal size of convolutional kernel may not be easily decided. While the filters with relatively larger size are capable of learning the spatial relationship across somewhat distant pixels in the image; the smaller filters focus on encoding the features in local patches of the input. To explore the effects of kernel size of convolutional filters, we design two variants of the RCAE: (a) *Model 1*: with a constant kernel shape of 3×3 for each convolutional layer; and (b) *Model 2*: where the dimensions of convolutional kernel have been gradually decreased from 9×9 to 3×3 in the encoder, whereas the decoder layers observe a gradual increase in the kernel sizes. Also the number of channels inside the network have been drastically increased to 64 to introduce the complexity needed for the task.

3.2 Training Procedure

Generation of targets: The first step in the training of the RCAE is the generation of the targets, *i.e.,* synthetically generated enhanced FV images that act as the reference outputs during training. Our training datasets have been manually annotated for vein patterns in the form of binary masks. We consider the enhanced image (target) as the linear combination of actual input image and the vein-annotated binary mask. If $\mathbf{x}$ is the input FV presentation with $\mathbf{h}$ as the binary mask depicting vein structure, then the target image, $\mathbf{y}$, is obtained as: $\mathbf{y} = \alpha\mathbf{x} + (1-\alpha)\mathbf{h}$, where α refers to the fixed weight parameter. It should be noted that $\mathbf{x}$, $\mathbf{h}$, and $\mathbf{y}$ have the same dimensions. Figure 3 shows the example of input presentation and its mask, along with the generated target.

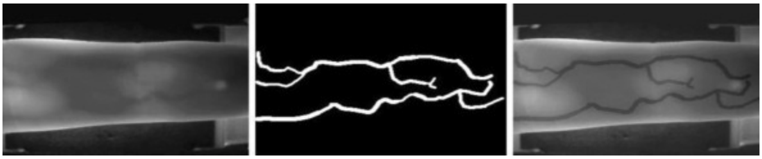


Fig. 3: Generation of target presentations: On the left, the original FV image; in the middle the binary annotation; and on the right the target image.

Data Augmentation: To deal with a possible lack of training data and generalization, we have incorporated data augmentation strategies during the training process of the RCAE model. We have considered the following four strategies for data augmentation: (a) random horizontal flip with probability $p = 50\%$, (b) random rotation with maximum angle $\theta = 2.5^\circ$, (c) random translation with maximum distance $d = 5\%$, and (d) random shear with maximum degrees $x = 5^\circ$.

Since the dimensions of the FV presentations across datasets may differ, it is necessary to resize the input to a fixed size (as determined during the training process). In our experiments, we have fixed the input size to 320×240 pixels. If the input FV presentation has different size, then it is re-scaled, in anti-aliased manner, after the enhancements to yield its original dimensions.

Loss Function: Since the FV images are quite distant from usual images in most common datasets, we have chosen a Mean Squared Error (MSE) loss ($\mathscr{L}_{\text{mse}}$) for the training RCAE. For the reconstructed image $\hat{\mathbf{y}}$, and the target (reference) image $\mathbf{y}$, the MSE loss is computed as $\mathscr{L}_{\text{mse}} = \frac{1}{N}(\mathbf{y}-\hat{\mathbf{y}})^{\mathsf{T}}.(\mathbf{y}-\hat{\mathbf{y}})$; where N is the number of pixels in the image $\mathbf{y}$.

For training the RCAE, we have chosen the Adam optimizer [KB17] with a learning rate of 0.001. In the beginning, the weights were initialized by random values normalized on a Gaussian centered around 0 with a bias of 0.05.

4 Experiments

We have implemented the experiments using PyTorch and Bob[2] frameworks with a focus on reproducible research. The python code and protocols to reproduce the experimental results are available publicly[3].

4.1 Datasets and Protocols

We demonstrate the efficacy of the proposed vein-enhancement RCAE on two publicly available FV datasets: UTFVP [Ro18] and SDUMLA [Lu13].[4]
The SDUMLA dataset consists of 634 masks, of medium quality FV presentations (320 × 240), from 636 identities (106 persons) collected in 6 sessions. The UTFVP dataset has 389 masks for high quality FV images with dimensions of (672 × 380).
As part of the verification experiments, it was necessary to define protocols for the use of the data in order to limit the correlation between the different sets for each database. In our experimental protocols, we ensure that the identities from training, validation, and testing of the RCAE do not overlap. We omit the details of protocols due to brevity of the space, but they can be obtained from the code repository.

4.2 Metrics for evaluation

Since the efficacy of the RCAE is measured through the FV recognition experiment, we have considered the False Match Rate (FMR) and False Non-Match Rate (FMNR) as the performance metrics. For performance comparison as well as for selection of score threshold, we have chosen the Half Total Error Rate (HTER) criteria which is computed as $HTER = 0.5 \times (FMR + FNMR)$. To evaluate the performance through the distributions of scores, the genuine (same identity presented) and the imposter (different identity), we have employed statistical tests as well. A *Cohen's d* test was used to measure the distance between the two empirical distributions to observe the impact of the preprocessing. This distance has been used as a comparison of the differentiation ability of the FV verification system. A higher value of the distance indicates a better separability of both distributions.

4.3 Experimental Results

Table 1 provides the results of verification experiments on the two datasets. For SDUMLA, a significant impact of the preprocessing can be seen through the performance of the verification system. For both models, the HTER has reduced by nearly 20%, and the *Cohen's distance* is higher by nearly 10% as compared to the corresponding values without the RCAE preprocessing. This second experiment was conducted to show the generalisation

[2] `https://www.idiap.ch/software/bob`

[3] `https://gitlab.idiap.ch/bob/bob.paper.biosig2021_deep_vein_enhancement`

[4] The vein annotations for both datasets were provided by the University of Salzburg.

		SDUMLA		UTFVP	
Model	Statistical test	development set	evaluation set	development set	evaluation set
w/o	HTER	15.4%	14.2%	0.2%	2.1%
	Cohen's d	5.0	6.8	15.5	15.2
Model 1	HTER	12.0%	9.8%	0.2%	1.3%
	Cohen's d	5.7	7.6	15.7	16.7
Model 2	HTER	12.5%	10.1%	0.2%	1.0%
	Cohen's d	5.6	7.6	15.7	16.2

Tab. 1: Results of the verification experiment on SDUMLA dataset and UTFVP dataset.

of the model to other databases, which means other sensors and other image quality. Similarly, the HTER on the evaluation set is lower for both RCAE models for the UTFVP dataset. The Cohen's d test also shows that the impact of the matching algorithm is higher with a preprocessing since the distance is higher for both models than without preprocessing. Figure 4 also highlights the overall improvement brought by the models for the verification on UTFVP, with lower *FNMR* at all *FMR* on the evaluation set. It may be inferred that the preprocessing has been able to generalize to other databases with success.

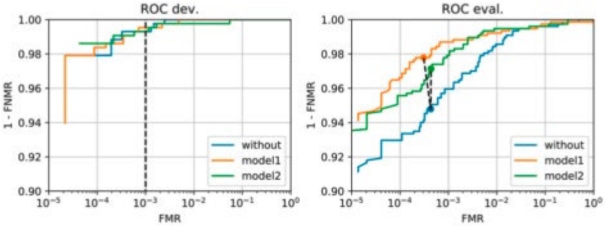


Fig. 4: ROC curves of the experiments on UTFVP: without RCAE preprocessing, model 1, and model 2.

5 Conclusion

In this work, we have proposed an enhancement method for FV images captured in NIR spectra. We have developed an RCAE model that can be integrated as the preprocessor into a biometric recognition pipeline. The purpose of the RCAE is to learn the prominent as well as subtle vein patterns in the image, and improve the quality of presentation, in terms of a better contrast. We have demonstrated that with the proposed preprocessing, the overall accuracy of the FV recognition has increased, as well as the separation between the distributions of recognition scores of genuine and imposter identities has also improved. For two publicly available FV datasets, our method has resulted in nearly 20% reduction in the average error in recognition.

The proposed RCAE enhances major FV structures, and also occasionally identifies the subtle vein patterns that might have been missed by human annotators. However, in some examples, it may lead to generate spurious vein patterns. We are presently working on improving the accuracy of enhanced vein patterns. We are also working on extending the application of RCAE beyond preprocessing tasks.

6 Acknowledgement

This work was supported by the Swiss Center for Biometrics Research & Testing, and the InnoSuisse project CANDY. Also this work was made possible thanks to Pr. Andreas Uhl from Salzburg University and his team for providing the vein annotations required for our experiments.

References

[Hu10] Huang, Beining; Dai, Yanggang; Li, Rongfeng; Tang, Darun; Li, Wenxin: Finger-vein authentication based on wide line detector and pattern normalization. pp. 1269–1272, Aug. 2010.

[HY20] Hou, B.; Yan, R.: Convolutional Autoencoder Model for Finger-Vein Verification. IEEE Transactions on Instrumentation and Measurement, 69(5):2067–2074, 2020.

[IS15] Ioffe, Sergey; Szegedy, Christian: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. pp. 448–456, 2015.

[KB17] Kingma, Diederik P.; Ba, Jimmy: Adam: A Method for Stochastic Optimization. 2017.

[KZ11] Kumar, Ajay; Zhou, Yingbo: Human identification using finger images. IEEE Transactions on Image Processing, 21(4):2228–2244, 2011.

[Li17] Liu, W.; Li, W.; Sun, L.; Zhang, L.; Chen, P.: Finger vein recognition based on deep learning. pp. 205–210, 2017.

[Li20] Li, Chongyi; Fu, Huazhu; Cong, Runmin; Li, Zechao; Xu, Qianqian: NuI-Go: Recursive Non-Local Encoder-Decoder Network for Retinal Image Non-Uniform Illumination Removal. 2020.

[LLP09] Lee, Eui Chul; Lee, Hyeon Chang; Park, Kang Ryoung: Finger vein recognition using minutia-based alignment and local binary pattern-based feature extraction. International Journal of Imaging Systems and Technology, 19(3):179–186, 2009.

[Lu13] Lu, Y.; Xie, S. J.; Yoon, S.; Wang, Z.; Park, D. S.: An available database for the research of finger vein recognition. 01:410–415, 2013.

[MNM04] Miura, Naoto; Nagasaka, Akio; Miyatake, Takafumi: Feature extraction of finger-vein patterns based on repeated line tracking and its application to personal identification. Machine Vision and Applications, 15:194–203, 10 2004.

[MNM07] Miura, Naoto; Nagasaka, Akio; Miyatake, Takafumi: Extraction of Finger-Vein Patterns Using Maximum Curvature Points in Image Profiles. IEICE - Trans. Inf. Syst., E90-D(8):1185–1194, August 2007.

[Pe13] Peng, Jialiang; Li, Qiong; Wang, Ning; Abd El-Latif, Ahmed; Niu, Xiamu: An Effective Preprocessing Method for Finger Vein Recognition. Proc SPIE, 8878:08–, 07 2013.

[Ro18] Rozendal, S.P.: Redesign of a finger vein scanner. February 2018.

[YS12] Yang, Jinfeng; Shi, Yihua: Finger–vein ROI localization and vein ridge enhancement. Pattern Recognition Letters, 33(12):1569–1579, 2012.

VeinPLUS+: A Publicly Available and Free Software Framework for Vein Recognition

Michael Linortner[1], Andreas Uhl[1]

Abstract: In this work a new and open source software framework for vein recognition is presented, which supports a complete tool chain to conduct biometric experiments. The software can be used out of the box on several publicly available databases and offers a convenient mechanism to configure the tool chain to fit the needs for individual experimental setups. It is implemented in C++ and works on Linux as well as on Windows and therefore offers an advantage over existing vein recognition libraries. The project's aim is to offer a valuable tool to conduct research in the field of vein recognition and importantly, supporting reproducible research.

Keywords: Biometrics, vein recognition, free software framework, VeinPLUS+, C++, software.

1 Introduction

Utilizing veins in fingers or hands as biometric trait, known under the term vein recognition, has become an emerging technology in both industry and science. To accomplish reproducible research outcome in the field of vein recognition or in biometrics in general, two important requirements need to be met: Firstly, experiments should be conducted on publicly available data sets and secondly, the software which produces the experimental results should be publicly accessible as well. Performing biometric experiments requires a software which provides several functionalities. It needs to manage the data set used in the experiment, to provide functions to extract features from biometric samples, and to compare them. To evaluate a recognition technique usually a set of comparison scores needs to be generated following a dedicated protocol which subsequently demands functionality to deduce common accepted performance parameters from these scores like receiver operating characteristic (ROC), detection error tradeoff (DET) or equal error rate (EER). There exist several publicly available data sets containing vein images of hands or fingers, but only two software frameworks consisting of a complete tool chain to run biometric experiments. In this work we propose a new platform independent, publicly available and free software framework suitable to conduct experiments and performance analysis of biometric vein recognition tasks. As it is meant to be a framework for research it meets several requirements crucial for such a type of software: it is platform independent and relies only on few external dependencies, which is important to set up the software quickly and easily; it can be used out of the box. The user configures a vein recognition system as desired, then calls the software, which loads the specified data set, processes each sample, generates a set of comparison scores following the specified protocol and delivers the results of a performance evaluation in the form of DET data.

[1] Department of Computer Sciences, University of Salzburg, Salzburg, Austria, {mlinortner, uhl}@cs.sbg.ac.at

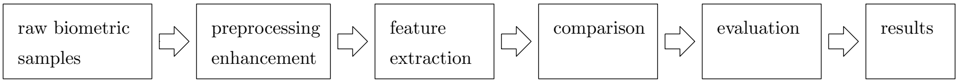


Fig. 1: Biometric tool chain

2 Prior work

Literature research revealed that there are only two software frameworks publicly available which cover a full vein recognition system with all necessary tools to conduct full research experiments out of the box.

One is “Bob”[2] a signal-processing and machine learning toolbox developed by the IDIAP research institute [An12, An17]. The framework is implemented in Python and its core functionality in C/C++. Bob runs on Linux and MacOS 64-bit, but not on Windows. The comprehensive toolbox includes a library for vein recognition providing the feature extraction methods repeated line tracking (RLT) [MNM04], maximum curvature (MC) [MNM07], principal curvature (PC) [Ch09], wide line detector (WLD) [Hu10] and several preprocessing methods, mostly for masking the finger region in a raw biometric sample.

A second framework is the “PLUS OpenVein Toolkit” (OVTK)[3] [KU19], which is specialized for vein recognition and also implements the full biometric extraction, comparison and evaluation tool chain. It is written in MATLAB code and therefore is platform independent, of course, relying on a MATLAB installation, which is a commercial non-free product. It implements the same feature extractions methods as Bob, which are based on a MATLAB implementation by Bram Ton, publicly available on MATLAB Central[4]. In comparison to Bob it provides several other and also more recent feature extraction and preprocessing methods. There is code available for certain vein recognition techniques, like the above mentioned MATLAB implementation for MC, or vein image enhancement like [BRG17] or for deep learning approaches[5], but they are no comprehensive framework supporting a complete biometric tool chain out of the box.

3 The Software

The software presented in this work is a comprehensive and flexible framework for conducting research experiments in the field of vein recognition. It is a platform independent implementation in C++, following the C++11 standard. The software is named Vein-PLUS+ where PLUS stands for Paris Lodron University Salzburg and PLUS+ refers to the implementation language.

The software supports a full biometric tool chain as depicted in figure 1: loading and managing data sets, preprocessing/enhancing samples, extracting features, creating comparison scores consisting of genuine and impostor comparisons following a specified protocol, evaluating the computed scores and producing results as DET.

[2] https://www.idiap.ch/software/bob/

[3] http://www.wavelab.at/sources/OpenVein-Toolkit/

[4] https://de.mathworks.com/matlabcentral/profile/authors/1836574

[5] https://github.com/ridvansalihkuzu/vein-biometrics

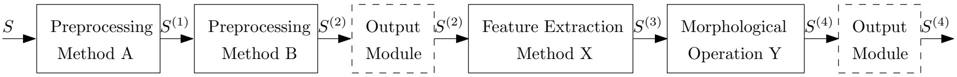
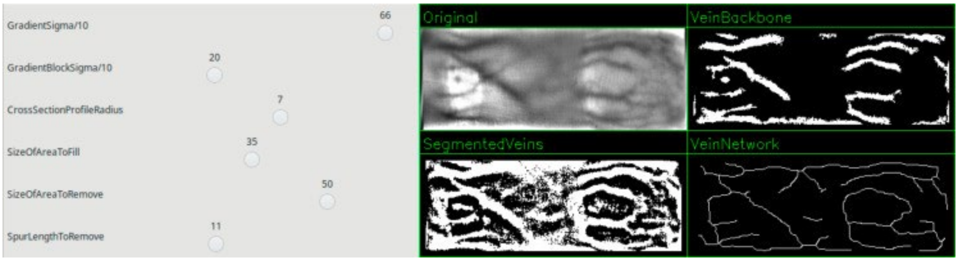


Fig. 2: Preprocessing and feature extraction pipeline example

The whole tool chain is configurable via settings specified in an XML structure and stored as file. Each supported preprocessing, feature extraction and feature comparison method is implemented as a separate module, the parameters of which are defined in the settings. Those modules can be arranged in any order, configurable in the settings file, to create an individual feature extraction pipeline and comparison process fitting the needs of the desired experimental setup without the need to change a line of code, therefore offering maximum flexibility. Figure 2 depicts an exemple configuration for a process pipeline consisting of two preprocessing methods A and B, a feature extraction method X and a morphological postprocessing method Y. Each biometric (vein) sample S from the data set is passed through the processing pipeline consisting of n modules, where $S^{(i)}$ is the processed sample and output of the i-th module. The final output $S^{(n)}$ is used by the comparison module to generate the comparison scores. As indicated in figure 2, at any position of the pipeline an output module can be inserted optionally which writes the samples' current processing state $S^{(i)}$ to files, providing the user a convenient mechanism to check the results after each processing step. Table 1 lists the implemented preprocessing and feature extraction methods. Further it is labeled whether the method is provided by Bob or the OVTK. Note that Bob and the OVTK provide additional preprocessing methods and the

	Method	VP+	OVTK	Bob
Preprocessing	Gaussian filter	✓	✓	✓
	Contrast limited adaptive histogram equalization (CLAHE)	✓	✓	✓
	Finger region masking [LLP09]	✓	✓	✓
	Background subtraction [KZ12]	✓	✗	✗
	Image normalization [Ki12]	✓	✗	✗
	High frequency emphasis filtering (HFE) [Zh09, Ha71]	✓	✓	✗
	Even-symmetric circular Gabor filters [ZY09]	✓	✓	✗
	Single scale retinex (SSR) with guided filter [Xi15]	✓	✗	✗
	Speeded-up adaptive contrast enhancement (SUACE) [BRG17]	✓	✓	✗
Feature extraction	Repeated line tracking (RLT) [MNM04]	✓	✓	✓
	Maximum curvature (MC) [MNM07]	✓	✓	✓
	Principal curvature (PC) [Ch09]	✓	✓	✓
	Wide line detector (WLD) [Hu10]	✓	✓	✓
	Gabor filter as in [KZ12] proposed	✓	✓	✗
	Anatomy structure analysis based vein extraction (ASAVE) [Ya17]	✓	✓	✗
	SIFT based vein recognition	✓	✓	✗
	Minutiae point extraction [LU21]	✓	✗	✗
	Spectral minutiae representation (SMR) [XV10]	✓	✗	✗

Tab. 1: Overview of the implemented methods in VeinPLUS+ (denoted as VP+). It is also marked whether a method is available in Bob or the OVTK.



GradientSigma/10 66

GradientBlockSigma/10 20

CrossSectionProfileRadius 7

SizeOfAreaToFill 35

SizeOfAreaToRemove 50

SpurLengthToRemove 11

Fig. 3: Helper GUI provided by VeinPLUS+. Exemplary the GUI for the ASAVE method is depicted. With the track bars each parameter can be changed and the effect is immediately visible. On the top left of the display the input image is shown, below the segmented veins and on the right side the two extracted feature components “vein backbone” and “vein network” of the ASAVE method.

OVTK also provides additional feature extraction approaches, other than the listed.
Besides the biometric tool chain two additional features are implemented in VeinPLUS+. When conducting experiments it is important to use optimal parameters for the configured tool chain in order to achieve good performance results. Often a grid search is executed to find such optimal parameters, meaning a combination of selected parameter values is evaluated. The software supports such an approach by automatically generating settings files. A settings file which configures the tool chain as desired is used as template. Each parameter, for which different values should be evaluated, is written into an instruction file and an individual list of values is assigned to them. Using the template settings file, the software automatically generates settings files for each value combination specified in the instruction file. The second feature is a helper tool, which provides a graphical user interface (GUI) for each preprocessing and feature extraction method where the method’s parameters can be changed and the effects on the input image are immediately visible on the screen. Figure 3 shows the GUI for the ASAVE method. This convenient tool allows to quickly check the outcome of a method and helps to understand a parameters impact on the outcome especially when several parameters influence each other.

VeinPLUS+ uses the OpenCV library[6] for image processing tasks and CLI11[7] for command line parsing. Therefore, the user needs to install OpenCV in order to use VeinPLUS+. CLI11 is included in the source code of the framework. VeinPLUS+ offers an option to use MatIO[8], a library witch enables reading and writing standard MATLAB files. The software has been developed using OpenCV 3.4.2 and has been tested on Linux with GCC 4.8 and GCC 8.3 as on Windows 10 with Microsoft (R) C/C++ Optimizing Compiler Version 19 (Visual Studio 2017).
The software can be downloaded from our webpage[9] and the source code is hosted on our gitlab server[10].

[6] https://opencv.org

[7] https://github.com/CLIUtils/CLI11

[8] https://github.com/tbeu/matio

[9] http://www.wavelab.at/sources/VeinPLUSplus/

[10] Link can be found on http://www.wavelab.at/sources/VeinPLUSplus/

4 Exemplary experiments

This section describes exemplary experiments for a basic workflow when using the VeinPLUS+ framework. The results are compared to the outcome produced by the OVTK.

4.1 Experimental setup

The experiments are conducted on two data sets from the publicly available "PLUSVein-FV3 Finger Vein Database"[11] [KPU18]. Two data sets are used, both captured under near-infrared (NIR) laser illumination, one showing the palmar, the other showing the dorsal view of a finger, denoted as Las-P and Las-D, respectively. The database has been chosen for the experiments because, additionally to the raw images, extracted region of interest (ROI) images are provided which are used throughout this work. The ROI extraction is a crucial step in vein recognition which can considerably influence the accuracy of the outcome. Using these ROI data helps to improve the reproducibility of the conducted experiments.

As the results produced by the VeinPLUS+ software are compared to the outcome of the OVTK two vein recognition methods are selected which are available in both frameworks. MC, which is a well established recognition method and ASAVE, a more recently published approach, have been selected. The same settings of the two methods have been used on both data sets. The tool chain has been configured as follows:

MC: In the first processing step CLAHE is applied on the input images, followed by applying even-symmetric circular Gabor filters [ZY09]. Afterwards the image is downsampled by a factor of two and the MC features are extracted. To generate the comparison scores the correlation based approach proposed in [MNM04] is used.

ASAVE: Gaussian smoothing is applied on the input image in the first step followed by CLAHE and downsampling by a factor of two. Subsequently, the ASAVE features are extracted and the comparison scores are computed by executing ASAVE's "integration comparison" algorithm [Ya17].

The comparison modules use the fingerprint verification contest (FVC)[12] protocol which results in 3600 genuine and 64620 impostor comparisons for the used data sets. The performance accuracy is evaluated in terms of EER. Both frameworks should produce the same EER values, but variations are expected because of the different implementations. Additionally, the time performance is measured and reported. The settings are available for download together with the VeinPLUS+ software and the experimental results can easily be reproduced, which is one of the important features of the software.

The experiments are executed on a PC with Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz and Debian GNU/Linux 10 (buster). The software has been compiled with GCC 8.3.0 on optimization level O3.

[11] http://wavelab.at/sources/PLUSVein-FV3/

[12] http://bias.csr.unibo.it/fvc2006/perfeval.asp

Data set	Method	VeinPLUS+				OVTK			
		EER(%)	t_e	t_c	t_t	EER(%)	t_e	t_c	t_t
Las-D	MC	0.25	46	185	232	0.28	529	135	709
Las-D	ASAVE	0.24	177	683	864	0.36	647	2461	3135
Las-P	MC	2.56	44	186	235	2.79	536	118	667
Las-P	ASAVE	3.39	175	681	858	3.31	606	2479	3143

Tab. 2: Accuracy (EER) and time performances (t_e: feature extraction time, t_c feature comparison time, t_t: total execution time in seconds) of the VeinPLUS+ framework in comparison to the OVTK.

4.2 Results

Table 2 lists both recognition as well as computational performance results produced by the experiments. As expected, both frameworks produce similar EER values with slight variations. Only for ASAVE on the Las-D data set the relative difference is larger than in the other cases. This can be explained with several divergences in the implementation. For example, OpenCV and MATLAB provide CLAHE, but is implemented differently, even the parameter range differs. For the ASAVE approach there is no original reference implementation available. OVTK implements the original proposed search method for the correlation-based part of the feature comparison procedure: first a horizontal shift is applied and subsequently a vertical shift, which in theory demands less computational effort than a full correlation. Exploiting optimized functionality, offered by OpenCV for example, a full correlation is executed faster in practice, which has been utilized by VeinPLUS+. Small differences in the output of each module in the processing pipeline may sum up and can make a noticeable difference in the final performance. In case of ASAVE on the Las-D data set the considerable difference can be explained that VeinPLUS+ performs a full correlation in the comparison stage and finds a more optimal solution than the implementation in the OVTK.

For the time performance evaluation the time t_e for extracting the features from all 1800 samples in the data set and the time t_c for computing 68220 comparison scores have been measured as well as the total execution time $t_t = t_e + t_c + t_\Delta$, where t_Δ includes the time for loading the data, executing the performance evaluation, etc. It can be observed that both feature extraction methods are executed faster in the VeinPLUS+ framework than in the OVTK, caused by more efficient data handling, function calls and loops in C++ compared to MATLAB. The correlation based feature comparison for MC is processed faster by the OVTK, exploiting the strength of MATLAB. For ASAVE VeinPLUS+ executes the comparisons faster than the OVTK, which can be reasoned by divergences in the implementation details of the algorithm as explained above. As the values for the total execution time show VeinPLUS+ completes the task for both recognition methods faster.

Note that the settings used for the experimental setup are not optimized to achieve the best possible accuracy. The utilized parameters are a compromise to manage that both frameworks process the data in a similar way in order to compare the performance values, again explained by differences in the implementation.

5 Conclusion

In this work a new platform independent software framework for vein recognition has been presented. It is the first free multi-platform vein recognition toolkit. The framework, named VeinPLUS+, comprises the complete tool chain to conduct biometric experiments and can be used out of the box on several publicly available databases for finger vein recognition. The software offers the user a flexible way to individually configure the tool chain to fitting the needs of a desired experimental setup. Sharing this configuration, in form of a small text-based (XML) file, allows perfect reproducibility of the given experiments. The framework is implemented in C++ and runs on Linux as well as on Windows. Therefore, it offers an advantage over the two existing frameworks Bob, which does not work on Windows, and the OpenVein Toolkit, which runs on MATLAB a commercial software package.

VeinPLUS+ is under continuous development with the goal of extending the functionality by adding new enhancement and vein recognition methods as well improving the performance. It is planned to add deep learning vein recognition methods and to parallelize the comparison process. As the software is open source the research community is encouraged to actively contribute to the project. The project's aim is to offer a tool for reproducible research.

6 Acknowledgment

This project was partly funded by the Austrian Science Fund FWF project "Advanced Methods and Applications for Fingervein Recognition" under grant No. P 32201-NBL.

References

[An12] Anjos, A.; Shafey, L. El; Wallace, R.; Günther, M.; McCool, C.; Marcel, S.: Bob: a free signal processing and machine learning toolbox for researchers. In: 20th ACM Conference on Multimedia Systems (ACMMM), Nara, Japan. October 2012.

[An17] Anjos, A.; Günther, M.; de Freitas Pereira, T.; Korshunov, P.; Mohammadi, A.; Marcel, S.: Continuously Reproducing Toolchains in Pattern Recognition and Machine Learning Experiments. In: International Conference on Machine Learning (ICML). August 2017.

[BRG17] Bandara, A. M. R. R.; Rajarata, K. A. S. H. K.; Giragama, P. W. G. R. M. P. B.: Super-efficient spatially adaptive contrast enhancement algorithm for superficial vein imaging. In: 2017 IEEE International Conference on Industrial and Information Systems (ICIIS). pp. 1–6, Dec 2017.

[Ch09] Choi, Joon Hwan; Song, Wonseok; Kim, Taejeong; Lee, Seung-Rae; Kim, Hee Chan: Finger vein extraction using gradient normalization and principal curvature. In: Image Processing: Machine Vision Applications II. volume 7251. International Society for Optics and Photonics, SPIE, pp. 359 – 367, 2009.

[Ha71] Hall, E.L.; Kruger, R.P.; Dwyer, S.J.; Hall, D.L.; Mclaren, R.W.; Lodwick, G.S.: A Survey of Preprocessing and Feature Extraction Techniques for Radiographic Images. IEEE Transactions on Computers, C-20(9):1032–1044, 1971.

[Hu10] Huang, B.; Dai, Y.; Li, R.; Tang, D.; Li, W.: Finger-Vein Authentication Based on Wide Line Detector and Pattern Normalization. In: 2010 20th International Conference on Pattern Recognition. pp. 1269–1272, Aug 2010.

[Ki12] Kim, Hwi-Gang; Lee, Eun Jung; Yoon, Gang-Joon; Yang, Sung-Dae; Lee, Eui Chul; Yoon, Sang Min: Illumination Normalization for SIFT Based Finger Vein Authentication. In: Advances in Visual Computing. Springer Berlin Heidelberg, pp. 21–30, 2012.

[KPU18] Kauba, Christof; Prommegger, Bernhard; Uhl, Andreas: The Two Sides of the Finger - An Evaluation on the Recognition Performance of Dorsal vs. Palmar Finger-Veins. In: Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG'18). pp. 1–8, 2018.

[KU19] Kauba, Christof; Uhl, Andreas: An Available Open-Source Vein Recognition Framework. In: Handbook of Vascular Biometrics, chapter 4, pp. 113–142. Springer Nature Switzerland AG, 2019.

[KZ12] Kumar, A.; Zhou, Y.: Human Identification Using Finger Images. IEEE Transactions on Image Processing, 21(4):2228–2244, April 2012.

[LLP09] Lee, Eui Chul; Lee, Hyeon Chang; Park, Kang Ryoung: Finger vein recognition using minutia-based alignment and local binary pattern-based feature extraction. International Journal of Imaging Systems and Technology, 19(3):179–186, 2009.

[LU21] Linortner, Michael; Uhl, Andreas: Towards Match-on-Card Finger Vein Recognition. In: Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security. IH&MMSec '21, Association for Computing Machinery, New York, NY, USA, pp. 87–92, 2021.

[MNM04] Miura, Naoto; Nagasaka, Akio; Miyatake, Takafumi: Feature extraction of finger-vein patterns based on repeated line tracking and its application to personal identification. Machine Vision and Applications, 15(4):194–203, Oct 2004.

[MNM07] Miura, Naoto; Nagasaka, Akio; Miyatake, Takafumi: Extraction of finger-vein patterns using maximum curvature points in image profiles. IEICE TRANSACTIONS on Information and Systems, 90(8):1185–1194, 2007.

[Xi15] Xie, Shan Juan; Lu, Yu; Yoon, Sook; Yang, Jucheng; Park, Dong Sun: Intensity Variation Normalization for Finger Vein Recognition Using Guided Filter Based Singe Scale Retinex. Sensors, 15(7):17089–17105, 2015.

[XV10] Xu, H.; Veldhuis, R. N. J.: Complex spectral minutiae representation for fingerprint recognition. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops. pp. 1–8, June 2010.

[Ya17] Yang, L.; Yang, G.; Yin, Y.; Xi, X.: Finger Vein Recognition with Anatomy Structure Analysis. IEEE Transactions on Circuits and Systems for Video Technology, pp. 1–1, 2017.

[Zh09] Zhao, J.; Tian, H.; Xu, W.; Li, X.: A New Approach to Hand Vein Image Enhancement. In: 2009 Second International Conference on Intelligent Computation Technology and Automation. volume 1, pp. 499–501, Oct 2009.

[ZY09] Zhang, J.; Yang, J.: Finger-Vein Image Enhancement Based on Combination of Gray-Level Grouping and Circular Gabor Filter. In: 2009 International Conference on Information Engineering and Computer Science. pp. 1–4, Dec 2009.

Rotation Tolerant Finger Vein Recognition using CNNs

Bernhard Promegger[1,2], Georg Wimmer[1,2], Andreas Uhl[1]

Abstract: Finger vein recognition deals with the recognition of subjects based on their venous pattern within the fingers. The majority of the available systems acquire the vein pattern using only a single camera. Such systems are susceptible to misplacements of the finger during acquisition, in particular longitudinal finger rotation poses a severe problem. Besides some hardware based approaches that try to avoid the misplacement in the first place, there are several software based solutions to counter fight longitudinal finger rotation. All of them use classical hand-crafted features. This work presents a novel approach to make CNNs robust to longitudinal finger rotation by training CNNs using finger vein images from varying perspectives.

Keywords: Finger vein recognition, longitudinal finger rotation, rotation tolerance, CNN.

1 Introduction

The performance of finger vein recognition systems suffers from environmental conditions (e.g. temperature and humidity) and deformations due to misplacement of the finger, typically including shifts, tilt, bending and longitudinal rotation. The influence of some of these misplacements can be reduced or even prevented completely either during acquisition by adding support structures for finger positioning or a correction during pre-processing, feature extraction or comparison. Especially longitudinal finger rotation is hard to avoid. In [PKU19], the authors showed that existing finger vein data sets contain longitudinal rotation to a non neglectable extend. By eliminating only longitudinal finger rotation (all other condition remain unchanged), they achieved performance increases of up to 350%. This indicates that longitudinal finger rotation is not only a problem in selected use cases, but a general problem in finger vein recognition. As finger vein recognition systems evolve towards contact less acquisition (e.g. [Ma17, KPU19]), problems due to finger misplacements will become more severe.

Longitudinal finger rotation is hard to counteract as it changes the positioning of the veins and their visibility due to a non-linear transformation. As can be seen in Fig. 1, the acquired vein pattern of a finger differs depending on its rotation. There is already some work on rotation detection and compensation for single-camera systems. Prommegger *et al.* [Pr19] analysed different approaches and showed that existing recognition systems, even when applying rotation compensation, cannot handle rotational distances of >30°. Others try to tackle the problem by acquiring the vein pattern from different perspectives (e.g. [PU19, Ka19]). The disadvantage of multi-camera systems are the increased cost and complexity.

[1] University of Salzburg, Department of Computer Science, Jakob Haringer Strasse 2, 5020 Salzburg, {bprommeg, gwimmer, uhl}@cs.sbg.ac.at

[2] These authors contributed equally

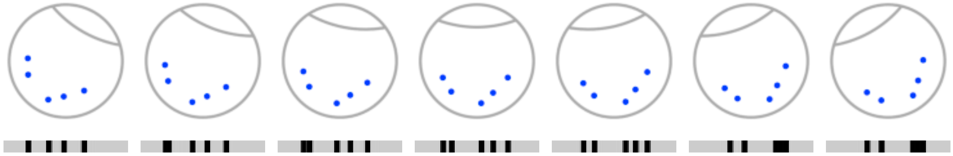


Fig. 1: Schematic finger cross section showing five veins (blue dots) rotated from -30° (left) to +30° (right) in 10° steps. The projection (bottom row) of the vein pattern changes depending on the rotation angle according to a non-linear transformation (originally published in [PKU18b])

Recently, finger vein recognition systems using convolutional neural networks (CNN) are getting more attention. These systems are either not designed to counter fight longitudinal rotation (e.g. [HLP17, WPU20]) or require the acquisition of finger vein images from multiple perspectives [Ka19]. Therefore, this paper is the first to present a CNN based rotation tolerant single camera finger vein recognition system. The proposed idea is to train CNNs using finger vein images from varying perspectives. There are two different sources for these images: (1) images that were actually taken from different perspectives and (2) augmented images using a novel approach that simulates longitudinal rotations during training. This way, the CNNs should learn to recognize the connection between images that come from different angles and thus to recognize the non-linear distortion caused by the rotation.

2 CNN Architectures

To demonstrate that our proposed approach to increase the rotation tolerance of CNNs is independent of the used CNN architecture and loss function, two different CNN architectures and loss functions are used in our experiments:

Squeeze-Net (SqNet) with triplet loss function: The advantage of the triplet loss compared to more common loss functions (e.g. SoftMax) is that the CNNs learn to group images of the same classes together in the CNN feature output space and separate images from different classes, instead of directly classifying images. So, contrary to common loss functions, CNNs can also be applied to images whose classes are not included in the training data. This property is crucial in biometric applications. The triplet loss using the squared Euclidean distance is defined as follows:

$$L(A,P,N) = \max(||f(A) - f(P)||^2 - ||f(A) - f(N)||^2 + \alpha, 0) \quad (1)$$

where A and P are two images from the same class (finger), N from a different one. α is a margin that is enforced between positive and negative pairs (in our case $\alpha = 1$), and $f(x)$ is the CNN output of an input image x. Same as in [WPU20], we employ hard triplet selection and the Squeeze-Net architecture.

DenseNet with SoftMax loss: A more common approach than using the triplet loss is to train a net with the common Soft-Max loss function and then use the net as feature extractor for evaluation by using the CNN activations of intermediate layers. This approach has already been applied in prior work (e.g. [HLP17]). As network architecture we employ the DenseNet-161. For evaluation, we remove the final layer and thereby get a 2208 dimensional feature vector output when feeding an image through the network.

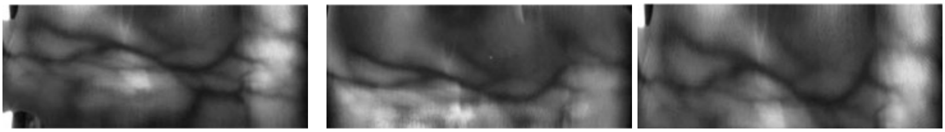


Fig. 2: ROI of sample images of the PLUSVein-FR after applying CPN. Left: palmar view (0°), middle: vein image captured at 45°, right: 45° artificially rotated version of the palmar image.

3 Training Data

As already described, the rotated versions of the training images are provided using two different approaches. In the first approach, the images are acquired at different rotation angles, while in the second approach the rotation is artificially generated using data augmentation. Using vein images that were actually captured from different perspectives for CNN training is of course more effort than generating the rotated versions with the help of data augmentation. It should be noted that rotated samples of the same finger are only needed for training. The actual angles of rotation of these samples do not necessarily have to be known, as long as the acquired samples cover the rotational range for which the recognition system should be tolerant. This can be achieved by e.g. placing the finger in different rotations on the existing single camera capturing device or by rotating the camera and illumination module around the finger as done for the employed data sets. This is certainly a reasonable expense for commercial products. Recognition is still applied using vein images from a single perspective.

All images for training and evaluation are normalized using *Circular Pattern Normalization* (CPN) [Pr19]. In principle CPN corresponds to a rolling of the finger surface assuming a circular finger shape. After this unrolling, longitudinal rotations correspond to shifts in the acquired images.

Finger Vein Images Captured from Different Perspectives: For the training of the CNNs, finger vein images acquired from different perspectives are used (the finger is rotated around its longitudinal axis). All images of a finger, regardless of the angle at which they were taken, are considered as the same class. As a result of this, the CNN should learn to recognize finger vein images independent of their perspective. The left and middle image in Fig. 2 show two such input images. The left image has been acquired from the palmar view (0°), the middle one from 45°. It is clearly visible that the vein pattern is vertically shifted and deformed in a non-linear manner due to the rotational difference.

CNN Training using Augmentation of Finger Vein Images: The augmented training data is generated from images acquired at the palmar view (0°). The height of a CPN image is $h_{CPN} = r \cdot \pi$, which is half the fingers perimeter with an assumed radius of r. The displacement (in pixels) that must be applied for a rotation of a defined rotation angle φ_{rotate} can be calculated as:

$$h_{shift}\left(\varphi_{rotate}\right) = \frac{2 \cdot r \cdot \pi \cdot \varphi_{rotate}}{360^\circ} = \frac{h_{CPN} \cdot \varphi_{rotate}}{180^\circ} \qquad (2)$$

For data augmentation in the rotational range of $\pm\varphi$, the height h of the input images is enlarged by twice the maximum shift $h = h_{CPN} + 2 \cdot h_{shift}\left(\varphi\right)$. Augmentation is applied by randomly cropping patches with height h_{CPN} (and original width) of the enlarged images, which corresponds to rotations in the range of $\pm\varphi$. The right image in Fig. 2 is an artificially rotated version (45°) of the original image at the left.

4 Experiments

Datasets: The datasets used for the experiments are the *PLUSVein Finger Rotation Dataset* (PLUSVein-FR) [PKU18a] and the *PROTECT Multimodal Dataset* (PMMDB) [Ga20]. Both datasets provide finger vein images acquired all around the finger (360° in steps of 1°) and have been acquired using the same sensor and the same acquisition protocol. In this work, only the perspectives in the range of ±45° around the palmar view are used. The PLUSVein-FR provides vein images from 63 different subjects with 4 fingers per subject and each finger is acquired 5 times per perspective, resulting in 1.260 finger vein images per perspective and 115.920 vein images in total. PMMDB is acquired from 29 subjects with 4 fingers per subject and either 5 or 10 images per finger (one or two sessions) for each perspective with a total of 102.987 finger vein images.

For *finger region detection* and *finger alignment*, an implementation that is based on [Lu13] is used. The *ROI extraction* differs from [Lu13]: Instead of cutting out a defined rectangle within the finger, the width of the finger is stretched to a fixed width and normalized using CPN. It is worth to note, that no image enhancement techniques (e.g. contrast enhancement) have been applied to the input images.

CNN Training: In order to study the influence of using training data from different longitudinal rotations on the rotation invariance of the CNNs, the range of rotation from which the training samples were taken was varied. The ranges are 0° (which corresponds to the training of a classical single-camera recognition system with images from palmar view) and ±5°, ±15°, ±30° and ±45° from the palmar view (0°). All experiments are executed using (1) images acquired at different rotations and (2) augmented images simulating different rotations for CNN training.

Training is performed for 400 epochs starting with a learning rate of 0.001 for SqNet and 0.005 for DenseNet. The learning rate is divided by 10 each 120 epochs. For both nets, training is performed on batches of 128 images. The images are resized to 224×224 pixels and normalized to zero mean and unit variance before feeding them into the CNNs. The two employed nets are pre-trained on the ImageNet database. In order to ensure a 100% separation of the training and evaluation data set, the training data was taken from the PMMDB, while for evaluation it was taken from PLUSVein-FR.

Evaluation Protocol: The EER is used to assess the recognition performance. The evaluation follows the test protocol of the FVC2004 [Ma04]. The employed similarity metric to measure the similarity between CNN feature outputs of different images (genuine and impostor scores) is derived from the Euclidean distance. To transform the Euclidean distance to a similarity metric, the Euclidean distances are inversed ($d \rightarrow 1/d$) and normalized so that the resulting similarity values range from zero to one.

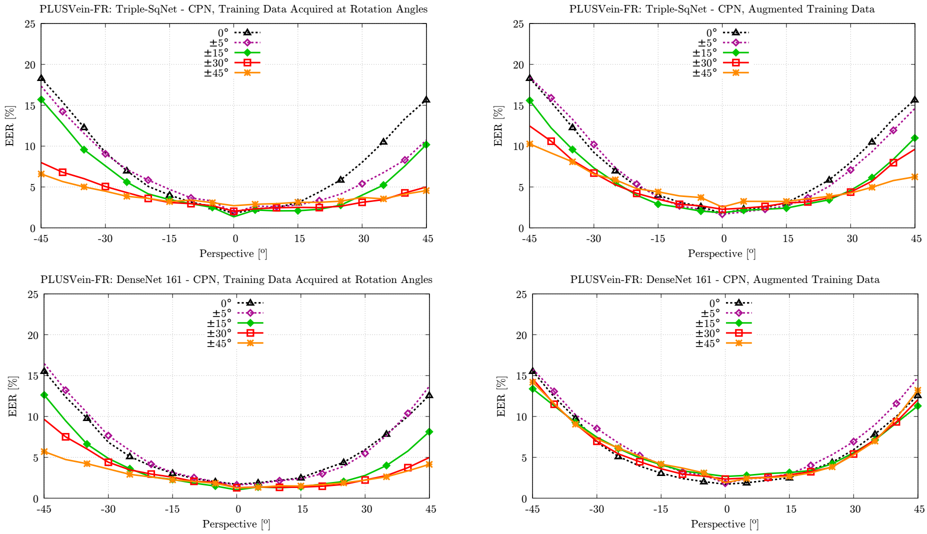


Fig. 3: Trend of the EER across different longitudinal rotations applying Triplet-SqNet and DenseNet-161 trained with different rotational ranges

Evaluation of the CNN's Rotation Invariance: For the evaluation of the rotation invariance of the CNNs, we apply networks trained using images from different rotational ranges (0°, ±5°, ±15°, ±30° and ±45°). For the evaluations, the vein images acquired at a certain rotation angle φ are compared to the ones acquired at the palmar view. φ is varied from -45° to 45°.

The trend of the EERs for Triplet-SqNet are shown in the top row of Figure 3. The left image holds the results for the experiments using training images actually acquired at different angles, whereas the right plot depicts the results for the augmented training images, where the images have been acquired at the palmar view and the rotation has been simulated as described in Section 3. The plots reveal that the proposed approach to train CNNs with vein images from different rotations works quite well. For Triplet-SqNet, the recognition rates of the reference evaluation (training only with images of the palmar view) drop rapidly for increasing rotational differences. With an increasing rotational range of the actually acquired training data, this decline becomes far less pronounced. For a training range of ±45°, the EER at the palmar view (0°) is approximately 3%. For the perspectives at +45° and -45° it is still around 6% for using training data acquired at different rotation angles. Training the CNN with augmented image data improves the results as well, but not to the same extent. For the training range of ±45°, this results in EERs below 10% at +45° and -45°.

The same evaluations have been executed for DenseNet-161 (bottom row of Figure 3). Training the DenseNet-161 using images acquired from larger rotational ranges improves the recognition results, and therefore also the CNNs invariance to longitudinal rotations. Training with larger rotational ranges leads to slightly smaller improvements compared to the reference setting (training images only taken from the palmar view) as for Triplet-

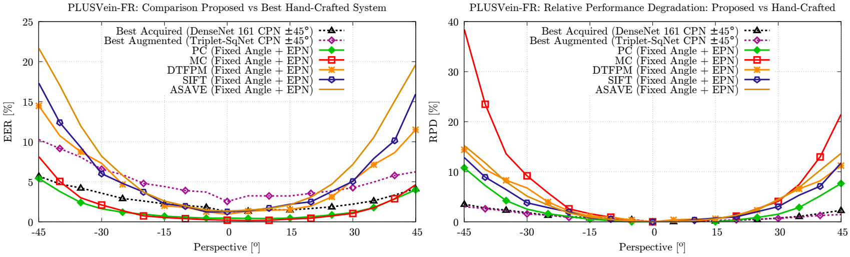


Fig. 4: Trend of the EER (left) and RPD (right) across different longitudinal rotations comparing the proposed systems with hand-crafted single perspective recognition systems

SqNet, but the performance of the reference settings is also noticeable better for DenseNet-161 over the whole range of ±45°. The results using augmented input data for DenseNet-161 show no clear improvements.

5 Discussions

In order to be able to quantify the performance of the proposed method, the best performing methods for actually acquired rotated training data (DenseNet-161 using CPN and ±45°) and for using augmented training data (Triplet-SqNet using CPN and ±45°) are compared to the best performing methods of a previous evaluation on longitudinal rotations in finger vein recognition [Pr19]. The comparison methods comprise *Principal Curvature* (PC) [Ch09], *Maximum Curvature* (MC) [MNM07], *Deformation Tolerant Feature-Point Matching* (DTFPM) [Ma16], a SIFT based approach [Qi13] and *Finger Vein Recognition With Anatomy Structure Analysis* (ASAVE) [Ya17]. The evaluations in [Pr19] showed, that the mentioned recognition schemes achieve their best results when combined with *Elliptic Pattern Normalization* (EPN) [Hu10] and the *Fixed Angle Approach* [Pr19] to counteract longitudinal finger rotation.

The results presented in the left plot of Fig. 4 indicate, that the performance of classical hand-crafted recognition systems is good if the samples contain little to no longitudinal rotations. With an increasing rotational distance between the probe and enrolment samples, the performance drops noticeable. The best performing classical systems are the simple vein pattern based approaches PC and MC. For more sophisticated approaches (DTFPM, SIFT and ASAVE), the absolute performance degradation due to longitudinal rotation is higher. In contrast, the EER of the proposed CNN based approaches are higher for smaller rotations, but the drop of the performance is lower when the rotation increases. The right plot of Fig. 4 should visualize this effect by plotting the relative performance degradation (RPD), calculated as RPD $= \frac{ERR_{rotated} - ERR_{palmar}}{ERR_{palmar}}$, of the different methods. It is obvious that the two CNNs using our proposed training strategies are most robust against longitudinal rotation as their drop in performance is the least.

Besides to the robustness against longitudinal rotation, the proposed CNN approach has some additional advantages over traditional hand-crafted solutions:

Pre-Processing: Most traditional finger vein recognition systems require different preprocessing steps (e.g. image enhancement) that have to be tailored to each data set. Apart from the ROI extraction, the approach presented in this article does not require any preprocessing except of resizing and normalization (which are standard preprocessing steps for CNNs and require hardly any computation time and do not require any adaption to different data sets).

Cost of Time: Once the CNNs are trained, executing a single comparison is very fast. On average, feature extraction takes 7 ms, a single comparison 0.01 ms. This is way faster as for hand-crafted approaches applying time consuming approaches to increase rotation invariance. Experiments in [Pr19] have shown that e.g. for PC with the rotation compensation scheme ”fixed angle” and EPN feature extraction takes just below 130 ms, and a comparison 2.4 ms.

6 Conclusions

In this article, we presented a novel CNN training strategy to increase the CNN’s tolerance against longitudinal finger rotation. It is the first CNN-based approach to achieve rotation tolerance on single camera finger vein recognition systems and it can be applied to any CNN, regardless of the used net architecture and loss function. We showed, that by training the CNNs using vein images acquired from different perspectives, the tolerance with respect to longitudinal finger rotation of the CNNs can be increased noticeable. For Triplet-SqNet, the same holds true if images acquired from a single perspective are artificially rotated into different perspectives for the training (data augmentation), but to a smaller extent.

Although the trained CNNs do not yet achieve the same baseline performance (when all samples are acquired from the same perspectives) as systems utilizing classic hand-crafted features, their tolerance against longitudinal finger rotation is exceptional good. The performance degradation caused by longitudinal finger rotation is noticeable lower for the trained CNNs compared to classical systems. Besides the CNN’s robustness to rotations, other advantages compared to classical systems are that CNNs do not need any special pre-processing (besides of the ROI extraction) and that a single biometric comparison is very fast.

References

[Ch09] Choi, J. H.; Song, W.; Kim, T.; Lee, S-R; Kim, H. C.: Finger vein extraction using gradient normalization and principal curvature. In: Image Processing: Machine Vision Applications II. volume 7251, pp. 1–9, 2009.

[Ga20] Galdi, C.; Boyle, J.; Chen, L.; Chiesa, V.; Debiasi, L.; Dugelay, J-L; Ferryman, J.; Grudzien, A.; Kauba, C.; Kirchgasser, S.; Kowalski, M.; Linortner, M.; Maik, P.; Michon, K.; Patino, L.; Prommegger, B.; Sequeira, A. F.; Szklarski, L.; Uhl, A.: PROTECT: Pervasive and UseR Focused BiomeTrics BordEr ProjeCT. IET Biometrics, September 2020.

[HLP17] Hong, H. G.; Lee, M. B.; Park, K. R.: Convolutional neural network-based finger-vein recognition using NIR image sensors. Sensors, 17(6):1297, 2017.

[Hu10] Huang, B.; Dai, Y.; Li, R.; Tang, D.; Li, W.: Finger-vein authentication based on wide line detector and pattern normalization. In: Pattern Recognition (ICPR), 2010 20th International Conference on. IEEE, pp. 1269–1272, 2010.

[Ka19] Kang, W.; Liu, H.; Luo, W.; Deng, F.: Study of a full-view 3D Finger Vein Verification Technique. IEEE Transactions on Information Forensics and Security, pp. 1–1, 2019.

[KPU19] Kauba, C.; Prommegger, B.; Uhl, A.: Combined Fully Contactless Finger and Hand Vein Capturing Device with a Corresponding Dataset. Sensors, 19(22)(5014):1–25, 2019.

[Lu13] Lu, Y.; Xie, S. J.; Yoon, S.; Yang, J.; Park, D. S.: Robust finger vein ROI localization based on flexible segmentation. Sensors, 13(11):14339–14366, 2013.

[Ma04] Maio, D.; Maltoni, D.; Cappelli, R.; Wayman, J. L.; Jain, A. K.: FVC2004: Third Fingerprint Verification Competition. In: ICBA. volume 3072 of LNCS. Springer Verlag, pp. 1–7, 2004.

[Ma16] Matsuda, Y.; Miura, N.; Nagasaka, A.; Kiyomiu, H.; Miyatake, T.: Finger-vein authentication based on deformation-tolerant feature-point matching. Machine Vision and Applications, 27(2):237–250, 2016.

[Ma17] Matsuda, Y.; Miura, N.; Nonomura, Y.; Nagasaka, A.; Miyatake, T.: Walkthrough-style multi-finger vein authentication. In: Proceedings of the IEEE International Conference on Consumer Electronics (ICCE'17). pp. 438–441, 2017.

[MNM07] Miura, N.; Nagasaka, A.; Miyatake, T.: Extraction of finger-vein patterns using maximum curvature points in image profiles. IEICE transactions on information and systems, 90(8):1185–1194, 2007.

[PKU18a] Prommegger, B.; Kauba, C.; Uhl, A.: Multi-Perspective Finger-Vein Biometrics. In: Proceedings of the IEEE 9th International Conference on Biometrics: Theory, Applications, and Systems (BTAS). Los Angeles, California, USA, 2018.

[PKU18b] Prommegger, Bernhard; Kauba, Christof; Uhl, Andreas: Longitudinal Finger Rotation - Problems and Effects in Finger-Vein Recognition. In: Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG'18). Darmstadt, Germany, 2018.

[PKU19] Prommegger, B.; Kauba, C.; Uhl, A.: On the Extent of Longitudinal Finger Rotation in Publicly Available Finger Vein Data Sets. In: Proceedings of the 12th IAPR/IEEE International Conference on Biometrics (ICB'19). Crete, Greece, pp. 1–8, 2019.

[Pr19] Prommegger, B.; Kauba, C.; Linortner, M.; Uhl, A.: Longitudinal Finger Rotation - Deformation Detection and Correction. IEEE Transactions on Biometrics, Behavior, and Identity Science, 1(2):123–138, 2019.

[PU19] Prommegger, B.; Uhl, A.: Rotation Invariant Finger Vein Recognition. In: Proceedings of the IEEE 10th International Conference on Biometrics: Theory, Applications, and Systems (BTAS). Tampa, Florida, USA, 2019.

[Qi13] Qin, H.; Qin, L.; Xue, L.; He, X.; Yu, C.; Liang, X.: Finger-Vein Verification Based on Multi-Features Fusion. Sensors, 13(11):15048–15067, Nov 2013.

[WPU20] Wimmer, G.; Prommegger, B.; Uhl, A.: Finger Vein Recognition and Intra-Subject Similarity Evaluation of Finger Veins using the CNN Triplet Loss. In: Proceedings of the 25th International Conference on Pattern Recognition (ICPR). pp. 1–7, 2020.

[Ya17] Yang, L.; Yang, G.; Yin, Y.; Xi, X.: Finger Vein Recognition with Anatomy Structure Analysis. IEEE Transactions on Circuits and Systems for Video Technology, pp. 1–1, 2017.

N-shot Palm Vein Verification Using Siamese Networks

Felix Marattukalam[1] , Waleed H. Abdulla[2] , Akshya Swain[3]

Abstract: The use of deep learning methods to extract vascular biometric patterns from the palm surface has been of interest among researchers in recent years. In many biometric recognition tasks, there is a limit in the number of training samples. This is because of limited vein biometric databases being available for research. This restricts the application of deep learning methods to design algorithms that can effectively identify or authenticate people for vein recognition. This paper proposes an architecture using Siamese neural network structure for few shot palm vein verification. The proposed network uses images from both the palms and consists of two sub-nets that share weights to identify a person. The architecture's performance was tested on the HK PolyU multi spectral palm vein database with limited samples. The results suggest that the method is effective since it has 91.9% precision, 91.1% recall, 92.2% specificity, 91.5% F1-Score, and 90.5% accuracy values.

Keywords: Palm Vein Verification, biometrics, Siamese neural network, few- shot learning.

1 Introduction

The need for contactless biometric systems have significantly increased due to the onset of the Covid-19 global pandemic. Although various *extrinsic* modalities like face, iris, and palmprint [Fe18] which are tangible part of the body are successfully being used, now there is a need for *intrinsic* systems like finger vein, hand vein, and palm vein [MA19] which are subcutaneous [Uh20] and not visible to the naked eye. The features in these systems are the veins which helps liveliness detection, and is slightly more robust to spoof attacks that has been a challenge among researchers [Uh20]. Palm vein systems, which is the focus of the present investigation, is an*intrinsic* biometric system and preferred due to the ease of interaction with palm vein scanners. A palm vein biometric system has to go through several stages: *acquisition*, *pre-processing*, *feature extraction*, *decision making*. Here, we briefly look into the functions of every stage. A vein scanner captures the palm vein image using a near infrared camera in the acquisition stage. The veins are visible to the camera when illuminated under infrared light with wavelengths of 760-800 nm [MA19]. The acquisition process suffers from issues due to uneven illumination. These issues are addressed in the pre-processing and feature extraction stages. The incoming images are cropped to a region of interest (ROI), essentially the palm region having maximum vein information to perform recognition. Then the ROI image is processed using image processing methods, and passed on to matching algorithms for decision making. The methods for matching can essentially be classified as traditional and deep learning methods. Since palm vein recognition is a classification problem and deep learning methods have been successful on such tasks, researchers are inclined towards its use.

[1] Faculty of Engineering, The University of Auckland, New Zealand, felix.marattukalam@auckland.ac.nz
[2] Faculty of Engineering, The University of Auckland, New Zealand, w.abdulla@auckland.ac.nz
[3] Faculty of Engineering, The University of Auckland, New Zealand, a.swain@auckland.ac.nz

However, the limitation that deep learning has for palm vein recognition is the need for massive databases with high quality labeled images [Th19]. This is often scarce in the case of palm vein recognition systems. Therefore, our work showcases the use of deep learning networks using limited samples for vein verification. The architecture used is inspired from the Siamese neural network structure, and specifically addresses the verification setting in the recognition system.

The contributions of this research paper are: 1) The advantages of Siamese neural network architecture is exploited for palm-vein recognition by sharing the information from both the palms. As a result, a unique Siamese neural network architecture is developed for palm vein verification 2) the proposed architecture is tested on the HK PolyU multi-spectral palm vein database [Zh09], and its performance is evaluated. The performance evaluation show that this Siamese neural network setting is effective for palm vein verification and useful for palm vein recognition systems.

2 Related work

Numerous methods have been proposed to extract and match vein patterns from vein images. These patterns are used for biometric recognition using different approaches. In this section a few methods are presented for sake of completeness.

The extraction methods can be categorised into subspace learning, local descriptor, vessel geometry, and deep learning. The recent inclination is in using Siamese networks to curb the need for large databases [Th19]. [ES14] elaborates how subspace learning uses obtained coefficients as unique features for recognition. Local descriptor approaches are better described in [XYY17]. A detailed comparison on vessel geometry is discussed in one of our previous works [MA20]. Finally, deep learning approaches such as convolutional neural network(CNN), deep belief network (DBN), and auto-encoders (AE) [Qi21] are used for feature extraction and subject recognition.

As highlighted already,deep learning in biometrics needs for large datasets . In one current public database, the images for each subject or class are limited to twelve [Qi21] and not accounting dynamic class change. This led to the need for alternate approaches. Some researchers proposed to augment the available data. They explored generative adversarial networks (GANs) with data augmentation to improve classification performance [Qi21]. This did achieve reasonable performance but did not solve the problem of system speed, data privacy and storage space associated with duplication of input data. Also, data augmentation can easily lead to overfitting. Researchers are inclined towards more effective methods like similarity learning and few-shot learning. [Sh21] discusses these latest approaches and proposes a few shot learning approach for palmprint recognition, which displays good accuracy. We propose in this paper a novel combined Siamese structure for palm vein verification.

3 Methodology

This section discusses the database used, the general Siamese neural network architecture, and the network structure implemented with the loss functions that have been used to evaluate the network performance.

3.1 Database

The palm vein image database used in our research is the HK PolyU Multispectral Palmprint and Palm Vein database (publicly available) [Zh09] released by Hong Kong Polytechnic University (PolyU) Biometric Research Centre. A Near Infrared Region (NIR) scanner is used to capture the palm vein images. The database released has images from 250 subjects (195 male and 55 female) 20-60 years of age. The total database comprises 6000 images from 500 palms collected in two separate illumination sessions. The images captured are of resolution 352 × 288 pixels.

3.2 Siamese Neural Networks

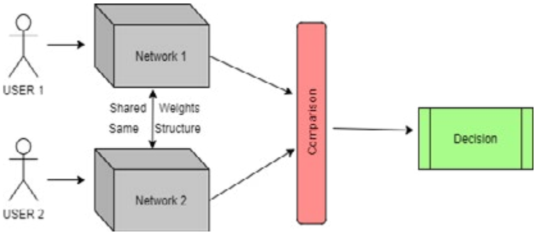


Fig. 1: A typical Siamese neural network structure for biometric system

Even though deep learning algorithms have proven their ability to produce exceptional results, the performance of the designed algorithm is often dependant on the number of data samples available to train the network [Sh21]. The performance of the network improves with the increasing number of data samples. In biometric systems, especially palm vein systems, suitably labelled datasets are not readily available. One-shot or few-shot learning is the appropriate approach when only a few training examples are available for the network to train. The few-shot learning approach uses Siamese neural network. As shown in Fig. 1, a typical Siamese neural network has two paths and aims to find the similarity between its inputs. It has identical parallel networks which share the same architecture and weights. Siamese neural networks was first proposed by Bromley et al [Br93] for signature verification in biometrics and is widely used in face verification tasks. Profound details about Siamese networks for image recognition are available in [Ch21].

3.3 Network Structure

This paper proposes to develop two identical networks that process two images simultaneously and compute the similarity or difference between the two images. If the images are from two different candidates, the network essentially needs to compute the similarity function and increase the distance between them.

Fig.2 shows the overview of the proposed network structure which is based on Siamese architecture. As introduced briefly in section 3.2, the network consists of multiple subnetworks having the same structure and share weights. Here, there are two sub-networks, subnet 1 and subnet 2. Each of them has sub subnets within them to process the input image. The left and right palm images pass through a spatial feature extractor with the convolutional neural network network structure shown in Table 1.

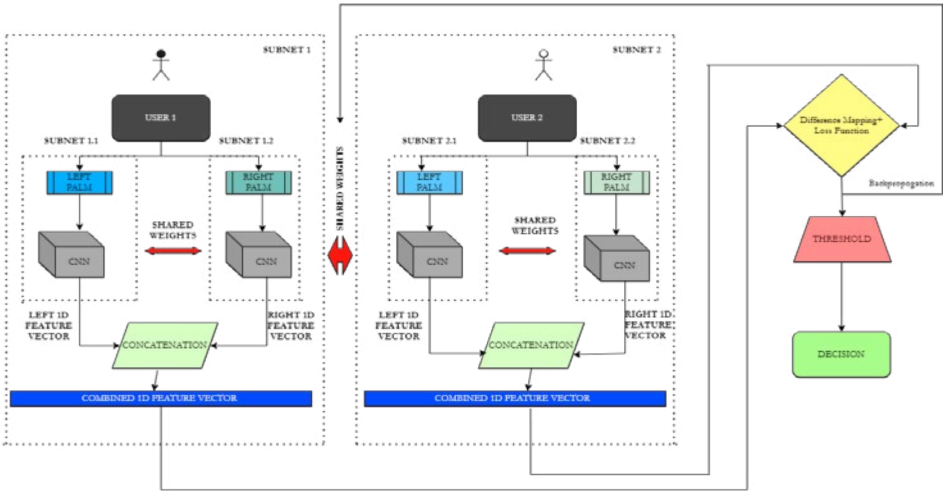


Fig. 2: Overview of the proposed network structure based on Siamese architecture. The red arrows indicate the weights being shared between sub-networks. The sub-networks outputs are 1-D feature vectors. The distance between the feature vectors are then calculated

Tab. 1: CNN feature extractor structure

	Input image	
Layer 1	Convolution 1, 64 x 3 x 3, Stride 1, Padding 0, ReLU	Batch Norm+Max Pool
Layer 2	Convolution 2, 64 x 3 x 3, Stride 1, Padding 0, ReLU	Batch Norm+Max Pool
Layer 3	Convolution 3, 64 x 3 x 3, Stride 1, Padding 1, ReLU	Batch Norm+Max Pool
Layer 4	Convolution 4, 64 x 3 x 3, Stride 1, Padding 1, ReLU	Batch Norm+Max Pool
Layer 5	Fully Connected, 1000 hidden units, ReLU	
Layer 6	Fully Connected, 128 hidden units, Sigmoid	
	Extracted Features	

Consider two users, where each user submits the left and right palm image to the network. Hence, the network receives four images in total, namely, x_1, x_2 and y_1, y_2. The spatial feature extractor network shown in Table 1 generates the feature embeddings $f(x_1)$ and $f(x_2)$ respectively, which are one-dimensional vectors of length 128×1. These vectors are then concatenated together to form $F(X)$. Similar process is followed to obtain $F(Y)$. Then the feature embeddings are subjected to a function E which computes the L_1 distance. The function is given by eqn (1):

$$E(X,Y) = d(X,Y) = ||F(X) - F(Y)|| \tag{1}$$

The function E will be smaller if the concatenated feature vector $F(X)$ is similar to $F(Y)$. This distance value is used to fine-tune the network weights using back propagation. A sigmoid activation function is used to convert the distance to probability P.

3.4 Loss function

Siamese networks classify the inputs into binary classes ie. ”1” being same inputs and ”0” being different. Contrastive loss and binary cross-entropy function loss are the two common loss options in binary classification.

- Contrastive loss: It requires pairs of input samples. The encoder is penalized by the loss function based on the class of the input image. If the input images are from the same class, the model produces similar feature embeddings. Mathematically it is given by eqn (2):

$$Loss = (1-y) * \frac{1}{2}(d)^2 + (y) * \frac{1}{2}[max(0, m-d)]^2 \tag{2}$$

  Here y is the actual label and will be zero when the embeddings of combined input images (left palm and right palm) and one if they are not same, d is the distance measure between the feature embeddings and the input images, m is the hyper parameter margin which is maximized if the input images are similar. If the input pairs are dissimilar and the distance d is greater than the margin *m*, no loss is incurred.

- Binary cross-entropy loss: It is also known as log loss and is used to calculate the classifier performance which is in the range between 0 and 1. If the predicted probability varies from actual class, the loss increases. Mathematically it is given by eqn (3):

$$Loss = -ylogp + (1-y)log(1-p)^2 \tag{3}$$

  Here y is the class label and p is the prediction probability. It is used to differentiate between similar and different images by providing the aggregate of positive and negative loss probability.

4 Experimental Results and Analysis

This section discusses about the results based on the experimental setting and the proposed architecture. The performance analysis of the results is done using the matrices namely, accuracy, precision, recall, specificity and F1-score.

4.1 Implementation

This is a k-way n shot classification problem. The dataset D with a data split of 70:30 was used. The training set contains n samples from k-classes adding upto $k \times n$ samples in the training dataset and a query set in the testing dataset. Predominantly there were only two classes i.e. genuine and imposter subjects, and hence, k=2 and n varying from one to five depending on the sample set considered from the existing database. The model was trained using batch of training tasks to ultimately categorize the image during the testing task. At the end of each epoch, the model parameters were updated through back-propagation as per the loss calculated.

The HK PolyU multispectral database consists of uniform images. We used the region of interest (ROI) images of resolution 128 × 128 pixels by using the method in [LXY16]. The database was prepared into small subsets of classes containing two, three, four and five ROI images for training depending on the value of n. This was subjected to the spatial feature extractor described section 3.3. The experiment was carried out using the Keras framework with Tensorflow on a NVIDIA GTX 2080 8GB GPU with i7 3.3 GHz processor supported with 16 GBs of RAM. The learning rate was set at 0.0001 and Adam Optimizer was used. The one dimensional feature vector extractor which uses the structure shown in Table 1 has fully connected layers having 128 hidden units followed by the sigmoid function.

The model was evaluated using the metrics: accuracy, precision, recall, specificity and F1- score. These parameters were preferred based on our previous study for comparison between SVM with CNN. These parameters are briefly summarised in Table 2.

Tab. 2: Performance metrics using confusion matrix

Accuracy (A)	**Precision (P)**	**Recall (R)**	**Specificity (S)**	**F1-Score**
$\frac{(TP+TN)}{(TP+TN+FP+FN)}$	$\frac{(TP)}{(TP+FP)}$	$\frac{(TP)}{(TP+FN)}$	$\frac{(TN)}{(TN+FP)}$	$\frac{(2\times TP)}{(2\times TP+FP+FN)}$

Here, the number of predictions where the classifier correctly predicts the positive class as positive is True Positive (TP), the negative class as negative is True Negative (TN), the negative class as positive is False Positive (FP), and the positive class as negative is False Negative (FN).

4.2 Results and Discussion

An ideal result for this experiment would be obtaining a classification accuracy of 100 %, precision, recall and specificity values of unity, and a 100 % F1 score. Based on literature it can be said that such results are rare for deep learning models with small datasets. The goal of this study specifically is to exploit the benefits of Siamese neural network in palm-vein verification by discussing its performance parameters and establish its use in an end-to-end palm vein recognition system.

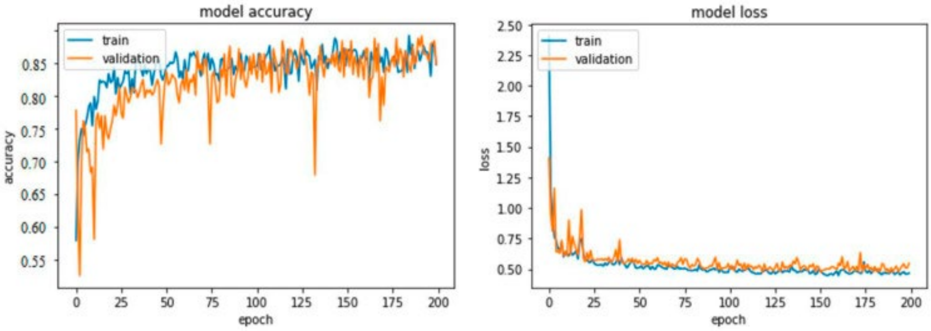


Fig. 3: Accuracy and loss plots: Contrastive loss k=2, n=5

Experiments were performed to evaluate the network performance for different k-way,n-shot learning iterations using both contrastive and cross-entropy losses for classification. However, the results graphically represented in fig.3 are for contrastive loss function as it was seen to be more effective than cross-entropy loss and is in line with what has been

Tab. 3: Results using contrastive loss for 2-way, n-shot settings using both palm images with n varying from 2 to 5.

Model	Accuracy	Recall	Precision	Specificity	F1-Score
k=2, n=2	0.862	0.867	0.874	0.881	0.871
k=2, n=3	0.881	0.885	0.892	0.899	0.889
k=2, n=4	0.892	0.897	0.906	0.911	0.903
k=2, n=5	0.905	0.911	0.919	0.922	0.915

reported in [Li18]. The results obtained in the experiment show that the model training and validation accuracy and loss using contrastive loss function were more stable and merged better than the plots generated for cross-entropy (even though cross-entropy used lower epochs than contrastive function). Also, Siamese neural networks use the principle of similarity between image pairs. As the experiment revolves around n-shot learning, the observations were based on how the model performance varied for different n values. As mentioned in section 4.1, the main results discussed here are for k=2 and n varying from 2 to 5.

Table 3 shows the performance metrics used. Here, contrastive loss was used and *n* varies from 2 to 5. The results show that the proposed network model performance metrics increase steadily as the number of shots/ support samples is varied sequentially. This is justified because the model takes maximum benefit of the increased number of available palm vein image pairs which helps to differentiate a similar image from non-similar ones. Adam optimizer was used along with dropout prevention techniques and relevant learning rate reduction classes in Keras to pause the training process as soon as stagnancy is detected thus reducing overfitting.

5 Conclusion

This paper discusses the dynamics and benefits of palm vein verification and proposes a state-of-the-art deep learning Siamese neural network that can be used in contactless biometric systems. This is achieved by integrating a k-way n-shot learning network model with contrastive loss and an optimized CNN feature encoder. The results highlight that the model for k=2, n-shot learning settings using contrastive loss function is effective. The best case amongst the experiments was the 5-shot learning setting that provides an accuracy of 90.5% in verifying the palm vein image with good recall (91.1%) and specificity (92.2%). These results are critical performance estimates in medical and biometric applications. The results obtained are promising owing to the fact that this model is trained with only a limited sample set of five samples from each palm for a given training class/subset.

References

[Br93] Bromley, Jane; Guyon, Isabelle; LeCun, Yann; Säckinger, Eduard; Shah, Roopak: Signature verification using a" siamese" time delay neural network. Advances in neural information processing systems, 6:737–744, 1993.

[Ch21] Chicco, Davide: Siamese neural networks: An overview. Artificial Neural Networks, pp. 73–94, 2021.

[ES14] Elnasir, Selma; Shamsuddin, Siti Mariyam: Proposed scheme for palm vein recognition based on Linear Discrimination Analysis and nearest neighbour classifier. In: 2014 International Symposium on Biometrics and Security Technologies (ISBAST). IEEE, pp. 67–72, 2014.

[Fe18] Fei, Lunke; Lu, Guangming; Jia, Wei; Teng, Shaohua; Zhang, David: Feature extraction methods for palmprint recognition: A survey and evaluation. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 49(2):346–363, 2018.

[Li18] Lian, Zheng; Li, Ya; Tao, Jianhua; Huang, Jian: Speech emotion recognition via contrastive loss under siamese networks. In: Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and First Multi-Modal Affective Computing of Large-Scale Multimedia Data. pp. 21–26, 2018.

[LXY16] Lin, Sen; Xu, Tianyang; Yin, Xinyong: Region of interest extraction for palmprint and palm vein recognition. In: 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). IEEE, pp. 538–542, 2016.

[MA19] Marattukalam, Felix; Abdulla, Waleed H: On palm vein as a contactless identification technology. In: 2019 Australian & New Zealand Control Conference (ANZCC). IEEE, pp. 270–275, 2019.

[MA20] Marattukalam, Felix; Abdulla, Waleed H: Segmentation of Palm Vein Images Using U-Net. In: 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, pp. 64–70, 2020.

[Qi21] Qin, Huafeng; El-Yacoubi, Mounim A; Li, Yantao; Liu, Chongwen: Multi-Scale and Multi-Direction GAN for CNN-Based Single Palm-Vein Identification. IEEE Transactions on Information Forensics and Security, 16:2652–2666, 2021.

[Sh21] Shao, Huikai; Zhong, Dexing; Du, Xuefeng; Du, Shaoyi; Veldhuis, Raymond NJ: Few-Shot Learning for Palmprint Recognition via Meta-Siamese Network. IEEE Transactions on Instrumentation and Measurement, 2021.

[Th19] Thapar, Daksh; Jaswal, Gaurav; Nigam, Aditya; Kanhangad, Vivek: PVSNet: Palm vein authentication siamese network trained using triplet loss and adaptive hard mining by learning enforced domain specific features. In: 2019 IEEE 5th International Conference on Identity, Security, and Behavior Analysis (ISBA). IEEE, pp. 1–8, 2019.

[Uh20] Uhl, Andreas; Busch, Christoph; Marcel, Sébastien; Veldhuis, Raymond: Handbook of vascular biometrics. Springer Nature, 2020.

[XYY17] Xi, Xiaoming; Yang, Lu; Yin, Yilong: Learning discriminative binary codes for finger vein recognition. Pattern Recognition, 66:26–33, 2017.

[Zh09] Zhang, David; Guo, Zhenhua; Lu, Guangming; Zhang, Lei; Zuo, Wangmeng: An online system of multispectral palmprint verification. IEEE transactions on instrumentation and measurement, 59(2):480–490, 2009.

Assessment of Sensor Ageing-Impact in Air Travelled Fingerprint Capturing Devices

Christof Kauba[1], Simon Kirchgasser[1], Robert Jöchl[1], Andreas Uhl[1]

Abstract: Biometric recognition performance is affected by many factors, like varying acquisition conditions or ageing related effects, commonly denoted as biometric template ageing. Image sensor ageing, being part of biometric template ageing and a sub-field of image and video forensics, leads to defective pixels due to cosmic radiation, depending on the altitude. So far, image sensor ageing has only been a peripheral target in fingerprint research. We investigate the impact of image sensor ageing on various fingerprint capturing devices, including optical, capacitive and thermal ones. We established a fingerprint ageing dataset utilising 10 capturing devices which travelled on an air-plane for 127 days (to increase the number of developed defects). By evaluating the samples captured prior to their travel and afterwards using several state-of-the-art fingerprint quality metrics as well as minutiae-based fingerprint recognition systems we quantify the effect of image sensor ageing on fingerprint recognition. Furthermore, by employing a defect detection technique we quantify the number of defects developed during that period.

Keywords: Fingerprint Recognition, Biometric Template Ageing, Fingerprint Sensor Ageing, Performance Evaluation, Quality Evaluation.

1 Introduction

Fingerprints (FP) are one of the most frequently used biometric modalities for authenticating people in many devices and applications in everyday life, e.g. in smartphones, laptops or border control. Fingerprint recognition systems are also deployed for long-term usage, e.g. for passports typically having a validity period of 10 years, during which time the biometric data (FP and face) is not updated and should be robust against ageing effects, which are referred to as “biometric template ageing”. According to the ISO/IEC 19795-1:2006 standard [IS06] these include biological subject ageing, changes in subject’s behaviour, changes in the acquisition conditions and ageing of the capturing device itself.

Many biometric capturing devices, especially optical FP sensors, contain an optical image sensor, which develops in-field defects in the form of isolated defective pixels over time. These defective pixels exhibit different characteristics than at manufacturing time and appear as point like, spiky shot noise in an output image. There is extensive literature about image sensor ageing related defects and the source causing these defects [Th07, Th08, Ch13, Du07, Le10, Le09, Le08]. In-field sensor defects are permanent, the number of defects increases linearly with time and inter-defect times follow an exponential distribution, indicating a constant defect rate, are randomly distributed over the sensor area, with no significant bias towards short or long distances, i.e. no defect clustering. Evaluation on various types of cameras over several years indicate that cosmic radiation, actually the neutrons of the cosmic rays are most likely causing these single pixel defects.

[1] Department of Computer Sciences, University of Salzburg, AUSTRIA, {ckauba, skirch, rjoechl, uhl}@cs.sbg.ac.at

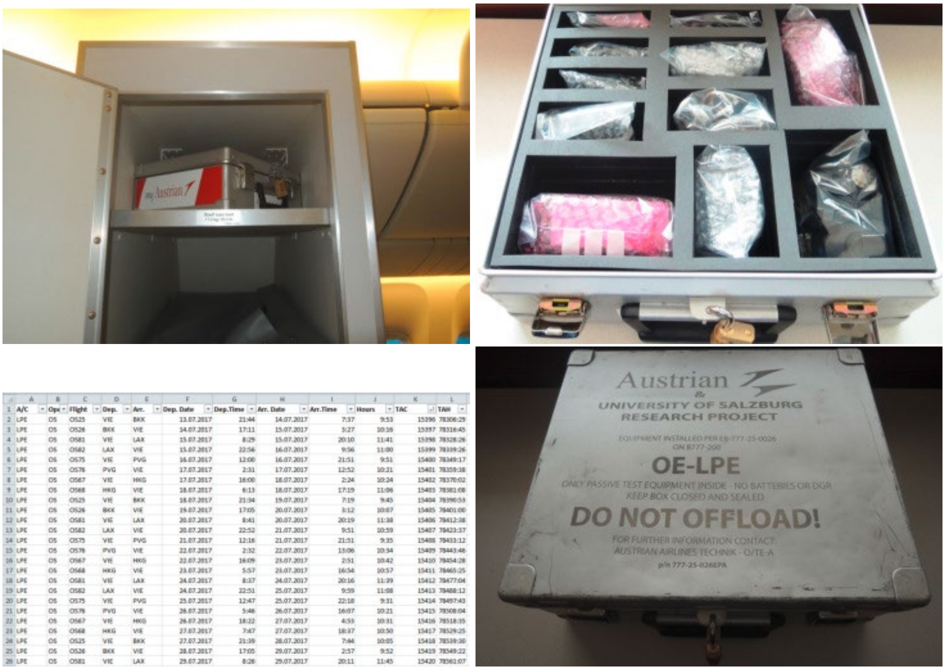
The question if and how image sensor ageing affects the performance of the whole biometric system is still ongoing research. Bergmller et al. [Be14] showed that the sensor ageing related pixel defects may have an influence on iris recognition. Kauba et al. did some work on finger veins [KU15b], hand veins [KU15a] and fingerprints [KU16] where they concluded that image sensor ageing is negligible in practical applications. However, they only simulated the in-field sensor defects rather than letting real devices “age”.
The focus of this work is on the effect of image sensor ageing in practical applications of FP recognition systems and thus, to find out which role it plays within the scope of biometric template ageing. In order to evaluate the impact of in-field sensor defects on fingerprint recognition, a FP ageing dataset using 10 different commercial-off-the-shelf FP capturing devices, including optical, capacitive and thermal ones, divided into two sessions, was established. The first session is the reference session. As the total flux of radiation depends on the altitude, which is about 300-400 times higher at 30,000 feet than on ground level, the capturing devices have been situated on board of a long-haul plane for 127 days to expose them to a higher cosmic ray radiation total flux. Afterwards the second capturing session with the same subjects was conducted. By comparing the samples of both sessions using 1) several FP quality metrics and 2) minutiae-based FP recognition systems, the impact of image sensor ageing induced single pixel defects on the biometric recognition performance can be quantified.
The rest of this work is organised as follows: Section 2 explains the experimental set-up, including details about the FP ageing dataset, the utilised FP capturing devices, the FP quality metrics, the FP recognition systems and the defect detection algorithm. The empirical analysis results are presented in Section 3 and Section 4 concludes this paper.

2 Experimental Set-up

The key aspect of the experimental analysis in order to quantify the effects introduced by FP sensor ageing is that the acquired FP samples were captured before (2 sessions named ’before’ and ’reference’) and after (1 session named ’after’) the utilised capturing devices have been exposed to a higher flux level of cosmic radiation.

2.1 Dataset

As mentioned in Section 1, most studies on image sensor ageing done so far observed corresponding effects after experimentally simulating the ageing process. Naturally, the increase of sensor defects causing detectable sensor ageing effects are introduced by cosmic radiation, which is depending on the altitude. To induce an accelerated ageing process, the capturing devices were exposed to a higher amount of cosmic ray total flux by situating them on a Boeing 777-200 long-haul air plane (the flux in 30,000 feet is about 300-400 times higher than at ground level) for about 127 days while it was travelling around the world. This corresponds to a sensor “age” of about 100 years after the travel (350 times the flux on ground level * 85% of the time at cruise altitude * 127 days / 365 days = 102.85 years). Figure 1 shows the box with the capturing devices and the location on the plane and an excerpt of the flight protocol.


myAustrian
Austrian
&
UNIVERSITY OF SALZBURG
RESEARCH PROJECT
EQUIPMENT INSTALLED PER EB 777-25-0026
ON B777-200
OE-LPE
ONLY PASSIVE TEST EQUIPMENT INSIDE - NO BATTERIES OR DGR
KEEP BOX CLOSED AND SEALED
DO NOT OFFLOAD!
FOR FURTHER INFORMATION CONTACT:
AUSTRIAN AIRLINES TECHNIK - O/TE-A
p/n 777-25-026EPA

	A	B	C	D	E	F	G	H	I	J	K	L
1	A/C	Ope	Flight	Dep.	Arr.	Dep. Date	Dep.Time	Arr. Date	Arr.Time	Hours	TAC	TAH
2	LPE	OS	OS25	VIE	BKK	13.07.2017	21:44	14.07.2017	7:37	9:53	15396	78306:29
3	LPE	OS	OS26	BKK	VIE	14.07.2017	17:11	15.07.2017	3:27	10:16	15397	78316:45
4	LPE	OS	OS81	VIE	LAX	15.07.2017	8:29	15.07.2017	20:10	11:41	15398	78328:26
5	LPE	OS	OS82	LAX	VIE	15.07.2017	22:56	16.07.2017	9:56	11:00	15399	78339:26
6	LPE	OS	OS75	VIE	PVG	16.07.2017	12:00	16.07.2017	21:51	9:51	15400	78349:17
7	LPE	OS	OS76	PVG	VIE	17.07.2017	2:31	17.07.2017	12:52	10:21	15401	78359:38
8	LPE	OS	OS67	VIE	HKG	17.07.2017	16:00	18.07.2017	2:24	10:24	15402	78370:02
9	LPE	OS	OS68	HKG	VIE	18.07.2017	6:13	18.07.2017	17:19	11:06	15403	78381:08
10	LPE	OS	OS25	VIE	BKK	18.07.2017	21:34	19.07.2017	7:19	9:45	15404	78390:53
11	LPE	OS	OS26	BKK	VIE	19.07.2017	17:05	20.07.2017	3:12	10:07	15405	78401:00
12	LPE	OS	OS81	VIE	LAX	20.07.2017	8:41	20.07.2017	20:19	11:38	15406	78412:38
13	LPE	OS	OS82	LAX	VIE	20.07.2017	22:52	21.07.2017	9:51	10:59	15407	78423:37
14	LPE	OS	OS75	VIE	PVG	21.07.2017	12:16	21.07.2017	21:51	9:35	15408	78433:12
15	LPE	OS	OS76	PVG	VIE	22.07.2017	2:32	22.07.2017	13:06	10:34	15409	78443:46
16	LPE	OS	OS67	VIE	HKG	22.07.2017	16:09	23.07.2017	2:51	10:42	15410	78454:28
17	LPE	OS	OS68	HKG	VIE	23.07.2017	5:57	23.07.2017	16:54	10:57	15411	78465:25
18	LPE	OS	OS81	VIE	LAX	24.07.2017	8:37	24.07.2017	20:16	11:39	15412	78477:04
19	LPE	OS	OS82	LAX	VIE	24.07.2017	22:51	25.07.2017	9:59	11:08	15413	78488:12
20	LPE	OS	OS75	VIE	PVG	25.07.2017	12:47	25.07.2017	22:18	9:31	15414	78497:43
21	LPE	OS	OS76	PVG	VIE	26.07.2017	5:46	26.07.2017	16:07	10:21	15415	78508:04
22	LPE	OS	OS67	VIE	HKG	26.07.2017	18:22	27.07.2017	4:53	10:31	15416	78518:35
23	LPE	OS	OS68	HKG	VIE	27.07.2017	7:47	27.07.2017	18:37	10:50	15417	78529:25
24	LPE	OS	OS25	VIE	BKK	27.07.2017	21:39	28.07.2017	7:44	10:05	15418	78539:30
25	LPE	OS	OS26	BKK	VIE	28.07.2017	17:05	29.07.2017	2:57	9:52	15419	78549:22
26	LPE	OS	OS81	VIE	LAX	29.07.2017	8:26	29.07.2017	20:11	11:45	15420	78561:07

Fig. 1: FP capturing devices situated on the plane: top left - location on the place, bottom left - flight protocol excerpt, top right - capturing devices stored in the aluminium box, bottom right - aluminium box top side.

10 different commercial off-the-shelf FP capturing devices, the optical Lumidigm V311, Lumidigm M311, DigitalPersona URU5160, Suprema RealScan G1, the capacitive Integrated Biometrics Columbo, Integrated Biometrics Curve, Zvetco Verifi P5000, DigitalPersona Eikon 710 Touch, Upek Eikon II Swipe and the thermal one Next Biometrics NB-3010-U were used to acquire the FP samples. In the following they are abbreviated as V311, M311, URU5160, G1, IBCol., IBCur., P5000, Touch, Swipe and NB-U, respectively. None of those devices has an electromagnetic shielding and they have not been stored in an isolated storage but inside an aluminium case. The 'before' and 'after' session contain FP samples from 59 subjects. All 10 fingers of each subject were acquired, 5 images per finger, which results in about 2800 samples for each capturing device per session in total. All samples exhibit an image resolution of 500 dpi and have been captured under the same environmental conditions to reduce other influencing factors to a minimum. These samples are a subset of a publicly available database and can be obtained from `http://wavelab.at/sources/PLUS-MSL-FP/`. The remaining third session, 'reference', has been captured before mounting the devices on the air plane similar like the session 'before'. Opposed to the other two sessions the aim was not to capture as much FP specific information, instead the focus was on capturing samples that contain as less FP structure as possible which was achieved by positional variations of the finger in order to capture as much as possible background information (see Figure 2). The uniform background helps to detect possible sensor defects (hot and stuck pixel) which are likely to be

overlaid by the presence of FP information. In total this ‘reference’ subset contains 8000 FP samples acquired from 2 subjects using their right index and middle finger. For each finger 200 samples were captured which results in 400 images per subject.

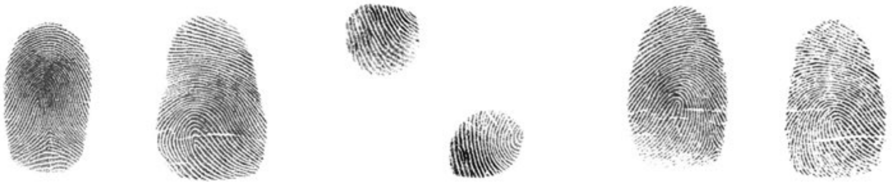


Fig. 2: FP samples contained in the dataset: first two - samples from ’before’, 3rd and 4th - samples from ’reference’ and 5th and 6h - samples from ’after’. Examples were acquired with IBColumbo.

2.2 Quality Measures, Recognition Systems and Defect Detection Algorithm

A three step analysis was performed to measure the impact on the biometric performance. At first, the possible impact on the FP samples’ quality is evaluated by applying several FP quality metrics including the current state-of-the-art metric NIST FP Image Quality 2.0 (NFIQ 2.0 - `https://github.com/usnistgov/NFIQ2`) as well as eleven quality metrics proposed by [OŠB16], namely frequency domain analysis (FDA), Gabor quality (GAB), Gabor-Shen quality (GSH), local clarity score (LCS), orientation flow (OFL), orientation certainty level (OCL), ridge-valley uniformity (RVU) and radial power spectrum (RPS). Second, the impact on the recognition performance is quantified in terms of the equal error rate (EER) by calculating 5,920 mated and 4,335,000 non-mated comparison scores for each capturing device and evaluated session, by employing two fingerprint systems: ANSI & ISO SDK developed by Innovatrics (`https://www.innovatrics.com`) and VeriFinger SDK 11.0 developed by Neurotechnology (`https://www.neurotechnology.com`). Finally, the developed in-field sensor defects are detected and quantified. In principle, in-field sensor defects can be localised with the help of calibration images (e.g. flat-field images). However, the acquisition of such calibration images is hardly possible with FP sensors (most of them only capture an image if they detect a finger on the sensor). Thus, the presence of in-field sensor defects must be determined from regular FP images. A simple approach to threshold the median filter residual variance, as suggested in [FG11] is utilised. A pixel is considered defective if $\sigma(\vec{r})^2 > t$, where $\vec{r}$ is a vector containing the median filter residual values of an arbitrary but fixed pixel over the set of images used for detection, t is a threshold defined by the average residual variance of all pixels plus the standard deviation times an adaptive weight w. The median filter is able to filter out a single point defect in the middle of a homogeneous image region, but it responds on structured image regions (like the ridges and valleys in FP images) as well. This raises the question of how reliable the defect detection method works with FP images.

To evaluate the reliability of the defect detection algorithm, synthetically created defects are embedded into the available FP images from the ‘after’ session using a sensor ageing simulation algorithm [KU16, Be14]. At first a defect map with uniformly distributed defects based on the defined parameters is created and then applied to the output images. The parameters are estimated based on the empirical formula proposed in [Ch13] as done in [KU16]. The resulting rate of 0.244 defects / (MP * year) was multiplied by 400 (due to the higher cosmic ray total flux at 30,000 feet) and $\mu = 0.15$ (exponential distribution) was adopted for this work (correct parameter settings are less relevant for evaluating the defect detection performance). The resulting defect maps are used as ground-truth and the precision $= \frac{TP}{TP+FP}$ is evaluated for the detection method parameter w. TP (True Positives) denotes the correctly detected embedded defects and FP (False Positives) represents additional defect candidates.

3 Experimental Results

In the following we present the quality and performance evaluation results as well as the sensor defect detection results together with a results discussion.

3.1 Quality Evaluation

	quality	NFIQ 2.0	FDA	GAB	GSH	LCS	OFL	OCL	RVU	RPS
before	mean	46.22	0.45	643.94	563.55	0.62	0.62	0.59	751.14	901.25
	median	50.00	0.45	668.39	699.50	0.62	0.63	0.51	672.50	689.06
	std	22.31	0.05	163.78	154.67	0.18	0.17	0.18	174.61	213.71
after	mean	46.00	0.45	649.28	568.22	0.61	0.61	0.58	757.38	908.72
	median	50.00	0.45	665.05	692.63	0.61	0.62	0.50	672.49	684.42
	std	22.65	0.05	165.11	155.93	0.17	0.16	0.18	176.03	226.49

Tab. 1: Mean, median and standard deviation (std) of quality metrics applied on the used FP samples.

The FP quality metrics evaluation results are presented in Table 1. The mean, median and std have been calculated over all quality values with no regard to a specific sensor type. Other results on this dataset showed that the quality depends on the capturing devices' choice and thus, sensor specific fluctuations are detectable. However, even if there are slight fluctuations between the two sessions (before and after), they are due to the used quality metric as all of them measure different characteristics of the FP. There are no consistent changes detectable in the FP quality. Hence, possibly existing sensor defects do not seem to have any influence on the quality of the FP samples in terms of FP quality.

3.2 Recognition Performance

The quality analysis of the FPs did not indicate an influence of the image sensor ageing. So the next step was to quantify the effect on the recognition performance. In Figure 3 the EER results of before and after session are presented. The trend is highly depending on the FP capturing device and similar for both recognition systems (with Verifinger achieving lower EERs than Innovatrics ANSI). For example using Touch and IBCol. an improved recognition performance can be reported for the 'after' session samples. VeriFinger's EER of 0.29% and 0.27% improved to 0.16% and 0.08%, respectively. On the other hand, for URU5160, IBCur. und NB-U, the EER for session 'after' is inferior to the 'before' case (before: 0.16%, 0.79% and after: 0.49%, 0.94% for Innovatrics ANSI). Once again, no general deterioration in the EER is discernible and thus, no influence of potential sensor defects with regard to the recognition performance can be detected.

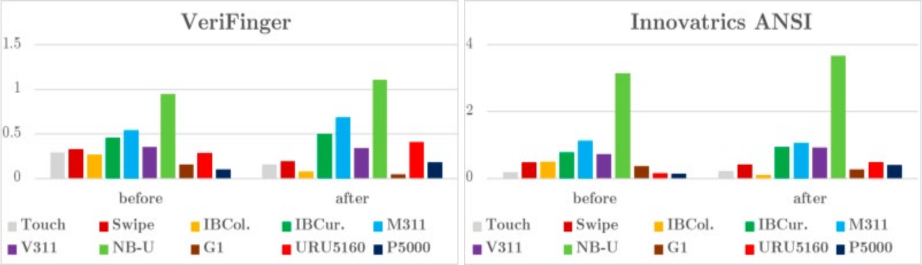


Fig. 3: Recognition performance - EER results in percent, left: VeriFinger, right: Innovatrics ANSI.

3.3 Defect Detection

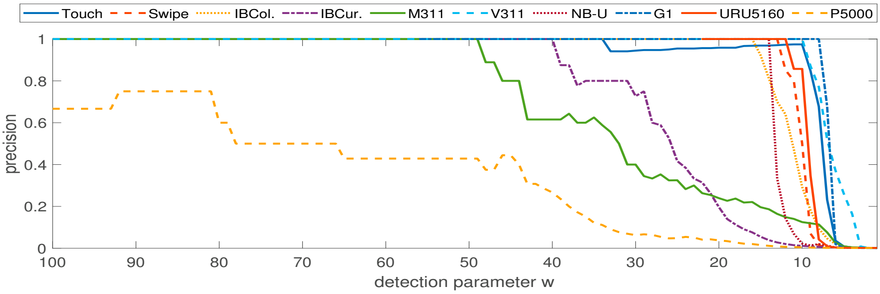
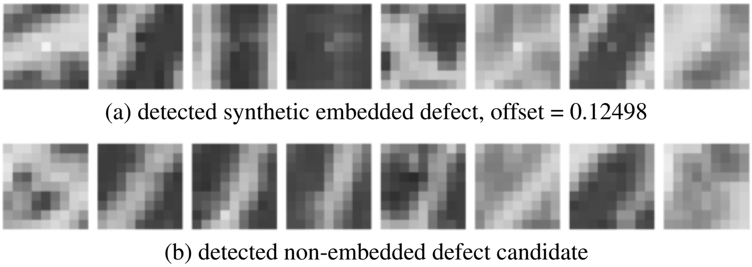


Fig. 4: Precision of the detected embedded defects over the detection parameter *w*.



(a) detected synthetic embedded defect, offset = 0.12498

(b) detected non-embedded defect candidate

Fig. 5: Example for the visual inspection of the detected non-embedded defect candidates.

As neither the FP quality nor the recognition performance evaluation revealed any considerable performance degradations, the next step was to verify if there are any in-field sensor defects detectable at all. By analysing the ’after’ session samples with the embedded defects, we can determine whether the detection method works with FP images and whether real defects are present in addition to the embedded defects. Figure 4 shows the precision of the detected embedded defects. As illustrated, all FP sensors (except the P5000) achieve a precision of 100% (i.e. only embedded defects are detected, $FP = 0$) for higher parameter values. This indicates that the used defect detection method can reliably detect point defects (such as the synthetic embedded ones), at least those with a high offset. In addition, a precision of 100% indicates that there are very likely no real defects with an offset similar to that of the detected synthetic defects. To determine whether the candidates found in addition to the embedded defects are real defects, a visual inspection was performed. This is illustrated in Figure 5, where image patches extracted around a detected defect candidate are compared to image patches (extracted from the same images) of a detected embedded defect. Based on the visual inspection, we can assume that there are no detectable (strong) point defects (for any of the evaluated FP sensors) present in the 'after' session samples. However, if weak (with a small offset) defects are present, they will not affect the FP quality or FP recognition performance. The complete absence of (strong) defects indicates that the evaluated FP sensors either do not develop any in-field defects or they are able to suppress these single pixel defects (e.g. by applying image processing techniques).

4 Conclusion

In order to analyse the potential impact of image sensor ageing on FP recognition systems in general and on FP capturing devices in particular, a FP ageing dataset using 10 different FP capturing devices, including optical, capacitive and thermal ones was established. The capturing devices were situated on an air plane for 127 days to expose them to a higher cosmic radiation flux. The samples acquired in the sessions before and after situating the devices on the plane were evaluated using several FP image quality metrics and two minutiae-based FP recognition systems. In addition, an image sensor defect detection technique was utilised to quantify the number of developed defects.

The results of the FP quality assessment and the FP recognition performance suggest that even the increased level of cosmic ray radiation flux had no impact on the FP samples, neither in terms of the FP quality nor in terms of the biometric recognition performance. Moreover, the applied defect detection did not even result in a single detectable defect in all of the FP capturing devices.

Hence, the FP capturing devices either do not develop any defects at all or they employ some kind of defect concealment (during image post-processing done by the capturing device itself). Hence, there is no influence on the biometric quality and recognition performance which confirms the synthetic evaluation results of Kauba et. al [KU16] that in practice image sensor ageing is not a problem for FP recognition. Thus, these FP devices can be employed as an access control system for the cockpit door on an aircraft.

Due to the limited space the study presented only a first insight. The future work will include a more detailed per sensor analysis (quality as well as recognition performance), relating the results to the type of sensor (optical, capacitive, thermal) and analysing a potential correlation of the sensor ageing effects with the type of sensor. A further exposure to radiation with a dosimeter in parallel to measure the exact amount of radiation as well as an exposure to a controlled radiation (e.g. medical radiation device) are planned.

Acknowledgements

We gratefully thank the Austrian Airlines for their cooperation, situating the FP capturing devices on one of their air planes for 127 days in 2017. Furthermore, the current study received funding from the FFG KIRAS project BioCapture under grant No. 873462.

References

[Be14] Bergmüller, Thomas; Debiasi, Luca; Uhl, Andreas; Sun, Zhenan: Impact of sensor ageing on iris recognition. In: Proceedings of the IAPR/IEEE International Joint Conference on Biometrics (IJCB'14). 2014.

[Ch13] Chapman, Glenn H; Thomas, Rohit; Koren, Zahava; Koren, Israel: Empirical formula for rates of hot pixel defects based on pixel size, sensor area, and ISO. In: IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics, p. 8 pages, 2013.

[Du07] Dudas, Jozsef; Wu, Linda M; Jung, Cory; Chapman, Glenn H; Koren, Zahava; Koren, Israel: Identification of in-field defect development in digital image sensors. In: Electronic Imaging 2007. International Society for Optics and Photonics, p. 8 pages, 2007.

[FG11] Fridrich, Jessica; Goljan, Miroslav: Determining approximate age of digital images using sensor defects. In (Memon, Nasir D.; Dittmann, Jana; Alattar, Adnan M.; III, Edward J. Delp, eds): Media Watermarking, Security, and Forensics III. volume 7880. International Society for Optics and Photonics, SPIE, pp. 49 – 59, 2011.

[IS06] ISO 19795-1:2006: Information technology–Biometric performance testing and reporting–Part 1: Principles and framework. Standard, International Organization for Standardization, 2006.

[KU15a] Kauba, Christof; Uhl, Andreas: Robustness Evaluation of Hand Vein Recognition Systems. In: Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG'15). Darmstadt, Germany, p. 8, 2015.

[KU15b] Kauba, Christof; Uhl, Andreas: Sensor Ageing Impact on Finger-Vein Recognition. In: Proceedings of the 8th IAPR/IEEE International Conference on Biometrics (ICB'15). Phuket, Thailand, pp. 1–8, May 2015.

[KU16] Kauba, Christof; Uhl, Andreas: Fingerprint Recognition under the Influence of Sensor Ageing. IET Biometrics, 4(6):245–255, 2016.

[Le08] Leung, Jenny; Dudas, Jozsef; Chapman, Glenn H; Koren, Zahava; Koren, Israel: Characterization of pixel defect development during digital imager lifetime. In: Electronic Imaging 2008. International Society for Optics and Photonics, pp. 1–12, 2008.

[Le09] Leung, Jenny; Chapman, Glenn H; Koren, Zahava; Koren, Israel: Statistical identification and analysis of defect development in digital imagers. In: IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics, pp. 1–12, 2009.

[Le10] Leung, Jenny; Chapman, Glenn H; Choi, Yong H; Thomas, Rohit; Koren, Zahava; Koren, Israel: Analyzing the Impact of ISO on Digital Imager Defects with an Automated Defect Trace Algorithm. In: Proc. SPIE. volume 7536, p. 8 pages, 2010.

[OŠB16] Olsen, Martin Aastrup; Šmida, Vladimír; Busch, Christoph: Finger image quality assessment features–definitions and evaluation. IET Biometrics, 5(2):47–64, 2016.

[Th07] Theuwissen, Albert JP: Influence of terrestrial cosmic rays on the reliability of CCD image sensors-Part 1: Experiments at room temperature. Electron Devices, IEEE Transactions on, 54(12):3260–3266, 2007.

[Th08] Theuwissen, Albert JP: Influence of Terrestrial Cosmic Rays on the Reliability of CCD Image Sensors-Part 2: Experiments at Elevated Temperature. Electron Devices, IEEE Transactions on, 55(9):2324–2328, 2008.

GI-Edition Lecture Notes in Informatics

P-1 Gregor Engels, Andreas Oberweis, Albert Zündorf (Hrsg.): Modellierung 2001.

P-2 Mikhail Godlevsky, Heinrich C. Mayr (Hrsg.): Information Systems Technology and its Applications, ISTA'2001.

P-3 Ana M. Moreno, Reind P. van de Riet (Hrsg.): Applications of Natural Lan-guage to Information Systems, NLDB'2001.

P-4 H. Wörn, J. Mühling, C. Vahl, H.-P. Meinzer (Hrsg.): Rechner- und sensorgestützte Chirurgie; Workshop des SFB 414.

P-5 Andy Schürr (Hg.): OMER – Object-Oriented Modeling of Embedded Real-Time Systems.

P-6 Hans-Jürgen Appelrath, Rolf Beyer, Uwe Marquardt, Heinrich C. Mayr, Claudia Steinberger (Hrsg.): Unternehmen Hochschule, UH'2001.

P-7 Andy Evans, Robert France, Ana Moreira, Bernhard Rumpe (Hrsg.): Practical UML-Based Rigorous Development Methods – Countering or Integrating the extremists, pUML'2001.

P-8 Reinhard Keil-Slawik, Johannes Magenheim (Hrsg.): Informatikunterricht und Medienbildung, INFOS'2001.

P-9 Jan von Knop, Wilhelm Haverkamp (Hrsg.): Innovative Anwendungen in Kommunikationsnetzen, 15. DFN Arbeitstagung.

P-10 Mirjam Minor, Steffen Staab (Hrsg.): 1st German Workshop on Experience Management: Sharing Experiences about the Sharing Experience.

P-11 Michael Weber, Frank Kargl (Hrsg.): Mobile Ad-Hoc Netzwerke, WMAN 2002.

P-12 Martin Glinz, Günther Müller-Luschnat (Hrsg.): Modellierung 2002.

P-13 Jan von Knop, Peter Schirmbacher and Viljan Mahni_ (Hrsg.): The Changing Universities – The Role of Technology.

P-14 Robert Tolksdorf, Rainer Eckstein (Hrsg.): XML-Technologien für das Semantic Web – XSW 2002.

P-15 Hans-Bernd Bludau, Andreas Koop (Hrsg.): Mobile Computing in Medicine.

P-16 J. Felix Hampe, Gerhard Schwabe (Hrsg.): Mobile and Collaborative Business 2002.

P-17 Jan von Knop, Wilhelm Haverkamp (Hrsg.): Zukunft der Netze –Die Verletzbarkeit meistern, 16. DFN Arbeitstagung.

P-18 Elmar J. Sinz, Markus Plaha (Hrsg.): Modellierung betrieblicher Informationssysteme – MobIS 2002.

P-19 Sigrid Schubert, Bernd Reusch, Norbert Jesse (Hrsg.): Informatik bewegt – Informatik 2002 – 32. Jahrestagung der Gesellschaft für Informatik e.V. (GI) 30.Sept.-3. Okt. 2002 in Dortmund.

P-20 Sigrid Schubert, Bernd Reusch, Norbert Jesse (Hrsg.): Informatik bewegt – Informatik 2002 – 32. Jahrestagung der Gesellschaft für Informatik e.V. (GI) 30.Sept.-3. Okt. 2002 in Dortmund (Ergänzungs-band).

P-21 Jörg Desel, Mathias Weske (Hrsg.): Promise 2002: Prozessorientierte Methoden und Werkzeuge für die Entwicklung von Informationssystemen.

P-22 Sigrid Schubert, Johannes Magenheim, Peter Hubwieser, Torsten Brinda (Hrsg.): Forschungsbeiträge zur "Didaktik der Informatik" – Theorie, Praxis, Evaluation.

P-23 Thorsten Spitta, Jens Borchers, Harry M. Sneed (Hrsg.): Software Management 2002 – Fortschritt durch Beständigkeit

P-24 Rainer Eckstein, Robert Tolksdorf (Hrsg.): XMIDX 2003 – XML-Technologien für Middleware – Middleware für XML-Anwendungen

P-25 Key Pousttchi, Klaus Turowski (Hrsg.): Mobile Commerce – Anwendungen und Perspektiven – 3. Workshop Mobile Commerce, Universität Augsburg, 04.02.2003

P-26 Gerhard Weikum, Harald Schöning, Erhard Rahm (Hrsg.): BTW 2003: Datenbanksysteme für Business, Technologie und Web

P-27 Michael Kroll, Hans-Gerd Lipinski, Kay Melzer (Hrsg.): Mobiles Computing in der Medizin

P-28 Ulrich Reimer, Andreas Abecker, Steffen Staab, Gerd Stumme (Hrsg.): WM 2003: Professionelles Wissensmanagement – Er-fahrungen und Visionen

P-29 Antje Düsterhöft, Bernhard Thalheim (Eds.): NLDB'2003: Natural Language Processing and Information Systems

P-30 Mikhail Godlevsky, Stephen Liddle, Heinrich C. Mayr (Eds.): Information Systems Technology and its Applications

P-31 Arslan Brömme, Christoph Busch (Eds.): BIOSIG 2003: Biometrics and Electronic Signatures, Proceedings of the Special Interest Group on Biometrics and Electronic Signatures,
24. July 2003 in Darmstadt, Germany

P-32 Peter Hubwieser (Hrsg.): Informatische Fachkonzepte im Unterricht – INFOS 2003

P-33 Andreas Geyer-Schulz, Alfred Taudes (Hrsg.): Informationswirtschaft: Ein Sektor mit Zukunft

P-34 Klaus Dittrich, Wolfgang König, Andreas Oberweis, Kai Rannenberg, Wolfgang Wahlster (Hrsg.): Informatik 2003 – Innovative Informatikanwendungen (Band 1)

P-35 Klaus Dittrich, Wolfgang König, Andreas Oberweis, Kai Rannenberg, Wolfgang Wahlster (Hrsg.): Informatik 2003 – Innovative Informatikanwendungen (Band 2)

P-36 Rüdiger Grimm, Hubert B. Keller, Kai Rannenberg (Hrsg.): Informatik 2003 – Mit Sicherheit Informatik

P-37 Arndt Bode, Jörg Desel, Sabine Rathmayer, Martin Wessner (Hrsg.): DeLFI 2003: e-Learning Fachtagung Informatik

P-38 E.J. Sinz, M. Plaha, P. Neckel (Hrsg.): Modellierung betrieblicher Informationssysteme – MobIS 2003

P-39 Jens Nedon, Sandra Frings, Oliver Göbel (Hrsg.): IT-Incident Management & IT-Forensics – IMF 2003

P-40 Michael Rebstock (Hrsg.): Modellierung betrieblicher Informationssysteme – MobIS 2004

P-41 Uwe Brinkschulte, Jürgen Becker, Dietmar Fey, Karl-Erwin Großpietsch, Christian Hochberger, Erik Maehle, Thomas Runkler (Edts.): ARCS 2004 – Organic and Pervasive Computing

P-42 Key Pousttchi, Klaus Turowski (Hrsg.): Mobile Economy – Transaktionen und Prozesse, Anwendungen und Dienste

P-43 Birgitta König-Ries, Michael Klein, Philipp Obreiter (Hrsg.): Persistance, Scalability, Transactions – Database Mechanisms for Mobile Applications

P-44 Jan von Knop, Wilhelm Haverkamp, Eike Jessen (Hrsg.): Security, E-Learning. E-Services

P-45 Bernhard Rumpe, Wofgang Hesse (Hrsg.): Modellierung 2004

P-46 Ulrich Flegel, Michael Meier (Hrsg.): Detection of Intrusions of Malware & Vulnerability Assessment

P-47 Alexander Prosser, Robert Krimmer (Hrsg.): Electronic Voting in Europe – Technology, Law, Politics and Society

P-48 Anatoly Doroshenko, Terry Halpin, Stephen W. Liddle, Heinrich C. Mayr (Hrsg.): Information Systems Technology and its Applications

P-49 G. Schiefer, P. Wagner, M. Morgenstern, U. Rickert (Hrsg.): Integration und Datensicherheit – Anforderungen, Konflikte und Perspektiven

P-50 Peter Dadam, Manfred Reichert (Hrsg.): INFORMATIK 2004 – Informatik verbindet (Band 1) Beiträge der 34. Jahrestagung der Gesellschaft für Informatik e.V. (GI), 20.-24. September 2004 in Ulm

P-51 Peter Dadam, Manfred Reichert (Hrsg.): INFORMATIK 2004 – Informatik verbindet (Band 2) Beiträge der 34. Jahrestagung der Gesellschaft für Informatik e.V. (GI), 20.-24. September 2004 in Ulm

P-52 Gregor Engels, Silke Seehusen (Hrsg.): DELFI 2004 – Tagungsband der 2. e-Learning Fachtagung Informatik

P-53 Robert Giegerich, Jens Stoye (Hrsg.): German Conference on Bioinformatics – GCB 2004

P-54 Jens Borchers, Ralf Kneuper (Hrsg.): Softwaremanagement 2004 – Outsourcing und Integration

P-55 Jan von Knop, Wilhelm Haverkamp, Eike Jessen (Hrsg.): E-Science und Grid Ad-hoc-Netze Medienintegration

P-56 Fernand Feltz, Andreas Oberweis, Benoit Otjacques (Hrsg.): EMISA 2004 – Informationssysteme im E-Business und E-Government

P-57 Klaus Turowski (Hrsg.): Architekturen, Komponenten, Anwendungen

P-58 Sami Beydeda, Volker Gruhn, Johannes Mayer, Ralf Reussner, Franz Schweiggert (Hrsg.): Testing of Component-Based Systems and Software Quality

P-59 J. Felix Hampe, Franz Lehner, Key Pousttchi, Kai Ranneberg, Klaus Turowski (Hrsg.): Mobile Business – Processes, Platforms, Payments

P-60 Steffen Friedrich (Hrsg.): Unterrichtskonzepte für inforrmatische Bildung

P-61 Paul Müller, Reinhard Gotzhein, Jens B. Schmitt (Hrsg.): Kommunikation in verteilten Systemen

P-62 Federrath, Hannes (Hrsg.): „Sicherheit 2005“ – Sicherheit – Schutz und Zuverlässigkeit

P-63 Roland Kaschek, Heinrich C. Mayr, Stephen Liddle (Hrsg.): Information Systems – Technology and ist Applications

P-64 Peter Liggesmeyer, Klaus Pohl, Michael Goedicke (Hrsg.): Software Engineering 2005

P-65 Gottfried Vossen, Frank Leymann, Peter Lockemann, Wolffried Stucky (Hrsg.): Datenbanksysteme in Business, Technologie und Web

P-66 Jörg M. Haake, Ulrike Lucke, Djamshid Tavangarian (Hrsg.): DeLFI 2005: 3. deutsche e-Learning Fachtagung Informatik

P-67 Armin B. Cremers, Rainer Manthey, Peter Martini, Volker Steinhage (Hrsg.): INFORMATIK 2005 – Informatik LIVE (Band 1)

P-68 Armin B. Cremers, Rainer Manthey, Peter Martini, Volker Steinhage (Hrsg.): INFORMATIK 2005 – Informatik LIVE (Band 2)

P-69 Robert Hirschfeld, Ryszard Kowalcyk, Andreas Polze, Matthias Weske (Hrsg.): NODe 2005, GSEM 2005

P-70 Klaus Turowski, Johannes-Maria Zaha (Hrsg.): Component-oriented Enterprise Application (COAE 2005)

P-71 Andrew Torda, Stefan Kurz, Matthias Rarey (Hrsg.): German Conference on Bioinformatics 2005

P-72 Klaus P. Jantke, Klaus-Peter Fähnrich, Wolfgang S. Wittig (Hrsg.): Marktplatz Internet: Von e-Learning bis e-Payment

P-73 Jan von Knop, Wilhelm Haverkamp, Eike Jessen (Hrsg.): "Heute schon das Morgen sehen"

P-74 Christopher Wolf, Stefan Lucks, Po-Wah Yau (Hrsg.): WEWoRC 2005 – Western European Workshop on Research in Cryptology

P-75 Jörg Desel, Ulrich Frank (Hrsg.): Enterprise Modelling and Information Systems Architecture

P-76 Thomas Kirste, Birgitta König-Riess, Key Pousttchi, Klaus Turowski (Hrsg.): Mobile Informationssysteme – Potentiale, Hindernisse, Einsatz

P-77 Jana Dittmann (Hrsg.): SICHERHEIT 2006

P-78 K.-O. Wenkel, P. Wagner, M. Morgenstern, K. Luzi, P. Eisermann (Hrsg.): Land- und Ernährungswirtschaft im Wandel

P-79 Bettina Biel, Matthias Book, Volker Gruhn (Hrsg.): Softwareengineering 2006

P-80 Mareike Schoop, Christian Huemer, Michael Rebstock, Martin Bichler (Hrsg.): Service-Oriented Electronic Commerce

P-81 Wolfgang Karl, Jürgen Becker, Karl-Erwin Großpietsch, Christian Hochberger, Erik Maehle (Hrsg.): ARCS´06

P-82 Heinrich C. Mayr, Ruth Breu (Hrsg.): Modellierung 2006

P-83 Daniel Huson, Oliver Kohlbacher, Andrei Lupas, Kay Nieselt and Andreas Zell (eds.): German Conference on Bioinformatics

P-84 Dimitris Karagiannis, Heinrich C. Mayr, (Hrsg.): Information Systems Technology and its Applications

P-85 Witold Abramowicz, Heinrich C. Mayr, (Hrsg.): Business Information Systems

P-86 Robert Krimmer (Ed.): Electronic Voting 2006

P-87 Max Mühlhäuser, Guido Rößling, Ralf Steinmetz (Hrsg.): DELFI 2006: 4. e-Learning Fachtagung Informatik

P-88 Robert Hirschfeld, Andreas Polze, Ryszard Kowalczyk (Hrsg.): NODe 2006, GSEM 2006

P-90 Joachim Schelp, Robert Winter, Ulrich Frank, Bodo Rieger, Klaus Turowski (Hrsg.): Integration, Informationslogistik und Architektur

P-91 Henrik Stormer, Andreas Meier, Michael Schumacher (Eds.): European Conference on eHealth 2006

P-92 Fernand Feltz, Benoît Otjacques, Andreas Oberweis, Nicolas Poussing (Eds.): AIM 2006

P-93 Christian Hochberger, Rüdiger Liskowsky (Eds.): INFORMATIK 2006 – Informatik für Menschen, Band 1

P-94 Christian Hochberger, Rüdiger Liskowsky (Eds.): INFORMATIK 2006 – Informatik für Menschen, Band 2

P-95 Matthias Weske, Markus Nüttgens (Eds.): EMISA 2005: Methoden, Konzepte und Technologien für die Entwicklung von dienstbasierten Informationssystemen

P-96 Saartje Brockmans, Jürgen Jung, York Sure (Eds.): Meta-Modelling and Ontologies

P-97 Oliver Göbel, Dirk Schadt, Sandra Frings, Hardo Hase, Detlef Günther, Jens Nedon (Eds.): IT-Incident Mangament & IT-Forensics – IMF 2006

P-98 Hans Brandt-Pook, Werner Simonsmeier und Thorsten Spitta (Hrsg.): Beratung in der Softwareentwicklung – Modelle, Methoden, Best Practices

P-99 Andreas Schwill, Carsten Schulte, Marco Thomas (Hrsg.): Didaktik der Informatik

P-100 Peter Forbrig, Günter Siegel, Markus Schneider (Hrsg.): HDI 2006: Hochschuldidaktik der Informatik

P-101 Stefan Böttinger, Ludwig Theuvsen, Susanne Rank, Marlies Morgenstern (Hrsg.): Agrarinformatik im Spannungsfeld zwischen Regionalisierung und globalen Wertschöpfungsketten

P-102 Otto Spaniol (Eds.): Mobile Services and Personalized Environments

P-103 Alfons Kemper, Harald Schöning, Thomas Rose, Matthias Jarke, Thomas Seidl, Christoph Quix, Christoph Brochhaus (Hrsg.): Datenbanksysteme in Business, Technologie und Web (BTW 2007)

P-104 Birgitta König-Ries, Franz Lehner, Rainer Malaka, Can Türker (Hrsg.)
MMS 2007: Mobilität und mobile Informationssysteme

P-105 Wolf-Gideon Bleek, Jörg Raasch, Heinz Züllighoven (Hrsg.)
Software Engineering 2007

P-106 Wolf-Gideon Bleek, Henning Schwentner, Heinz Züllighoven (Hrsg.)
Software Engineering 2007 – Beiträge zu den Workshops

P-107 Heinrich C. Mayr, Dimitris Karagiannis (eds.)
Information Systems Technology and its Applications

P-108 Arslan Brömme, Christoph Busch, Detlef Hühnlein (eds.)
BIOSIG 2007: Biometrics and Electronic Signatures

P-109 Rainer Koschke, Otthein Herzog, Karl-Heinz Rödiger, Marc Ronthaler (Hrsg.)
INFORMATIK 2007
Informatik trifft Logistik
Band 1

P-110 Rainer Koschke, Otthein Herzog, Karl-Heinz Rödiger, Marc Ronthaler (Hrsg.)
INFORMATIK 2007
Informatik trifft Logistik
Band 2

P-111 Christian Eibl, Johannes Magenheim, Sigrid Schubert, Martin Wessner (Hrsg.)
DeLFI 2007:
5. e-Learning Fachtagung Informatik

P-112 Sigrid Schubert (Hrsg.)
Didaktik der Informatik in Theorie und Praxis

P-113 Sören Auer, Christian Bizer, Claudia Müller, Anna V. Zhdanova (Eds.)
The Social Semantic Web 2007
Proceedings of the 1st Conference on Social Semantic Web (CSSW)

P-114 Sandra Frings, Oliver Göbel, Detlef Günther, Hardo G. Hase, Jens Nedon, Dirk Schadt, Arslan Brömme (Eds.)
IMF2007 IT-incident management & IT-forensics
Proceedings of the 3rd International Conference on IT-Incident Management & IT-Forensics

P-115 Claudia Falter, Alexander Schliep, Joachim Selbig, Martin Vingron and Dirk Walther (Eds.)
German conference on bioinformatics GCB 2007

P-116 Witold Abramowicz, Leszek Maciszek (Eds.)
Business Process and Services Computing
1st International Working Conference on Business Process and Services Computing BPSC 2007

P-117 Ryszard Kowalczyk (Ed.)
Grid service engineering and manegement
The 4th International Conference on Grid Service Engineering and Management GSEM 2007

P-118 Andreas Hein, Wilfried Thoben, Hans-Jürgen Appelrath, Peter Jensch (Eds.)
European Conference on ehealth 2007

P-119 Manfred Reichert, Stefan Strecker, Klaus Turowski (Eds.)
Enterprise Modelling and Information Systems Architectures
Concepts and Applications

P-120 Adam Pawlak, Kurt Sandkuhl, Wojciech Cholewa, Leandro Soares Indrusiak (Eds.)
Coordination of Collaborative Engineering - State of the Art and Future Challenges

P-121 Korbinian Herrmann, Bernd Bruegge (Hrsg.)
Software Engineering 2008
Fachtagung des GI-Fachbereichs Softwaretechnik

P-122 Walid Maalej, Bernd Bruegge (Hrsg.)
Software Engineering 2008 - Workshopband
Fachtagung des GI-Fachbereichs Softwaretechnik

P-123 Michael H. Breitner, Martin Breunig, Elgar Fleisch, Ley Pousttchi, Klaus Turowski (Hrsg.)
Mobile und Ubiquitäre Informationssysteme – Technologien, Prozesse, Marktfähigkeit
Proceedings zur 3. Konferenz Mobile und Ubiquitäre Informationssysteme (MMS 2008)

P-124 Wolfgang E. Nagel, Rolf Hoffmann, Andreas Koch (Eds.)
9th Workshop on Parallel Systems and Algorithms (PASA)
Workshop of the GI/ITG Speciel Interest Groups PARS and PARVA

P-125 Rolf A.E. Müller, Hans-H. Sundermeier, Ludwig Theuvsen, Stephanie Schütze, Marlies Morgenstern (Hrsg.)
Unternehmens-IT:
Führungsinstrument oder Verwaltungsbürde
Referate der 28. GIL Jahrestagung

P-126 Rainer Gimnich, Uwe Kaiser, Jochen Quante, Andreas Winter (Hrsg.)
10th Workshop Software Reengineering (WSR 2008)

P-127 Thomas Kühne, Wolfgang Reisig, Friedrich Steimann (Hrsg.)
Modellierung 2008

P-128 Ammar Alkassar, Jörg Siekmann (Hrsg.)
Sicherheit 2008
Sicherheit, Schutz und Zuverlässigkeit
Beiträge der 4. Jahrestagung des Fachbereichs Sicherheit der Gesellschaft für Informatik e.V. (GI)
2.-4. April 2008
Saarbrücken, Germany

P-129 Wolfgang Hesse, Andreas Oberweis (Eds.)
Sigsand-Europe 2008
Proceedings of the Third AIS SIGSAND European Symposium on Analysis, Design, Use and Societal Impact of Information Systems

P-130 Paul Müller, Bernhard Neumair, Gabi Dreo Rodosek (Hrsg.)
1. DFN-Forum Kommunikationstechnologien Beiträge der Fachtagung

P-131 Robert Krimmer, Rüdiger Grimm (Eds.)
3rd International Conference on Electronic Voting 2008
Co-organized by Council of Europe, Gesellschaft für Informatik and E-Voting. CC

P-132 Silke Seehusen, Ulrike Lucke, Stefan Fischer (Hrsg.)
DeLFI 2008:
Die 6. e-Learning Fachtagung Informatik

P-133 Heinz-Gerd Hegering, Axel Lehmann, Hans Jürgen Ohlbach, Christian Scheideler (Hrsg.)
INFORMATIK 2008
Beherrschbare Systeme – dank Informatik
Band 1

P-134 Heinz-Gerd Hegering, Axel Lehmann, Hans Jürgen Ohlbach, Christian Scheideler (Hrsg.)
INFORMATIK 2008
Beherrschbare Systeme – dank Informatik
Band 2

P-135 Torsten Brinda, Michael Fothe, Peter Hubwieser, Kirsten Schlüter (Hrsg.)
Didaktik der Informatik –
Aktuelle Forschungsergebnisse

P-136 Andreas Beyer, Michael Schroeder (Eds.)
German Conference on Bioinformatics
GCB 2008

P-137 Arslan Brömme, Christoph Busch, Detlef Hühnlein (Eds.)
BIOSIG 2008: Biometrics and Electronic Signatures

P-138 Barbara Dinter, Robert Winter, Peter Chamoni, Norbert Gronau, Klaus Turowski (Hrsg.)
Synergien durch Integration und Informationslogistik
Proceedings zur DW2008

P-139 Georg Herzwurm, Martin Mikusz (Hrsg.)
Industrialisierung des Software-Managements
Fachtagung des GI-Fachausschusses Management der Anwendungsentwicklung und -wartung im Fachbereich Wirtschaftsinformatik

P-140 Oliver Göbel, Sandra Frings, Detlef Günther, Jens Nedon, Dirk Schadt (Eds.)
IMF 2008 - IT Incident Management & IT Forensics

P-141 Peter Loos, Markus Nüttgens, Klaus Turowski, Dirk Werth (Hrsg.)
Modellierung betrieblicher Informationssysteme (MobIS 2008)
Modellierung zwischen SOA und Compliance Management

P-142 R. Bill, P. Korduan, L. Theuvsen, M. Morgenstern (Hrsg.)
Anforderungen an die Agrarinformatik durch Globalisierung und Klimaveränderung

P-143 Peter Liggesmeyer, Gregor Engels, Jürgen Münch, Jörg Dörr, Norman Riegel (Hrsg.)
Software Engineering 2009
Fachtagung des GI-Fachbereichs Softwaretechnik

P-144 Johann-Christoph Freytag, Thomas Ruf, Wolfgang Lehner, Gottfried Vossen (Hrsg.)
Datenbanksysteme in Business, Technologie und Web (BTW)

P-145 Knut Hinkelmann, Holger Wache (Eds.)
WM2009: 5th Conference on Professional Knowledge Management

P-146 Markus Bick, Martin Breunig, Hagen Höpfner (Hrsg.)
Mobile und Ubiquitäre Informationssysteme – Entwicklung, Implementierung und Anwendung
4. Konferenz Mobile und Ubiquitäre Informationssysteme (MMS 2009)

P-147 Witold Abramowicz, Leszek Maciaszek, Ryszard Kowalczyk, Andreas Speck (Eds.)
Business Process, Services Computing and Intelligent Service Management
BPSC 2009 · ISM 2009 · YRW-MBP 2009

P-148 Christian Erfurth, Gerald Eichler, Volkmar Schau (Eds.)
9th International Conference on Innovative Internet Community Systems
I^{2}CS 2009

P-149 Paul Müller, Bernhard Neumair, Gabi Dreo Rodosek (Hrsg.)
2. DFN-Forum Kommunikationstechnologien
Beiträge der Fachtagung

P-150 Jürgen Münch, Peter Liggesmeyer (Hrsg.)
Software Engineering
2009 - Workshopband

P-151 Armin Heinzl, Peter Dadam, Stefan Kirn, Peter Lockemann (Eds.)
PRIMIUM
Process Innovation for Enterprise Software

P-152 Jan Mendling, Stefanie Rinderle-Ma, Werner Esswein (Eds.)
Enterprise Modelling and Information Systems Architectures
Proceedings of the 3rd Int‘l Workshop EMISA 2009

P-153 Andreas Schwill, Nicolas Apostolopoulos (Hrsg.)
Lernen im Digitalen Zeitalter
DeLFI 2009 – Die 7. E-Learning Fachtagung Informatik

P-154 Stefan Fischer, Erik Maehle Rüdiger Reischuk (Hrsg.)
INFORMATIK 2009
Im Focus das Leben

P-155 Arslan Brömme, Christoph Busch, Detlef Hühnlein (Eds.)
BIOSIG 2009:
Biometrics and Electronic Signatures Proceedings of the Special Interest Group on Biometrics and Electronic Signatures

P-156 Bernhard Koerber (Hrsg.)
Zukunft braucht Herkunft
25 Jahre »INFOS – Informatik und Schule«

P-157 Ivo Grosse, Steffen Neumann, Stefan Posch, Falk Schreiber, Peter Stadler (Eds.)
German Conference on Bioinformatics 2009

P-158 W. Claupein, L. Theuvsen, A. Kämpf, M. Morgenstern (Hrsg.)
Precision Agriculture
Reloaded – Informationsgestützte Landwirtschaft

P-159 Gregor Engels, Markus Luckey, Wilhelm Schäfer (Hrsg.)
Software Engineering 2010

P-160 Gregor Engels, Markus Luckey, Alexander Pretschner, Ralf Reussner (Hrsg.)
Software Engineering 2010 – Workshopband
(inkl. Doktorandensymposium)

P-161 Gregor Engels, Dimitris Karagiannis Heinrich C. Mayr (Hrsg.)
Modellierung 2010

P-162 Maria A. Wimmer, Uwe Brinkhoff, Siegfried Kaiser, Dagmar Lück-Schneider, Erich Schweighofer, Andreas Wiebe (Hrsg.)
Vernetzte IT für einen effektiven Staat
Gemeinsame Fachtagung Verwaltungsinformatik (FTVI) und Fachtagung Rechtsinformatik (FTRI) 2010

P-163 Markus Bick, Stefan Eulgem, Elgar Fleisch, J. Felix Hampe, Birgitta König-Ries, Franz Lehner, Key Pousttchi, Kai Rannenberg (Hrsg.)
Mobile und Ubiquitäre Informationssysteme
Technologien, Anwendungen und Dienste zur Unterstützung von mobiler Kollaboration

P-164 Arslan Brömme, Christoph Busch (Eds.)
BIOSIG 2010: Biometrics and Electronic Signatures Proceedings of the Special Interest Group on Biometrics and Electronic Signatures

P-165 Gerald Eichler, Peter Kropf,
Ulrike Lechner, Phayung Meesad,
Herwig Unger (Eds.)
10th International Conference on
Innovative Internet Community Systems
(I^2CS) – Jubilee Edition 2010 –

P-166 Paul Müller, Bernhard Neumair,
Gabi Dreo Rodosek (Hrsg.)
3. DFN-Forum Kommunikationstechnologien
Beiträge der Fachtagung

P-167 Robert Krimmer, Rüdiger Grimm (Eds.)
4th International Conference on
Electronic Voting 2010
co-organized by the Council of Europe,
Gesellschaft für Informatik and
E-Voting.CC

P-168 Ira Diethelm, Christina Dörge,
Claudia Hildebrandt,
Carsten Schulte (Hrsg.)
Didaktik der Informatik
Möglichkeiten empirischer
Forschungsmethoden und Perspektiven
der Fachdidaktik

P-169 Michael Kerres, Nadine Ojstersek
Ulrik Schroeder, Ulrich Hoppe (Hrsg.)
DeLFI 2010 - 8. Tagung
der Fachgruppe E-Learning
der Gesellschaft für Informatik e.V.

P-170 Felix C. Freiling (Hrsg.)
Sicherheit 2010
Sicherheit, Schutz und Zuverlässigkeit

P-171 Werner Esswein, Klaus Turowski,
Martin Juhrisch (Hrsg.)
Modellierung betrieblicher
Informationssysteme (MobIS 2010)
Modellgestütztes Management

P-172 Stefan Klink, Agnes Koschmider
Marco Mevius, Andreas Oberweis (Hrsg.)
EMISA 2010
Einflussfaktoren auf die Entwicklung
flexibler, integrierter Informationssysteme
Beiträge des Workshops
der GI-Fachgruppe EMISA
(Entwicklungsmethoden für Informationssysteme und deren Anwendung)

P-173 Dietmar Schomburg,
Andreas Grote (Eds.)
German Conference on Bioinformatics
2010

P-174 Arslan Brömme, Torsten Eymann,
Detlef Hühnlein, Heiko Roßnagel,
Paul Schmücker (Hrsg.)
perspeGKtive 2010
Workshop „Innovative und sichere
Informationstechnologie für das
Gesundheitswesen von morgen“

P-175 Klaus-Peter Fähnrich,
Bogdan Franczyk (Hrsg.)
INFORMATIK 2010
Service Science – Neue Perspektiven für
die Informatik
Band 1

P-176 Klaus-Peter Fähnrich,
Bogdan Franczyk (Hrsg.)
INFORMATIK 2010
Service Science – Neue Perspektiven für
die Informatik
Band 2

P-177 Witold Abramowicz, Rainer Alt,
Klaus-Peter Fähnrich, Bogdan Franczyk,
Leszek A. Maciaszek (Eds.)
INFORMATIK 2010
Business Process and Service Science –
Proceedings of ISSS and BPSC

P-178 Wolfram Pietsch, Benedikt Krams (Hrsg.)
Vom Projekt zum Produkt
Fachtagung des GI-
Fachausschusses Management der
Anwendungsentwicklung und -wartung
im Fachbereich Wirtschafts-informatik
(WI-MAW), Aachen, 2010

P-179 Stefan Gruner, Bernhard Rumpe (Eds.)
FM+AM`2010
Second International Workshop on
Formal Methods and Agile Methods

P-180 Theo Härder, Wolfgang Lehner,
Bernhard Mitschang, Harald Schöning,
Holger Schwarz (Hrsg.)
Datenbanksysteme für Business,
Technologie und Web (BTW)
14. Fachtagung des GI-Fachbereichs
„Datenbanken und Informationssysteme“
(DBIS)

P-181 Michael Clasen, Otto Schätzel,
Brigitte Theuvsen (Hrsg.)
Qualität und Effizienz durch
informationsgestützte Landwirtschaft,
Fokus: Moderne Weinwirtschaft

P-182 Ronald Maier (Hrsg.)
6th Conference on Professional
Knowledge Management
From Knowledge to Action

P-183 Ralf Reussner, Matthias Grund, Andreas
Oberweis, Walter Tichy (Hrsg.)
Software Engineering 2011
Fachtagung des GI-Fachbereichs
Softwaretechnik

P-184 Ralf Reussner, Alexander Pretschner,
Stefan Jähnichen (Hrsg.)
Software Engineering 2011
Workshopband
(inkl. Doktorandensymposium)

P-185 Hagen Höpfner, Günther Specht,
Thomas Ritz, Christian Bunse (Hrsg.)
MMS 2011: Mobile und ubiquitäre
Informationssysteme Proceedings zur
6. Konferenz Mobile und Ubiquitäre
Informationssysteme (MMS 2011)

P-186 Gerald Eichler, Axel Küpper,
Volkmar Schau, Hacène Fouchal,
Herwig Unger (Eds.)
11th International Conference on
Innovative Internet Community Systems
(I^{2}CS)

P-187 Paul Müller, Bernhard Neumair,
Gabi Dreo Rodosek (Hrsg.)
4. DFN-Forum Kommunikationstechnologien, Beiträge der Fachtagung
20. Juni bis 21. Juni 2011 Bonn

P-188 Holger Rohland, Andrea Kienle,
Steffen Friedrich (Hrsg.)
DeLFI 2011 – Die 9. e-Learning
Fachtagung Informatik
der Gesellschaft für Informatik e.V.
5.–8. September 2011, Dresden

P-189 Thomas, Marco (Hrsg.)
Informatik in Bildung und Beruf
INFOS 2011
14. GI-Fachtagung Informatik und Schule

P-190 Markus Nüttgens, Oliver Thomas,
Barbara Weber (Eds.)
Enterprise Modelling and Information
Systems Architectures (EMISA 2011)

P-191 Arslan Brömme, Christoph Busch (Eds.)
BIOSIG 2011
International Conference of the
Biometrics Special Interest Group

P-192 Hans-Ulrich Heiß, Peter Pepper, Holger
Schlingloff, Jörg Schneider (Hrsg.)
INFORMATIK 2011
Informatik schafft Communities

P-193 Wolfgang Lehner, Gunther Piller (Hrsg.)
IMDM 2011

P-194 M. Clasen, G. Fröhlich, H. Bernhardt,
K. Hildebrand, B. Theuvsen (Hrsg.)
Informationstechnologie für eine
nachhaltige Landbewirtschaftung
Fokus Forstwirtschaft

P-195 Neeraj Suri, Michael Waidner (Hrsg.)
Sicherheit 2012
Sicherheit, Schutz und Zuverlässigkeit
Beiträge der 6. Jahrestagung des
Fachbereichs Sicherheit der
Gesellschaft für Informatik e.V. (GI)

P-196 Arslan Brömme, Christoph Busch (Eds.)
BIOSIG 2012
Proceedings of the 11th International
Conference of the Biometrics Special
Interest Group

P-197 Jörn von Lucke, Christian P. Geiger,
Siegfried Kaiser, Erich Schweighofer,
Maria A. Wimmer (Hrsg.)
Auf dem Weg zu einer offenen, smarten
und vernetzten Verwaltungskultur
Gemeinsame Fachtagung
Verwaltungsinformatik (FTVI) und
Fachtagung Rechtsinformatik (FTRI)
2012

P-198 Stefan Jähnichen, Axel Küpper,
Sahin Albayrak (Hrsg.)
Software Engineering 2012
Fachtagung des GI-Fachbereichs
Softwaretechnik

P-199 Stefan Jähnichen, Bernhard Rumpe,
Holger Schlingloff (Hrsg.)
Software Engineering 2012
Workshopband

P-200 Gero Mühl, Jan Richling, Andreas
Herkersdorf (Hrsg.)
ARCS 2012 Workshops

P-201 Elmar J. Sinz Andy Schürr (Hrsg.)
Modellierung 2012

P-202 Andrea Back, Markus Bick,
Martin Breunig, Key Pousttchi,
Frédéric Thiesse (Hrsg.)
MMS 2012:Mobile und Ubiquitäre
Informationssysteme

P-203 Paul Müller, Bernhard Neumair,
Helmut Reiser, Gabi Dreo Rodosek (Hrsg.)
5. DFN-Forum Kommunikationstechnologien
Beiträge der Fachtagung

P-204 Gerald Eichler, Leendert W. M.
Wienhofen, Anders Kofod-Petersen,
Herwig Unger (Eds.)
12th International Conference on
Innovative Internet Community Systems
(I2CS 2012)

P-205 Manuel J. Kripp, Melanie Volkamer,
Rüdiger Grimm (Eds.)
5th International Conference on Electronic
Voting 2012 (EVOTE2012)
Co-organized by the Council of Europe,
Gesellschaft für Informatik and E-Voting.CC

P-206 Stefanie Rinderle-Ma,
Mathias Weske (Hrsg.)
EMISA 2012
Der Mensch im Zentrum der Modellierung

P-207 Jörg Desel, Jörg M. Haake,
Christian Spannagel (Hrsg.)
DeLFI 2012: Die 10. e-Learning
Fachtagung Informatik der Gesellschaft
für Informatik e.V.
24.–26. September 2012

P-208 Ursula Goltz, Marcus Magnor,
Hans-Jürgen Appelrath, Herbert Matthies,
Wolf-Tilo Balke, Lars Wolf (Hrsg.)
INFORMATIK 2012

P-209 Hans Brandt-Pook, André Fleer, Thorsten
Spitta, Malte Wattenberg (Hrsg.)
Nachhaltiges Software Management

P-210 Erhard Plödereder, Peter Dencker,
Herbert Klenk, Hubert B. Keller,
Silke Spitzer (Hrsg.)
Automotive – Safety & Security 2012
Sicherheit und Zuverlässigkeit für
automobile Informationstechnik

P-211 M. Clasen, K. C. Kersebaum, A.
Meyer-Aurich, B. Theuvsen (Hrsg.)
Massendatenmanagement in der
Agrar- und Ernährungswirtschaft
Erhebung - Verarbeitung - Nutzung
Referate der 33. GIL-Jahrestagung
20. – 21. Februar 2013, Potsdam

P-212 Arslan Brömme, Christoph Busch (Eds.)
BIOSIG 2013
Proceedings of the 12th International
Conference of the Biometrics
Special Interest Group
04.–06. September 2013
Darmstadt, Germany

P-213 Stefan Kowalewski,
Bernhard Rumpe (Hrsg.)
Software Engineering 2013
Fachtagung des GI-Fachbereichs
Softwaretechnik

P-214 Volker Markl, Gunter Saake, Kai-Uwe
Sattler, Gregor Hackenbroich, Bernhard Mit
schang, Theo Härder, Veit Köppen (Hrsg.)
Datenbanksysteme für Business,
Technologie und Web (BTW) 2013
13. – 15. März 2013, Magdeburg

P-215 Stefan Wagner, Horst Lichter (Hrsg.)
Software Engineering 2013
Workshopband
(inkl. Doktorandensymposium)
26. Februar – 1. März 2013, Aachen

P-216 Gunter Saake, Andreas Henrich,
Wolfgang Lehner, Thomas Neumann,
Veit Köppen (Hrsg.)
Datenbanksysteme für Business,
Technologie und Web (BTW) 2013 –
Workshopband
11. – 12. März 2013, Magdeburg

P-217 Paul Müller, Bernhard Neumair, Helmut
Reiser, Gabi Dreo Rodosek (Hrsg.)
6. DFN-Forum Kommunikationstechnologien
Beiträge der Fachtagung
03.–04. Juni 2013, Erlangen

P-218 Andreas Breiter, Christoph Rensing (Hrsg.)
DeLFI 2013: Die 11 e-Learning
Fachtagung Informatik der Gesellschaft
für Informatik e.V. (GI)
8. – 11. September 2013, Bremen

P-219 Norbert Breier, Peer Stechert,
Thomas Wilke (Hrsg.)
Informatik erweitert Horizonte
INFOS 2013
15. GI-Fachtagung Informatik und Schule
26. – 28. September 2013

P-220 Matthias Horbach (Hrsg.)
INFORMATIK 2013
Informatik angepasst an Mensch,
Organisation und Umwelt
16. – 20. September 2013, Koblenz

P-221 Maria A. Wimmer, Marijn Janssen,
Ann Macintosh, Hans Jochen Scholl,
Efthimios Tambouris (Eds.)
Electronic Government and
Electronic Participation
Joint Proceedings of Ongoing Research of
IFIP EGOV and IFIP ePart 2013
16. – 19. September 2013, Koblenz

P-222 Reinhard Jung, Manfred Reichert (Eds.)
Enterprise Modelling
and Information Systems Architectures
(EMISA 2013)
St. Gallen, Switzerland
September 5. – 6. 2013

P-223 Detlef Hühnlein, Heiko Roßnagel (Hrsg.)
Open Identity Summit 2013
10. – 11. September 2013
Kloster Banz, Germany

P-224 Eckhart Hanser, Martin Mikusz, Masud
Fazal-Baqaie (Hrsg.)
Vorgehensmodelle 2013
Vorgehensmodelle – Anspruch und
Wirklichkeit
20. Tagung der Fachgruppe
Vorgehensmodelle im Fachgebiet
Wirtschaftsinformatik (WI-VM) der
Gesellschaft für Informatik e.V.
Lörrach, 2013

P-225 Hans-Georg Fill, Dimitris Karagiannis,
Ulrich Reimer (Hrsg.)
Modellierung 2014
19. – 21. März 2014, Wien

P-226 M. Clasen, M. Hamer, S. Lehnert,
B. Petersen, B. Theuvsen (Hrsg.)
IT-Standards in der Agrar- und
Ernährungswirtschaft Fokus: Risiko- und
Krisenmanagement
Referate der 34. GIL-Jahrestagung
24. – 25. Februar 2014, Bonn

P-227 Wilhelm Hasselbring,
Nils Christian Ehmke (Hrsg.)
Software Engineering 2014
Fachtagung des GI-Fachbereichs
Softwaretechnik
25. – 28. Februar 2014
Kiel, Deutschland

P-228 Stefan Katzenbeisser, Volkmar Lotz,
Edgar Weippl (Hrsg.)
Sicherheit 2014
Sicherheit, Schutz und Zuverlässigkeit
Beiträge der 7. Jahrestagung des
Fachbereichs Sicherheit der
Gesellschaft für Informatik e.V. (GI)
19. – 21. März 2014, Wien

P-229 Dagmar Lück-Schneider, Thomas
Gordon, Siegfried Kaiser, Jörn von
Lucke,Erich Schweighofer, Maria
A.Wimmer, Martin G. Löhe (Hrsg.)
Gemeinsam Electronic Government
ziel(gruppen)gerecht gestalten und
organisieren
Gemeinsame Fachtagung
Verwaltungsinformatik (FTVI) und
Fachtagung Rechtsinformatik (FTRI)
2014, 20.-21. März 2014 in Berlin

P-230 Arslan Brömme, Christoph Busch (Eds.)
BIOSIG 2014
Proceedings of the 13th International
Conference of the Biometrics Special
Interest Group
10. – 12. September 2014 in
Darmstadt, Germany

P-231 Paul Müller, Bernhard Neumair,
Helmut Reiser, Gabi Dreo Rodosek
(Hrsg.)
7. DFN-Forum
Kommunikationstechnologien
16. – 17. Juni 2014
Fulda

P-232 E. Plödereder, L. Grunske, E. Schneider,
D. Ull (Hrsg.)
INFORMATIK 2014
Big Data – Komplexität meistern
22. – 26. September 2014
Stuttgart

P-233 Stephan Trahasch, Rolf Plötzner, Gerhard
Schneider, Claudia Gayer, Daniel Sassiat,
Nicole Wöhrle (Hrsg.)
DeLFI 2014 – Die 12. e-Learning
Fachtagung Informatik
der Gesellschaft für Informatik e.V.
15. – 17. September 2014
Freiburg

P-234 Fernand Feltz, Bela Mutschler, Benoît
Otjacques (Eds.)
Enterprise Modelling and Information
Systems Architectures
(EMISA 2014)
Luxembourg, September 25-26, 2014

P-235 Robert Giegerich,
Ralf Hofestädt,
Tim W. Nattkemper (Eds.)
German Conference on
Bioinformatics 2014
September 28 – October 1
Bielefeld, Germany

P-236 Martin Engstler, Eckhart Hanser,
Martin Mikusz, Georg Herzwurm (Hrsg.)
Projektmanagement und
Vorgehensmodelle 2014
Soziale Aspekte und Standardisierung
Gemeinsame Tagung der Fachgruppen
Projektmanagement (WI-PM) und
Vorgehensmodelle (WI-VM) im
Fachgebiet Wirtschaftsinformatik der
Gesellschaft für Informatik e.V., Stuttgart
2014

P-237 Detlef Hühnlein, Heiko Roßnagel (Hrsg.)
Open Identity Summit 2014
4.–6. November 2014
Stuttgart, Germany

P-238 Arno Ruckelshausen, Hans-Peter
Schwarz, Brigitte Theuvsen (Hrsg.)
Informatik in der Land-, Forst- und
Ernährungswirtschaft
Referate der 35. GIL-Jahrestagung
23. – 24. Februar 2015, Geisenheim

P-239 Uwe Aßmann, Birgit Demuth, Thorsten
Spitta, Georg Püschel, Ronny Kaiser
(Hrsg.)
Software Engineering & Management
2015
17.-20. März 2015, Dresden

P-240 Herbert Klenk, Hubert B. Keller, Erhard
Plödereder, Peter Dencker (Hrsg.)
Automotive – Safety & Security 2015
Sicherheit und Zuverlässigkeit für
automobile Informationstechnik
21.–22. April 2015, Stuttgart

P-241 Thomas Seidl, Norbert Ritter,
Harald Schöning, Kai-Uwe Sattler,
Theo Härder, Steffen Friedrich,
Wolfram Wingerath (Hrsg.)
Datenbanksysteme für Business,
Technologie und Web (BTW 2015)
04. – 06. März 2015, Hamburg

P-242 Norbert Ritter, Andreas Henrich,
Wolfgang Lehner, Andreas Thor,
Steffen Friedrich, Wolfram Wingerath
(Hrsg.)
Datenbanksysteme für Business,
Technologie und Web (BTW 2015) –
Workshopband
02. – 03. März 2015, Hamburg

P-243 Paul Müller, Bernhard Neumair, Helmut
Reiser, Gabi Dreo Rodosek (Hrsg.)
8. DFN-Forum
Kommunikationstechnologien
06.–09. Juni 2015, Lübeck

P-244 Alfred Zimmermann,
Alexander Rossmann (Eds.)
Digital Enterprise Computing
(DEC 2015)
Böblingen, Germany June 25-26, 2015

P-245 Arslan Brömme, Christoph Busch ,
Christian Rathgeb, Andreas Uhl (Eds.)
BIOSIG 2015
Proceedings of the 14th International
Conference of the Biometrics Special
Interest Group
09.–11. September 2015
Darmstadt, Germany

P-246 Douglas W. Cunningham, Petra Hofstedt,
Klaus Meer, Ingo Schmitt (Hrsg.)
INFORMATIK 2015
28.9.-2.10. 2015, Cottbus

P-247 Hans Pongratz, Reinhard Keil (Hrsg.)
DeLFI 2015 – Die 13. E-Learning
Fachtagung Informatik der Gesellschaft
für Informatik e.V. (GI)
1.–4. September 2015
München

P-248 Jens Kolb, Henrik Leopold, Jan Mendling
(Eds.)
Enterprise Modelling and Information
Systems Architectures
Proceedings of the 6th Int. Workshop on
Enterprise Modelling and Information
Systems Architectures, Innsbruck, Austria
September 3-4, 2015

P-249 Jens Gallenbacher (Hrsg.)
Informatik
allgemeinbildend begreifen
INFOS 2015 16. GI-Fachtagung
Informatik und Schule
20.–23. September 2015

P-250 Martin Engstler, Masud Fazal-Baqaie,
Eckhart Hanser, Martin Mikusz,
Alexander Volland (Hrsg.)
Projektmanagement und
Vorgehensmodelle 2015
Hybride Projektstrukturen erfolgreich
umsetzen
Gemeinsame Tagung der Fachgruppen
Projektmanagement (WI-PM) und
Vorgehensmodelle (WI-VM) im
Fachgebiet Wirtschaftsinformatik
der Gesellschaft für Informatik e.V.,
Elmshorn 2015

P-251 Detlef Hühnlein, Heiko Roßnagel,
Raik Kuhlisch, Jan Ziesing (Eds.)
Open Identity Summit 2015
10.–11. November 2015
Berlin, Germany

P-252 Jens Knoop, Uwe Zdun (Hrsg.)
Software Engineering 2016
Fachtagung des GI-Fachbereichs
Softwaretechnik
23.–26. Februar 2016, Wien

P-253 A. Ruckelshausen, A. Meyer-Aurich,
T. Rath, G. Recke, B. Theuvsen (Hrsg.)
Informatik in der Land-, Forst- und
Ernährungswirtschaft
Fokus: Intelligente Systeme – Stand der
Technik und neue Möglichkeiten
Referate der 36. GIL-Jahrestagung
22.-23. Februar 2016, Osnabrück

P-254 Andreas Oberweis, Ralf Reussner (Hrsg.)
Modellierung 2016
2.–4. März 2016, Karlsruhe

P-255 Stefanie Betz, Ulrich Reimer (Hrsg.)
Modellierung 2016 Workshopband
2.–4. März 2016, Karlsruhe

P-256 Michael Meier, Delphine Reinhardt,
Steffen Wendzel (Hrsg.)
Sicherheit 2016
Sicherheit, Schutz und Zuverlässigkeit
Beiträge der 8. Jahrestagung des
Fachbereichs Sicherheit der
Gesellschaft für Informatik e.V. (GI)
5.–7. April 2016, Bonn

P-257 Paul Müller, Bernhard Neumair, Helmut
Reiser, Gabi Dreo Rodosek (Hrsg.)
9. DFN-Forum
Kommunikationstechnologien
31. Mai – 01. Juni 2016, Rostock

P-258 Dieter Hertweck, Christian Decker (Eds.)
Digital Enterprise Computing (DEC 2016)
14.–15. Juni 2016, Böblingen

P-259 Heinrich C. Mayr, Martin Pinzger (Hrsg.)
INFORMATIK 2016
26.–30. September 2016, Klagenfurt

P-260 Arslan Brömme, Christoph Busch,
Christian Rathgeb, Andreas Uhl (Eds.)
BIOSIG 2016
Proceedings of the 15th International
Conference of the Biometrics Special
Interest Group
21.–23. September 2016, Darmstadt

P-261 Detlef Rätz, Michael Breidung, Dagmar
Lück-Schneider, Siegfried Kaiser, Erich
Schweighofer (Hrsg.)
Digitale Transformation: Methoden,
Kompetenzen und Technologien für die
Verwaltung
Gemeinsame Fachtagung
Verwaltungsinformatik (FTVI) und
Fachtagung Rechtsinformatik (FTRI) 2016
22.–23. September 2016, Dresden

P-262 Ulrike Lucke, Andreas Schwill,
Raphael Zender (Hrsg.)
DeLFI 2016 – Die 14. E-Learning
Fachtagung Informatik
der Gesellschaft für Informatik e.V. (GI)
11.–14. September 2016, Potsdam

P-263 Martin Engstler, Masud Fazal-Baqaie,
Eckhart Hanser, Oliver Linssen, Martin
Mikusz, Alexander Volland (Hrsg.)
Projektmanagement und
Vorgehensmodelle 2016
Arbeiten in hybriden Projekten: Das
Sowohl-als-auch von Stabilität und
Dynamik
Gemeinsame Tagung der Fachgruppen
Projektmanagement (WI-PM) und
Vorgehensmodelle (WI-VM) im
Fachgebiet Wirtschaftsinformatik
der Gesellschaft für Informatik e.V.,
Paderborn 2016

P-264 Detlef Hühnlein, Heiko Roßnagel,
Christian H. Schunck, Maurizio Talamo
(Eds.)
Open Identity Summit 2016
der Gesellschaft für Informatik e.V. (GI)
13.–14. October 2016, Rome, Italy

P-265 Bernhard Mitschang, Daniela
Nicklas,Frank Leymann, Harald
Schöning, Melanie Herschel, Jens
Teubner, Theo Härder, Oliver Kopp,
Matthias Wieland (Hrsg.)
Datenbanksysteme für Business,
Technologie und Web (BTW 2017)
6.–10. März 2017, Stuttgart

P-266 Bernhard Mitschang, Norbert Ritter,
Holger Schwarz, Meike Klettke, Andreas
Thor, Oliver Kopp, Matthias Wieland
(Hrsg.)
Datenbanksysteme für Business,
Technologie und Web (BTW 2017)
Workshopband
6.–7. März 2017, Stuttgart

P-267 Jan Jürjens, Kurt Schneider (Hrsg.)
Software Engineering 2017
21.–24. Februar 2017, Hannover

P-268 A. Ruckelshausen, A. Meyer-Aurich,
W. Lentz, B. Theuvsen (Hrsg.)
Informatik in der Land-, Forst- und
Ernährungswirtschaft
Fokus: Digitale Transformation –
Wege in eine zukunftsfähige
Landwirtschaft
Referate der 37. GIL-Jahrestagung
06.–07. März 2017, Dresden

P-269 Peter Dencker, Herbert Klenk, Hubert
Keller, Erhard Plödereder (Hrsg.)
Automotive – Safety & Security 2017
30.–31. Mai 2017, Stuttgart

P-270 Arslan Brömme, Christoph Busch,
Antitza Dantcheva, Christian Rathgeb,
Andreas Uhl (Eds.)
BIOSIG 2017
20.–22. September 2017, Darmstadt

P-271 Paul Müller, Bernhard Neumair, Helmut
Reiser, Gabi Dreo Rodosek (Hrsg.)
10. DFN-Forum Kommunikations-
technologien
30. – 31. Mai 2017, Berlin

P-272 Alexander Rossmann, Alfred
Zimmermann (eds.)
Digital Enterprise Computing
(DEC 2017)
11.–12. Juli 2017, Böblingen

P-273 Christoph Igel, Carsten Ullrich,
Martin Wessner (Hrsg.)
BILDUNGSRÄUME
DeLFI 2017
Die 15. e-Learning Fachtagung Informatik
der Gesellschaft für Informatik e.V. (GI)
5. bis 8. September 2017, Chemnitz

P-274 Ira Diethelm (Hrsg.)
Informatische Bildung zum Verstehen
und Gestalten der digitalen Welt
13.–15. September 2017, Oldenburg

P-275 Maximilian Eibl, Martin Gaedke (Hrsg.)
INFORMATIK 2017
25.–29. September 2017, Chemnitz

P276 Alexander Volland, Martin Engstler,
Masud Fazal-Baqaie, Eckhart Hanser,
Oliver Linssen, Martin Mikusz (Hrsg.)
Projektmanagement und
Vorgehensmodelle 2017
Die Spannung zwischen dem Prozess
und den Menschen im Projekt
Gemeinsame Tagung der Fachgruppen
Projektmanagement und
Vorgehensmodelle im Fachgebiet
Wirtschaftsinformatik der
Gesellschaft für Informatik e.V.
in Kooperation mit der Fachgruppe
IT-Projektmanagement der GPM e.V.,
Darmstadt 2017

P-277 Lothar Fritsch, Heiko Roßnagel,
Detlef Hühnlein (Hrsg.)
Open Identity Summit 2017
5.–6. October 2017, Karlstad, Sweden

P-278 Arno Ruckelshausen,
Andreas Meyer-Aurich, Karsten Borchard,
Constanze Hofacker, Jens-Peter Loy,
Rolf Schwerdtfeger,
Hans-Hennig Sundermeier, Helga Floto,
Brigitte Theuvsen (Hrsg.)
Informatik in der Land-, Forst- und
Ernährungswirtschaft
Referate der 38. GIL-Jahrestagung
26.–27. Februar 2018, Kiel

P-279 Matthias Tichy, Eric Bodden,
Marco Kuhrmann, Stefan Wagner,
Jan-Philipp Steghöfer (Hrsg.)
Software Engineering und Software
Management 2018
5.–9. März 2018, Ulm

P-280 Ina Schaefer, Dimitris Karagiannis,
Andreas Vogelsang, Daniel Méndez,
Christoph Seidl (Hrsg.)
Modellierung 2018
21.–23. Februar 2018, Braunschweig

P-281 Hanno Langweg, Michael Meier, Bernhard
C. Witt, Delphine Reinhardt (Hrsg.)
Sicherheit 2018
Sicherheit, Schutz und Zuverlässigkeit
25.–27. April 2018, Konstanz

P-282 Arslan Brömme, Christoph Busch,
Antitza Dantcheva, Christian Rathgeb,
Andreas Uhl (Eds.)
BIOSIG 2018
Proceedings of the 17th International
Conference of the Biometrics Special
Interest Group
26.–28. September 2018
Darmstadt, Germany

P-283 Paul Müller, Bernhard Neumair, Helmut
Reiser, Gabi Dreo Rodosek (Hrsg.)
11. DFN-Forum Kommunikations-
technologien
27.–28. Juni 2018, Günzburg

P-284 Detlef Krömker, Ulrik Schroeder (Hrsg.)
DeLFI 2018 – Die 16. E-Learning
Fachtagung Informatik
10.–12. September 2018, Frankfurt a. M.

P-285 Christian Czarnecki, Carsten Brockmann,
Eldar Sultanow, Agnes Koschmider,
Annika Selzer (Hrsg.)
Workshops der INFORMATIK 2018 -
Architekturen, Prozesse, Sicherheit und
Nachhaltigkeit
26.–27. September 2018, Berlin

P-286 Martin Mikusz, Alexander Volland, Martin
Engstler, Masud Fazal-Baqaie, Eckhart
Hanser, Oliver Linssen (Hrsg.)
Projektmanagement und
Vorgehensmodelle 2018
Der Einfluss der Digitalisierung auf
Projektmanagementmethoden und
Entwicklungsprozesse
Düsseldorf 2018

P-287 A. Meyer-Aurich, M. Gandorfer, N. Barta,
A. Gronauer, J. Kantelhardt, H. Floto (Hrsg.)
Informatik in der Land-, Forst- und Ernährungswirtschaft
Fokus: Digitalisierung für landwirtschaftliche Betriebe in kleinstrukturierten Regionen – ein Widerspruch in sich?
Referate der 39. GIL-Jahrestagung
18.–19. Februar 2019, Wien

P-288 Arno Pasternak (Hrsg.)
Informatik für alle
18. GI-Fachtagung
Informatik und Schule
16.-18. September 2019 in Dortmund

P-289 Torsten Grust, Felix Naumann, Alexander Böhm, Wolfgang Lehner, Jens Teubner, Meike Klettke, Theo Härder, Erhard Rahm, Andreas Heuer, Holger Meyer (Hrsg.)
Datenbanksysteme für Business, Technologie und Web (BTW 2019)
4.–8. März 2019 in Rostock

P-290 Holger Meyer, Norbert Ritter, Andreas Thor, Daniela Nicklas, Andreas Heuer, Meike Klettke (Hrsg.)
Datenbanksysteme für Business, Technologie und Web (BTW 2019)
Workshopband
4.–8. März 2019 in Rostock

P-291 Michael Räckers, Sebastian Halsbenning, Detlef Rätz, David Richter,
Erich Schweighofer (Hrsg.)
Digitalisierung von Staat und Verwaltung
Gemeinsame Fachtagung Verwaltungsinformatik (FTVI) und Fachtagung Rechtsinformatik (FTRI) 2019
6.–7. März 2019 in Münster

P-292 Steffen Becker, Ivan Bogicevic,Georg Herzwurm, Stefan Wagner (Hrsg.)
Software Engineering and Software Management 2019
18.–22. Februar 2019 in Stuttgart

P-293 Heiko Roßnagel, Sven Wagner, Detlef Hühnlein (Hrsg.)
Open Identity Summit 2019
28.–29. März 2019
Garmisch-Partenkirchen

P-294 Klaus David, Kurt Geihs, Martin Lange, Gerd Stumme (Hrsg.)
INFORMATIK 2019
50 Jahre Gesellschaft für Informatik – Informatik für Gesellschaft
23.–26. September 2019 in Kassel

P-295 Claude Draude, Martin Lange, Bernhard Sick (Hrsg.)
INFORMATIK 2019
50 Jahre Gesellschaft für Informatik – Informatik für Gesellschaft
Workshop-Beiträge
23.–26. September 2019 in Kassel

P-296 Arslan Brömme, Christoph Busch,
Antitza Dantcheva, Christian Rathgeb,
Andreas Uhl (Eds.)
BIOSIG 2019
Proceedings of the 18th International Conference of the Biometrics Special Interest Group
18.–20. September 2019
Darmstadt, Germany

P-297 Niels Pinkwart, Johannes Konert (Hrsg.)
DELFI 2019 –Die 17. Fachtagung Bildungstechnologien
16.–19. September 2019 in Berlin

P-298 Oliver Linssen, Martin Mikusz,
Alexander Volland, Enes Yigitbas,
Martin Engstler, Masud Fazal-Baqaie,
Marco Kuhrmann (Hrsg.)
Projektmanagement und Vorgehensmodelle 2019 –Neue Vorgehensmodelle in Projekten – Führung, Kulturen und Infrastrukturen im Wandel
1Gemeinsame Tagung der Fachgruppen Projektmanagement (WI-PM), Vorgehensmodelle (WI-VM) und Software Produktmanagement (WI-ProdM) im Fachgebiet Wirtschaftsinformatik der Gesellschaft für Informatik e.V.
in Kooperation mit der Fachgruppe IT-Projektmanagement der GPM e.V.,
Lörrach 2019

P-299 M. Gandorfer, A. Meyer-Aurich, H. Bernhardt, F. X. Maidl, G. Fröhlich, H. Floto (Hrsg.)
Informatik in der Land-, Forst- und Ernährungswirtschaft
Fokus: Digitalisierung für Mensch, Umwelt und Tier
Referate der 40. GIL-Jahrestagung
17.–18. Februar 2020,
Campus Weihenstephan

P-300 Michael Felderer, Wilhelm Hasselbring, Rick Rabiser, Reiner Jung (Hrsg.)
Software Engineering 2020
24.–28. Februar 2020
Innsbruck, Austria

P-301 Delphine Reinhardt, Hanno Langweg, Bernhard C. Witt, Mathias Fischer (Hrsg.)
Sicherheit 2020
Sicherheit, Schutz und Zuverlässigkeit
17.–20. März 2020, Göttingen

P-302 Dominik Bork, Dimitris Karagiannis, Heinrich C. Mayr (Hrsg.)
Modellierung 2020
19.–21. Februar 2020, Wien

P-303 Peter Heisig, Ronald Orth, Jakob Michael Schönborn, Stefan Thalmann (Hrsg.)
Wissensmanagement in digitalen Arbeitswelten: Aktuelle Ansätze und Perspektiven
18.–20.03.2019, Potsdam

P-304 Heinrich C. Mayr, Stefanie Rinderle-Ma, Stefan Strecker (Hrsg.)
40 Years EMISA
Digital Ecosystems of the Future: Methodology, Techniques and Applications
May 15.–17. 2019
Tutzing am Starnberger See

P-305 Heiko Roßnagel, Christian H. Schunck, Sebastian Mödersheim, Detlef Hühnlein (Hrsg.)
Open Identity Summit 2020
26.–27. May 2020, Copenhagen

P-306 Arslan Brömme, Christoph Busch, Antitza Dantcheva, Kiran Raja, Christian Rathgeb, Andreas Uhl (Eds.)
BIOSIG 2020
Proceedings of the 19th International Conference of the Biometrics Special Interest Group
16.–18. September 2020
International Digital Conference

P-307 Ralf H. Reussner, Anne Koziolek, Robert Heinrich (Hrsg.)
INFORMATIK 2020
Back to the Future
28. September – 2. Oktober 2020, Karlsruhe

P-308 Raphael Zender, Dirk Ifenthaler, Thiemo Leonhardt, Clara Schumacher (Hrsg.)
DELFI 2020 –
Die 18. Fachtagung Bildungstechnologien der Gesellschaft für Informatik e.V.
14.–18. September 2020
Online

P-309 A. Meyer-Aurich, M. Gandorfer, C. Hoffmann, C. Weltzien, S. Bellingrath-Kimura, H. Floto (Hrsg.)
Informatik in der Land-, Forst- und Ernährungswirtschaft
Referate der 41. GIL-Jahrestagung
08.–09. März 2021, Leibniz-Institut für Agrartechnik und Bioökonomie e.V., Potsdam

P-310 Anne Koziolek, Ina Schaefer, Christoph Seidl (Hrsg.)
Software Engineering 2021
22.–26. Februar 2021,
Braunschweig/Virtuell

P-311 Kai-Uwe Sattler, Melanie Herschel, Wolfgang Lehner (Hrsg.)
Datenbanksysteme für Business, Technologie und Web (BTW 2021)
Tagungsband
13.–17. September 2021,
Dresden

P-312 Heiko Roßnagel, Christian H. Schunck, Sebastian Mödersheim (Hrsg.)
Open Identity Summit 2021
01.–02. Juni 2021, Copenhagen

P-313 Ludger Humbert (Hrsg.)
Informatik – Bildung von Lehrkräften in allen Phasen
19. GI-Fachtagung Informatik und Schule
8.–10. September 2021 Wuppertal

P-315 Arslan Brömme, Christoph Busch,
Naser Damer, Antitza Dantcheva,
Marta Gomez-Barrero, Kiran Raja,
Christian Rathgeb, Ana F. Sequeira,
Andreas Uhl (Eds.)
BIOSIG 2021
Proceedings of the 20th International
Conference of the Biometrics
Special Interest Group
15.–17. September 2021
International Digital Conference

P-316 Andrea Kienle, Andreas Harrer,
Jörg M. Haake, Andreas Lingnau
(Hrsg.)
DELFI 2021
Die 19. Fachtagung Bildungstechnologien
der Gesellschaft für Informatik e.V.
13.–15. September 2021
Online8.–10. September 2021

The titles can be purchased at:

Köllen Druck + Verlag GmbH
Ernst-Robert-Curtius-Str. 14 · D-53117 Bonn
Fax: +49 (0)228/9898222
E-Mail: druckverlag@koellen.de

Gesellschaft für Informatik e.V. (GI)

publishes this series in order to make available to a broad public recent findings in informatics (i.e. computer science and information systems), to document conferences that are organized in co-operation with GI and to publish the annual GI Award dissertation.

Broken down into
- seminars
- proceedings
- dissertations
- thematics

current topics are dealt with from the vantage point of research and development, teaching and further training in theory and practice. The Editorial Committee uses an intensive review process in order to ensure high quality contributions.

The volumes are published in German or English.

Information: http://www.gi.de/service/publikationen/lni/

ISSN 1617-5468
ISBN 978-3-88579-709-8

The proceedings of the BIOSIG 2021 include scientific contributions of the annual international conference of the Biometrics Special Interest Group (BIOSIG) of the Gesellschaft für Informatik (GI). Due to the pandemic situation the conference was held as a digital conference, 15.-17. September 2021. The advances of biometrics research and new developments in the core biometric application field of security have been presented and discussed by international biometrics and security professionals.