# GI-Edition

## Lecture Notes in Informatics

Naser Damer, Marta Gomez-Barrero,
Kiran Raja, Christian Rathgeb, Ana F. Sequeira,
Massimiliano Todisco, Andreas Uhl (Eds.)

# BIOSIG 2023

## Proceedings of the 22nd International Conference of the Biometrics Special Interest Group

20.–22. September 2023,
Darmstadt, Germany

Proceedings

339

Naser Damer, Marta Gomez-Barrero, Kiran Raja, Christian Rathgeb, Ana F. Sequeira, Massimiliano Todisco, Andreas Uhl (Eds.): BIOSIG 2023 – 22nd International Conference of the Biometrics Special Interest Group

GESELLSCHAFT
FÜR INFORMATIK

Naser Damer, Marta Gomez-Barrero,
Kiran Raja, Christian Rathgeb, Ana F. Sequeira,
Massimiliano Todisco, Andreas Uhl (Eds.)

# BIOSIG 2023

## Proceedings of the 22nd International Conference of the Biometrics Special Interest Group

**20.-22. September 2023**
**Darmstadt, Germany**

Gesellschaft für Informatik e.V. (GI)

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

Volume P-339

ISBN 978-3-88579-733-3
ISSN 1617-5468

## Volume Editors

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

**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
Teknologivegen 22, 2816 Gjøvik
*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*

**Massimiliano Todisco**
EURECOM, Sophia Antipolis,
450 route des Chappes
06410 Biot Sophia Antipolis
France
*todisco@eurecom.fr*

**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

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 bio-metrics and security professionals from industry, science, representatives of the national governmental bodies and European institutions who are working in these areas.

The BIOSIG 2023 international 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 2023 is once again technically co-sponsored by the Institute of Electrical and Electronics Engineers (IEEE) and is enriched with satellite workshops by the TeleTrust Biometric Working Group and the European Association for Biometrics. BIOSIG 2023 is held again in Darmstadt, Germany.

The international program committee accepted full scientific papers (18 out of 61 submissions) strongly according to the LNI guidelines (acceptance rate ~30%) within a scientific double-blinded review process of at minimum five reviews per paper. All pa-pers were formally restricted for the digital proceedings up to 12 pages for regular re-search contributions including an oral presentation and up to 10 pages for further con-ference 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 2023 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, 20th September 2023

Naser Damer (Fraunhofer IGD), Marta Gomez-Barrero (Hochschule Ansbach), Kiran Raja (NTNU), Christian Rathgeb (Hochschule Darmstadt), Ana F. Sequeira (INESC TEC), Massimiliano Todisco (EURECOM), Andreas Uhl (University of Salzburg)

## Chairs

General Chair
Marta Gomez-Barrero, Hochschule Ansbach, Germany
Program Chairs
Kiran Raja, Norwegian University of Science and Technology NTNU, Norway
Christian Rathgeb, Hochschule Darmstadt, Germany
Massimiliano Todisco, EURECOM, France
Andreas Uhl, University of Salzburg, Austria

Publication Chair
Naser Damer, Fraunhofer IGD, 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
Siri Lorenz, CAST e.V., Darmstadt, Germany

## Programm Committee

Abe Narishige
Ajay Kumar
Ana F. Sequeira
Andrea Atzori
Andreas Uhl
Andreas Wolf
Angel M. Gomez
Annalisa Franco
Antonio M. Peinado
Biying Fu
Chiara Galdi
Christian Rathgeb
Cunjian Chen
Davide Maltoni
Emanuele Maiorana
Fernando Alonso-Fernandez
Gian Luca Marcialis
Günter Schumacher
Heiko Roßnagel
Ines Domingues
Jaime Cardoso
James Wayman
Jan Niklas Kolf
Jannis Priesnitz
Jennifer Williams
Juan Tapia
Julien Bringer
Kiran Raja
Lazaro Janier Gonzalez-Soler
Luuk Spreeuwers
Manuel Günther
Marcel Grimmer
Maria De Marsico
Marta Gomez-Barrero
Massimiliano Todisco
Massimiliano Todisco
Matteo Ferrara
Mauro Falcone
Meiling Fang
Naser Damer
Nicholas Evans
Norbert Jung
Olaf Henniger
Patrick Bours
Paulo Lobato Correia
Pawel Drozdowski
Peter Peer
Philipp Terhörst
Pia Bauspieß
Rohan Kumar Das
Ruben Tolosana
Xin Wang

# 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

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 2023 – Biometrics Special Interest Group

"**2023 International Conference of the Biometrics Special Interest Group**"
20th -22nd September 2023

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 Identifi-cation (UID) in India require high accuracy and 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 2023 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 2023 – Further Conference Contributions……..……**

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# Towards Contactless Fingerprint Presentation Attack Detection using Algorithms from the Contact-based Domain

Jannis Priesnitz[1], Roberto Casula[2], Christian Rathgeb[1], Gian Luca Marcialis[2], Christoph Busch[1]

**Abstract:**

In this work, we investigate whether contact-based fingerprint Presentation Attack Detection (PAD) methods can generalize to the contactless domain. To this end, we selected a state-of-the-art patch-based fingerprint PAD algorithm which achieved high detection performance in the contact-based domain and adapted it for contactless fingerprints. We train and test the method using three contactless fingerprint databases and evaluate its generalization capabilities using Leave-One-Out (LOO) protocols. Further, we acquired a new PAD database and use it in a cross-database evaluation. The adopted method shows low error rates in most scenarios and can generalize to unseen contactless presentation attacks.

**Keywords:** Contactless fingerprint recognition, security, presentation attack detection, generalizability

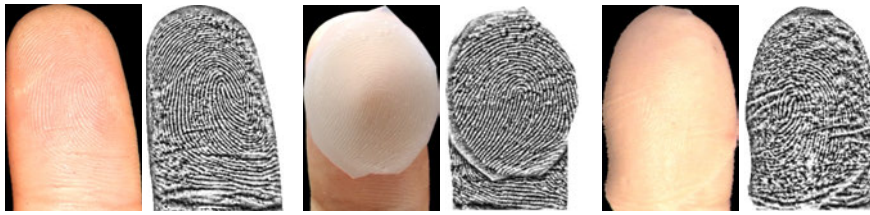
## 1 Introduction

Presentation Attack Detection (PAD) is of utmost importance to ensure the operational security of biometric systems. Like for many other biometric characteristics, various PAD algorithms are proposed for fingerprint recognition systems. PAD methods are designed to reliably detect artificial replicas, *i.e.* Presentation Attack Instruments (PAIs), which can be made of various materials. Most common PAI species are made of gelatin, silicone, different glues, playdoh or latex [SB14]. State-of-the-art methods mainly rely on machine learning algorithms like Convolutional Neural Networks (CNNs). CNNs are known for their good generalization capabilities to data which is not included in the training set.

Complementary to contact-based schemes, contactless fingerprint recognition has established itself as a more comfortable alternative. Contactless fingerprint technologies enable the recognition of individuals without any contact between a capture device surface and a fingertip [Pr21b, YZH21]. Contactless capture devices typically have a higher user acceptance, especially when multiple users interact with one single device.

Like most biometric characteristics, contactless fingerprint recognition is vulnerable to Presentation Attacks (PAs). Here, PAI species similar to contact-based setups can be used

[1] da/sec - Biometrics and Internet Security Research Group, Hochschule Darmstadt, Schöfferstraße 9, 64295 Darmstadt, Germany, firstname.lastname@h-da.de
[2] PRA Lab, University of Cagliari, Via Marengo, 3, 09123 Cagliari, Italy, firstname.lastname@unica.it



(a) bona fide (b) PA (wood glue) COLFISPOOF (c) PA (overlay) HDA

Fig. 1: Fingerprint images with pre-processed samples captured using a contactless capture device. (a) bona fide, (b) low-quality PA, (c) high-quality PA.

to launch PAs [Ko23]. Additionally, printout attacks have a high potential since the color of the PAI can be precisely adjusted to that of bona fide samples. Figure 1 shows a bona fide sample and two types of PAs captured using a contactless capturing device. Since similar PAIs can be exploited to attack contact-based as well as contactless capturing devices, it is assumed that the same concepts can reliably detect PAs.

Contact-based fingerprint PAD is a well-studied research area. Comprehensive overviews discuss relevant aspects of the research area [Ga23, Mi23, Ka21a]. PAD methods are basically categorized into two types: hardware-based methods employing specific sensors for obtaining supplementary information and software-based methods which utilize features extracted from the captured fingerprint images. For contact-based fingerprint recognition, software-based methods are more popular because they are generally applicable and more flexible to deploy. Various methods have been proposed to detect contact-based PAs. Feature-based methods use either holistic features or local features [Gr15]. Holistic features can be, *e.g.* global texture properties [AS06]. Local features, *e.g.* texture patches, can be processed with hand-crafted feature extraction methods combined with basic machine learning-based classifiers like Support Vector Machines (SVMs) or CNNs [Gr15, Go21].

In comparison to hand-crafted feature extraction approaches, CNNs generally require less pre-processing and generalize better to unseen data [Ya17, Mi23, Af20]. For fingerprint PAD, general purpose CNNs designed for object detection can be adapted to the binary classification task [Ng18]. The data preparation for these CNNs is characterized by two distinct approaches: one which considers the whole sample for PAD [Pa19, Ge20] and one which considers patches extracted during the pre-processing [CJ19, Hu18, CCJ18]. These patches can be randomly selected [PB17] or arranged around minutiae points [CCJ18].

Latest research suggests vision transformers for fingerprint PAD [Ra23]. This work compares various proposed methods with Data-Efficient Image Transformers (DeiT) and reports a significant improvement of detection performance in their experimental setup.

Contactless PAD methods are less comprehensively studied. Published works suggest analyzing the reflection properties of different PAI species [SBB13], hand-crafted feature extractors with a Support Vector Machine (SVM) [Ta16, Wa18] and CNN-based methods [Pu23]. Further, deep fusion strategies to combine deep representations obtained from different color spaces are studied [MV22].

The majority of contactless fingerprint capturing devices capture color images, which are further processed into contrast-enhanced grayscale images, *c.f.* Figure 1. For this reason, PAD methods could analyze color properties. However, it is assumed that color-based PAD schemes are not generalizing well to unseen data. *E.g.* a PAD scheme which is trained on PAIs which have a different color compared to the bona fide samples might learn the color difference and, hence, fail on unseen PAIs which exhibit a color more similar to skin color. Also, color-based PAD schemes might be biased to the skin color over-represented in the training set. A study in skin color-based bias for contactless fingerprint recognition supports this hypothesis [BND22]. Furthermore, obviously PAIs comprising a realistic skin color can be assembled as shown in Figure 1 (c), which might easily fool color-based PAD methods. For this reason, a PAD method which operates on pre-processed gray-scale image is a vital alternative.

As can be seen from the related work, patch-based fingerprint PAD using CNNs represent the state-of-the-art in this research area. Compared to others, this method also incorporates several advantages for contactless fingerprint PAD. Compared to PAD schemes which analyze the full image, no cropping or resizing is needed to present the sample to the PAD algorithm. In contrast to methods which extract random patches from the image, only the most relevant part around minutiae are extracted.

To the best of the author's knowledge, no study has yet been conducted which benchmarks PAD algorithms from the contact-based domain on contactless data. In this work, we evaluate if a contact-based fingerprint PAD algorithm can detect pre-processed contactless fingerprint PAs. For this reason, we modify the SpoofBuster method proposed by Chugh *et al*. [CCJ18] to the special requirements of contactless fingerprint PAD and evaluate its suitability to detect contactless fingerprints. The SpoofBuster PAD algorithm represents a state-of-the-art CNN-based contact-based fingerprint PAD method that extracts and classifies local features extracted from texture patches. We consider COLFISPOOF [Ko23] as a publicly available contactless fingerprint PA database together with three bona fide databases to re-train the adapted version of the SpoofBuster algorithm. To do so, we train the method using a baseline and four LOO protocols proposed in [Ko23].

Further, we acquired a new PAD database, the UniCa-HDA PAD database, to test our proposal on more challenging data. Here, we produce thin overlays which are precisely adjusted to human skin color. To benchmark our proposal in a real-world scenario, we conduct a training on COLFISPOOF combined with the bona fide databases and test on the UniCa-HDA databases.

Our results show high detection accuracy. In our first experiment, the APCER at a BPCER of 1.00% range between 0.00% and 0.33%. The algorithm is found to generalize well to PAs unseen in during the training (6.75% APCER at a BPCER of 5.00%), which is showcased in the newly acquired PA database.

The rest of the paper is structured as follows: Section 2 presents the considered PAD methods. Section 3 describes the experimental setup. The results are discussed in Section 4. Finally, Section 5 concludes.

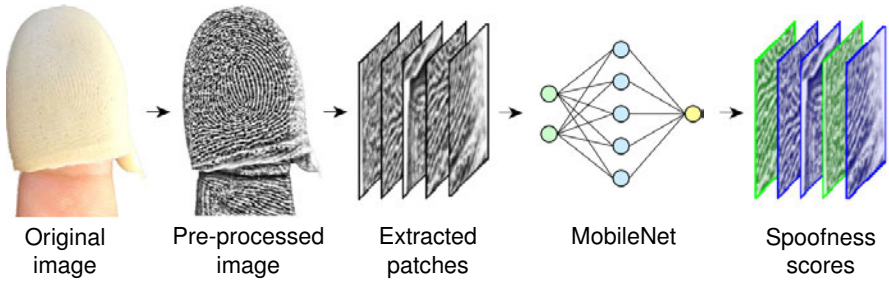



Fig. 2: Overview on the proposed workflow.

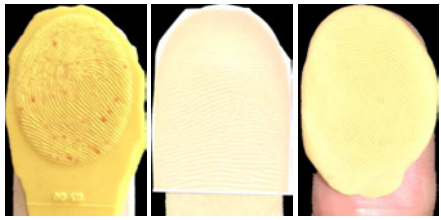
## 2 Patch-based Contactless Fingerprint PAD

Patch-based fingerprint PAD offers several advantages for contact-based fingerprints. Extracting patches around minutiae ensure that only patches that contain a fingerprint characteristic are processed. This is also favorable for contactless samples in which regions of low contrast could confuse the PAD method, *c.f.* Figure 1 (c). Further, minutiae provide particular relevant area for PAD. *E.g.* ridge endings can be caused by borders of an overlay, or bifurcations / islands could stem from air bubbles or dirt particles in silicone materials. Figures 3a and 3c illustrate these properties for contactless PAIs. That is, this approach is expected to also be beneficial in the contactless domain.

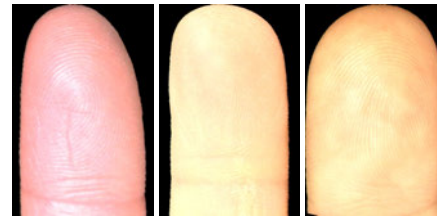
Since, the extracted minutiae are not used for recognition rather than extracting regions of interest from the fingerprint sample, any feature extractor independent of the recognition accuracy is generally applicable. However, minutiae extractors which provide a minutia quality assessment provide a proper option for reducing the number of extracted patches to high-quality minutia. Hence, the MINDCT method [Wa07] which includes a minutia quality assessment is considered in this work.

The number of extracted minutiae candidates is typically around 50, but varies due to many factors like fingerprint quality, finger ID, capturing device fidelity or feature extractor fidelity. Chugh et al. [CJ19] propose two strategies to reduce the number of corresponding patches. First, they cluster minutiae and extract patches around minutiae centers and second, they use a feature extractor which provides a quality score for each extracted minutiae. Here, the authors define a static threshold to consider minutiae of high quality only. We introduce a further patch reduction strategy for segmented contactless fingerprints by analyzing the proportion of connected background pixels in an extracted patch. Minutiae located at the border region of a fingerprint contain many background pixels and, thus, contain a lower amount of information. For this reason, we compute the proportion of background pixels and discard all patches which are above a pre-defined threshold.

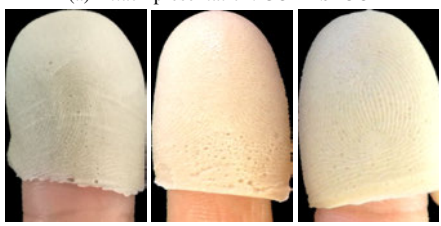
The considered patch size is a crucial part of the algorithm. Small patches might contain too little spatial information to achieve an accurate prediction, whereas an extraction of too big patches lead to a large overlapping areas and, hence, redundant information. Furthermore, the CNN might have challenges to disseminating the relevant information from



(a) Attack presentation: COLFISPOOF



(b) Bona fide: HDA



(c) Attack presentation: UniCa-HDA

(d) Bona fide: ISPFDv1

Fig. 3: Example finger images of the used datasets. It should be noted that only samples of the ISPFDv1 database are presented. Samples of the ISPFDv2 database look similar.

big patches. Another considerable option is to upscale the extract patches to emphasize fine details. The original method proposes a so-called patch alignment. In this process, the patches are extracted according to the rotation angle of the minutia and then rotated into an upright position. This ensures that all patches contain a horizontally aligned ridge pattern, which might be easier to process.

Any CNN-based method with a modified last fully connected layers to support binary classifications is generally suited to distinguish between bona fide samples and PAs. Here, sensitivity to fine-grained texture patterns is beneficial for an accurate classification. The considered MobileNetv1 architecture outputs a Presentation Attack (PA) score in the range [0.0, 1.0] for every patch. All PA scores are averaged to the final PAD result of the tested sample.

In summary, the following main adaptations are applied to the original SpoofBuster method by Chugh et al. [CJ19]:

- **Patch angle and alignment:** We do not implement a patch rotation based on the minutiae angle, rather than extracting aligned patches. In our experiments, no improvement in terms of detection accuracy due to patch alignment was observed.
- **Patch number reduction:** To reduce the number of patches, we consider a combination of minutiae quality and background thresholding. The minutiae quality threshold excludes minutiae with a quality score below 0.25 (in a range [0, 1]) whereas the background threshold excludes patches with more than 10% white background pixels.

Tab. 1: Number of bona fide and attack presentations for each database. It should be noted that the UniCa-HDA database is only used for testing.

| Database | BF Samples | PA Samples |
|---|---|---|
| COLFISPOOF | – | 7,200 |
| HDA | 1,069 | – |
| ISPFDv1 | 4,029 | – |
| ISPFDv2 | 16,175 | – |
| UniCa-HDA | 2,040 | 1,512 |

- **Patch size:** We use a patch size of $112\times112$ pixels instead of $96\times96$ pixels as in the original approach from Chugh [CJ19]. This has two reasons: firstly, the acquired contactless fingerprints are in our setup of larger size compared to contact-based ones, so that the patch-size needs to be increased. Secondly, due to the lack of rotation, we avoid protruding edges and, for this reason, it is possible to increase the patch size.

All further settings such as the usage of MINDCT and the MobileNetv1 training parameters remain the same as in the original SpoofBuster algorithm.

In summary, patch-based PAD schemes offer various options for optimizing and fine-tuning with good potential for a general purpose fingerprint PAD methods. However, finding proper settings which are suited best for the operational scenario can be a challenge.

## 3 Experimental Setup

We test the algorithm using the publicly available COLFISPOOF database and three bona fide databases, both version of the IIITD SmartPhone Fingerphoto Database (ISPFD) [Sa15, Ma20a] and the HDA data set [Pr22]. In addition, we conduct a test on the newly captured UniCa-HDA database comprising bona fide samples and PAs. Here, the same capturing setup like in the HDA database was used, but a different environmental scenario was chosen. Along with the newly acquired bona fide samples, PAs were prepared and captured. A thin overlay is produced using a Body Double material, which then is turned inside out after the material has solidified. To make the PAI as realistic as possible, additional color was added to the PA material. This process makes the PAI appear as similar as possible to the subject's real skin color, *c.f.* Fig. 3 (c). The overlays are then captured in the same setup as the bona fide fingerprints. Example images of every database are presented in Figure 3 and an overview of the database properties is given in Table 1.

All samples are pre-processed in the same way, using the contactless fingerprint pre-processing proposed in [Pr22]. The conducted pre-processing composed of a deep-learning-based fingertip segmentation [Pr21a], gray-scale conversions, rotation, normalization and contrast enhancement aligns with the state-of-the-art. It has proven to be suitable in a general in several recognition workflows, *e.g.* [Ka21b, Ma20b].

Tab. 2: APCERs for a fixed BPCER of 1.00% and D-EERs for using COLFISPOOF together with different bona fide databases.

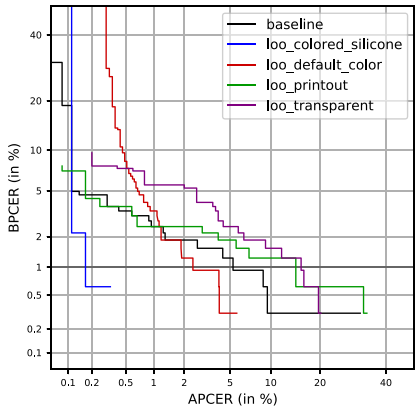
| | **HDA** | | **ISPFDv1** | | **ISPFDv2** | |
|---|---|---|---|---|---|---|
| | APCER | D-EER | APCER | D-EER | APCER | D-EER |
| Baseline | 5.28 | 1.86 | 0.03 | 0.64 | 0.83 | 0.92 |
| LOO colored silicone | 0.17 | 0.32 | 0.08 | 0.68 | 0.00 | 0.39 |
| LOO default color | 2.41 | 1.85 | 0.00 | 0.54 | 0.00 | 0.43 |
| LOO printout | 14.50 | 2.49 | 0.33 | 0.68 | 3.92 | 2.42 |
| LOO transparent | 15.6 | 3.75 | 0.00 | 0.64 | 0.40 | 0.60 |
| Average LOO | 8.17 | 2.10 | 0.10 | 0.64 | 1.08 | 0.86 |
| Std dev LOO | ± 8.01 | ± 1.43 | ± 0.16 | ± 0.07 | ± 1.90 | ± 1.06 |

For our evaluation, we use six protocols. The baseline scenario randomly splits the samples of the bona fide and PAs into training (30%), validation (20%), and test (50%) partitions. These non-overlapping partitions ensure that PAD algorithms are tested on samples which are not considered during training and validation and, thus, guarantee a fair evaluation. Four more advanced LOO protocols we do not split samples of each PAI rather than groups of PAIs in subject disjoint training, validation and test sets. This method evaluates the generalization capabilities to PAIs which have not been seen during the training. For more detailed information, the reader is referred to the original COLFISPOOF protocol [Ko23]. A further protocol benchmarks the generalization capabilities to new high-quality PAs. Here, we train and validate on the entire COLFISPOOF database in combination with one of the bona fide databases (60% train, 40% validation) and test on the UniCa-HDA database. This highlights the generalization capabilities in two ways: first the generalization to a new type of PA and second the generalization from PAs contained in the COLFISPOOF database to those from a newly captured database. This is particularly interesting, since the PAs in the COLFISPOOF database were generated based on synthetic ridge line patterns while in the newly captured database, PAs are generated from real fingerprints. This experiment is referred to as cross-database experiment.

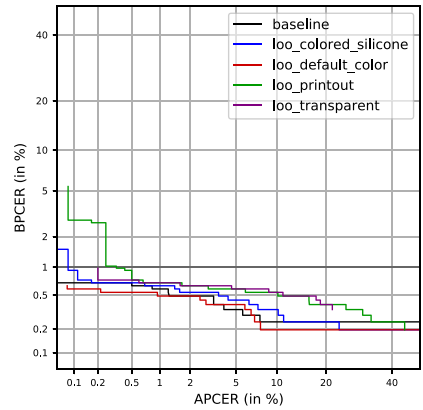
We use the Attack Presentation Classification Error Rate (APCER) vs. Bona fide Attack Presentation Classification Error Rate (BPCER) metric standardized in ISO/IEC 30107-3 [IS23] and the Detection Equal Error Rate (D-EER) metric to report the results of our experiments. To make our work comparable to others, we fix the BPCER at an operation point of 1% for the baseline as well as the LOO experiments and 5% for the cross-database test and report the corresponding APCER.

## 4 Results

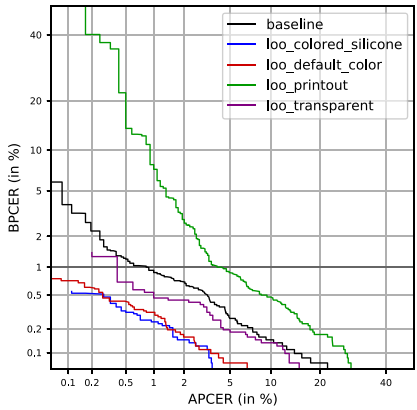
Table 2 gives an overview of APCERs at a BPCER of 1% and D-EERs for the baseline and LOO experiments, whereas Figure 4 (a – c) shows the corresponding DET plots. The results show that the selected contact-based fingerprint PAD algorithm with the incorporated adaptations is able to detect contactless fingerprint PAs. For the best performing ISPFDv1 database, the obtained APCERs are between 0.00% and 0.33%. Except printout



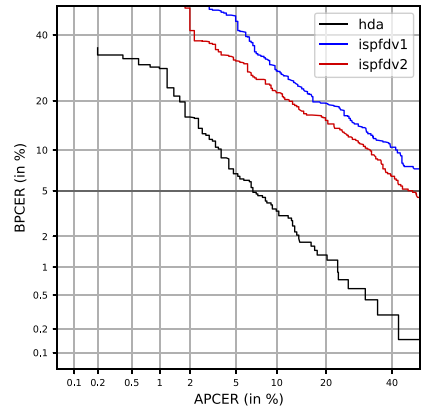

(a) HDA & COLFISPOOF




(b) ISPFDv1 & COLFISPOOF




(c) ISPFDv2 & COLFISPOOF




(d) UniCa-HDA

Fig. 4: DET curves obtained on the considered databases using the baseline and LOO protocols (a – c) and the cross-database test (d).

attacks, the method is also accurately detects PAs when trained and tested on the ISPFDv2 database. The HDA database performs worse, which is most likely caused be the low number of training data.

From the LOO experiments it might be concluded that some materials are easier to detect, *e.g.* colored silicone, whereas others are harder, *e.g.* printouts. Printouts have the worst detection accuracy, which might be caused by the high contrast and clear ridge pattern and should be investigated further. Nevertheless, the selected method has in general a high generalization capability to unseen PAIs.

Tab. 3: Cross-database PAD: APCERs for a fixed BPCER of 5.00% and D-EERs for training on COLFISPOOF together with different bona fide databases, *e.g.* HDA, ISPFDv1, ISPFDv2, and tested on UniCa-HDA.

| Training | Test | APCER | D-EER |
|---|---|---|---|
| HDA & COLFISPOOF | UniCa-HDA | 6.75 | 5.99 |
| ISPFDv1 & COLFISPOOF | UniCa-HDA | 56.15 | 19.42 |
| ISPFDv2 & COLFISPOOF | UniCa-HDA | 45.63 | 16.65 |

This finding is also supported by the cross-database experiment, which is presented in Table 3 and Figure 4 (d). Here we see that the algorithm generalizes to a certain extent to completely unseen PAIs and a new capturing environment. In this experiment, we achieved a D-EER of 5.99% and an APCER of 6.75% at a BPCER of 5.00% for the HDA database. These results indicate that the algorithm generalizes well to completely unseen PAI materials, real fingerprint characteristics (from synthetic ones) and a new capturing environment. However, it should be noted that the algorithm cannot generalize to bona fide samples captured using a different capturing setup as can be seen from Table 3. As can be seen, training on one of the ISPFD databases and testing on the UniCa-HDA setup leads to poor results.

## 5 Conclusion

In this work, we have investigated the possibilities of adapting a state-of-the-art PAD mechanism for contact-based fingerprints to the contactless domain. We showed that the SpoofBuster method can be adapted to a contactless scenario by a parameter refinement and a training on contactless databases. Extensive tests showcase the effectiveness and performance of the methods. The cross-database experiments show high generalization capabilities to new capturing setups, but also highlighted its limitations.

Our results indicate that the considered method is well-suited also to detect PAs in the contactless domain. Most notably, the algorithm detects PAIs which are unseen during the training with high confidence. Furthermore, the experiments on the newly acquired UniCa-HDA database show good potential for generalizing from synthetic PAs to PAs from real subjects. However, more experiments and bigger databases are required to assess the overall system performance.

Future work could be focused on extending the presented approach to a PAD method that works for both, contactless and contact-based fingerprints. If one PAD algorithm can be used for samples from various capture devices maintenance, effort could be reduced.

## 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

[Af20] Afonso Pereira, Joao; Sequeira, Ana F.; Pernes, Diogo; Cardoso, Jaime S.: A robust fingerprint presentation attack detection method against unseen attacks through adversarial learning. In: 2020 International Conference of the Biometrics Special Interest Group (BIOSIG). pp. 1–5, 2020.

[AS06] Abhyankar, Aditya; Schuckers, Stephanie: Fingerprint liveness detection using local ridge frequencies and multiresolution texture analysis techniques. In: 2006 international conference on image processing. IEEE, pp. 321–324, 2006.

[BND22] Berti, Aeddon; Nasrabadi, Nasser; Dawson, Jeremy: Investigating the Impact of Demographic Factors on Contactless Fingerprint Interoperability. In: International Conference of the Biometric Special Interest Group (BIOSIG). LNI. GI, pp. 1–8, September 2022.

[CCJ18] Chugh, Tarang; Cao, Kai; Jain, Anil K.: Fingerprint Spoof Buster: Use of Minutiae-Centered Patches. IEEE Transactions on Information Forensics and Security, 13(9):2190–2202, 2018.

[CJ19] Chugh, Tarang; Jain, Anil K.: Fingerprint Presentation Attack Detection: Generalization and Efficiency. In: 2019 International Conference on Biometrics (ICB). pp. 1–8, 2019.

[Ga23] Galbally, Javier; Fierrez, Julian; Cappelli, Raffaele; Marcialis, Gian Luca: Introduction to Presentation Attack Detection in Fingerprint Biometrics. In: Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment, pp. 3–15. Springer, 2023.

[Ge20] George, Anjith; Mostaani, Zohreh; Geissenbuhler, David; Nikisins, Olegs; Anjos, André; Marcel, Sébastien: Biometric Face Presentation Attack Detection With Multi-Channel Convolutional Neural Network. IEEE Transactions on Information Forensics and Security, 15:42–55, 2020.

[Go21] Gonzalez-Soler, Lazaro Janier; Gomez-Barrero, Marta; Chang, Leonardo; Pérez-Suárez, Airel; Busch, Christoph: Fingerprint presentation attack detection based on local features encoding for unknown attacks. IEEE Access, 9:5806–5820, 2021.

[Gr15] Gragnaniello, Diego; Poggi, Giovanni; Sansone, Carlo; Verdoliva, Luisa: An investigation of local descriptors for biometric spoofing detection. IEEE transactions on information forensics and security, 10(4):849–863, 2015.

[Hu18] Hussein, Mohamed E.; Spinoulas, Leonidas; Xiong, Fei; Abd-Almageed, Wael: Fingerprint Presentation Attack Detection Using A Novel Multi-Spectral Capture Device and Patch-Based Convolutional Neural Networks. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS). pp. 1–8, 2018.

[IS23] ISO/IEC JTC1 SC37 Biometrics: . ISO/IEC 30107-3. Information Technology - Biometric presentation attack detection - Part 3: Testing and Reporting. International Organization for Standardization, 2023.

[Ka21a] Karampidis, Konstantinos; Rousouliotis, Minas; Linardos, Euangelos; Kavallieratou, Ergina: A comprehensive survey of fingerprint presentation attack detection. Journal of Surveillance, Security and Safety, 2(4):117–161, 2021.

[Ka21b] Kauba, Christof; Söllinger, Dominik; Kirchgasser, Simon; Weissenfeld, Axel; Fernández Domínguez, Gustavo; Strobl, Bernhard; Uhl, Andreas: Towards using police officers' business smartphones for contactless fingerprint acquisition and enabling fingerprint comparison against contact-based datasets. Sensors, 21(7):2248, 2021.

[Ko23] Kolberg, Jascha; Priesnitz, Jannis; Rathgeb, Christian; Busch, Christoph: COLFISPOOF: A New Database for Contactless Fingerprint Presentation Attack Detection Research. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 653–661, 2023.

[Ma20a] Malhotra, Aakarsh; Sankaran, Anush; Vatsa, Mayank; Singh, Richa: On matching fingerselfies using deep scattering networks. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2(4):350–362, 2020.

[Ma20b] Malhotra, Aakarsh; Sankaran, Anush; Vatsa, Mayank; Singh, Richa: On Matching Finger-Selfies Using Deep Scattering Networks. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2(4):350–362, 2020.

[Mi23] Micheletto, Marco; Orrù, Giulia; Casula, Roberto; Yambay, David; Marcialis, Gian Luca; Schuckers, Stephanie: Review of the Fingerprint Liveness Detection (LivDet) competition series: from 2009 to 2021. Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment, pp. 57–76, 2023.

[MV22] Marasco, Emanuela; Vurity, Anudeep: Late Deep Fusion of Color Spaces to Enhance Finger Photo Presentation Attack Detection in Smartphones. Applied Sciences, 12(22), 2022.

[Ng18] Nguyen, Thi Hai Binh; Park, Eunsoo; Cui, Xuenan; Nguyen, Van Huan; Kim, Hakil: fPADnet: Small and Efficient Convolutional Neural Network for Presentation Attack Detection. Sensors, 18(8), 2018.

[Pa19] Park, Eunsoo; Cui, Xuenan; Nguyen, Thi Hai Binh; Kim, Hakil: Presentation Attack Detection Using a Tiny Fully Convolutional Network. IEEE Transactions on Information Forensics and Security, 14(11):3016–3025, 2019.

[PB17] Pala, Federico; Bhanu, Bir: Deep triplet embedding representations for liveness detection. Deep Learning for Biometrics, pp. 287–307, 2017.

[Pr21a] Priesnitz, J.; Rathgeb, C.; Buchmann, N.; Busch, C.: Deep Learning-Based Semantic Segmentation for Touchless Fingerprint Recognition. In: Proc. Intl. Conf. Pattern Recognition (ICPR) (Workshops). 2021.

[Pr21b] Priesnitz, Jannis; Rathgeb, Christian; Buchmann, Nicolas; Busch, Christoph; Margraf, Marian: An overview of touchless 2D fingerprint recognition. EURASIP Journal on Image and Video Processing, 2021(1):1–28, 2021.

[Pr22] Priesnitz, J.; Huesmann, R.; Rathgeb, C.; Buchmann, N.; Busch, C.: Mobile Contactless Fingerprint Recognition: Implementation, Performance and Usability Aspects. Sensors, 22(3), 2022.

[Pu23] Purnapatra, Sandip; Miller-Lynch, Conor; Miner, Stephen; Liu, Yu; Bahmani, Keivan; Dey, Soumyabrata; Schuckers, Stephanie: , Presentation Attack Detection with Advanced CNN Models for Noncontact-based Fingerprint Systems, 2023.

[Ra23] Raja, Kiran; Ramachandra, Raghavendra; Venkatesh, Sushma; Gomez-Barrero, Marta; Rathgeb, Christian; Busch, Christoph: Vision Transformers for Fingerprint Presentation Attack Detection. In (Marcel, Sébastien; Fierrez, Julian; Evans, Nicholas, eds): Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment. Springer Nature Singapore, Singapore, pp. 17–56, 2023.

[Sa15] Sankaran, A.; Malhotra, A.; Mittal, A.; Vatsa, M.; Singh, R.: On smartphone camera based fingerphoto authentication. In: 2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS). pp. 1–7, 2015.

[SB14] Sousedik, Ctirad; Busch, Christoph: Presentation attack detection methods for fingerprint recognition systems: a survey. Iet Biometrics, 3(4):219–233, 2014.

[SBB13] Stein, Chris; Bouatou, Vincent; Busch, Christoph: Video-based fingerphoto recognition with anti-spoofing techniques with smartphone cameras. In: International Conference of the Biometric Special Interest Group (BIOSIG). pp. 1–12, 2013.

[Ta16] Taneja, Archit; Tayal, Aakriti; Malhorta, Aakarsh; Sankaran, Anush; Vatsa, Mayank; Singh, Rieha: Fingerphoto spoofing in mobile devices: a preliminary study. In: IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS). IEEE, pp. 1–7, 2016.

[Wa07] Watson, Craig I; Garris, Michael D; Tabassi, Elham; Wilson, Charles L; McCabe, R Michael; Janet, Stanley; Ko, Kenneth: User's guide to NIST biometric image software (NBIS). 2007.

[Wa18] Wasnik, Pankaj; Ramachandra, Raghavendra; Raja, Kiran; Busch, Christoph: Presentation Attack Detection for Smartphone Based Fingerphoto Recognition Using Second Order Local Structures. In: 2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS). pp. 241–246, 2018.

[Ya17] Yambay, David; Becker, Benedict; Kohli, Naman; Yadav, Daksha; Czajka, Adam; Bowyer, Kevin W; Schuckers, Stephanie; Singh, Richa; Vatsa, Mayank; Noore, Afzel et al.: LivDet iris 2017—Iris liveness detection competition 2017. In: 2017 IEEE International Joint Conference on Biometrics (IJCB). IEEE, pp. 733–741, 2017.

[YZH21] Yin, Xuefei; Zhu, Yanming; Hu, Jiankun: A Survey on 2D and 3D Contactless Fingerprint Biometrics: A Taxonomy, Review, and Future Directions. IEEE Open Journal of the Computer Society, 2:370–381, 2021.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# On the Impact of Tattoos on Hand Recognition

Lázaro J. González-Soler[1], Kacper M. Zyla[2], Christian Rathgeb[1], Daniel Fischer[1]

**Abstract:** From Native Americans, who used tattoos as a way of seducing the opposite sex, to prisoners in the last century, who were identified by tattooed numbers, tattoos have been used for many years for a variety of purposes. Nowadays, tattoos express affiliation or beliefs and can therefore serve as complementary information to identify individuals. To support forensic investigations, hand-based biometrics have emerged as a promising technology to recognise individuals. As several statistics have reported an increase in the use of tattoos on hands, in this paper, we investigate the impact of tattoos on the performance of state-of-the-art hand recognition systems. To this end, we first propose a method for generating semi-synthetic tattooed hands. A benchmark is then performed for tattooed and non-tattooed hands. Experimental results computed on a freely available database showed that, although in some cases the use of tattoos assists hand recognition, the observed trend is a deterioration of recognition accuracy, indicating the sensitivity of hand recognition systems to tattoos.

**Keywords:** Hand recognition, hand tattoos, semi-synthetic tattoos, forensics.

## 1 Introduction

The use of tattoos has experienced broad popularity over the years. According to a report by the National Institute of Standards and Technology (NIST) (Tatt-C) [NG15], one-fifth of US adults have at least one tattoo, ranking the US population as the third most tattooed in the world, after Italy and Sweden. These numbers are constantly growing as it has been shown by a research survey conducted in 2019[3] that also revealed that 34% of tattoos within the US population are on the hands or wrists. Tattooed hands are a useful indicator to identify individuals who are members of a gang or criminal organisation, thus being an important field of interest for forensic investigators [MJJ12].

The anatomy and appearance of the hand define an emerging biometric characteristic. Due to the broad development and great success of deep neural networks (DNNs) traditional handcrafted approaches have been recently replaced. DNNs are based on powerful architectures that are able to learn highly discriminative features from hand images. In 2019, Afifi [Af19] proposed a public database of hand images together with a convolutional neural network (CNN) for gender classification and subject identification. Following the above idea, more recent CNN-based techniques have exploited either attention mechanisms [Ba22a, Ba22b] or Vision-Transformers [Eb22] to improve upon the baseline

[1] da/sec - Biometrics and Security Research Group, Hochschule Darmstadt, Germany,
{lazaro-janier.gonzalez-soler;christian.rathgeb;daniel.fischer}@h-da.de
[2] Technical University of Denmark, Denmark, s202617@student.dtu.dk
[3] `https://comparecamp.com/tattoo-statistics/`

identification performance reported in [Af19]. These techniques have achieved high recognition performance in closed-set identification scenarios where the subject identity being searched for is known to be found within the enrolled references.

To the best of our knowledge, the evaluation of the impact of tattoos on hand recognition has not been addressed so far. In this work, we evaluate how the use of tattoos in the dorsal area of the hand affects the identification performance of state-of-the-art hand recognition systems. Given the lack of public databases including tattooed hands, an algorithm to synthetically blend tattoos on the dorsum of the hand is additionally proposed.

The remainder of this article is organised as follows: Sect. 2 provides work related to hand recognition. In Sect. 3, an overview of the proposed pipeline for blending tattoos on the dorsum of the hand is introduced. The experimental setup including a summary of the database used, together with the evaluated systems is presented in Sect. 4. The experimental results are discussed in Sect 5. Conclusions are finally drawn in Sect. 6.

## 2 Related Work

Recently, the use of the hand in forensic investigations has gained interest in detecting both wanted criminals and missing victims. The latest hand recognition systems map whole hand images acquired in the visible spectrum into a latent representation using DNNs. Afifi [Af19] introduced an annotation-rich hand database (referred to in the scientific literature as 11K Hands) consisting of 11,076 high-quality hand images of 190 subjects. Utilising this database, Afifi proposed a dual-stream CNN-based algorithm whose recognition performance values (i.e., Identification Rate (IRs) ranging from 94% to 97% for the palmar and dorsal area, respectively) provided a first benchmark for future forensic investigations. Following the above idea, Baisa *et al.* [Ba22a] recently proposed a dual-stream CNN approach based on attention mechanisms that learns both global and local features of the hand image. The experimental results reported an IR at Rank-1 of around 95% on 11K Hands [Af19]. Baisa *et al.* [Ba22b] extended the above architecture by including an extra stream and incorporating both channel and spatial attention modules. An improvement in terms of recognition performance of around 3 percentage points (i.e., IR = 98.05%) was attained on the right palm images compared to the results obtained in [Ba22a], (i.e., IR = 95.83%). In the same study, other CNNs were evaluated for hand recognition, e.g., ABD-Net [Ch19] and RGA-Net [Zh20b], resulting in similar recognition performance to the system in [Ba22a]. Finally, Ebrahimian *et al.* [Eb22] evaluated the feasibility of using Vision Transformers for hand recognition, resulting in an IR of 99.4% on a small set consisting of 30% of the images in 11K Hands. Despite the results obtained by the techniques described above, there is still a lack of evaluations including images of tattooed hands.

## 3 Generation of Tattooed Hands

Inspired by the approach in [GSRF23] which shows the feasibility of using semi-synthetic tattoos for the segmentation of real tattoos, a generation method for blending realistic

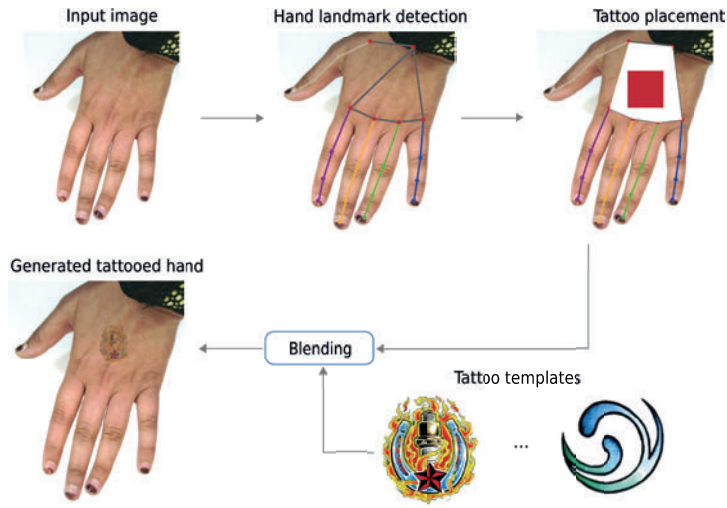



Fig. 1: Overview of the tattoo blending pipeline.

tattoos on the dorsal area of the hand is proposed. Fig. 1 shows a general overview which consists of three main steps:

*i*) Detection of the 20 hand landmarks using Google's MediaPipe algorithm [Zh20a].

*ii*) Selection of a suitable area based on a subset of the landmark coordinates (i.e., 0, 1, 5, 9, 13, 17) which define the hand's dorsal area.

*iii*) Blending the tattoo onto the corresponding image at the selected area.

To correctly place the tattoo template on the hand image, a suitable area is selected after detecting the hand landmarks with Google's MediaPipe. In contrast to the approach in [GSRF23], which computes the tattoo position based on the segmentation map of the input image, the proposed method finds the dorsal area defined by a subset of landmark coordinates (i.e., the white region in Fig. 1). Then, a skin coordinate $\mathbf{x}' = (x, y)$ is randomly selected where the rectangle formed by the edges of the area, with $\mathbf{x}'$ as the top-left corner, is at least half the size of the tattoo. This way, tattoo visibility is ensured in the generated image. The tattoo is then placed at the position $\mathbf{x}'$ and blended on the hand image by multiplying the tattoo layer with the hand image. In a previous step, the tattoo template at hand is randomly resized to a factor calculated over the size of the white region. Areas of the tattoo which end up outside the skin are cut out. The final image is made more realistic by adjusting the tattoo colour, applying a Gaussian blur and reducing its opacity. To generate the images, the tattoo templates proposed in [Ib22] were used.

Fig. 2: Examples of 11K Hands images (first row) and their respective tattooed hands (second row).

## 4 Experimental Setup

The goal of the experimental evaluation is twofold: *i*) investigate the impact of the use of tattooed hands on the recognition performance of current hand recognition systems for the closed-set scenario (i.e., the same subjects participate in both enrolment and biometric transactions) and *ii*) analyse the extent to which blending tattoos in the dorsal area of the hand increases false match rates in an open-set scenario (i.e., searched subjects are potentially not enrolled in the gallery). To reach our goals, we define four experimental protocols: *i*) non-tattooed reference and probe in a closed-set scenario, *ii*) non-tattooed reference and tattooed probe in a closed-set scenario, *iii*) tattooed reference and probe in a closed-set scenario, and *iv*) non-tattooed reference and tattooed probe in an open-set scenario.

### 4.1 Databases

The 11K Hands [Af19] database is an extensive collection of over 11 thousand images of hands collected in 2019. The samples in this database come from subjects of different ethnicities and ages ranging from 18 to 75 years. It contains the dorsal and palmar side images of the hands of 190 subjects on a white background (see Fig. 2, first row). For the evaluation, the database is partitioned as in [Ba22a] and images of hands with accessories are excluded to avoid bias to external variables. Therefore, the right and left dorsal subsets comprise 143 and 146 identities, respectively. For each case, half of the identities are used for training the algorithms and the remaining ones are employed to compute the

identification performance, i.e., 72:71 for the right dorsal and 73:73 for the left dorsal, respectively.

To evaluate the impact of using tattoos on hand recognition performance, 10 tattooed versions of each hand image of the respective evaluation subset are generated using the proposed method described in Sect. 3, resulting in 21,030 generated images, i.e., 10,420 for right dorsal and 10,610 for left dorsal (see Fig. 2, second row). It is worth noting that not all images used as biometric transactions in the dorsal evaluation were processed, due to a failure in the detection of their landmarks. In the final evaluation of the tattoos, 33 out of 71 identities are considered for the right dorsal and 32 out of 73 identities for the left dorsal.

### 4.2 Implementation Details

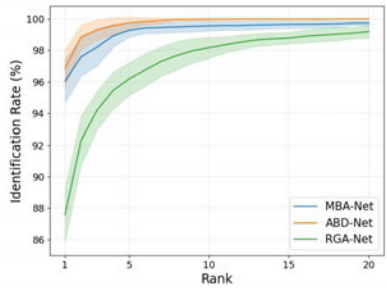
In the evaluation, state-of-the-art hand-based recognition systems are employed, i.e., MBA-Net [Ba22b], ABD-Net[Ch19], and RGA-Net [Zh20b]. All algorithms are implemented in PyTorch [Pa19] and trained utilising an NVidia A100 Tensor Core GPU with 40GB DRAM. For the training and testing of the systems, the parameters indicated in the corresponding publications were considered. The image size was set to $256 \times 256$ pixels for ABD-Net [Ch19] and RGA-Net [Zh20b] and $356 \times 356$ for MBA-Net [Ba22b]. In addition, the networks were initialised with their pre-trained weights in ImageNet [De09] and optimised on 70 epochs using the Adam optimiser. As indicated in [Zh20b], the RGA-Net architecture was trained for 600 epochs. In each case, the best-performing weights are selected from a subset of the training set.

## 5 Results and Discussion

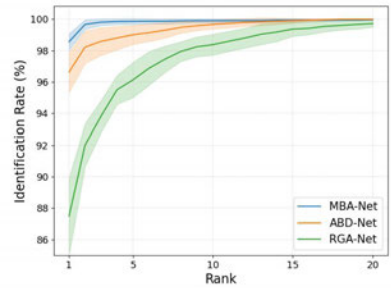
As mentioned in Sect. 3, the use of tattoos on the hand has recently gained popularity. In this section, the results of the impact of the use of tattooed hands on the recognition performance of the systems evaluated for closed-set (Sect. 5.1) and open-set (Sect. 5.2) scenarios respectively are presented.

### 5.1 Closed-set Scenario

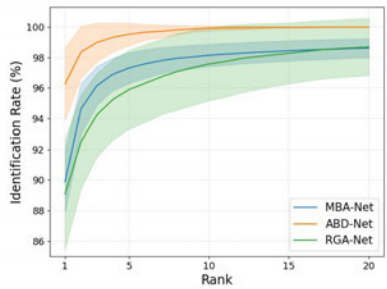
Fig. 3 shows the identification rates for non-tattooed (row 1), tattooed only on the probe (row 2), and tattooed on both reference and probe (row 3) hands for left and right dorsal hand images. To compute IRs at different rank values, we split the database into 10 disjoint sets of enrolment and biometric transactions, each time randomly selecting one sample per subject for enrolment and the remaining samples for identification transactions. Then the mean and standard deviation (std) are reported. For biometric transactions of tattooed hand images, we enrolled either a non-tattooed (row 2) or tattooed (row 3) reference from the same probe subject. To simulate a real scenario, reference and probe hand images were generated using the same tattoo template in the latter case (row 3).



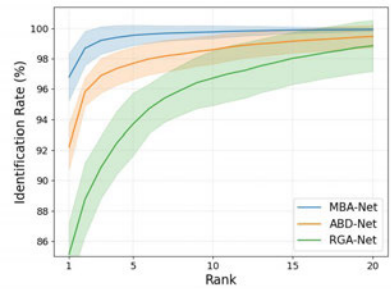

(a) Non-tattooed left dorsal on reference and probe.



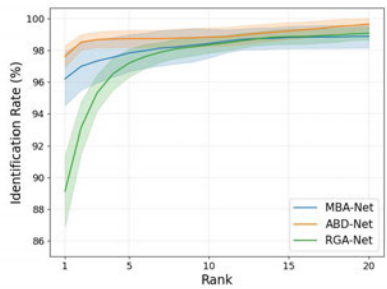

(b) Non-tattooed right dorsal on reference and probe.



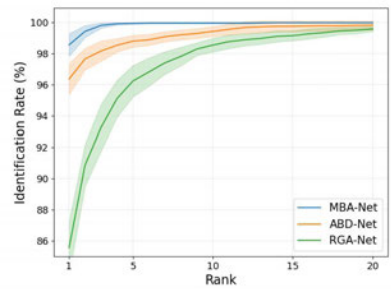

(c) Tattooed left dorsal on the probe.




(d) Tattooed right dorsal on the probe.




(e) Tattooed left dorsal on reference and probe.




(f) Tattooed right dorsal on reference and probe.

Fig. 3: CMC curves reported by the evaluated systems on non-tattooed (3a,3b), tattooed only on probe (3c,3d), and tattooed on reference and probe (3e,3f) images from 11K Hands.

Note that, similar IRs across different rank values are achieved for right and left dorsal subsets. While MBA-Net reports, on average, the best IR at Rank-1 for non-tattooed right dorsal images, i.e., IRs $\geq$ 98%, ABD-Net yields the best IRs for the respective left dorsal subset at the same rank. Those approaches (i.e., MBA-Net and ABD-Net) also achieve an IR $\geq$ 99.9% in the Rank-5, indicating that the searched identities of transactions are

retrieved by the system with almost 100% success in the top 5 positions of the candidate list. From a forensic point of view, high IRs for Ranks higher than 1 are still interesting, as the number of possible suspects is reduced. Regarding RGA-Net, lower performance in terms of IRs can be observed, i.e., an IR around 86% is obtained at Rank-1.

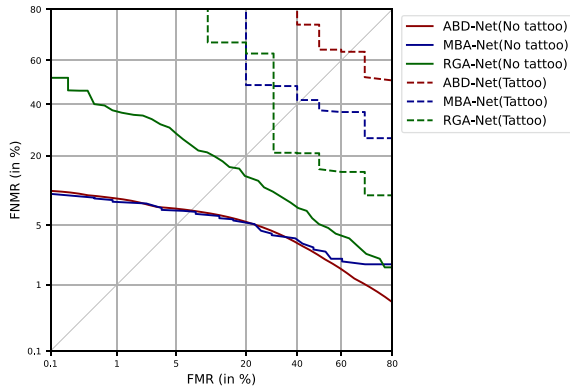
Comparing the results in Fig. 3a and 3b, all networks report on average a performance deterioration for tattooed hands, see Fig. 3c and 3d: the IRs decrease down to 97% and 92% in Rank-1 for right and left dorsal sets for the best-performing approaches (i.e., MBA-Net and ABD-Net). Furthermore, the std values increase regarding the ones depicted for non-tattooed hands. This deterioration in recognition performance is due to the fact that the features calculated by both architectures describe mainly textural details. Therefore, they are prone to fail on tattooed hands. In contrast to MBA-Net and ABD-Net, RGA-Net obtains on average similar results for tattooed and non-tattooed hands, i.e., IRs in around 87% and 88% for right and left dorsal images, respectively. However, compared to the other methods, this technique obtains the worst std values for subjects with tattooed hands.

Finally, note that the biometric performance yielded by the networks when both reference and biometric transactions contain tattooed hands (see Fig. 3e and 3f) is similar to that of non-tattooed hands in Fig. 3a and 3b. Observe that the use of tattoos can positively assist hand recognition in most cases: an improvement in the IR values for the left dorsal can be perceived compared to the baseline (e.g., 89% vs. 87.50% at Rank-1 for RGA-Net). Note also that the results reported for the left dorsal differ from those for the right dorsal. This is because the tattoo templates were randomly selected by subject and hand; the same subject could therefore have different tattoos on the left and right hand. A direct result of these observations is focused on the use of images of tattooed hands to train the algorithms. Thus, the performance shown in Fig. 3c and 3d can be significantly improved.

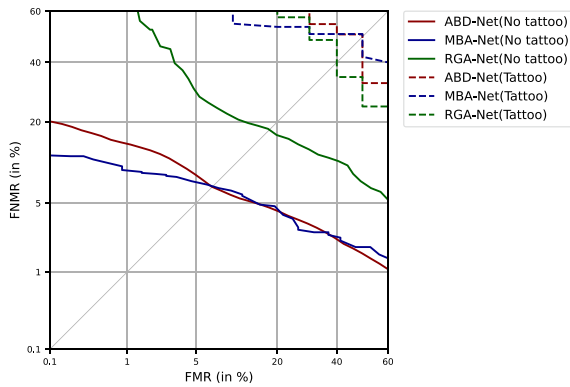
### 5.2 Open-set Scenario

The identification performance of the available hand-based methods is also reported in Fig. 4 for an open-set scenario. To compute mated and non-mated comparisons, we perform a 10-fold cross-validation evaluation. Thus, each time, the subjects belonging to the validation fold at hand are employed for computing the non-mated comparisons, while the remaining subjects from the other subsets are used for the mated comparisons. For the assessment of the impact of the tattoos, the non-tattooed subjects in the validation fold in question are replaced by the same subjects with tattooed hands.

Similar to the results in Fig. 3, MBA-Net achieves the best performance (dark blue thick lines), resulting in a FNIR = 12.03% and 10.00% for a high-security threshold, i.e., FPIR = 0.1% on the right and left dorsal, respectively: 1 out of 1000 non-mated transactions is accepted, while at most 12 out of 100 mated transactions are rejected by the recognition system. Note that ABD-Net yields better performance than MBA-Net on the left dorsal for the closed-set scenario (see Fig. 3a). However, the latter outperforms the ABD-Net in the open-set scenario for small FPIR values, i.e., high-security thresholds. It can be




(a) Left dorsal




(b) Right dorsal

Fig. 4: DET curves for left and right dorsal images.

also observed that the use of tattooed hands significantly affects the performance of the architectures: FNIR values at a FPIR=0.1% are above 60% and 80% respectively for right and left dorsal, indicating the sensitivity of the current hand recognition systems to tattooed hands. Finally, we note that RGA-Net is less sensible to tattooed hands than the other approaches. This is due to some attention mechanisms which leverage both texture and shape properties.

## 6 Conclusions

In this work, we evaluated the impact of tattoos on the biometric performance of hand recognition systems. To do that, an approach which synthetically blends tattoo templates on real hand images, from the 11K Hands database, is proposed. In essence, this synthetic generator first computes the reference points of the input hand image to draw a

random position where to place the tattoo. Using the 11K Hands database, which contains around 11,000 images, 10 tattooed samples per image were generated. A benchmark of the most competitive hand recognition systems was then established. Experimental results on a closed-set scenario showed, on average, a decrease in the performance of the schemes, as well as a high standard deviation indicating their sensitivity to these tattooed hands. In addition, a most challenging evaluation on an open-set setup reported a significant performance deterioration for high-security thresholds: FNIR values increased from roughly 10% to over 60% for an FPIR = 0.1% when the tattooed hands were presented to the systems as non-mated transactions. One solution that emerges from this work is the use of the proposed method as a data augmentation strategy to generate tattooed hands that can be considered as training data for recognition systems. In this way, the performance of hand recognition systems against tattooed hands can also be improved. In addition, further research may also propose inpainting techniques to remove tattoos from hands.

## Acknowledgements

This research work has been partially funded by the Hessian Ministry of the Interior and Sport in the course of the Bio4ensics project and 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

[Af19] Afifi, M.: 11K Hands: gender recognition and biometric identification using a large dataset of hand images. Multimedia Tools and Applications, 2019.

[Ba22a] Baisa, N.; Williams, B.; Rahmani, H.; Angelov, P.; Black, S.: Hand-based person identification using global and part-aware deep feature representation learning. In: Proc. Intl. Conf. on Image Processing Theory, Tools and Applications (IPTA). pp. 1–6, 2022.

[Ba22b] Baisa, N.; Williams, B.; Rahmani, H.; Angelov, P.; Black, S.: Multi-Branch with Attention Network for Hand-Based Person Recognition. In: Proc. Intl. Conf. on Pattern Recognition (ICPR). pp. 727–732, 2022.

[Ch19] Chen, T.; Ding, S.; Xie, J.; Yuan, Y.; Chen, W.; Yang, Y.; Ren, Z.; Wang, Z.: ABD-Net: Attentive but diverse person re-identification. In: Proc. Intl Conf. on Computer Vision (ICCV). pp. 8351–8361, 2019.

[De09] Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proc. Intl. Conf. on Computer Vision and Pattern Recognition (CVPR). Ieee, pp. 248–255, 2009.

[Eb22] Ebrahimian, Z.; Mirsharji, S.; Toosi, R.; Akhaee, M.: Automated Person Identification from Hand Images using Hierarchical Vision Transformer Network. In: Proc. Intl. Conf. on Computer and Knowledge Engineering (ICCKE). pp. 398–403, 2022.

[GSRF23] Gonzalez-Soler, L. J.; Rathgeb, C.; Fischer, D.: Semi-synthetic Data Generation for Tattoo Segmentation. In: Proc. Intl. Workshop on Biometrics and Forensics (IWBF). pp. 1–6, 2023.

[Ib22] Ibsen, M.; Rathgeb, C.; Drozdowski, P.; Busch, C.: Face Beneath the Ink: Synthetic Data and Tattoo Removal with Application to Face Recognition. Applied Sciences, 12(24), 2022.

[MJJ12] Mun, J.; Janigo, K.; Johnson, K.: Tattoo and the self. Clothing and Textiles Research Journal, 30(2):134–148, 2012.

[NG15] Ngan, M.; Grother, P.: Tattoo recognition technology-challenge (Tatt-C): an open tattoo database for developing tattoo recognition research. In: Proc. Intl. Conf. on Identity, Security and Behavior Analysis (ISBA 2015). pp. 1–6, 2015.

[Pa19] Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Advances in Neural Information Processing Systems 32. pp. 8024–8035, 2019.

[Zh20a] Zhang, F.; Bazarevsky, V.; Vakunov, A.; Tkachenka, A.; Sung, G.; Chang, C.; Grundmann, M.: Mediapipe hands: On-device real-time hand tracking. arXiv preprint arXiv:2006.10214, 2020.

[Zh20b] Zhang, Z.; Lan, C.; Zeng, W.; Jin, X.; Chen, Z.: Relation-aware global attention for person re-identification. In: Proc. Intl. Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 3186–3195, 2020.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# A RISE-based explainability method for genuine and impostor face verification

Naima Bousnina[1], João Ascenso[2], Paulo Lobato Correia[2], Fernando Pereira[2]

**Abstract:** Heat Map (HM)-based explainable Face Verification (FV) has the goal to visually interpret the decision-making of black-box FV models. Despite the impressive results, state-of-the-art FV explainability methods based on HMs mainly address genuine verification by generating visual explanations that reveal the similar face regions which most contributed for acceptance decisions. However, the similar face regions may not be the unique critical regions for the model decision, notably when rejection decisions are performed. To address this issue, this paper proposes a more complete FV explainability method, providing meaningful HM-based explanations for both genuine and impostor verification and associated acceptance and rejection decisions. The proposed method adapts the RISE algorithm for FV to generate Similarity Heat Maps (S-HMs) and Dissimilarity Heat Maps (D-HMs) which offer reliable explanations to all types of FV decisions. Qualitative and quantitative experimental results show the effectiveness of the proposed FV explainability method beyond state-of-the-art benchmarks.

**Keywords:** Face verification explainability, Genuines and impostors, Explainability heat maps

## 1 Introduction

The field of Face Verification (FV) has demonstrated remarkable performance advances since the adoption of deep learning technology. However, alongside their impressive performance, FV models remain black-box tools with complex and unintuitive decision-making processes. Therefore, it became critical to understand and explain their decision-making to further improve their performance and make this technology more acceptable by the society at large.

The idea underlying FV explainability is to develop reliable methods that offer insights into why two face images have/have not been matched. Various explainability methods have been proposed to create post-hoc explanations for black-box FV models which are explanations created while not interfering with the model working process itself. Generally, post-hoc explainability methods may be grouped into model-agnostic and model-specific methods [Moln19]. While the former can be plugged into any FV model to explain its behavior, the latter are designed for a single or a specific type of FV models. Both categories may generate different types of explainability outputs, notably

[1] Instituto de Telecomunicações, Lisbon, Portugal, naima.bousnina@lx.it.pt
[2] Instituto de Telecomunicações – Instituto Superior Técnico – Universidade de Lisboa, Lisbon, Portugal, { joao.ascenso, paulo.correia, fp}@lx.it.pt

face features relevance and saliency/heat maps (HMs).

Several available works [Mery22] [MeMo22] have explored the approach of explaining FV decisions with post-hoc and model-agnostic explainability methods using HMs. Those works typically define FV explainability as a way to highlight similar face regions which the FV model believes contribute the most for acceptance decisions when genuine face pairs are processed. However, explaining a FV decision does not simply mean highlighting the critical similar face regions in the probe-gallery pair. In fact, dissimilar face regions are also critical to offer good explanations, notably when rejection decisions are performed. In this context, it is beneficial from an explainability point of view that an HM-based FV explainability method creates two types of HMs to explain the various possible FV decisions, notably Similarity Heat Maps (S-HMs) to highlight the similar face regions when the FV model believes the probe-gallery pair belongs to the same individual, and Dissimilarity Heat Maps (D-HMs) to highlight the dissimilar face regions when the FV model believes the probe-gallery pair belongs to different individuals.

In this context, this paper proposes a post-hoc and model-agnostic RISE-based FV explainability method (FV-RISE) to explain the decision-making of any FV model using HMs, thus without accessing or modifying the inner architecture of the FV model; a key novelty of the proposed method is that it addresses both genuine and impostor verification attempts, as well as acceptance and rejection decisions. The proposal is based on the Randomized Input Sampling for Explanation (RISE) tool, originally designed to estimate the pixels' importance in the context of object classification tasks, by applying random masks to the input image and using the output class probabilities as weights to compute a HM as a weighted sum of the masks. The choice of the RISE tool is motivated by the adoption of the pixel-wise perturbation-based approach, with the potential to generate more spatially accurate explainability heat maps. In the proposed FV-RISE, the created FV explanation depends on the decision, notably whether an acceptance or rejection verification decision is made. The decision-making is explained using a single HM according to the type of decision, notably the S-HM is used when a True/False Acceptance decision is performed, while the D-HM is used when a True/False Rejection decision is performed. The experimental results not only show that FV-RISE is able to offer reliable explanations for all four possible FV cases (genuine/ impostor and acceptance/rejection), but it also offers better FV explanations for the cases that state-of-the-art methods are able to explain.

The remainder of this paper is organized as follows. A brief review of the stat-of-the-art on HM-based FV explainability is provided in Section 2. Section 3 proposes the novel FV-RISE explainability method. Section 4 reports and analyses the obtained experimental results. Finally, Section 5 concludes the paper and discusses the potential directions for future work.

## 2 Related work

Recently, HM-based explainability methods have been largely used to explain the decision-making of black-box FV models. Generally, these methods may be classified into two categories. The first category requires access to the intrinsic architecture or the model gradient information [YTLS19] [ZhYC21] [CaBy18]. For example, Yin *et al.* [YTLS19] proposed a spatial activation diversity loss to learn more structured face features to encourage filters to capture more discriminative visual cues and push the interpretable representations to be more discriminative. [ZhYC21] introduced a novel explainability method for deep metric learning and their applications, e.g. face recognition, person re-id, image retrieval, etc. The key idea of this method is to use a point-to-point activation response technique to decompose the HM, targeting to uncover the relationship between each activated region between the probe and gallery pair. [CaBy18] explored the use of network attention and contrastive network attention for visualizing discriminative features for face recognition; this work demonstrated through the hiding game technique that excitation backpropagation best identifies the face regions contributing to a correct recognition.

The second category generates the HMs by performing random perturbations, e.g. noise, occlusion, etc., on the input/probe images and measures the perturbations' impact on the FV model performance [MeMo22] [Mery22] [KTHR23]. For example, Mery and Morris [MeMo22] proposed the so-called AVeraGe (AVG) explainability method which consists of six sub-methods to generate six different similarity HMs. Four of these HMs are generated by removing and aggregating relevant face regions and measuring the individual contributions of these regions as well as in collaboration, while the other two HMs are combinations of the first four HMs. [Mery22] proposed the so-called MinPlus method which consists of six sub-methods to generate six similarity HMs using a similar removal and aggregation idea as in [MeMo22]. Both AVG and MinPlus methods differ in three main characteristics, notably: *i)* the way the first four HMs are generated; *ii)* the way the first four HMs are combined; and *iii)* in contrast to AVG designed for FV only, MinPlus may be applied to any face analysis task. The experimental results demonstrated that MinPlus generates better HMs and showed promising results when compared with state-of-the-art explainability methods. Knoche *et al.* [KTHR23] followed Mery's [MeMo22] strategy to design three explainability methods, each generating a different HM, highlighting similar and dissimilar face regions contributing to the FV decision. While these three explainability methods generate HMs highlighting only the higher degree similarities areas using one single color, the proposed FV-RISE method generates similarity and dissimilarity HMs highlighting areas with varying degrees of similarity/dissimilarity using a color code ranging from blue to red.

## 3 Proposed FV-RISE explainability method

The key idea behind the proposed FV-RISE method is to generate S-HMs and D-HMs to

explain the FV decisions according to the type of decision performed, notably acceptance or rejection, for both genuine and impostor verification attempts. More precisely, S-HMs and D-HMs are used to explain the acceptance and rejection cases, respectively, regardless of true or false FV decisions being performed. The FV-RISE method is inspired by the random-masking approach used by RISE [PeDS18] to explain object classification, with the novelty that FV-RISE applies that approach to the FV task. Fig. 1 shows the overall architecture of the proposed FV-RISE method which main modules are explained in the next subsections.

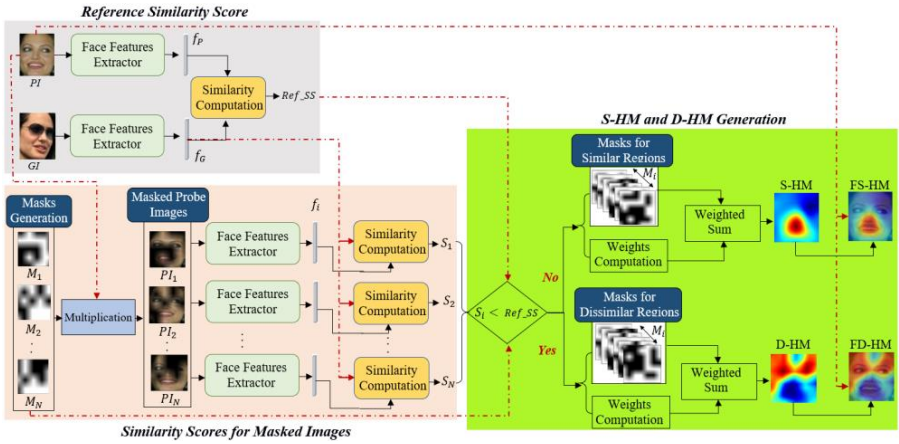



Fig. 1: FV-RISE explainability method overall architecture.

### 3.1 Reference similarity score and similarity scores for masked images

As shown in the grey area in Fig. 1, the FV-RISE process starts by feeding the probe and gallery images (*PI, GI*) into the *Face Features Extractor* to obtain their face features, $f_P$ and $f_G$. Then, the reference similarity score (*Ref_SS*) for the FV decision is computed for the selected similarity metric.

As a second step (see the pink area in Fig. 1), the RISE masking technique [PeDS18] is adopted to randomly generate $N$ masks, $M_i$, $i \in \{1, 2, \ldots N\}$, with values between 0 and 1. In summary the mask generation process is performed through three key steps, notably:

1. Generation of $N$ binary masks with $w \times h$ resolution smaller than the $W \times H$ probe image resolution. The percentages of 0 and 1 values in each generated mask is defined by setting the mask pixels to 0 with a probability $p$ and to 1 with a probability $(1-p)$.
2. Up sampling the generated masks to a resolution of $(w+1)C_W \times (h+1)C_H$ using bilinear interpolation, where $C_W \times C_H = (W/w) \times (H/h)$.

3. Cropping each up sampled mask to a $W \times H$ resolution. The start of the cropping is set with uniformly random mask pixels' locations going from (0,0) up to ($C_W$, $C_H$) to get masks with $W \times H$ resolution with values between 0 and 1.

The generated masks are basically used to pixel-wise perturb random regions of the face and measure the impact of masking randomized regions on the FV model's performance. Each of the generated $N$ masks is pixel-wise multiplied by the probe image, *PI,* to produce $N$ masked probe images $PI_i$, $i \in \{1, 2, \ldots N\}$. Each of the generated masked probe images is fed into the *Face Features Extractor* to capture its corresponding face features, $f_i$, $i \in \{1, 2, \ldots N\}$. Finally, the similarity scores $S_i$, $i \in \{1, 2, \ldots N\}$ are computed between the unmasked gallery image *GI* and each of the masked probe images $PI_i$, $i \in \{1, 2, \ldots N\}$. The similarity score computed between each masked probe image and the gallery image reveals the importance of the masked face region, when compared to the reference similarity score *Ref_SS*, for the FV decision.

### 3.2 S-HM and D-HM generation

Once the similarity scores $S_i$, $i \in \{1, 2, \ldots N\}$ are obtained, a set of steps are performed to generate S-HM and D-HM (see the green block in Fig. 1), notably: *i)* the similarity scores $S_i$, $i \in \{1, 2, \ldots N\}$ are sequentially compared with the reference similarity score *Ref_SS*. If $S_i$ is smaller than *Ref_SS,* the masked face region is considered important for an acceptance decision, meaning that it corresponds to a similar face region in the probe-gallery pair; otherwise, it is considered non-important for an acceptance decision, meaning that it is a dissimilar face region in the probe-gallery pair. Using this comparison, the $N$ generated masks are categorized into masks for similar regions and masks for dissimilar regions. *ii)* The absolute difference between *Ref_SS* and each $S_i$ is computed as the similarity weight for the corresponding mask; the higher this similarity weight, the more confident is the FV model on its decision, thus changing the HM pixels' color from blue colors to red ones. *iii)* The weighted sum of the masks multiplied by the corresponding similarity weights is computed for similar and dissimilar face regions. Finally, the generated S-HM and D-HM are superimposed over the original probe image luminance to generate the so-called Facial S-HMs (FS-HM) and Facial D-HMs (FD-HM), respectively, which allow a better visualization of the FV model decision HM explanation.

The FV decision is explained using one single HM (S-HM or D-HM), depending on the type of decision. More precisely, the S-HM is used to explain True Acceptance (genuine) or False Acceptance (impostor) decisions, since it highlights the face regions contributing most for a FV positive decision (high verification similarity), while D-HM is used to explain True Rejection (impostor) or False Rejection (genuine) decisions, since it highlights the face regions contributing most for a FV negative decision (low verification similarity).

# 4 Experimental assessment

This section reports the quantitative and qualitative performance assessment for the proposed FV-RISE explainability method, whenever possible in comparison with state-of-the-art explainability benchmarks.

## 4.1 Datasets and benchmarks

The example test face images used in this paper are selected from three face recognition datasets, notably LFW [RMBL08], CPLFW [ZhDe18], and Webface-Occ [HWWJ21], in order to assess the behavior of the FV model in differently challenging scenarios, such as partial face occlusion and head pose variation. In addition, a subset of the LFW [RMBL08] dataset with 1000 matching pairs of images is used for the quantitative explainability performance.

The proposed FV-RISE explainability method is compared with four explainability benchmarks from the literature, notably LIME [RiSG16], AVG [MeMo22], Grad-CAM [SCDV17] and MinPlus [Mery22].

For the LIME, AVG, and MinPlus benchmarks, the software implementations available in the Google Colab material [Mery23] used in [Mery22] were adopted, whereas for Grad-CAM, the software implementation version available in the GitHub project [Zile23] of the [ZhYC21] 'official' implementation was adopted, providing the adapted Grad-Cam approach for face verification task.

## 4.2 Experimental set-up

While the proposed explainability method is agnostic to the FV model and similarity metric, the experiments are performed using the ArcFace face recognition model trained with the MS1MV2 dataset built with images from the MS1M dataset [GZHH16]. Additionally, the cosine similarity is used as similarity metric.

The face areas in the images from the selected face recognition datasets are detected using the RetinaFace face detector [DGVK20] and cropped using the facial area coordinates provided by the RetinaFace detector. In addition, to satisfy the image resolution constraint imposed by ArcFace face verification model, the cropped face images are resized to a spatial resolution of $W \times H = 112 \times 112$ pixels.

In the masks' generation stage, the number of generated masks is set to $N=10000$ to obtain more accurate similarity and dissimilarity heat maps. In addition, the conducted experiments demonstrated that an appropriate probability value is $p = 0.1$. Furthermore, the masks are initially generated with a resolution of $w \times h = 5 \times 5$ before bilinear interpolation.

### 4.3 Qualitative explainability performance

This section presents qualitative explainability results for four types of FV decisions, notably depending on whether a genuine or impostor face is processed, and an acceptance or rejection FV decision is made.

#### 1. Genuine case: True Acceptances and False Rejections

Fig. 2 shows the HMs generated using the proposed FV-RISE method for the genuine pair examples, notably (F)S-HMs for True Acceptance decisions (left) and (F)D-HMs for False Rejection decisions (right). Regarding the True Acceptance decisions (left), it can be observed that the S-HMs duly highlight the face regions that the FV model believes to be similar in the probe-gallery pair, and thus have contributed to a True Acceptance decision. For example, in the second row (left), the FV model finds that the mouth, nose and eyes regions are similar enough to classify the probe-gallery pair as the same individual.

For the False Rejection decisions (right), it can be observed that the D-HMs highlight the face regions that the FV model finds to be dissimilar in the probe-gallery pair, and made it fail to identify the pair as the same individual. For example, in the first row (right), the FV model finds that mostly all the face regions – except the subject right eye region – are dissimilar, and thus have contributed to a False Rejection (particularly, his left face side, which is missing information in the gallery image).

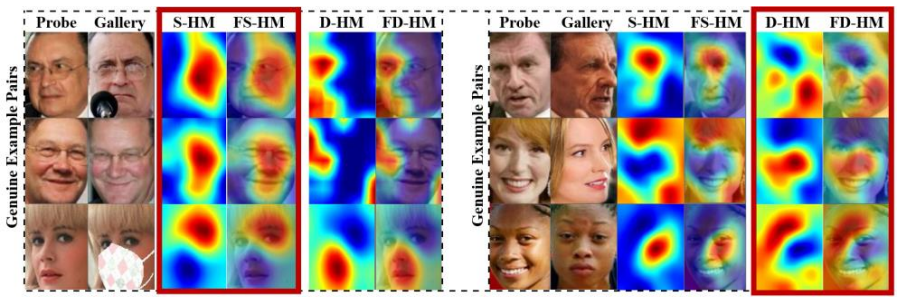



Fig. 2: HM explanations (red rectangle) for genuine pairs: True Acceptance (left) and False Rejections (right) face verification decision examples.

#### 2. Impostors case: True Rejections and False Acceptances

Fig. 3 shows the HMs generated using the FV-RISE method for the impostor pair examples, notably (F)D-HMs for the True Rejection decisions (left) and (F)S-HMs for False Acceptance decisions (right). Regarding the True Rejection decisions (left), it can be observed that the D-HMs highlight the face regions that the FV model finds dissimilar in the probe-gallery pair, which leads to correctly classify the pair as different individuals. For example, in the first row (left), the FV model finds that the eyes and the

mouth regions are dissimilar enough in the probe-gallery pair to correctly make a rejection decision.

Regarding the False Acceptance decision (right), it can be observed that the S-HMs highlight the face regions that the FV model finds similar in the probe-gallery pair to mistakenly classify the pair as being from the same individual. For example, the FV model finds that the nose and eyes regions (resp. the nose region) are similar enough for the first-row pair (resp. second-row pair) to falsely make an acceptance decision.

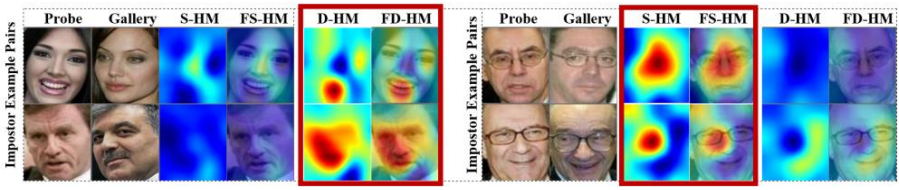



Fig. 3: HM explanations (red rectangle) for impostor pairs: True Rejections (left) and False Acceptance (right) face verification decision examples.

Fig. 4 shows a comparison of the FV-RISE method with four explainability benchmarks. Since the explainability benchmarks do not produce D-HMs, the comparison is performed with S-HMs only. Fig. 4 shows that the proposed FV-RISE method outperforms the considered explainability benchmarks by generating more accurate S-HMs, notably by focusing on the important face regions in a more compact way even when occluded gallery images are processed, which better explain the FV decisions. While the AVG and MinPlus methods also succeed in providing good explanations, their S-HMs highlight a larger part of the face making them less precise. In addition, the comparison with the LIME and Grad-Cam methods adapted for FV demonstrates that both benchmarks generate less meaningful explainability heat maps.

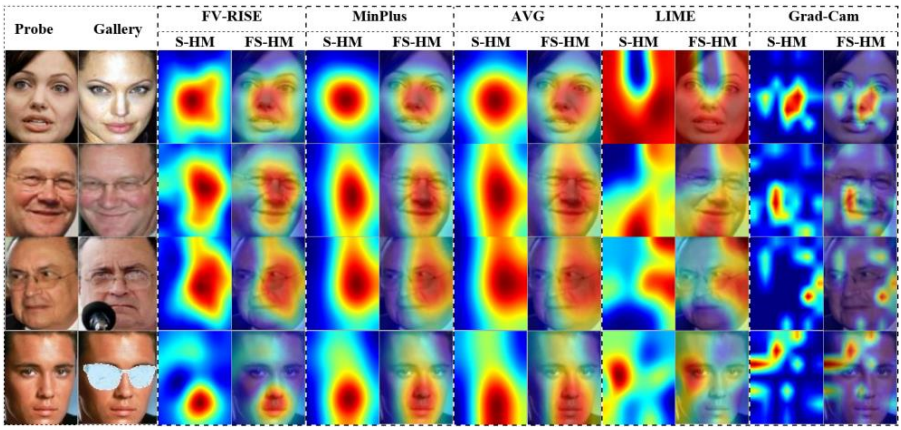



Fig. 4: (F)S-HMs comparison for five FV explainability methods including four benchmarks.

### 4.4 Quantitative explainability performance

This section reports the quantitative explainability performance assessment for the FV-RISE method, notably in comparison with the four considered benchmarks, using the Recall metric when the faces are manipulated through the so-called *deletion and insertion* processes proposed in [PeDS18] and adapted for the FV task.

Since the benchmarks do not produce D-HMs, and thus cannot deal with impostors' cases, a fair comparison requires that the comparison is performed using only the S-HMs. This motivates the selection of the Recall metric which is computed as the ratio between the number of True Acceptances and the sum of the number of True Acceptances with the number of False Rejections. The deletion and insertion processes manipulate the probe image by sequentially deleting or inserting the most important pixels in the probe image and measuring the impact on the FV model's performance. Specifically, the deletion process starts with the original probe image and successively masks/removes a growing percentage of pixels with the highest S-HM importance in the original probe image; the similarity score for the new probe-gallery pair is computed for each percentage of important pixels removed. On the other hand, the insertion process starts with a black image, and successively inserts a growing percentage of pixels with the highest S-HM importance in the starting black image; the similarity score for the new probe-gallery pair is computed for each percentage of important pixels inserted.

The intuition behind this performance assessment method is that the S-HM reliably highlights the most important face regions with an importance score for the FV decision with the selected FV model. In this context, the faster the Recall drops/rises for the deletion and insertion metrics, respectively, the more accurate are the S-HMs in terms of explainability power.

For the experiments reported in this subsection, the deletion and insertion processes are conducted on a subset of the LFW dataset with 1000 genuine pairs and the Recall for the FV model is measured for the whole subset for each percentage of pixels deleted or inserted. Fig. 5 shows the Recall for FV-RISE and four benchmarks, notably for the deletion and insertion processes on the left and right, respectively. The analysis of the Recall curve on the left shows that the FV model's Recall decreases more rapidly for FV-RISE than for the benchmarks while removing the most important pixels; in the same direction, the analysis of the Recall curve on the right shows that the FV model's Recall increases more rapidly for the FV-RISE than for the benchmarks while inserting the most important pixels. Both behaviors allow concluding that the proposed FV-RISE method generates the most accurate explainability S-HMs for the used LFW subset. It can also be observed that MinPlus shows a good performance among the four considered state-of-the-art explainability benchmarks.

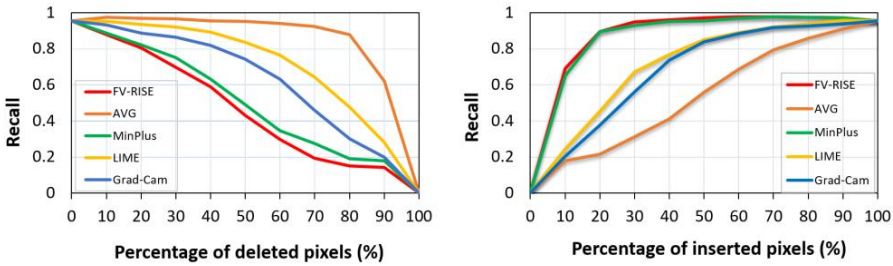



Fig. 5: Recall versus percentage of manipulated pixels: deleted (left) and inserted pixels (right).

## 5 Conclusions and future work

This paper adapts the RISE algorithm to propose a novel FV explainability method which is able to explain any type of FV decision. More specifically, the proposed FV-RISE method generates S-HMs and D-HMs to explain the True/False Acceptance and True/False Rejection decisions, respectively. The experimental results show that the proposed method generates qualitatively reliable FV decision explanations for any FV case; moreover, it is quantitatively shown that the FV-RISE S-HMs-based explanations are more accurate when compared with relevant state-of-the-art explainability benchmarks.

A potential future direction would be to evaluate and compare the proposed FV-RISE method performance with multiple FV models and similarity metrics. Additionally, the proposed FV-RISE method can be adapted to also explain face identification decisions.

## Acknowledgements

This work has been partially supported by the European CHIST-ERA program via the French National Research Agency (ANR) within the XAIface project (grant agreement CHIST-ERA-19-XAI-011) and FCT/MEC under the project UID/50008/2020.

## References


[CaBy18] Castanon, G.; Byrne, J.: Visualizing and Quantifying Discriminative Features for Face Recognition. In: IEEE International Conference on Automatic Face & Gesture Recognition, Xi'an, China, 2018.

[DGVK20] Deng, J. et.al.: RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 2020.

[GZHH16] Guo, Y. et.al.: MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In: Leibe, B.; Matas, J.; Sebe, N.; Welling, M. (Hrsg.): Computer Vision – ECCV 2016, Lecture Notes in Computer Science. Bd. 9907. Cham: Springer International Publishing, Amsterdam, The Netherlands, 2016.

[HWWJ21] Huang, B. et.al.: When Face Recognition Meets Occlusion: A New Benchmark. In: IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto, ON, Canada, 2021.

[KTHR23] Knoche, M. et.al.: Explainable Model-Agnostic Similarity and Confidence in Face Verification. In: IEEE/CVF Winter Conference on Applications of Computer Vision Workshops. Waikoloa, HI, USA, 2023.

[MeMo22] Mery, D.; Morris, B.: On Black-Box Explanation for Face Verification. In: IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, HI, USA, 2022.

[Mery22] Mery, D.: True Black-Box Explanation in Facial Analysis. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New Orleans, LA, USA, 2022.

[Mery23] Mery D., MinPlus XAI Facial Analysis, https://colab.research.google.com/drive/1AL2aEEyZOWJTyTaspFQcry_1g0E4b4x5?usp=sharing, Stand: 19.08.2023.

[Moln19] Molnar, C.: Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2end Edition, Independently Published, 2019.

[PeDS18] Petsiuk, D.; Das, A.; Saenko, K.: RISE: Randomized Input Sampling for Explanation of Black-box Models. arXiv:1806.07421v3 (2018).

[RiSG16] Ribeiro, M. et.al.: Why Should I Trust You?: Explaining the Predictions of Any Classifier. arXiv:1602.04938v3 (2016).

[RMBL08] Huang, G. B. et.al.: 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. Marseille, France, 2008.

[SCDV17] Selvaraju, R. R. et.al.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: IEEE International Conference on Computer Vision. Venice, Italy, 2017.

[YTLS19] Yin, B. et.al.: Towards Interpretable Face Recognition. In: IEEE/CVF International Conference on Computer Vision. Seoul, South Korea, 2019.

[ZhDe18] Zheng, T.; Deng, W.: Cross-pose LFW: A Database for Studying Cross-pose Face Recognition in Unconstrained Environments. Beijing University of Posts and Telecommunications, Technical Report, 2018.

[ZhYC21] Zhu, S.; Yang, T.; Chen, C.: Visual Explanation for Deep Metric Learning. In: IEEE Transactions on Image Processing, vol. 30, S. 7593-7607, 2019.

[Zile23] Zilence J., Visual Explanation for Deep Metric Learning Official Implementation, https://github.com/Jeff-Zilence/Explain_Metric_Learning, Stand: 19.08.2023.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# Unified Face Image Quality Score based on ISO/IEC Quality Components

Praveen Kumar Chandaliya, Kiran Raja, Raghavendra Ramachandra, Christoph Busch [1]

**Abstract:** Face image quality assessment is crucial in the face enrolment process to obtain high-quality face images in the reference database. Neglecting quality control will adversely impact the accuracy and efficiency of face recognition systems, resulting in an image captured with poor perceptual quality. In this work, we present a holistic combination of 21 component quality measures proposed in "ISO/IEC CD 29794-5" and identify the varying nature of different measures across different datasets. The variance is seen across both capture-related and subject-related measures, which can be tedious for validating each component metric by a human observer when judging the quality of the enrolment image. Motivated by this observation, we propose an efficient method of combining quality components into one unified score using a simple supervised learning approach. The proposed approach for predicting face recognition performance based on the obtained unified face image quality assessment (FIQA) score was comprehensively evaluated using three datasets representing diverse quality factors. We extensively evaluate the proposed approach using the Error-vs-Discard Characteristic (EDC) and show its applicability using five different FRS. The evaluation indicates promising results of the proposed approach combining multiple component scores into a unified score for broader application in face image enrolment in FRS.

**Keywords:** Biometrics, ISO/IEC face quality components, Face recognition system, Face image quality assessment.

## 1 Introduction

Owing to their convenience, unobtrusiveness, and enhanced performance, Face Recognition Systems (FRS) have become widely adopted in recent years for applications such as forensic investigations and border controls. As these systems have become a cornerstone element in our security infrastructure, their reliability is very important. The performance of facial recognition systems depends on the quality of the images presented to them. Reference face images of higher quality are expected to support better recognition performance, and poor-quality images can degrade the performance of these systems on all tasks. Assessing quality itself remains a challenge [Me22, CAN23]. Several studies in the existing literature have considered dealing with low-quality images and developing robust FRS to account for low quality [CY23]. While this is a positive aspect of technology advancement, enrolment systems need high-quality images for different use cases. For instance, a high-quality face image is needed in the passport application process, which a human expert (e.g., passport issuing officer) can also use to verify the identity and confirm the pre-set quality standards. Furthermore, a low-quality face image can lead to incorrect decisions

[1] Norwegian University of Science and Technology (NTNU), Gjøvik, Norway
{praveen.k.chandaliya, kiran.raja, raghavendra.ramachandra, christoph.busch}@ntnu.no

owing to perceptible face regions (occlusion or bad illumination) [Sc22, CSN20]. ISO/IEC

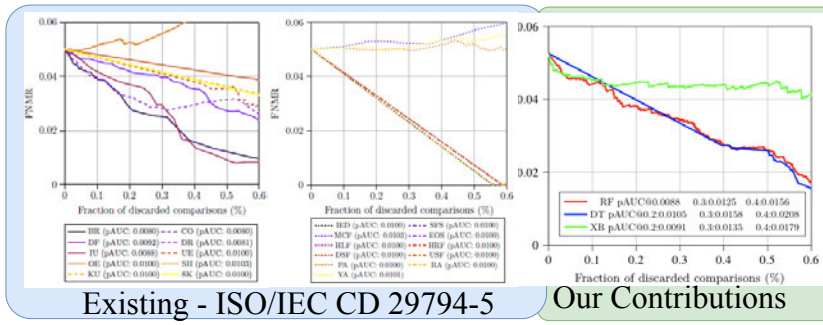



Fig. 1: Contributions of this work in unifyed the ISO/IEC CD 29794-5 quality components

29794-5:2023 [IS23] is intended to standardise face image quality measures and therefore categorizes factor-specific measures as subject-related or capture-related. ISO/IEC 29794-5:2023 [IS23] proposes independent quality components to assess different aspects of a face image such as Brightness (BR), Constrast (CO), Defocus (DF), Dynamic Range (DR), Illumination Uniformity (IU), Under-exposure (UE), Over-exposure (OE), Sharpness (SH), Kurtosis (KU), Skewness (SK), Inter eye distance (IED), Single face subject (SFS), Mouth closed (MCF), Eyes open (EOS), Horizontal left shift face (HLF), Horizontal right shift face (HRF), Vertical down shift face (DSF), Vertical upward shift face (USF), Pitch angle (PA), Roll angle (RA), Yaw angle (YA). However, recent deep learning systems provide a holistic quality measure for face images. Thus, developing a unified quality score for ISO/IEC CD 29794-5 makes it possible for operational systems to obtain one score (e.g., similar to NFIQ for fingerprints [Ta21] or to compare it against a unified score of DL-based systems. In this work, we combined the quality components of ISO/IEC CD 29794-5 to a single quality score using well-tested machine learning techniques (MLT) such as Random Forest (RF), Decision Tree (DT), and XGBoost (XB). We demonstrated that RF, DT, and XGBoost can be used to obtain a unified score for different databases, FRS, and FIQA measures. We assert the validity of the idea by evaluating it on three different public face datasets such as Labeled Faces in the Wild (LFW) [Hu07], XQLFW [KHR21], and color FERET [Ph98] to cover different kind of use cases. The contributions of this work are summarized as follows:

- a new approach towards a unified face image quality score using ISO/IEC 29794-5 that include 10 capture-related measures and 11 subject-related measures as shown in Figure 1.
- an extensive evaluation of the proposed method for obtaining a unified score as a predictor for the FRS performance using the Error-vs-Discard Characteristic (EDC).
- the evaluation is demonstrated on three diverse datasets with various quality factors using state-of-the-art supervised and unsupervised FIQA and diverse FRS.

In the rest of the paper, we present a set of related works in Section 2. Section 3 presents a detailed method description. Section 4 presents an experimental setup and evaluation.

Section 5 presented the results. Section 6 discusses the limitations of this work. In the final Section 7, the conclusion is discussed.

## 2 Related Work

While there exist a number of works for estimating the quality of face images, in this section, we present the most relevant works related to our work. A recent survey of this comprehensive picture can be found in [Sc22]. A set of standards has been proposed to ensure face image quality by constraining capture requirements, such as ISO/IEC 39794-5 [IS19] and ICAO 9303 [In21]. Assessment of face images quality is typically divided into capture-related measures that are affected by external circumstances caused by the capture device (such as brightness, illumination, and motion blur) and subject-related measures (such as facial expression, pose, and occluded facial parts) [Sc22].

FIQA approaches that include supervised learning algorithms based on human or artificially constructed quality labels have become increasingly popular because of their performance [BVS13, ZZL17, KGV20, RM14, Wa17]. The utilized algorithms include cumulative distribution with an SVM-based approach [BVS13], Spearman and Kendall rank-order correlation coefficient-based learning [ZZL17], Gaussian function-based de-focus, and motion blur intensity [KGV20]. Wasnik et al.[Wa17] examined FIQA in the context of smartphone-based FR, evaluating eight FIQAs specifications and proposed a vertical edge density FIQA for lighting and pose symmetry.

However, human perception may not always correlate with the details sought by the FRS and utility values derived from comparison scores. They rely on an error-prone labeling approach and require large-scale training datasets. SER-FIQ [Te20] is an unsupervised learning-based method that measures the face recognition model-specific quality by comparing the output embeddings of several randomly chosen sub-networks without requiring any ground truth quality score training labels. Supervised learning-based MagFace approach from Meng et al.[Me21] integrates FIQA within the FRS. This approach works by extending ArcFace [De19] training loss, changing the angular margin to a magnitude-aware angular margin, and adding magnitude regularization. Another supervised learning-based CR-FIQA is a recent face image quality assessment method introduced by Boutros et al.[Bo23], which estimates the quality of a facial image by predicting its relative classifiability. The classifiability of an image is measured based on the location of the feature representation in the angular space with respect to its class center and the nearest negative-class center. However, these methods incur additional computational costs or network blocks, which complicates their use in conventional face systems. So far, research on FIQA has directly utilized the FRS model during FIQA model inference without FIQA model training on ground truth quality scores. On the other hand, hybrid FRS/FIQA approaches simultaneously train FRS and FIQA as part of a single integrated framework, generating both face recognition and quality assessment output during inference.

### 2.1 ISO/IEC 29794-5 and Unified FIQA Score

While holistic FIQA scores are based on DL methods, ISO/IEC 29794-5 and its reference implementation OFIQ will typically be used in operational settings for various purposes, such as enrolment into national ID databases. Further, looking into each qaulity component can be tedious, and obtaining a baseline score decreases the labor involved in rejecting an image or in understanding the component on which the captured subject has to act. Thus, there is a need to combine the component measures into a unified score, making it compatible with DL-based FIQA without explicit ground truth annotation. A possible solution is to use the quality score provided by DL systems as the ground truth to create good and bad classes corresponding to the face quality vector obtained from different component measures. Using the ground truth provided by the DL-based FIQA, a classifier can be trained to obtain a unified score from the component measures.

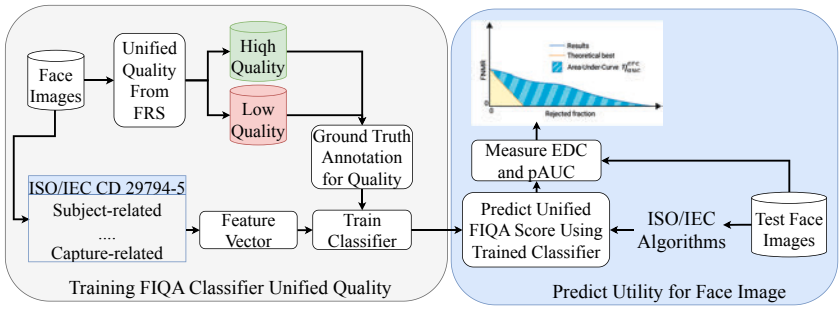



Fig. 2: Proposed approach for overall quality assessment.

## 3 Proposed Method

We propose an approach parallel to NFIQ2.0 to estimate the quality of face images. Fig 2 shows the steps in calculating the overall quality score for an input face image using 21 face image quality components for ISO/IEC 29794-5:2023. The first step is uncontrolled face detection using an InsightFace-based RetinaFace [De20]. In the second step, alignment, cropping, and resizing were performed using InsightFace-based ArcFace [De19]. The aligned image is further used to obtain 21 quality measures from ISO/IEC CD 29794-5, and can be represented as $F_{qc} = \{f_1, f_2 \dots f_{21}\}$ for all independent components. Each native quality measure is further normalized to a uniform range of 0 to 100 where a low value represents poor quality and a high value represents better image quality. However, the estimated component measures cannot individually indicate the quality or utility of a face image. Therefore, we used an auxiliary FRS that can estimate face quality. The obtained quality score from a DL-based system was further binned as good quality and bad quality for ground truth annotation in training a supervised classifier. For the sake of simplicity, we consider a hard threshold of normalized quality score of 70 to decide as good quality or bad quality. In particular, we used the FIQA algorithms CR-FIQA, MagFace,

and SER-FIQ independently and all of them can provide a unified score. The $F_{qc}$ consisting of 21 component measures was then estimated for the test set to establish the utility of the face image using the trained classifier. The utility was further measured using EDC and pAUC for various FRS such as FaceNet, ArcFace, PFE, MagFace, and ElasticFace.

## 4 Experimental setup and Performance Evaluation Metrics

We analyzed the performance of our proposed fusion method using three state-of-the-art FIQA models: CR-FIQA [Bo23] which adds regression networks to the recognition models for learning identity quality; MagFace [Me21] which associates the magnitude of face embeddings with face quality; and SER-FIQ [Te20] which employs several sub-networks of a recognition model to generate quality scores. We utilized pretrained models to extract embeddings for our analysis [Bo23, Me21, Te20]. Further, we conduct all our face-quality assessment experiments on LFW [Hu07], XQLFW [KHR21], and color FERET [Ph98] publicly available datasets that represent varying quality and diversity to study the generalization of the proposed approach.

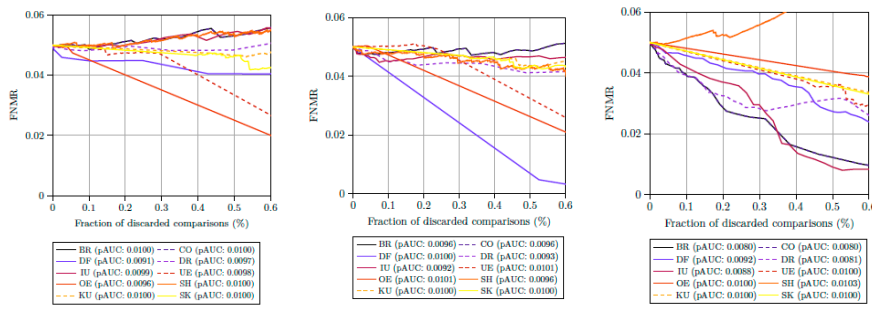
### 4.1 Evaluation metrics

To evaluate the face quality assessment algorithm performance, we employ the "Error versus Discard Characteristic" (EDC) standardized by ISO/IEC 29794-1[3] and the consequent partial Area Under the Curve (pAUC) values are reported [Sc23]. Furthermore, the EDC curves are plotted at a fixed FMR 0.1% as recommended for border control operations by Frontex [4]. EDC curves measure the performance of a given FRS when the percentage of the lowest-quality face images is discarded. Because discarding a large portion of all images is not a practical application scenario, we are typically interested in lower discard rates. Therefore, we report the partial area under the curve (pAUC) of the EDC at a discard rate of 20% for an FNMR of 0.05 starting error [Sc23].

## 5 Experiments and Results

In the following section, we present the results of our experiments conducted to investigate: (i) The performance of the implemented 21 ISO/IEC 29794-5 capture and subject-related measures was evaluated using three datasets: LFW, XQLFW, and color FERET. (ii) Performance analysis between CR-FIQA, MagFace, and SR-FIQ techniques and the impact of ArcFace, FaceNet, PFE, MagFace, and ElasticFace FRS models with the proposed FIQA quality measure fusion approach on three datasets. The evaluation of FIQA algorithms depends on face verification error rates. To evaluate the generalization of the methods, we investigate how well the quality components are generalising for five different state-of-the-art FRS to report the verification performance at different discard fractions

[3] https://www.iso.org/standard/79519.html
[4] Best practice technical guidelines for automated border control (abc) systems




(a) LFW (b) XQLFW (c) color FERET

Fig. 3: EDC and pAUC score at 0.2 discard rate on different datasets result of ISO capture quality components.

to inspect the generalizability of FIQA over FRS. We choose FaceNet [SKP15] based on softmax loss, Probabilistic Face Embeddings (PFE) [SJ19] based on Gaussian distribution, ArcFace[De19]. In addition, we also analyzed MagFace [Me21] and ElasticFace [Bo22] as both are based on adaptive angular marginal loss. For each of the five FRS, the images were prepossessed, as described in the corresponding reference. The embedding was extracted from the last layer of each model, and cosine similarity was used to generate comparison scores for face verification experiments.

### 5.1 ISO/IEC CD 29794-5 Quality Measures

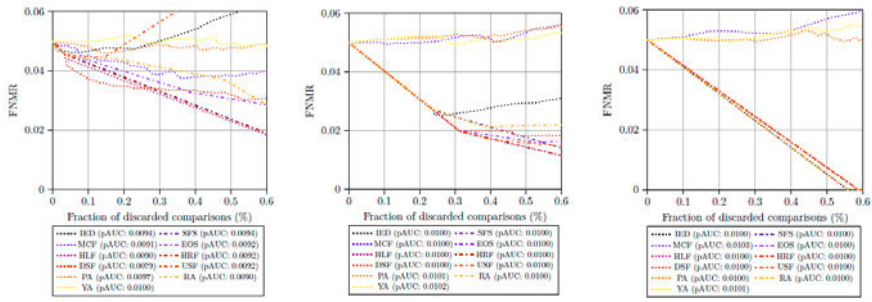
First, we report the performance of each of the 21 ISO/IEC 29794-5 quality measures, which include both capture- and subject-related measures using EDC curves and pAUC, as shown in Fig 3 (capture related) and Fig 4 (subject related) for the LFW, XQLFW, and color FERET datasets, respectively. We make the following observations by analyzing the independent component measures.

### 5.2 Capture related measures

- For LFW in Fig 3a Skewness (SK) and Kurtosis (KU) FNMR decreases slowly and drops around the 50% discard rate. Brightness (BR), Contrast (CO), Illumination Uniformity (IU), Dynamic Range (DR), and Sharpness (SH) demonstrate the same behavior where the error rate constantly increases, which shows that on the LFW dataset a non-correlation to utility. The error rates for Under-exposure (UE), Over-exposure (OE), and Defocus (DF) have a steady decrease showing correlation to the measure as a utility indicator.
- For XQLFW We can observe that the FNMR for DF, OE, and UE decrease steadily indicating the utility as shown in Fig 3b. The error rates of BR, IU, Kurtosis KU, and CO have the similar characteristic and remain constant regardless of the discard

rate indicating no correlation to utility from these measures for XQLFW. SK, SH and DR demonstrate the same behaviour where the error rate increases slowly and drops around the 60% discard rate.

- In Fig 3c, we can see for colorFERET that the FNMR of BR, CO, DF, DR, IU, UE, OE, KU, SK brightness and variance show the same behavior and they decrease quite steadily as the discard rate increases indicating a good correlation as a utility predictor on color FERET dataset. However, the FNMR for SH increased after 20% discard rate, indicating no correlation for utility on the color FERET dataset.




(a) LFW (b) XQLFW (c) color FERET

Fig. 4: EDC and pAUC score at 0.2 discard rate on different datasets result of ISO subject quality components.

### 5.3 Subject related measures

The EDC plots for subject-related measures are provided in Fig 4 for LFW, XQLFW, and color FERET dataset, respectively. We make the following observations as noted below:

- For LFW we can observe the FNMR values for Vertical Down Shift Face (DSF) and Mouth Closed (MCF) remain constant after a slight drop at the beginning indicating a weak correlation to utility. The error rate remains relatively unchanged for Inter eye distance (IED), Vertical upward shift face (USF), Yaw angle (YA), and Pitch angle (PA) before a steady increase after a discard ratio of 20% indicating no strong correlation with utility. However, the FNMR for Roll angle (RA), Single face subject (SFS), Horizontal right shift face (HRF), and Eyes open (EOS) components steadily decrease indicating a strong relationship between utility and FNMR in the case of LFW dataset.
- For the XQLFW dataset we can further observe the FNMR for MCF, YA, and DSF increase steadily as the proportion of discarded images increase indicating no strong correlation with utility for XQLFW dataset as shown in Fig 4b. However, a moderate correlation can be observed for IED upto 30% discard and no significant relation beyond. In addition, RA, DSF, USF, SFS, MCF, EOS, the error rate decreases sharply, demonstrating a strong correlation between quality components and utility.

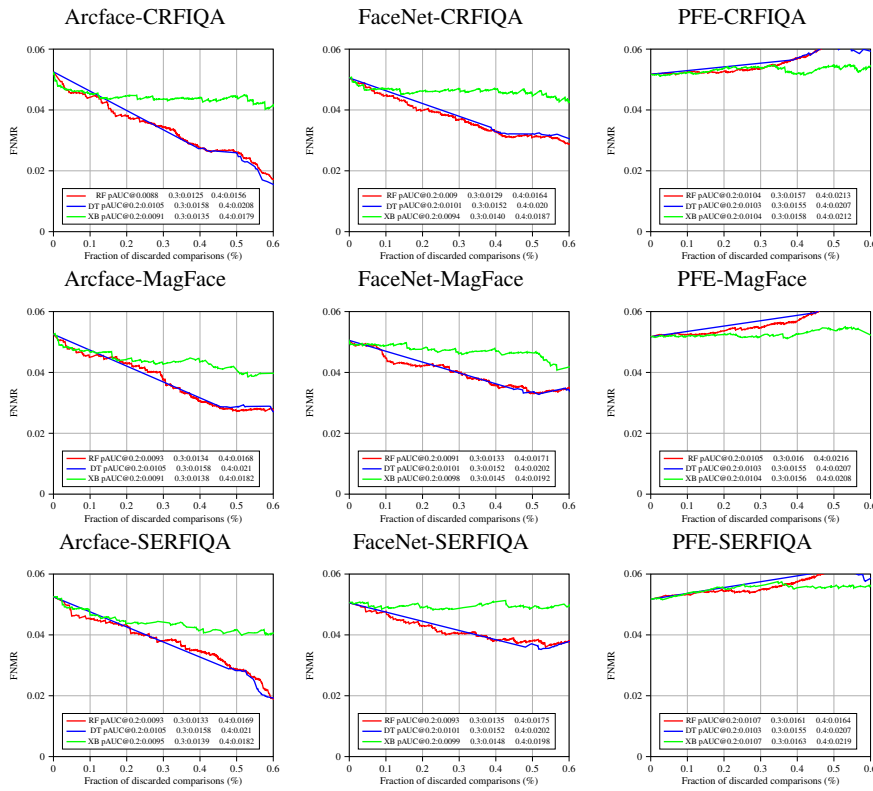



Fig. 5: Performance evaluation of ArcFace, FaceNet, and PFE (left to right) on CRFIQA, MagFace and SERFIQ (top to bottom): EDC and pAUC scores at 0.2, 0.3, and 0.4 discard rates on colorferet dataset.

- Fig 4c shows that the FNMR of the RA, HRF, USF, MCF, SFS, EOS, and IED have similar characteristic upto 40% discard rate before an increase indicating no correlation to utility for color FERET dataset. YA, PA, EOS, HLF, and DSF components on the other hand decrease the FNMR with increased discard ration suggesting a strong correlation of these component measures with utility.

### 5.4 Results on Unified Score based on Quality Components

As noted previously, we observe varying impact of component quality measures on different datasets. We therefore present the results of our proposed approach of unifying the component scores to one unified score as shown in Fig 5 for ColorFeret dataset [5]. Fig 5 illustrates the EDC curves obtained using three different FRS using three different FIQA. Based on the obtained results, we make the following observations:

- On a general note, we observe that the proposed approach of unifying scores using Decision Tree (DT) and Random Forest (RF) decreases the FNMR with increasing discard ratio. All evaluated supervised methods appear highly effective as a sharp decline in the FNMR can be seen with increasing reject rates with CR-FIQA method performing best across different FRS.
- We further observe ArcFace provides a consistent and low average pAUC score on the color FERET dataset across different FIQA while FaceNet follows a similar trend but relatively lower in performance as compared to ArcFace as FRS.
- While we note the pAUC value of the proposed approach as 0.0088 at 20% discard rate, certain quality measures outperform quality ground truth estimated using CR-FIQA when used with ArcFace the ColorFERET dataset.
- While ArcFace and FaceNet perform well on ColorFERET dataset, PFE FRS tends to perform relatively poorly indicating the need for further investigations.

## 6 Limitations of our work

Our approach generally scales well on estimating unified score from component quality measures. However, we note that newer FRS architectures such as MagFace and ElasticCosFace do not contribute in decreasing the FNMR with an increased discard ratio demanding further investigations (See Fig 6 in the supplementary material). In similar lines, Commercial-Off-The-Shelf (COTS) FRS have not been studied in this work. Further, certain inconsistencies in the trend can be observed LFW and XQLFW datasets (illustrated in the supplementary section) as compared to ColorFERET dataset. The inconsistencies can lead to highly inaccurate predictions of quality and this will be investigated in the future works.

[5] Due to the page constraints, we illustrate the result of LFW and XQLFW datasets in the supplementary section.

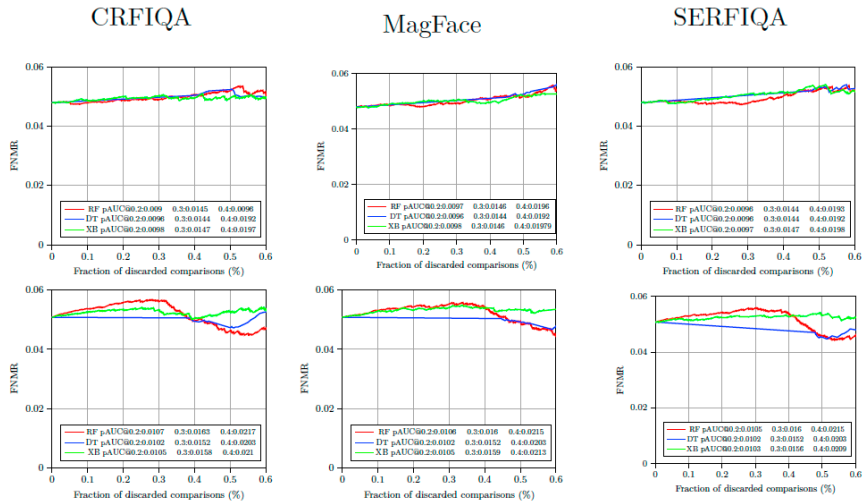



Fig. 6: Limitation of our approach on CRFIQA, MagFace and SERFIQ (left to right) on ElasticFace and MagFace (top to bottom): EDC and pAUC scores at 0.2, 0.3, and 0.4 discard rates on colorferet dataset

## 7 Conclusion

While the quality components proposed in ISO/IEC CD 29794-5 can measure different quality aspects, it is tedious for a human observer to analyze different values. We presented an efficient method for an unified FIQA score using 21 different component measures proposed in ISO/IEC CD 29794-5. The obtained score, can act as a predictor of FRS performance. The experiments conducted on three different datasets using five different FRS indicate a promising method as it can be observed performance using EDC. However, the invariance of the proposed approach to some recent deep-learning-based FRS architectures remains an open research question and will be studied in future works.

## 8 Acknowledgement

This work was supported by the European Union's Horizon 2020 Research and Innovation Program under Grant 883356. We thank Torsten Schlett for sharing the FIQA framework and giving insightful comments on this work.

## References

[Bo22] Boutros, Fadi; Damer, Naser; Kirchbuchner, Florian; Kuijper, Arjan: ElasticFace: Elastic Margin Loss for Deep Face Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 1578–1587, June 2022.

[Bo23] Boutros, Fadi; Fang, Meiling; Klemt, Marcel; Fu, Biying; Damer, Naser: CR-FIQA: Face Image Quality Assessment by Learning Sample Relative Classifiability. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5836–5845, June 2023.

[BVS13] Bharadwaj, Samarth; Vatsa, Mayank; Singh, Richa: Can holistic representations be used for face biometric quality assessment? In: 2013 IEEE International Conference on Image Processing. pp. 2792–2796, 2013.

[CAN23] Chandaliya, Praveen Kumar; Akhtar, Zahid; Nain, Neeta: Longitudinal Analysis of Mask and No-Mask on Child Face Recognition. In: Proceedings of the Thirteenth Indian Conference on Computer Vision, Graphics and Image Processing. volume 57, 2023.

[CSN20] Chandaliya, Praveen Kumar; Sinha, Aditya; Nain, Neeta: ChildFace: Gender Aware Child Face Aging. In: 2020 International Conference of the Biometrics Special Interest Group (BIOSIG). pp. 1–5, 2020.

[CY23] Chen, Zehao; Yang, Hua: L2RT-FIQA: Face Image Quality Assessment via Learning-to-Rank Transformer. In (Zhai, Guangtao; Zhou, Jun; Yang, Hua; Yang, Xiaokang; An, Ping; Wang, Jia, eds): Digital Multimedia Communications. Springer Nature Singapore, Singapore, pp. 270–285, 2023.

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

[De20] Deng, Jiankang; Guo, Jia; Ververas, Evangelos; Kotsia, Irene; Zafeiriou, Stefanos: RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5202–5211, 2020.

[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.

[In21] International Civil Aviation Organization: , Machine Readable Passports – Part 1 – Introduction. http://www.icao.int/publications/Documents/9303_p1_cons_en.pdf, 2021.

[IS19] ISO/IEC JTC1 SC37 Biometrics: . ISO/IEC 39794-5:2019 Information technology - Extensible biometric data interchange formats - Part 5: Face image data. International Organization for Standardization, 2019.

[IS23] ISOIEC: ISO/IEC 29794-5 Information technology — Biometric sample quality — Part 5: Face image data. ISO/IEC CD 29794-5, pp. 1–62, 2023.

[KGV20] Kumar, Himanshu; Gupta, Sumana; Venkatesh, K. S.: Simultaneous Estimation of Defocus and Motion Blurs From Single Image Using Equivalent Gaussian Representation. IEEE Transactions on Circuits and Systems for Video Technology, 30(10):3571–3583, 2020.

[KHR21] Knoche, Martin; Hoermann, Stefan; Rigoll, Gerhard: Cross-Quality LFW: A Database for Analyzing Cross- Resolution Image Face Recognition in Unconstrained Environments. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). pp. 1–5, 2021.

[Me21] Meng, Qiang; Zhao, Shichao; Huang, Zhida; Zhou, Feng: MagFace: A universal representation for face recognition and quality assessment. 2021.

[Me22] Mendez-Llanes, Nelson; Castillo-Rosado, Katy; Méndez-Vázquez, Heydi; Tistarelli, Massimo: On the Use of Local Fixations and Quality Measures for Deep Face Recognition. IEEE Transactions on Biometrics, Behavior, and Identity Science, 4(2):150–162, 2022.

[Ph98] Phillips, P.Jonathon; Wechsler, Harry; Huang, Jeffery; Rauss, Patrick J.: The FERET database and evaluation procedure for face-recognition algorithms. Image and Vision Computing, 16(5):295–306, 1998.

[RM14] Rui Min, Abdenour Hadid, Jean-Luc Dugelay: Efficient Detection of Occlusion prior to Robust Face Recognition. The Scientific World Journal, pp. 1–110, 2014.

[Sc22] Schlett, Torsten; Rathgeb, Christian; Henniger, Olaf; Galbally, Javier; Fierrez, Julian; Busch, Christoph: Face Image Quality Assessment: A Literature Survey. ACM Comput. Surv., 54(10s), sep 2022.

[Sc23] Schlett, Torsten; Rathgeb, Christian; Tapia, Juan; Busch, Christoph: , Considerations on the Evaluation of Biometric Quality Assessment Algorithms, 2023.

[SJ19] Shi, Yichun; Jain, Anil K.: Probabilistic Face Embeddings. 2019.

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

[Ta21] Tabassi, E.; Olsen, M.; Bausinger, O.; Busch, C.; Figlarz, A.; Fiumara, G.; Henniger, O.; Merkle, J.; Ruhland, T.; Schiel, C.; Schwaiger, M.: NIST Interagency Report 8382. NIST Interagency Report 8382, National Institute of Standards and Technology, July 2021.

[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: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. IEEE, pp. 5650–5659, 2020.

[Wa17] Wasnik, Pankaj; Raja, Kiran B.; Ramachandra, Raghavendra; Busch, Christoph: Assessing face image quality for smartphone based face recognition system. In: 2017 5th International Workshop on Biometrics and Forensics (IWBF). pp. 1–6, 2017.

[ZZL17] Zhang, Lijun; Zhang, Lin; Li, Lida: Illumination Quality Assessment for Face Images: A Benchmark and a Convolutional Neural Networks Based Model. In: Neural Information Processing. Springer International Publishing, pp. 583–593, 2017.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# Assessing the Human Ability to Recognize Synthetic Speech in Ordinary Conversation

Daniel Prudký[1], Anton Firc[2], Kamil Malinka[3]

**Abstract:** This work assesses the human ability to recognize synthetic speech (deepfake). This paper describes an experiment in which we communicated with respondents using voice messages. We presented the respondents with a cover story about testing the user-friendliness of voice messages while secretly sending them a pre-prepared deepfake recording during the conversation. We examined their reactions, knowledge of deepfakes, or how many could correctly identify which message was deepfake. The results show that none of the respondents reacted in any way to the fraudulent deepfake message, and only one retrospectively admitted to noticing something specific. On the other hand, a voicemail message that contained a deepfake was correctly identified by 83.9% of respondents after revealing the nature of the experiment. Thus, the results show that although the deepfake recording was clearly identifiable among others, no one reacted to it. In summary, we show that the human ability to recognize voice deepfakes is not at a level we can trust. It is very difficult for people to distinguish between real and fake voices, especially if they do not expect them.

**Keywords:** deepfake, synthetic speech, artificial intelligence, cybersecurity, deepfake detection

## 1 Introduction

Mirsky and Lee [ML21] define a deepfake simply as a *"Believable media generated by a deep neural network."* A more extensive definition says it is media created by artificial intelligence (AI), specifically using deep neural networks through deep learning (DL) methods. In their production, artificial intelligence merges combines, replaces, or overlays features of the media to create new fake representations of things that never happened. This media can be practically unnoticeable from authentic ones. Deepfake technology brings many benefits, it can be used for entertainment, but it can also be used for revenge porn, bullying, spreading fake news, political sabotage and more [FM22, FMH23, We19].

Nowadays, these fake media are reaching a stage where they are not even recognizable by machines, let alone humans, who may not even be aware of the existence of such threats in today's digital world. Moreover, within audio, it is no longer just about English models. Many multi-language tools for creating voice deepfakes are being developed, and they can appear in almost any language.

There have been many attack scenarios in which deepfakes have been used. For example, they could be attacks targeting specific individuals or institutions in the form of vishing

[1] Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic, xprudk08@stud.fit.vut.cz
[2] Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic, ifirc@fit.vut.cz
[3] Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic, malinka@fit.vut.cz

or widespread disinformation to spread propaganda, and so on. People should be able to defend themselves against this spread of fraudulent information and media, and they should know how to verify such things and how to deal with them. But we don't know if we will ever be able to do that.

A recently widespread method is vishing, derived from the two words defining it: voice and phishing. It is a version of phishing in which identity theft is carried out using voice devices such as the telephone, voice assistant, etc. Its use is described by Firc et al. [FMH23]. The authors point out that one such attack happened in 2019 when a fraudster using deepfakes created a transaction of almost $250,000. The CEO of an energy company thought he was talking to his boss on the phone, and when the caller asked him for an urgent transfer of this money, the victim did not hesitate and sent the money, believing he was completing a task from his boss. There are many cases like this today. The same article says that vishing was reported by 69% of companies in 2021, a big increase from 2020, when 54% of companies reported it. Spoofing is also very often associated with this scam, giving the scam much more credibility. For example, the fraudster can call the victim from the real phone number of the person they are playing. Spoofing can also be used with the phone number of a bank, the police, etc. In this way, the attackers try to exert authority on the victim, who is then more likely to disclose the required information in fear.

Therefore, this work aims to assess the human ability to recognize synthetic speech. There have been several attempts to assess whether people can distinguish a deepfake from a real one. However, these experiments first introduced participants to the deepfake problem before exposing them to deepfakes. Their results are quite variable and vary mainly depending on the methodology. For example, voice deepfakes have been tested in a survey by Müller et al. [MPW22], who report that the accuracy of identifying a deepfake and a genuine recording is 80%. In contrast, our approach first exposes the respondents to deepfakes and then asks if they noticed anything unusual or if they can identify the deepfake set among other sets.

The whole experiment is hidden behind a cover story of testing the usability of voice messaging. Respondents play the game *Two Truths One Lie*. They receive 5 voice messages from the narrator, each containing three facts about a selected country. One of these facts is incorrect, and the respondent's task is to identify the incorrect fact and report it back (using the voice message). This setup simulates communication using voice messages only. One of these sets was pre-prepared as a deepfake recording of the narrator's voice. At the end of the experiment, each respondent was sent a questionnaire asking about their knowledge of and attitude towards deepfakes, if they observed anything unusual during the conversation, and ultimately revealed the true nature of the experiment and asked if they could now identify the deepfake set. The work described in this paper results from a previously completed bachelor's thesis [Pr23].
The main contributions might be stated as follows:

- We assess the human ability to recognize deepfakes in the Czech language.
- We show that people cannot distinguish real and deepfake speech in common conversation.

## 2 Related work

The research that deals with detecting voice deepfakes by humans is described in the scientific article by Müller et al. [MPW22]. The authors focused on the ratio of the success rate of deepfakes detection by humans and artificial intelligence. The experiment compared human and machine detection capabilities, using a game-based challenge in which the respondent always played a recording and then determined whether it was fake or real. They made the same decision with machine learning models. For the experiment, the authors used the ASVspoof 2019 dataset, created for the ASVspoof 2019 Challenge, which aims to test Automatic Speaker Verification (ASV) systems resistant to spoofing attacks. Through the experiment, the authors found that the human ability to recognize deepfake and real recordings reaches 80%. Further, the experiment found that recordings created using TTS fooled humans much more than voice-conversion or waveform concatenation systems. The authors believe it could be because it used GAN as the waveform generator. Other interesting results are that native speakers handled recognition better than non-native speakers. At the same time, the level of IT experience did not affect performance, and people's ability to recognize deepfakes decreases with age. It is also interesting to note that people learned very quickly, and as the article says, after the first ten rounds, the success rate improved from 67% to 80%, but promptly stabilised at those levels and did not improve.

Other works focus mainly on deepfakes in the form of images and videos. The success rate of respondents in these experiments varies depending on the methodology and dataset used in the experiment itself. For experiments with deepfake images, success rates for better deepfakes and deepfakes with poorer image quality range roughly between 58-70%, while for poorer deepfakes, respondents have been close to the 90% success rate [Gr21, Go23, Ro19]. In terms of videos, success rates again depended on the quality, and for better quality and harder-to-detect deepfakes, the success rate dropped to the 20% mark, while for lower quality fakes, the success rate reached over 80% [KM20, Gr22, Ta21]. A paper by Tahir et al. [Ta21] describes the training of people in identifying deepfakes, and through a more sophisticated analysis of people's behaviour in identifying deepfakes and other parameters, they were able to develop training procedures that increased the success rate of the trained group by 33%. On average, we're talking about a deepfakes detection success rate of roughly 60-65%, depending on multiple factors.

In all former experiments, the participants knew they were exposed to deepfakes and, therefore, might have targeted it. This is where our research differs very fundamentally from others. Another significant difference is the execution in the Czech language.

## 3 Experiment

The design of the experiment is inspired by Matyáš et al. [Ma08], who propose using a cover story. Moreover, unlike other works, respondents do not know they must reveal deepfakes. Thus, our goal is to create a realistic attack scenario in which we change a real voice, which respondents know and do not consider suspicious, to a deepfake and try to see if they notice this change.

The experiment was conducted in the Czech Republic; therefore, all communication was in the Czech language. This is also related to producing deepfake voices in the Czech language. While most models and tools are suited for the English language, we show the feasibility of other languages that require individual approaches to training and using the speech synthesis models.

### 3.1 Research questions

For the whole experiment, we have identified three main research questions:

**RQ1: Are humans able to identify deepfake recording during casual conversation?**
We are interested in whether people notice during the interview that they have received a computer-generated recording and how they react to it.

**RQ2: Are humans able to detect a deepfake recording among genuine ones?**
We want to determine whether people can retrospectively identify which of the messages in a conversation was a deepfake recording.

**RQ3: What is people's awareness of deepfake technology?**
Given that victim knowledge of deepfakes is critical to detecting these scams, we are interested in how many had heard of the technology or were actively interested in it and what is their experience with deepfakes.

### 3.2 Experiment execution

To synthesize the deepfake set, we use YourTTS [Ca22] in the voice conversion setting. This decision was motivated mainly by the easy access to the tool via a demo on Google Colab and the fact that we possess a version with a trained model in the Czech language. After synthesis, we improved the quality of this set using post-processing. We removed the noise added during creation and smoothed out the frayed phonemes by cutting out the part of the recording where the phonemes resonated. We also adjusted the pitch of the voice. The test run revealed a significant difference in background noise between real (directly spoken) and deepfake (played by speakers) utterances. To diminish this difference and force the participants to focus on the spoken content instead of the background noise, we played brown noise as the background for all the real utterances.

Next, we performed a quality assessment of the synthesized set. We used an evaluation inspired by the *Mean Opinion Score (MOS)* subjective listening test method described by Loizou [Lo11]. We played the recording to 12 experts working with deepfakes regularly. Therefore, we expect their knowledge about deepfake recordings' qualities. Each expert rated the quality on a scale of 1 (poor) to 5 (excellent). The final mean score was 3.0; therefore, the recording qualitatively corresponds to the rating "*Fair*".

As previously mentioned, the experiment was hidden behind a cover story. Participants were presented with simple facts about countries in the form of the *Two Truths One Lie* game. All communication took place within the WhatsApp chat, using voice messages.

Each conversation starts with a brief introduction presenting the pre-prepared cover story, explaining the rules of the experiment, explaining the rules of the game and reminding the respondents that whenever they encounter anything unordinary, they should report it. This is important for our experiment because we need them to report any concerns (mainly about the deepfake set). It is also important for us to get them used to the narrator's voice and to listen to it. We then gradually send them voice messages containing the sets of facts for the game. The respondents listen to these sets and reply with voice messages as well. This way, we send five sets (voice messages), including one pre-prepared deepfake set. If any respondent raises any suspicion or questions about the deepfake set, we refer them directly to the questionnaire. Otherwise, after completing all five sets, we send the respondent a link to the final questionnaire to complete. This questionnaire first collects information about the attitude and knowledge of deepfakes and whether the respondent noticed anything unusual during the experiment (detected the deepfake set). Finally, the questionnaire discloses the true nature of the experiment and that one of the sets is a deepfake and asks the respondents to identify it. When creating the questionnaire, it was important to determine the correct sequence of questions so that the questions could not influence those yet to follow.

## 4 Results

During the experiment, we collected 31 responses. In terms of gender, 71% of respondents were male and 29% were female. The age of the respondents ranges from 18 to 46, but 80% of the values are less or equal to 23, and the average age is about 22.39 years. In focus on the field of work, IT has the highest representation, with 41.9% of respondents. The next common field is education with 19.4%, law and healthcare with 6.5%, and other fields like machinery, marketing, military, art, etc.

All of the research questions have been answered:

**RQ1: Are humans able to identify deepfake recording during casual conversation?**
No one reacted to the deepfake at all during the conversation. One respondent even asked to repeat this set, yet he continued and answered the question as the others did without noticing.

Only one respondent mentioned anything specific about deepfakes before being revealed the true nature of the experiment. This gives us a deepfake detection success rate of 3.2%. 13 respondents mentioned a lower quality of this recording; however, we cannot consider this as successful identification of the deepfake set.

Finally, a third of the respondents told us after the experiment or in their text responses in the questionnaire that the possibility of a fraudulent recording did not occur to them during the interview, and they focused primarily on the content and the correct answer, stating that they considered the lower quality to be normal. These results are summarized in Tab. 1.

| **Reaction during conversation** | |
|---|---|
| Reacted | 0% |
| **Described unnatural things from the conversation** | |
| Poorer audio quality | 41.9% |
| Deepfake sign | 3.2% |

Tab. 1: RQ1 summary.

**RQ2: Are humans able to detect a deepfake recording among genuine ones?**
After revealing that one of the sets is a deepfake, 83.9% of all respondents correctly identified this set. Respondents who marked the deepfake set, along with it other options, are not counted as successful. Counting these responses as successful would result in 96.8% respondents identifying the deepfake set.

54.8% of respondents justify selecting the deepfake set because it was different to others. The second most-stated reason was the lower quality compared to real recordings, as mentioned by 29% of respondents. Finally, the third most-stated reason is the presence of typical deepfake artefacts, mentioned by 22.6% of respondents. Some respondents gave a combination of stated reasons. These results are summarized in Tab. 2.

| **Identify deepfake set** | |
|---|---|
| Marked | 96.8% |
| Correctly identify | 83.9% |
| **Justification for identification** | |
| Different from the others | 54.8% |
| Lower quality than others | 29% |
| Deepfake sign | 22.6% |

Tab. 2: RQ2 summary.

**RQ3: What is people's awareness of deepfake technology?**
Respondents had a choice of three options, 16.1% of respondents answered, "I've never heard of deepfakes", 64.5% answered, "I've heard of deepfakes before", and 19.4% answered, "I'm actively interested in deepfakes". Where they heard about deepfakes is variable but can still be classified into several groups. More than a quarter of people (25.8%) said that they heard about deepfakes on social media, mainly in some informative videos, articles, etc. One respondent said to encounter deepfake videos of politicians on TikTok. Consistently, 19.4% of people wrote that they heard about them on the internet, nothing more specific, or that they heard about them and did not specify where, or tried to create them themselves, which were mainly people in the IT environment. In summary, 83.9% of the participants have at least heard of deepfakes, mainly from social media and informative videos.

Respondents were also asked before and after the experiment how confident they were that they would detect voice deepfakes. They were asked to express this confidence on a scale of 1 (not confident) to 5 (extremely confident). The mean before the experiment was 2.29, and 2.94 after. A total of 51.6% of respondents increased this value, while 45.2% did not

| **Heard of deepfakes** | |
|---|---|
| Heard of them | 64.5% |
| Actively interested | 19.4% |
| Never heard of them | 16.1% |
| **Where they heard about them** | |
| Social media | 25.8% |
| Internet | 19.4% |
| Not specify | 19.4% |
| Create them themselves | 19.4% |
| Never heard of them | 16.1% |

Tab. 3: RQ3 summary.

change it, and only 3.2% decreased it. Younger respondents mainly increased the value of their certainty.

Additionally, after completing the experiment, 74.2% of the respondents said they were surprised by the quality of today's voice deepfake in the Czech language.

### 4.1 Limitations

The major problem of the experiment was the quality of the recordings because of the artificial noise in the background. And although when we played the recordings back (on iPhone 11), the noise was minimal, and we could understand everything without any problems, many people reached back saying that the quality of the recordings was really bad mainly because of the noise. We suppose it depends on the device on which the respondent listened to the recordings; some devices can reduce the noise, while others can't. Poor quality and noise was also the most common thing that respondents identified as odd about the conversation. In total, 13 respondents mentioned the lowered quality.

### 4.2 Results discussion

Related work evaluating human ability often reports more than 60% success rate. The success rate of deepfake detection in our scenario is 3.2%, which is quite different. It is thus important to say that our approach is fundamentally different from the other works. Considering the case where respondents did know they were presented with deepfakes, the success rate of 83.9% is comparable to other research in this field.

These results give an interesting observation. During the conversation, no one responded to deepfakes, but when directly asked to identify the deepfake set, almost every respondent correctly identified it. Many respondents admitted to us that they didn't notice anything on the first listen. Still, when they listened a second time and focused on finding the computer-generated voice, they were immediately sure which one it was. There may

be several reasons for this, but we lean towards something similar to a psychological phenomenon called *The Monkey Business Illusion* [SC10], which states that if people focus on one thing, they are more prone to overlook another, in their opinion, less important things. In our case, it was the answers to the questions and the sound quality. People focused on the right answers and therefore ignored the difference in the voice recordings. However, when we told them to focus on quality and find the deepfake, they detected it easily. These results thus demonstrate how crucial role the knowledge of deepfakes plays in their correct identification and that the education of the broad public on this topic is inevitable.

## 5 Conclusions

This work has shown that the human ability to recognize voice deepfakes is not at a level we can trust. It is very difficult for people to distinguish between real and fake voices, especially if they are not expecting them. The human ability to detect deepfakes is largely influenced by the fact that people don't think about the voice they are listening to, are used to poor-quality audio conversations, and focus primarily on the content of the message.

It is evident that people without any knowledge of deepfakes cannot reliably identify deepfake recordings in conversation. Combined with the Czech language, we show this problem is general and poses a significant threat to society. Moreover, after revealing the presence of a deepfake set, most respondents could identify it. However, this identification was caused by a difference in audio quality or muffled sound compared to the real sets. It is thus important to address these imperfections in future and assess what role the audio quality play in the detection process.

## Acknowledgements

This work was supported by Fakulta Informačních Technologií, Vysoké Učení Technické v Brně [FIT-S-23-8151].

## References

[Ca22] Casanova, Edresson; Weber, Julian; Shulby, Christopher; Junior, Arnaldo Candido; Gölge, Eren; Ponti, Moacir Antonelli: , YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone, February 2022. arXiv:2112.02418 [cs, eess].

[FM22] Firc, Anton; Malinka, Kamil: The Dawn of a Text-Dependent Society: Deepfakes as a Threat to Speech Verification Systems. In: Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing. SAC '22, Association for Computing Machinery, New York, NY, USA, p. 1646–1655, 2022.

[FMH23] Firc, Anton; Malinka, Kamil; Hanáček, Petr: Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors. Heliyon, 9(4):e15090, April 2023.

[Go23] Godage, Sankini Rancha; Lovasdaly, Froy; Venkatesh, Sushma; Raja, Kiran; Ramachandra, Raghavendra; Busch, Christoph: Analyzing Human Observer Ability in Morphing Attack Detection -Where Do We Stand? IEEE Transactions on Technology and Society, pp. 1–1, 2023.

[Gr21] Groh, Matthew; Epstein, Ziv; Obradovich, Nick; Cebrian, Manuel; Rahwan, Iyad: Human detection of machine-manipulated media. Communications of the ACM, 64(10):40–47, October 2021.

[Gr22] Groh, Matthew; Epstein, Ziv; Firestone, Chaz; Picard, Rosalind: Deepfake detection by human crowds, machines, and machine-informed crowds. Proceedings of the National Academy of Sciences, 119(1):e2110013119, 2022.

[KM20] Korshunov, Pavel; Marcel, Sébastien: , Deepfake detection: humans vs. machines, September 2020. arXiv:2009.03155 [cs, eess].

[Lo11] Loizou, Philipos C: Speech quality assessment. Multimedia analysis, processing and communications, pp. 623–654, 2011.

[Ma08] Matyas, Vashek; Krhovjak, Jan; Kumpost, Marek; Cvrcek, Daniel: Authorizing Card Payments with PINs. Computer, 41:64 – 68, 03 2008.

[ML21] Mirsky, Yisroel; Lee, Wenke: The Creation and Detection of Deepfakes. ACM Computing Surveys, 54(1):1–41, January 2021.

[MPW22] Müller, Nicolas M.; Pizzi, Karla; Williams, Jennifer: Human Perception of Audio Deepfakes. In: Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia. pp. 85–91, October 2022. arXiv:2107.09667 [cs, eess].

[Pr23] Prudký, Daniel: Assessing the Human Ability to Recognize Synthetic Speech. Bachelor's thesis, Brno University of Technology, Brno, Czech republic, 2023. https://www.vut.cz/en/students/final-thesis/detail/140541.

[Ro19] Rossler, Andreas; Cozzolino, Davide; Verdoliva, Luisa; Riess, Christian; Thies, Justus; Nießner, Matthias: , FaceForensics++: Learning to Detect Manipulated Facial Images, August 2019. arXiv:1901.08971 [cs].

[SC10] Simons, Daniel J; Chabris, Christopher F: The monkey business illusion. Cognition, 119(1):23–32, 2010.

[Ta21] Tahir, Rashid; Batool, Brishna; Jamshed, Hira; Jameel, Mahnoor; Anwar, Mubashir; Ahmed, Faizan; Zaffar, Muhammad Adeel; Zaffar, Muhammad Fareed: Seeing is Believing: Exploring Perceptual Differences in DeepFake Videos. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. ACM, May 2021.

[We19] Westerlund, Mika: The Emergence of Deepfake Technology: A Review. Technology Innovation Management Review, 9:40–53, 11/2019 2019.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# Synthetic Latent Fingerprint Generation Using Style Transfer

Amol S. Joshi[1], Ali Dabouei[1], Nasser M. Nasrabadi[2], Jeremy Dawson[2]

**Abstract:** Limited data availability is a challenging problem in the latent fingerprint domain. Synthetically generated fingerprints are vital for training data-hungry neural network-based algorithms. Conventional methods distort clean fingerprints to generate synthetic latent fingerprints. We propose a simple and effective approach using style transfer and image blending to synthesize realistic latent fingerprints. Our evaluation criteria and experiments demonstrate that the generated synthetic latent fingerprints preserve the identity information from the input contact-based fingerprints while possessing similar characteristics as real latent fingerprints. Additionally, we show that the generated fingerprints exhibit several qualities and styles, suggesting that the proposed method can generate multiple samples from a single fingerprint.

**Keywords:** Latent fingerprints, Synthetic latent fingerprint generation, Style transfer.

## 1 Introduction

Fingerprints left on a surface unintentionally, also called latent fingerprints, play a vital role as evidence in forensic investigations. Unfortunately, these fingerprints are not readily viable for matching and recognition purposes. Due to the unconstrained environment at a crime scene and the complex acquisition process of latent fingerprints, they are notoriously indispensable to pre-processing, such as segmentation, enhancement, and feature extraction. Recent latent fingerprint pre-processing algorithms based on neural networks [Ta17, NCJ18, Da18, LQ20, ZYH23] require larger datasets for training. However, the collection of latent fingerprints is an expensive and cumbersome task. Table 1 summarizes the latent fingerprint datasets widely used for training and evaluating latent enhancement algorithms. The fingerprints in these datasets are deposited under controlled or uncontrolled conditions and lifted from various surfaces. NIST SD-302 dataset contains a large number of latent fingerprints, but not all of them have mated fingerprints. Moreover, the latent fingerprints in this dataset are substantially distinct from other datasets. Samples from these datasets are provided in Figure 1. Combining these datasets for training pre-processing algorithms may introduce a class imbalance. Further, it is essential to use real latent fingerprints to evaluate these methods.

This scarcity of data leads to the need for the generation of synthetic latent fingerprints that can be used to train the models so that real data can be utilized for a fair evaluation.

[1] West Virginia University, Lane Department of Computer Science and Electrical Engineering, Morgantown, West Virginia, USA, {asj00003, ad0046}@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

Tab. 1: Summary of released latent fingerprint datasets in the literature.

| Dataset | # of samples | # of surfaces | Availability |
|---|---|---|---|
| NIST SD-27 [GM00] | 258 | N/A | No |
| NIST SD-302 [Fi19] | 10000 | 30 | Yes |
| IIITD [Sa11] | 1016 | 2 | Yes |
| IIITD SLF [SVS12] | 720 | 1 | Yes |
| MOLF [SVS15] | 4400 | 1 | Yes |
| MSLFD [Sa15] | 551 | 8 | Yes |

With more synthetic images, these latent fingerprints need to possess certain characteristics. It is crucial to have identity features such as meaningful ridge structure, fingerprint shape, and minutiae points and noise features such as noisy backgrounds, surface, texture, etc. Many latent fingerprint pre-processing algorithms resort to a naive approach of blending a sensor-collected fingerprint with a noisy background to mimic a latent fingerprint [Da18, LQ20, HQL20]. Zhu et al. [ZYH23] extend the weighted combination approach by applying plastic distortion [CMM01] on high-quality rolled fingerprints. This image-blending approach preserves the identity but fails to generate realistic latent fingerprints.

Another method of generating synthetic fingerprints involves CycleGAN [Zh17], which uses Generative Adversarial Networks(GAN) to transform images from one domain to another. Authors in [ÖSA22, WJ23] trained CycleGAN to transform slap/rolled fingerprints into latent fingerprints. However, these methods have limited style generation capacity. Wyzykowski and Jain [WJ23] use multiple CycleGAN models to generate multiple styles. This might be inconvenient if latent fingerprints with more styles and qualities are required. Nonetheless, these approaches generate pairs of latent and sensor-collected fingerprints, which is ideal for training algorithms that use image-to-image translation for latent fingerprint pre-processing.

Our goal is to generate multiple styles of latent fingerprints using a single model. When lifted from surfaces like paper, cardboard, ceramic tiles, etc., latent fingerprints exhibit different characteristics than those lifted from plastic and metallic surfaces. Additionally, the interaction between the surface and the subject causes uneven ridge densities and orientations. Therefore, the synthetic fingerprint generator must be trained to learn these variations for generating fingerprints of different styles. To this aim, we pose latent generation as a style transfer task from latent fingerprints to sensor-collected fingerprints. The primary task is to transform the ridge patterns in sensor-collected fingerprints into the ridge patterns in real latent fingerprints. This can be achieved by learning the distribution of latent fingerprints and fusing the distribution parameters with the source fingerprints. We use adaptive instance normalization [HB17] to infuse the learned parameters of the latent fingerprint domain while reconstructing the fingerprints. Further, we blend these transformed ridge patterns with noisy backgrounds to manifest a similar distribution to the real latent fingerprints. For brevity, we will refer to sensor-collected fingerprints as fingerprints.

The generated latent fingerprints represent fingerprints lifted from different surfaces. Our contributions are three-fold:

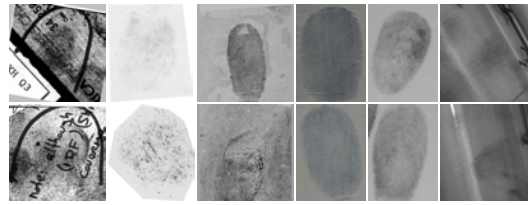



Fig. 1: Latent fingerprints from datasets listed in Table 1.

- We propose a simple and effective method that considers different surfaces and qualities while generating synthetic latent fingerprints.
- Our proposed method is flexible to generate multiple styles of the same fingerprint while preserving the underlying identity information.
- Our evaluation experiments demonstrate the similarities between synthetic and real latent fingerprints.

The paper is organized as follows; first, we discuss related work in Section 2. The proposed method is described in Section 3 followed by a discussion on experiments and results in Section 4. Finally, Section 5 concludes the paper.

## 2 Related Work

Many studies have been conducted to generate synthetic fingerprints. Before the advent of neural networks, hand-crafted feature-based approaches were developed to generate fingerprints. Capelli et al. [CMM02] used fingerprint shape, directional map, density map, and ridge patterns to create a master fingerprint. Further, they apply distortion, noise, and ridge variations to generate variants of the same master fingerprint. Zhao et al. [Zh12] proposed an approach based on statistics of fingerprint features such as type, size, ridge orientation, minutiae, and singular points. After generating a master print using the features, they apply non-linear plastic distortion and rigid transformations to get variants of the same fingerprint. Recent works typically use GANs to generate synthetic fingerprints [Bo18, MA18, Ba21]. These methods focus on training GAN to learn the distribution of real fingerprints and generate synthetic fingerprints that contain the necessary identity information.

Style transfer is a way to learn to map the style of an image onto the contents of another image. Neural network-based style transfer is also explored in the image synthesis task [Me20, Zh20, Ly23]. Men et al. [Me20] developed a person image synthesis algorithm that encodes attributes such as pose, head, base, clothes, etc. The style code is then injected into the AdaIN [HB17] features during decoding. Authors in [Zh20, Ly23] proposed region adaptive normalization to control the style encoding in different image patches. This allows more flexibility to generate images with fine details. Despite these works, to the best of our knowledge, latent fingerprint synthesis has yet to be attempted with style transfer.

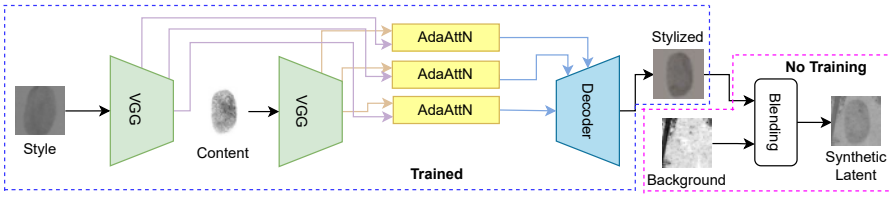



Fig. 2: Architecture of the proposed method. The style transfer network is trained using real latent fingerprints and is marked by the blue box, whereas the image blending does not involve training and is represented in the magenta box.

## 3 Methodology

A widely adopted conventional approach to generating synthetic latent fingerprints applies noise to good-quality fingerprints and blends them with noisy backgrounds. It uses the equation below:

$$I_{latent} = \alpha \times I_{fingerprint} + (1-\alpha) \times I_{noise}. \quad (1)$$

However, in real-world scenarios, the latent fingerprints are lifted from multiple surfaces under unforeseeable environments. Depending on the nature of the surface and the action that caused the fingerprint to be left on the surface, the latent fingerprints exhibit different styles. As a result, using the blending method naively with good-quality fingerprints may not represent the distribution of real latent fingerprints.

We propose learning the noise and distortions in ridge patterns acquired from multiple surfaces and transferring them to fingerprints to mimic the real latent fingerprints. To this aim, we devise a simple and efficient approach involving style transfer and image blending. Further, section 3.1 illustrates the style transfer network, and section 3.2 discusses image blending. Figure 2 illustrates the network architecture.

### 3.1 Style Transfer

The style transfer module is responsible for extracting style from a latent fingerprint $F_s$ and fusing it with the content fingerprint $F_c$ during the reconstruction phase. We use AdaAttN [Li21] to learn the style of latent fingerprints and transform the fingerprint ridges to have a similar style. The style transfer network uses an encoder $E(.)$ to extract content and style embeddings. The extracted embeddings are then passed to the AdaAttN block, which adaptively transfers the style statistics to the content embeddings. The style transfer network uses several layers of pre-trained VGG-19 [SZ14] model to obtain embeddings with different spatial sizes and image characteristics. The AdaAttN block uses the attention mechanism to compute the weighted mean and variance map. Then, adaptive normalization proposed by Huang et al. [HB17] is used to obtain the stylized features. Finally, the decoder reconstructs the stylized features to a synthetic latent fingerprint $F_{cs}$.

The decoding at the end of the style transfer network may produce a random texture pattern to satisfy the objective function. Therefore, to preserve the identity information and the

ridge structure of the fingerprints, we use another encoder $V(.)$ trained to extract features helpful in matching two fingerprints. The embeddings of the fingerprint $F_c$ and the stylized fingerprint $F_{cs}$ enforce the identity constraint during training. We trained the style transfer network using the following objective function:

$$\mathcal{L} = \lambda_g \mathcal{L}_{gs} + \lambda_l \mathcal{L}_{lf} + \lambda_i \mathcal{L}_{id}, \tag{2}$$

where $\mathcal{L}_{gs}$ is a global style loss computed between the mean and standard deviation of embeddings of $F_s$ and $F_{cs}$ extracted using $E(.)$. $\mathcal{L}_{lf}$ is a local feature loss that minimizes the distance between features of $E(F_{cs})$, $E(F_c)$, and $E(F_s)$. Lastly, $\mathcal{L}_{id}$ is an identity constraint between $V(F_{cs})$ and $V(F_c)$. We use mean squared error to calculate the loss terms. We empirically set a value of 1.0 for $\lambda_i$, whereas the default values of 3.0 for $\lambda_g$ and 10.0 for $\lambda_l$ are used during training.

### 3.2 Image Blending

The output of the style transfer network is distorted ridge patterns that appear similar to the ridge patterns in real latent fingerprints. However, we can profusely notice the noisy backgrounds and textured patterns in real latent fingerprints. Therefore, we incorporate the image blending from Eq. 1 to generate realistic latent fingerprints. We replace $I_{fingerprint}$ by the output of the style transfer network and consider several background images cropped from real latent fingerprints as $I_{noise}$. During all the experiments, we set $\alpha$ between 0.3 to 0.8. This combination of style transfer network and image blending presents the flexibility to manipulate the style, quality, surface, and background of the generated fingerprints without retraining the network. Further, the synthetic fingerprints and the corresponding content fingerprints are spatially consistent. Therefore, the spatial features extracted from the fingerprint can be used as a target while training a neural network for latent fingerprint pre-processing. Later in section 4.3, we discuss the effect of blending noisy background with the output of the style transfer network.

## 4 Experiments

In section 4.1, we discuss the datasets used and generated for the evaluation experiments. Later, we describe the evaluation criteria and results in Section 4.1 and Section 4.2, respectively.

### 4.1 Datasets

Training the style transfer network requires fingerprints as content images and latent fingerprints as style images. Therefore, we combined fingerprints from MOLF and MSLFD datasets totaling 12,444 for training and 600 for evaluation. For the style images, we used 4,400 latent fingerprints from MOLF, which has fingerprints lifted from ceramic tile

[SVS15]. Additionally, we included 170 latent fingerprints from two different surfaces from the MSLFD dataset. Further, we create pairs of latent fingerprints and content fingerprints such that they belong to the same finger of the same subject. This aids the identity preservation constraint in the objective function. We use latent fingerprints from IIITD and SLF datasets as the style images during evaluation experiments.

Once the style transfer network is trained, we generate 600 synthetic latent fingerprints. Finally, we create two sets, Synthetic-1 and Synthetic-2, using backgrounds from different surfaces and textures. The Synthetic-1 dataset represents latent fingerprints lifted from plain surfaces such as ceramic tiles and cardboard. In contrast, the Synthetic-2 dataset comprises latent fingerprints lifted from plastic and paper surfaces with printed text.

### 4.2 Evaluation Criteria

For evaluating a synthetic data generator, measuring the similarities between the synthetic and real data is imperative. We use various aspects of fingerprints for comparing the characteristics of real and generated latent fingerprints. First, we use quality distribution as a metric to demonstrate the similarity. To this aim, we use NFIQ 2.0 [Ta21] to obtain the quality scores of latent fingerprints. The second metric is the similarity between the data distribution of real and synthetic fingerprints. We use t-Distributed Stochastic Neighbor embeddings (t-SNE) [vdMH08] to showcase the distribution of multiple datasets to compare with the synthetic fingerprints. t-SNE uses high-dimensional feature embeddings of size 512 and reduces the dimensionality to generate two components to visualize the distribution.

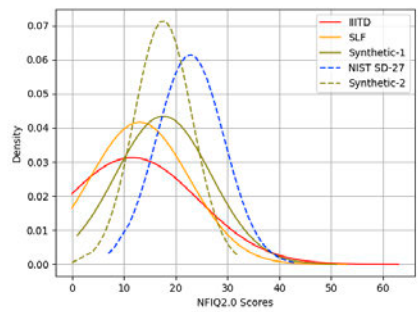



Fig. 3: NFIQ 2.0 quality score distribution of multiple datasets. Solid and dashed lines represent datasets with different surfaces and styles.

Further, we study minutiae points to analyze the realistic nature of synthetic fingerprints. This analysis helps determine if the synthetic latent fingerprints have meaningful patterns and genuine minutiae points. We use the Verifinger SDK v10.0 [Ne] to extract minutiae and perform matching experiments. Lastly, we analyze the matching score distribution of genuine pairs consisting of synthetic latent and corresponding mated fingerprints. Due to the noisy and distorted nature of latent fingerprints, the recognition accuracy is relatively

low compared to fingerprint matching. Comparing the matching scores of mated pairs helps estimate if the synthetic fingerprints are challenging enough for the matchers to extract features.

### 4.3 Results

Determining the quality of latent fingerprints is crucial in matching and recognition scenarios. Due to the complex acquisition process of latent fingerprints, they often exhibit poor quality scores. In Figure 3, we compare if the generated synthetic latent fingerprints have a similar quality score distribution with the real data. The latent fingerprints in IIITD and SLF datasets have a wide range of quality scores, whereas the NIST SD-27 dataset has a smaller range due to the arbitrary texture patterns and highly distorted ridge patterns. The plot suggests the closeness of quality levels among Synthetic-1, IIITD, and SLF datasets. Similarly, curves for NIST SD-27 and Synthetic-2 datasets also match each other. Next, we plot t-SNE to demonstrate the overlapping distribution of real and synthetic fingerprints. Figure 4 provides the distribution for multiple datasets of various styles. Note that in Figure 4(a), the data points for the Synthetic-1 dataset are congregated in two regions. This behavior is due to the limited style references used to transform the ridge patterns during synthetic generation. At the same time, the arbitrary noise patterns in the real latent fingerprints make the distribution widespread. Regardless, both plots show evidence of the embeddings of the synthetic and real latent fingerprints in the high-dimensional space corresponding with datasets of respective styles. Further, this suggests that our proposed method can generate realistic latent fingerprints with real latent fingerprint characteristics.

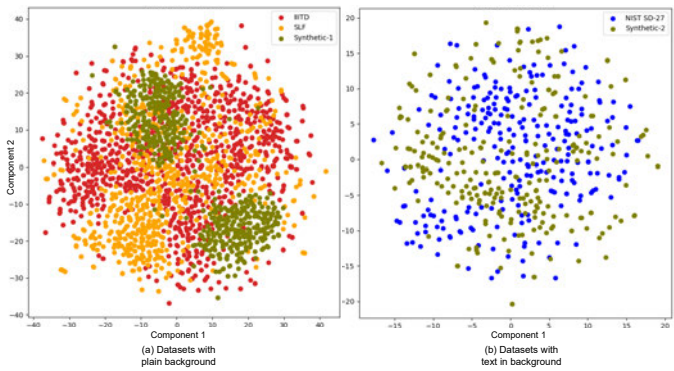



Fig. 4: t-SNE distribution of multiple datasets. Plot (a) represents datasets with plain backgrounds from surfaces like ceramic tiles and cards. Plot (b) represents latent fingerprints lifted from plastic and paper with text in the background.

Despite similar quality and t-SNE distributions, the synthetic latent fingerprints should represent some identity. Ideally, a synthetic latent fingerprint should have the same identity as the source fingerprint used as input to the style transfer network. Figure 5 demonstrates

the identity similarity between the synthetic latent and input fingerprint. It shows detected and correctly matched minutiae, suggesting that the proposed method preserves critical features such as the ridge structure and minutiae points. Further, the figure indicates the ability of the proposed method to generate multiple synthetic samples from the same fingerprint with varying quality and styles.

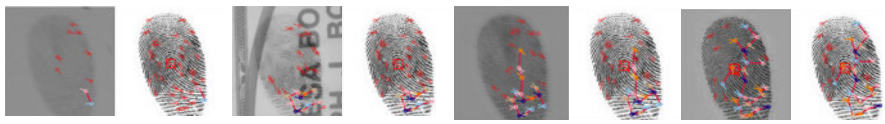



Fig. 5: Generated synthetic latent fingerprints of various styles and qualities. In each pair, the image on the left is the latent fingerprint, and on the right is the corresponding input fingerprint. The number mentioned at the bottom of each pair represents the matching score obtained using the Verifinger SDK v10.0. See Section 4.3 for additional details.

To investigate the importance of the style transfer network and compare it with the naive approach of image blending used in [Da18, LQ20, HQL20], we generated a set of synthetic latent fingerprints without using the style transfer network. We applied speckle noise to the fingerprints and blended them with noisy backgrounds. Then, we conducted a matching experiment with genuine pairs from this dataset. In Table 2, we compare the mean, standard deviation, and median of matching scores for genuine pairs of latent fingerprints generated by our method and the real latent fingerprint dataset. A significant difference between the distribution parameters shows that a weighted combination of a distorted fingerprint and noisy background is insufficient to model realistic latent fingerprints. The matcher can easily recognize the fingerprint despite the background noise.

Tab. 2: The mean, standard deviation, and median of matching scores for genuine pairs belonging to different latent fingerprint datasets. VeriFinger SDK v10.0 was used to obtain the matching scores.

| Latent dataset | Mean | Standard deviation | Median |
|---|---|---|---|
| W/o style transfer | 609.3629 | 381.131 | 572.0 |
| Ours | 89.1046 | 101.1472 | 43.5 |
| Real | 63.5454 | 47.9998 | 54.0 |

## 5 Conclusion

We proposed a simple and effective approach to synthetic latent fingerprint generation. We showed that the naive approximation of latent fingerprints inadequately represents real latent fingerprints. We revised it and proposed an algorithm to generate realistic latent fingerprints using a style transfer network to exploit the style features of real latent fingerprints and transform the ridge structure to appear as a latent fingerprint. Further, the stylized ridges are blended with noisy backgrounds for a better representation of real latent fingerprints. Our evaluation with various metrics suggests that the proposed method reliably generates latent fingerprints of various styles and qualities while preserving identity information.

## References


[Ba21] Bahmani, Keivan; Plesh, Richard; Johnson, Peter; Schuckers, Stephanie; Swyka, Timothy: High fidelity fingerprint generation: Quality, uniqueness, and privacy. In: 2021 IEEE International Conference on Image Processing (ICIP). IEEE, pp. 3018–3022, 2021.

[Bo18] Bontrager, Philip; Roy, Aditi; Togelius, Julian; Memon, Nasir; Ross, Arun: Deepmasterprints: Generating masterprints for dictionary attacks via latent variable evolution. In: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS). IEEE, pp. 1–9, 2018.

[CMM01] Cappelli, Raffaele; Maio, Dario; Maltoni, Davide: Modelling Plastic Distortion in Fingerprint Images. In: International Conference on Advances in Pattern Recognition. 2001.

[CMM02] Cappelli, R.; Maio, D.; Maltoni, D.: Synthetic fingerprint-database generation. In: 2002 International Conference on Pattern Recognition. volume 3, pp. 744–747 vol.3, 2002.

[Da18] Dabouei, Ali; Kazemi, Hadi; Iranmanesh, Seyed Mehdi; Dawson, Jeremy; Nasrabadi, Nasser M et al.: ID preserving generative adversarial network for partial latent fingerprint reconstruction. In: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS). IEEE, pp. 1–10, 2018.

[Fi19] Fiumara, Gregory; Flanagan, Patricia; Grantham, John; Ko, Kenneth; Marshall, Karen; Schwarz, Matthew; Tabassi, Elham; Woodgate, Bryan; Boehnen, Christopher: , NIST Special Database 302: Nail to Nail Fingerprint Challenge, 2019-12-11 2019.

[GM00] Garris, Michael; McCabe, R.: , NIST Special Database 27 Fingerprint Minutiae From Latent and Matching Tenprint Images, 2000-06-01 2000.

[HB17] Huang, Xun; Belongie, Serge: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510, 2017.

[HQL20] Huang, Xijie; Qian, Peng; Liu, Manhua: Latent fingerprint image enhancement based on progressive generative adversarial network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 800–801, 2020.

[Li21] Liu, Songhua; Lin, Tianwei; He, Dongliang; Li, Fu; Wang, Meiling; Li, Xin; Sun, Zhengxing; Li, Qian; Ding, Errui: Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6649–6658, 2021.

[LQ20] Liu, Manhua; Qian, Peng: Automatic segmentation and enhancement of latent fingerprints using deep nested unets. IEEE Transactions on Information Forensics and Security, 16:1709–1719, 2020.

[Ly23] Lyu, Yueming; Chen, Peibin; Sun, Jingna; Peng, Bo; Wang, Xu; Dong, Jing: DRAN: detailed region-adaptive normalization for conditional image synthesis. IEEE Transactions on Multimedia, 2023.

[MA18] Minaee, Shervin; Abdolrashidi, Amirali: Finger-GAN: Generating realistic fingerprint images using connectivity imposed GAN. arXiv preprint arXiv:1812.10482, 2018.

[Me20] Men, Yifang; Mao, Yiming; Jiang, Yuning; Ma, Wei-Ying; Lian, Zhouhui: Controllable person image synthesis with attribute-decomposed gan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5084–5093, 2020.

[NCJ18] Nguyen, Dinh-Luan; Cao, Kai; Jain, Anil K: Robust minutiae extractor: Integrating deep networks and fingerprint domain knowledge. In: 2018 International Conference on Biometrics (ICB). IEEE, pp. 9–16, 2018.

[Ne] Neurotechnology Inc., https://www.neurotechnology.com/verifinger.html: .

[ÖSA22] Öztürk, Halil İbrahim; Selbes, Berkay; Artan, Yusuf: Minnet: Minutia patch embedding network for automated latent fingerprint recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1627–1635, 2022.

[Sa11] Sankaran, Anush; Dhamecha, Tejas I.; Vatsa, Mayank; Singh, Richa: On matching latent to latent fingerprints. In: 2011 International Joint Conference on Biometrics (IJCB). pp. 1–6, 2011.

[Sa15] Sankaran, Anush; Agarwal, Akshay; Keshari, Rohit; Ghosh, Soumyadeep; Sharma, Anjali; Vatsa, Mayank; Singh, Richa: Latent fingerprint from multiple surfaces: Database and quality analysis. In: 2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS). pp. 1–6, 2015.

[SVS12] Sankaran, Anush; Vatsa, Mayank; Singh, Richa: Hierarchical fusion for matching simultaneous latent fingerprint. In: 2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS). pp. 377–382, 2012.

[SVS15] Sankaran, Anush; Vatsa, Mayank; Singh, Richa: Multisensor optical and latent fingerprint database. IEEE access, 3:653–665, 2015.

[SZ14] Simonyan, Karen; Zisserman, Andrew: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[Ta17] Tang, Yao; Gao, Fei; Feng, Jufu; Liu, Yuhang: FingerNet: An unified deep network for fingerprint minutiae extraction. In: 2017 IEEE International Joint Conference on Biometrics (IJCB). IEEE, pp. 108–116, 2017.

[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, 2021-07-13 04:07:00 2021.

[vdMH08] van der Maaten, Laurens; Hinton, Geoffrey: Visualizing Data using t-SNE. Journal of Machine Learning Research, 9(86):2579–2605, 2008.

[WJ23] Wyzykowski, André Brasil Vieira; Jain, Anil K: Synthetic latent fingerprint generator. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 971–980, 2023.

[Zh12] Zhao, Qijun; Jain, Anil K.; Paulter, Nicholas G.; Taylor, Melissa: Fingerprint image synthesis based on statistical feature models. In: 2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS). pp. 23–30, 2012.

[Zh17] Zhu, Jun-Yan; Park, Taesung; Isola, Phillip; Efros, Alexei A: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232, 2017.

[Zh20] Zhu, Peihao; Abdal, Rameen; Qin, Yipeng; Wonka, Peter: Sean: Image synthesis with semantic region-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5104–5113, 2020.

[ZYH23] Zhu, Yanming; Yin, Xuefei; Hu, Jiankun: Fingergan: a constrained fingerprint generation scheme for latent fingerprint enhancement. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# DEFT: A new distance-based feature set for keystroke dynamics

Nuwan Kaluarachchi,[1] Sevvandi Kandanaarachchi, [2] Kristen Moore,[3] Arathi Arakala[4]

**Abstract:**

Keystroke dynamics is a behavioural biometric utilised for user identification and authentication. We propose a new set of features based on the distance between keys on the keyboard, a concept that has not been considered before in keystroke dynamics. We combine flight times, a popular metric, with the distance between keys on the keyboard and call them as Distance Enhanced Flight Time features (DEFT). This novel approach provides comprehensive insights into a person's typing behaviour, surpassing typing velocity alone. We build a DEFT model by combining DEFT features with other previously used keystroke dynamic features. The DEFT model is designed to be device-agnostic, allowing us to evaluate its effectiveness across three commonly used devices: desktop, mobile, and tablet. The DEFT model outperforms the existing state-of-the-art methods when we evaluate its effectiveness across two datasets. We obtain accuracy rates exceeding 99% and equal error rates below 10% on all three devices.

**Keywords:** Keystroke dynamics, Continuous authentication, Desktop, Mobile, Tablet, Multi-Device, Feature optimisation, Key pair distances

## 1 Introduction

Keystroke dynamics refers to the systematic analysis of the pattern of key press and release on a physical or virtual keyboard by an individual. The information distilled from the typing patterns enables user identification and authentication. Thus, it has proven to be an effective behavioural biometric modality. One notable advantage of keystroke dynamics is its suitability for continuous authentication, as it involves an ongoing behaviour performed by an individual throughout a typing session. This continuous nature makes it well-suited for establishing and maintaining user authentication in various contexts.

Most keystroke dynamics studies in the last five years used temporal features **(TEMP)** [KBH22, Ac21, Ya21b, Ya21a, Av21, Ki20, KK20, Ay19]. TEMP features compute unigraph (key hold time), digraph (flight times) and trigraph (presses of three consecutive keys) attributes from typing behaviours. Alsultan et al. [AWW17] used non-conventional

[1] Mathematical Sciences, STEM College, RMIT University, Melbourne, Australia | CSIRO's Data61,Melbourne, Australia nuwan.kaluarachchi@student.rmit.edu.au | nuwan.kaluarachchi@data61.csiro.au
[2] CSIRO's Data61, Melbourne, Australia, sevvandi.kandanaarachchi@data61.csiro.au
[3] CSIRO's Data61, Melbourne, Australia, kristen.moore@data61.csiro.au
[4] Mathematical Sciences, STEM College, RMIT University, Melbourne, Australia, arathi.arakala@rmit.edu.au

features **(NC)** such as average backspace and negative flight times over a user typing session. Al-Saraireh and AlJa'afreh [ASA23] and Belman and Phoha [BP20a] used flight times of commonly typed keypairs **(CKP)** as features. While either TEMP, NC, or CKP features have been used individually in previous studies, their combined effect has not been explored.

Our study introduces a new set of features that considers the distance between key pairs when considering flight times. We call these Distance Enhanced Flight Time **(DEFT)** features. We combine DEFT with the previously studied TEMP, NC and CKP features (DEFT Model) and investigate the performance of the combined feature for continuous authentication. We show that using DEFT features significantly improves authentication performance on desktop, mobile and tablet compared to existing approaches. In addition, most studies only focus on free text. However, we frequently type usernames, passwords and email addresses, which come under the fixed text category. So in this study, we present our results with fixed and free text-typing datasets to incorporate the mixed nature of typing. Furthermore, while most device authentication studies are limited to a single device, with a handful of studies using two devices [BP20b], we demonstrate the broad applicability of this new set of features on three commonly used devices: desktop, mobile and tablet.

## 2 Related Work

Over the last 5 years, various feature types and classifiers have been tested for user identification and authentication with keystroke dynamics. While a detailed review of this work is out of scope, we briefly review (see Table 1) some insights gained from this study and explain how it has impacted our research. As seen from Table 1, most studies compute features using flight times, hold times, words per minute and error rate. In contrast, Alsultan et al. [AWW17] consider a novel set of features that capture the backspace, negative flight times and shift key usage in a session. A key takeaway from Table 1 is that none of the recent studies has combined these different types of features to construct an optimal set of features for user identification and authentication.

We combine these different types of features, and additionally, we develop a new set of features based on the physical distance between keyboard keys, which we call Distance Enhanced Flight Time (DEFT) features. The intuition behind these features is that flight and hold times depend on the distance between keys and the use of one or both hands. For example, if we consider keys typed by a single hand, we expect a normal user to take more time between two keys that are further apart, such as 'A' and 'T', compared to the time taken for two keys that are closer such as 'A' and 'S'. Subsequently, we employ feature selection strategies to determine whether the DEFT features contribute to discriminating between users.

| Year [Ref] | Hold Times | Flight Times | WPM | Error Rate | NegUD | NegUU | Shift Usage | Cpslck Usage |
|---|---|---|---|---|---|---|---|---|
| 2022 [KBH22] | X | X | | | | | | |
| 2022 [Ac21] | X | X | | | | | | |
| 2021 [Ya21b] | X | X | | | | | | |
| 2021 [Ya21a] | X | X | | | | | | |
| 2021 [Av21] | | X | X | X | | | | |
| 2020 [Ki20] | X | X | | | | | | |
| 2020 [Lu20] | X | X | | | | | | |
| 2020 [BP20a] | X | X | | | | | | |
| 2020 [KK20] | X | X | | | | | | |
| 2019 [Ay19] | X | X | | | | | | |
| 2019 [Wu19] | X | X | | | | | | |
| 2017 [AWW17] | | | X | X | X | X | X | X |
| 2017 [MB17] | X | X | | | | | | |

Tab. 1: Comparison of feature types used in recent keystroke dynamic studies. The majority of studies used Temporal features, while only one study used only non-conventional features for keystroke dynamics. None of the studies used a combination of all these feature types.

# 3 Methodology

## 3.1 Datasets

We use the publicly available SU-AIS BB-MAS (Syracuse University and Assured Information Security Behavioral Biometrics Multi-device and multi-Activity data from Same users) dataset [K.19] as our main dataset. This dataset plays a vital role in advancing the field of biometrics, addressing the existing gap in collecting data from individuals across multiple devices. The BB-MAS dataset captures multiple behavioural biometric modalities from three different devices, including swiping, keystroke, and gait dynamics. The data was collected over two months from 117 participants. However, missing data in the collection process leaves us with only 116 users for each device. A comprehensive account of the dataset and the data collection process can be found in [Be19].

We evaluate our newly introduced DEFT features using another well-known keystroke dynamic dataset, the Buffalo dataset [SCU16], to validate its efficacy. This dataset consists of keystroke data collected from 148 users using 4 different types of keyboards. The keystroke patterns from 75 users were captured by their typing on the same type of keyboard in 3 distinct sessions. Keystroke patterns from the other 73 users were acquired by typing on three keyboards in the 3 different sessions.

### 3.2 Distance Enhanced Flight Times (DEFT) features

DEFT features are computed under the assumption that the distance between the keys affects the flight time. The distance is computed using the spatial separation of the keys on the keyboard (see Figure 1). For example, we expect the flight times between keys, such as 'A,S', 'S,D' and 'W,E', to be similar (generated from a single probability distribution) due to the equal distance between the respective keys. The distance between the keys in each of those 3 pairs is 1. We consider the average flight time for all such key pairs, i.e., the average flight time (per user) across all distance 1 key pairs. Similarly, we calculate the average flight time for distances 0, 2, and 3 key pairs. In this manner, the DEFT features augment the flight times (TEMP features) by incorporating the distance between keys.

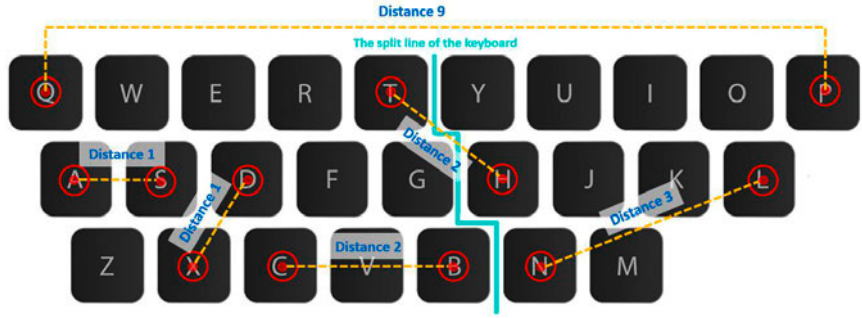



Fig. 1: Calculation of the distance between the key pairs on the keyboard. <A,S> is a distance 1 digraph while <T,H> is a distance 2 digraph and <N,L> is a distance 3 digraph. The longest distance between a key pair is distance 9. The blue line separates the keyboard into left and right sides, which helps identify keys typed by an individual's left or right hand.

Figure 1 demonstrates the distances between some key pairs. The calculated key pair distances range from zero to nine, with zero indicating the pressing of the same key and nine denoting the longest distance between two keys, specifically, 'Q' and 'P'. Generally, people use both hands for typing; they use the left hand to type keys on the left and the right hand for keys on the right. We separate the left and right sides of the keyboard as demarcated by the blue line in Figure 1. If both keys of the digraph are on the left side, we indicate it with LL. If both keys are on the right side, it is indicated with RR; if both digraph keys span either side of the keyboard, we denote it by LR.

Next, we combine the distance between the keys with the flight times. In the study, we try to identify the typing patterns of a single user for each hand. Therefore, using the standard QWERTY finger placement, we consider the maximum distance between keys typed by a single hand to be three units. Consequently, flight times for distances ranging from zero to three on both the left and right sides of the keyboard are selected for analysis. We focus on LL and RR key pairs and disregard LR key pairs. We compute the average of each of the four flight times, F1, F2, F3 and F4, as declared by Belman and Phoha in [BP20a, BP20b] grouping by the distance between the keys for each hand. Thus, we have 32 ( $4 \times 4 \times 2$)

new features ranging from F1_distance_0_LL to F4_distance_3_RR. These are the DEFT features, and they capture the average distance flight times per distance for a given user.

Unlike existing text-based features in the keystroke dynamics, which mainly revolve around temporal variations, including instances of key press and release, temporal gaps, and the frequency of such events, an aspect formerly unexplored involves the measurement of spatial separation between pairs of keys. Addressing this gap, our study introduces a novel approach by incorporating key pair distances in conjunction with the flight times of said key pairs. This innovative amalgamation enhances the depth and breadth of our analysis, ushering in a more comprehensive understanding of keystroke dynamics.

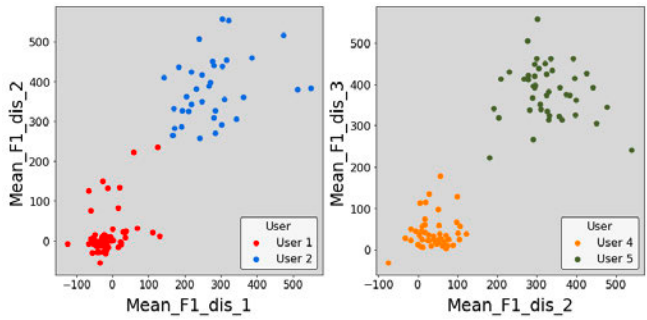



Fig. 2: The pairwise DEFT features of four different users for key pairs typed by the left hand on desktop devices. The average flight 1 (F1) timing for the two users for distance 1 key pairs and distance 2 key pairs is shown in the left figure, and the average flight 1 (F1) timing for two different users for distance 2 key pairs and distance 3 key pairs is shown by the right figure. The figures imply the validity of separating the flight times based on the key pair distances for user identification. Each point of a certain colour represents a user sample.

Figure 2 shows the pairwise distribution of three DEFT features for four different users. The results for these two examples illustrate distinct user clusters, which illustrates that the features (F1_distance_1_LL, F1_distance_2_LL and F1_distance_3_LL) are user dependent and can help user identification and authentication. As the distances between keys and flight times can vary depending on the type and size of the device, such as desktop, mobile or tablet, we adopt a solution that yields relative values for the distances. This is achieved by normalising the distances through division by a single key size, ensuring that the derived distances remain independent of the device being used. This device-agnostic approach can accurately capture and compare the relative distances between keys, irrespective of device type or screen size variations.

# 4 Results and Discussion

## 4.1 Feature selection

We compute TEMP, NC, CKP and DEFT features for each user sample in the BBMAS dataset, resulting in an expanded set of keystroke dynamic features. We found that the most frequently typed key pairs in the dataset were 'T,H', 'I,S', 'H,E', 'A,P', 'L,E' and 'C,O'. When calculating the flight times for TEMP, CKP and DEFT features, we use a simple filter to detect and remove the high or low time differences in keystroke dynamics. We eradicate any instances of a time difference of more than five seconds. We assume these scenarios happen by pauses, getting instructions during the data collection or recording issues. After calculating the TEMP, NC, CKP and DEFT features, we get the average values of each feature for each sample. To identify the most discriminative features from the expanded feature set, we employ the Random Forest (RF) classifier, specifically tuned for multi-class classification described by Ayotte [Ay19]. To conduct our analysis, we split the dataset into a training set comprising 70% of the user samples and a testing set comprising the remaining 30%.This process is carried out separately for desktop, mobile, and tablet devices. This procedure identifies 37, 41, and 42 discriminative features for each device, respectively. Table 2 shows the types and number of the shortlisted features for all three devices. Notably, the table reveals that the DEFT features exhibit the highest occurrence rate among the shortlisted features, accounting for more than 50% of the selected features in the case of mobile and tablet devices. This finding demonstrates the effectiveness of the DEFT features in capturing discriminative information necessary for keystroke authentication.

| Feature Category | Desktop | Mobile | Tablet |
|---|---|---|---|
| **DEFT** | **17** | **25** | **23** |
| CKP | 9 | 8 | 10 |
| TEMP | 6 | 6 | 6 |
| NC | 5 | 2 | 3 |
| Total | 37 | 41 | 42 |

Tab. 2: The category breakdown of features with the highest discriminative characteristics selected by the Random Forest classifier. More than 45% of the selected features on desktop and 50% of the selected features on mobile and tablet are DEFT features, showing the dominance of DEFT features in the feature list.

## 4.2 Authentication Framework

After selecting the discriminative features, we build a binary classifier for each user. A user has about 50 samples from their keystroke data and 6000 samples from other users' keystroke data. Each user's sample is 100 keystrokes long and comprises a vector of the features shortlisted by the feature selection stage. We use stratified five-fold cross-validation with the Extreme Gradient Boost (XGB) classifier. Due to the extreme class imbalance of

the mated compared to non-mated samples, we use the Synthetic Minority Oversampling Technique (SMOTE) [Ch02] to oversample the genuine user's class in each fold.

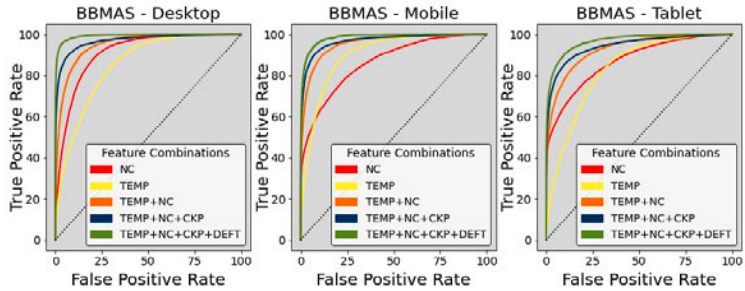



Fig. 3: The ROC curve representation of the authentication performance of different combinations of keystroke features for different devices. In all three devices, the keystroke dynamic performance of the models increased by adding the DEFT features.

First, we perform an ablation study to validate the discriminative power of our new DEFT features. Figure 3 shows ROC (Receiver Operating Characteristics) curves for five feature combinations: NC, TEMP, TEMP + NC, TEMP + NC + CKP, TEMP + NC + CKP + DEFT. The feature selection process, as described in Section 4.1, is only used for the TEMP + NC + CKP and TEMP + NC + CKP + DEFT combinations. The final curated set of features after adding DEFT achieves the best performance for all devices. We work with this set of features for the remainder of the paper, which we name the DEFT model.

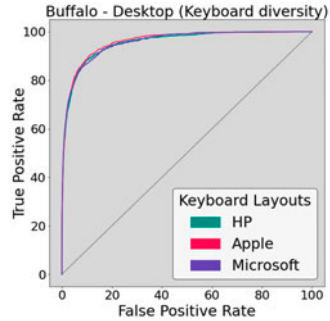



Fig. 4: The ROC curve representation of the authentication performance of the DEFT model on different desktop keyboard types in the Buffalo dataset. The proximity of the curves demonstrates equivalent performance for the three keyboard types.

Next, we test if there is a variation in the performance of the DEFT model when the keyboard type changes. In particular, we want to test if disparities in performance arise in the distance-based features when confronted with different keyboard types. The ROC curves on testing the three keyboard types: An HP wireless keyboard, a Microsoft ergonomic keyboard

and an Apple wireless keyboard. The results are shown in Figure 4. This demonstrates that the DEFT model remains notably consistent across different keyboard configurations.

We next compare our model with leading studies on desktop [AWW17, BP20a], mobile [ASA23, BP20a] and tablet [BP20a]. Alsultan et al. [AWW17] collected NC features using their own dataset of 30 users. Al-Saraireh and AlJa'afreh [ASA23] analysed CKP features from 54 mobile data users of BBMAS. Belman and Phoha [BP20a] examined user identification across desktops, mobiles, and tablets using CKP features and context-based multi-class classifiers on only 20 users of BBMAS. In order to compare [BP20a] with our approach, we convert their multi-class classification to a binary classification problem by selecting the highest-performing context-based classifiers. Although these results are published, the experimental settings in each study are different, including training and testing splits, sample sizes, and evaluation metrics. In our comparison, we use the same experimental setting for all methods. Specifically, we consider all 116 users available in the BBMAS dataset for three devices and all 75 users of the Buffalo dataset for desktops. We plot ROC curves for each model by getting the average of each fold in the cross-validation for all users for each device to get a complete picture of the studies under different thresholds. Figure 5 shows the results, demonstrating that the DEFT model performs better than other models for all three devices of BBMAS and desktop devices of the Buffalo dataset.

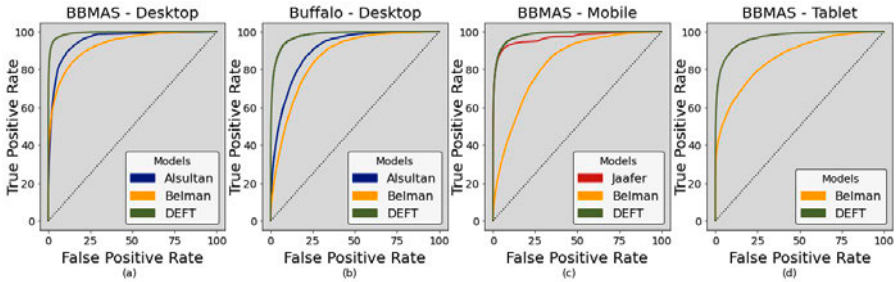



Fig. 5: Comparison of the authentication performance of the DEFT model against existing keystroke dynamics models across 3 devices and 2 datasets. (a) The comparison of desktop models on BBMAS. (b) The comparison of desktop models on the Buffalo dataset. (c) The comparison of mobile models on BBMAS. (d)The comparison of tablet models on BBMAS. The DEFT model outperforms other models for all three devices.

Table 3 summarises the results of our model and other compared state of art models using the BBMAS on keystroke dynamics for the three devices under different performance metrics. As implied by the table, our DEFT model has the highest accuracy, F1 score and AUC-ROC. The model's EER (Equal Error Rate) is 3.8% 6.6% and 9.8% for the desktop, mobile and tablets.

To ensure the reproducibility of our research, we have made relevant code snippets [3] available online.

[3] https://github.com/NuwanYasanga/DEFT

| Device | Model | Accuracy | EER | F1 | AUC-ROC |
|---|---|---|---|---|---|
| Desktop | Alsultan | 95.5 (0.25) | 10.9 (0.63) | 52.1 (1.48) | 95.6 (0.40) |
| | Belman | 97.9 (0.03) | 15.1 (0.49) | 60.9 (0.98) | 92.9 (0.29) |
| | DEFT | **99.6 (0.02)** | **3.8 (0.42)** | **77.5 (1.1)** | **99.3 (0.07)** |
| Mobile | Jaafer | 99.4 (0.02) | 6.7 (0.64) | 58.6 (1.2) | 98.2 (0.24) |
| | Belman | 88.7 (0.24) | 29.8 (0.77) | 46.9 (0.29) | 83.1 (0.87) |
| | DEFT | **99.4 (0.01)** | **6.6 (0.51)** | **65.6 (1.2)** | **98.4 (0.23)** |
| Tablet | Belman | 96.0 (0.04) | 40.4 (0.59) | 47.5 (0.9) | 86.4 (0.80) |
| | DEFT | **99.3 (0.02)** | **9.8 (0.65)** | **61.8 (1.1)** | **96.7 (0.42)** |

Tab. 3: Summary of the performance of the state-of-the-art keystroke models for the three devices using the BBMAS dataset. All the models followed the same experimental setting with their own features and classifiers. Our DEFT model performed better in all three devices for all performance metrics. The table proves the validation of using DEFT features for keystroke dynamics. All values are in percentages, and parenthesis values are the standard deviation

One limitation that arose in our analysis is a significant class imbalance issue within the test datasets, where the imposter class significantly outnumbered the genuine user class by a factor of over 100 in most cases. We don't employ any oversampling or undersampling techniques for the testing set, as we oversample the training dataset. It is important to note that inherent biases towards the imposter class primarily drove low F1 scores. Due to the scarcity of genuine user samples within the test set, even a slight deviation from the expected behaviour of genuine users may lead to misclassification as an imposter sample. This limitation is the ground truth in biometric models, with evaluation encompassing the entirety of 116 users within the BBMAS dataset and the complete cohort of 75 users in the Buffalo dataset. Importantly, this analysis is executed without the application of any data filtration or modification to the test dataset, preserving its original structure and characteristics intact.

The results presented in Table 3 and Figures 3, 5 indicate that the combined pool of features (TEMP + NC+ CKP + DEFT) in the DEFT model is the most discriminative, resulting in significant performance improvements in all three devices. These results highlight that the DEFT features played a key role in capturing the keystroke dynamics for all three devices.

## 5 Conclusion

This paper introduces DEFT, a new set of distance-based features for keystroke dynamics. By combining DEFT with existing features, we constructed a pool of features called a DEFT model that achieved improved authentication performance for desktop, tablet and mobile devices. Noting that DEFT features accounted for approximately 50% of the discriminative feature set identified by a feature selection process, we demonstrated the utility and broad applicability of DEFT features across devices and across datasets. Our comprehensive analysis identified that spatial features play a significant role in user discrimination and can

be effectively employed for continuous user authentication. In future research, we aim to extend the application of DEFT features in cross-device authentication by applying deep transfer learning techniques.

## References

[Ac21] Acien, Alejandro; Morales, Aythami; Monaco, John V; Vera-Rodriguez, Ruben; Fierrez, Julian: TypeNet: Deep learning keystroke biometrics. IEEE Transactions on Biometrics, Behavior, and Identity Science, 4(1):57–70, 2021.

[ASA23] Al-Saraireh, Jaafer; AlJa'afreh, Mohammad Rasool: Keystroke and swipe biometrics fusion to enhance smartphones authentication. Computers & Security, 125:103022, 2023.

[Av21] Aversano, Lerina; Bernardi, Mario Luca; Cimitile, Marta; Pecori, Riccardo: Continuous authentication using deep neural networks ensemble on keystroke dynamics. PeerJ Computer Science, 7:e525, 2021.

[AWW17] Alsultan, Arwa; Warwick, Kevin; Wei, Hong: Non-conventional keystroke dynamics for user authentication. Pattern Recognition Letters, 89:53–59, 2017.

[Ay19] Ayotte, Blaine; Banavar, Mahesh K; Hou, Daqing; Schuckers, Stephanie: Fast and accurate continuous user authentication by fusion of instance-based, free-text keystroke dynamics. In: 2019 International Conference of the Biometrics Special Interest Group (BIOSIG). IEEE, pp. 1–6, 2019.

[Be19] Belman, Amith K; Wang, Li; Iyengar, SS; Sniatala, Pawel; Wright, Robert; Dora, Robert; Baldwin, Jacob; Jin, Zhanpeng; Phoha, Vir V: Insights from BB-MAS–A Large Dataset for Typing, Gait and Swipes of the Same Person on Desktop, Tablet and Phone. arXiv preprint arXiv:1912.02736, 2019.

[BP20a] Belman, Amith K; Phoha, Vir V: Discriminative power of typing features on desktops, tablets, and phones for user identification. ACM Transactions on Privacy and Security (TOPS), 23(1):1–36, 2020.

[BP20b] Belman, Amith K; Phoha, Vir V: DoubleType: Authentication using relationship between typing behavior on multiple devices. In: 2020 International Conference on Artificial Intelligence and Signal Processing (AISP). IEEE, pp. 1–6, 2020.

[Ch02] Chawla, Nitesh V; Bowyer, Kevin W; Hall, Lawrence O; Kegelmeyer, W Philip: SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16:321–357, 2002.

[K.19] K. Belman, Amith; Wang, Li; S. Iyengar, Sundaraja; Sniatala, Pawel; Wright, Robert; Dora, Robert; Baldwin, Jacob; Jin, Zhanpeng; V. Phoha, Vir: SU-AIS BB-MAS (Syracuse University and Assured Information Security - Behavioral Biometrics Multi-device and multi-Activity data from Same users) Dataset, 2019.

[KBH22] Kasprowski, Pawel; Borowska, Zaneta; Harezlak, Katarzyna: Biometric Identification Based on Keystroke Dynamics. Sensors, 22(9):3158, 2022.

[Ki20] Kiyani, Anum Tanveer; Lasebae, Aboubaker; Ali, Kamran; Rehman, Masood Ur; Haq, Bushra: Continuous user authentication featuring keystroke dynamics based on robust recurrent confidence model and ensemble learning approach. IEEE Access, 8:156177–156189, 2020.

[KK20] Kim, Junhong; Kang, Pilsung: Freely typed keystroke dynamics-based user authentication for mobile devices based on heterogeneous features. Pattern Recognition, 108:107556, 2020.

[Lu20] Lu, Xiaofeng; Zhang, Shengfei; Hui, Pan; Lio, Pietro: Continuous authentication by free-text keystroke based on CNN and RNN. Computers & Security, 96:101861, 2020.

[MB17] Mondal, Soumik; Bours, Patrick: Person identification by keystroke dynamics using pairwise user coupling. IEEE Transactions on Information Forensics and Security, 12(6):1319–1329, 2017.

[SCU16] Sun, Yan; Ceker, Hayreddin; Upadhyaya, Shambhu: Shared keystroke dataset for continuous authentication. In: 2016 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, pp. 1–6, 2016.

[Wu19] Wu, Tong; Zheng, Kangfeng; Wu, Chunhua; Wang, Xiujuan; Xu, Guangzhi: User identification by keystroke dynamics based on feature correlation analysis and feature optimization. In: 2019 IEEE 5th International Conference on Computer and Communications (ICCC). IEEE, pp. 40–46, 2019.

[Ya21a] Yang, Haitian; Sun, Degang; Wang, Yan; Zhu, He; Li, Ning; Huang, Weiqing: FKTAN: Fusion Keystroke Time-Textual Attention Networks for Continuous Authentication. In: 2021 IEEE Symposium on Computers and Communications (ISCC). IEEE, pp. 1–6, 2021.

[Ya21b] Yang, Lulu; Li, Chen; You, Ruibang; Tu, Bibo; Li, Linghui: TKCA: a timely keystroke-based continuous user authentication with short keystroke sequence in uncontrolled settings. Cybersecurity, 4(1):1–16, 2021.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# Benchmarking fixed-length Fingerprint Representations across different Embedding Sizes and Sensor Types

Tim Rohwedder[1], Dailé Osorio-Roig[2], Christian Rathgeb[1], Christoph Busch[1]

**Abstract:** Traditional minutiae-based fingerprint representations consist of a variable-length set of minutiae. This necessitates a more complex comparison causing the drawback of high computational cost in one-to-many comparison. Recently, deep neural networks have been proposed to extract fixed-length embeddings from fingerprints. In this paper, we explore to what extent fingerprint texture information contained in such embeddings can be reduced in terms of dimension, while preserving high biometric performance. This is of particular interest, since it would allow to reduce the number of operations incurred at comparisons. We also study the impact in terms of recognition performance of the fingerprint textural information for two sensor types, *i.e.* optical and capacitive. Furthermore, the impact of rotation and translation of fingerprint images on the extraction of fingerprint embeddings is analysed. Experimental results conducted on a publicly available database reveal an optimal embedding size of 512 feature elements for the texture-based embedding part of fixed-length fingerprint representations. In addition, differences in performance between sensor types can be perceived. The source code of all experiments presented in this paper is publicly available at `https://github.com/tim-rohwedder/fixed-length-fingerprint-extractors`, so our work can be fully reproduced.

**Keywords:** Fingerprint recognition, fixed-length representation, computational workload reduction, deep templates.

## 1 Introduction

Fingerprint recognition has been indispensable for decades in law enforcement and border control and the technology has been extended to numerous commercial applications. Recent market trends suggest that the popularity of fingerprint biometrics will increase further in the coming years [Sk22], leading to broad deployment. This may further lead to higher workloads coupled with long transaction times.

The most commonly used fingerprint representation is based on minutiae. It is accurate and provides good interpretability of the ridge pattern of the fingerprint. Despite their popularity, minutiae-based representations lead to certain drawbacks, *e.g.* variable length in terms of the number of minutiae and unordered feature vectors (*i.e.* representation). A comparison of two minutiae sets commonly involves the determination of mated minutae pairs. This procedure can turn out to be computationally expensive, resulting in a complexity of $O(n^2)$ [Ma22, Chapter 4]. This computational complexity also limits the use of

[1] Hochschule Darmstadt, Germany, tim.rohwedder@stud.h-da.de
[2] dasec - Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany, {daile.osorio-roig;christian.rathgeb;christoph.busch}@h-da.de

minutiae-based fingerprint representations in combination with biometric template protection (BTP) schemes (*e.g.* homomorphic encryption [Ba23]) which results in a very high workload. Furthermore, their usability and interoperability are limited in feature-level multimodal fusion systems with other popular types of biometric characteristics (*e.g.* face and iris) that use floating-point values in their representations (*e.g.* [Bo22a, Ba22, Bo22b]).

In recent years, biometric technologies have been combined with deep learning approaches because of their capabilities to extract distinctive features, *i.e.* embeddings, that allow high recognition performance [Li17, Wa18, WD21]. In particular, texture-based representation has been of interest for many types of biometric characteristics. Extracted texture information can easily be dimensionally reduced without sacrificing biometric performance (*e.g.* search of the intrinsic dimensionality [GBJ19] for face templates). In contrast to minutiae-based representations, the embeddings extracted by deep neural networks are usually of fixed length and, thus, can be successfully combined with BTP schemes and other types of biometric characteristics in a multimodal system.

Recently, the extraction of texture-based fixed-length fingerprint representations has been proposed in different deep learning-based works [ECJ19, Gr22]. Engelsma *et al.* [ECJ19] proposed a Deep Neural Network (DNN) called DeepPrint that learns both minutiae and texture representations through multi-task learning. To evaluate the feasibility of the proposed approach, several experiments on some publicly available databases, *e.g.* FVC 2004 DB1 A [Ma04], resulting in high recognition performance including scanned rolled fingerprints in NIST SD4 [WW92] and NIST SD14 [Wa93] databases, were conducted by the authors. Despite the results achieved, a proper evaluation of this system based on different types of capture devices remains missing; only optical sensors are assessed by the authors. In addition, there is still a lack of comprehensive research on the extent to which fingerprint texture representations can be dimensionally reduced without impairing recognition performance. Motivated by the above fact, this work explores the trade-off between dimensionality reduction and biometric performance for the competitive fixed-length representation extractor DeepPrint for data from optical and capacitive sensors.

The remainder of this paper is organised as follows: Sect. 2 briefly introduces related works. In Sect. 3, the considered method for extracting fixed-length fingerprint representation is explained in detail. Sect. 4 presents the experimental setup and the achieved results are summarised in Sect. 5. Final remarks are outlined in Sect. 6.

## 2 Related work

The introduction of DNNs in biometrics in the last decade has led to the development of powerful face recognition systems which have replaced previously deployed schemes (*e.g.* [De19]). Those architectures allow to derive fixed-length representations which contain the most significant facial traits representing the captured subject. In order to extend such scientific works, few articles have studied the feasibility of learning fixed-length fingerprint embeddings via DNNs [ECJ19, Gr22]. Engelsma *et al.* [ECJ19] proposed a scheme for learning texture-based fixed-length fingerprint representations. Using the domain knowl-

edge injection of the minutiae map, the DNN approach produces a texture-based embedding of 192 components which reports competitive results in comparison with the one yielded by traditional minutiae-based techniques. Following this idea, Takahashi *et al.* [Ta20] included additional tasks on the multi-task learning framework proposed by Engelsma *et al.* [ECJ19]. Subsequently, Grosz and Jain [GJ22] introduced attention mechanisms within the DNN based on re-alignment strategies on local embeddings that refined the process of global embedding extraction. Afterwards, Grosz *et al.* [Gr22] showed that the minutiae-based domain knowledge combined with vision transformers increased the biometric performance of the work in [GJ22].

In general, previous approaches have focused on computing fixed-length discriminative representations from the fingerprint that achieve similar or superior biometric performance compared to traditional minutiae-based systems. Fixed-length representations can be easily combined with BTP schemes or used in a multi-modal pipeline and can therefore be deployed in privacy-protecting authentication systems. This may result in significant differences in terms of texture appearance. Therefore, fixed-length representations must be robust to these sensor type variations. However, so far, the robustness towards different sensors has not yet been studied for fixed-length fingerprint representations.

## 3 Deep fixed-length fingerprint representation

### 3.1 Dimensionality reduction via DNN classifiers

Generally, DNN classifier architectures can be used for dimensionality reduction. In our work, we consider the scenario of a classification problem with a large number of classes[1]. A DNN classifier is used to predict the probability of each possible class in the training dataset. As illustrated in Fig. 1, the architecture of such a network can typically be decomposed into four components: input data, main network, last layer representation $L$, and output $Z$, where $L$ and $Z$ are fully connected with a weight matrix $W$. While the main network can inhibit a highly complex structure, it always fulfills the task of reducing the

[1] In our case, each class corresponds to a distinct finger.

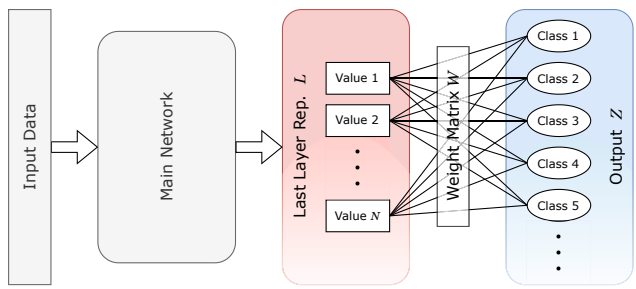



Fig. 1: Abstract representation of a typical DNN classifier architecture.

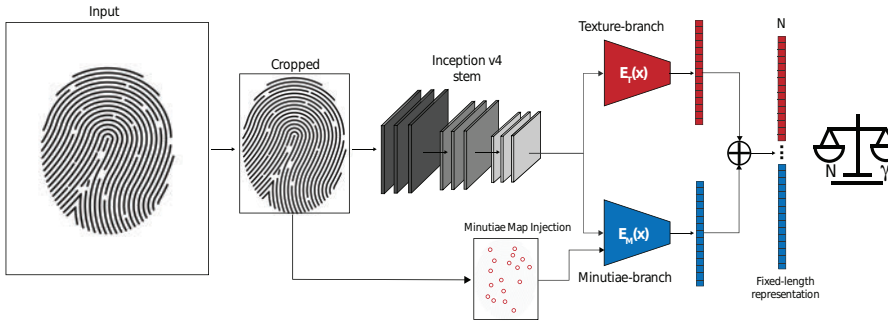



Fig. 2: Conceptual overview of the DeepPrint architecture. It consists of two branches; the upper branch $E_T(x)$ represents the fingerprint texture representation while the other one $E_M(x)$ is fed with the minutiae maps (*i.e.* minutiae coordinates and angles) to learn a compact minutiae representation. The final fixed-length representation comprises the concatenation of $E_T(x)$ and $E_M(x)$.

input data to the vector $L$. The probability weight of a class $i$ is then defined by $Z_i = \vec{L} \cdot \vec{W}_i$, where $\vec{W}_i$ denotes the $i$-th row of $W$, indicating the intra-class and inter-class similarity for the same and different classes. Thus, we can view $L$ as a dimensionality-reduced representation of the input data, which contains only the features most relevant for calculating class similarity. The dimension $N$ of this representation is determined by the number of neurons in $L$.

One should note, that the network only learns to extract representations of the classes present in the training data. In order to generalize to classes not present in the training dataset, the number of training classes must be sufficiently high and the unknown classes must be fundamentally similar to the classes seen during training. For a fingerprint dataset consisting of thousands of subjects, we can assume that this condition is fulfilled.

### 3.2 DeepPrint architecture

Fig. 2 shows the overview of the DNN-based scheme used in this paper for learning the fixed-length fingerprint representation. As mentioned in Sect. 1, we selected and implemented the competitive DeepPrint approach in [ECJ19]. This system consists of two main branches. A cropped fingerprint image is initially processed by the stem of the Inception v4 architecture (henceforth referred to as "stem"). Then the first branch $E_T(x)$, consisting of the remaining Inception v4 layers, performs the primary learning task of predicting a finger identity label. The second branch $E_M(x)$ also predicts the subject identity but it has a side task of detecting minutiae locations and orientations via the use of an AutoEncoder [ECJ19]. Thus, we guide this branch of the network to extract representations influenced by the fingerprint minutiae. In addition to these tasks, center loss [We16] is applied to both branches. The parameters of the stem are shared between the minutiae detection and representation learning branches. Finally, the embedding vectors computed by the two branches are concatenated into the last layer before the output class probabilities. In the

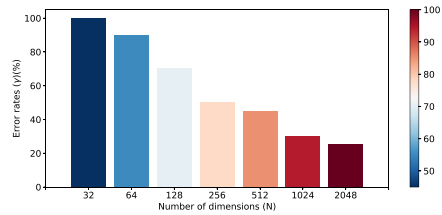



Fig. 3: Relation of different values of $N$ (dimensions of embeddings) with respect to the biometric performance $\gamma$ and the number of operations $O$ performed at a single comparison. Consider that 100% of $O$ represents the total of operations done by a fixed-length representation containing *e.g.* $N = 2{,}048$ floating points. In addition, $\gamma$ can represent some measure of evaluation in a biometric system (*e.g.* FMR at 0.1%).

original DeepPrint architecture, the size $N$ of this layer, which provides the fingerprint representation as explained in Sect. 3.1, was fixed at 192 dimensions. To train the network, we chose similar hyperparameters and loss functions as [ECJ19].

A drawback of our scheme is that the alignment step (prior to cropping the input data) was not considered. However, we investigated the effect of rotation and translation (see Fig. 6a and Fig. 6b) of the fingerprint image on fixed-length representation for the two concatenated branches on different types of sensors. Also, as part of this work, we experimented with various embedding sizes, *i.e.* $N = \{32, 64, 128, 256, 512, 1024, 2048\}$, and performed an ablation study of the complementary information provided by each branch. Theoretically, it is expected that values of $N$ have a significant impact on the biometric performance, while gaining workload (number of operations), as shown in Fig. 3.

## 4 Experimental setup

In this section, we describe the most relevant components in the training setup of the fixed-length extractor (Sect. 4.1) and databases used, while Sect. 4.2 describes metrics and protocols employed in the evaluation.

### 4.1 Training dataset

Although this work follows the architecture proposed by [ECJ19], there are some other considerations such as the training database, and pre-processing steps that differ from the original work. In order to conduct the experimental analyses of this paper, a synthetic database for training the fixed-length approach was created. Note that the original database [YJ15] used in [ECJ19] for training is not public. For the construction of the training database, synthetic fingerprint images together with the respective minutiae maps were generated by the framework SFinGe [Ca04]. In particular, 40,000 samples stemming from 4,000 unique fingers were generated. To improve the generalisation capability of the network over real fingerprint images, 200 subjects from the MCYT database [Or03], which are equivalent to 200·10 = 2,000 fingerprint instances, are selected and mixed into the set of synthetic fingerprint images forming a total of 40,000 + 48,000 samples. The

(a) Original sample (b) Data augmentation (c) Enhanced image

Fig. 4: Example of a synthetic fingerprint image generated by SFinGe with the respective pre-processing steps.

| **Databases** | **Sensors** | **Training** | | **Testing** | |
|---|---|---|---|---|---|
| | | #Subjects | #Samples | #Subjects | #Samples |
| Synthetic | Optical | 4,000 | 4,000×10 | - | - |
| MCYT [Or03] | Optical | 200×10 | 2,000×12 | 130×10 | 1,300×12 |
| | Capacitive | 200×10 | 2,000×12 | 130×10 | 1,300×12 |

Tab. 1: Summary of the training and testing sets employed in this work.

variability of the different types of sensors together with the synthetic images generated is expected to contribute to a robust extraction of distinctive features. Note that, in our experiments, each fingerprint is considered a different biometric instance and is therefore assigned a different class in training time. The fingerprint images were cropped and resized to 299×299 pixels as done in [ECJ19]. In addition, a data augmentation step based on the rotation, shifting, and variation of the brightness and contrast was randomly introduced. To enhance the image quality, a pre-processing step was considered based on Gabor wavelet transformation [Ka10]. Fig. 4 visualises examples of this virtual database composed with their corresponding pre-processing steps.

The remaining 130 different subjects from MCYT [Or03] resulting 1,300 identities are used to evaluate the CNN-based approach. Tab. 1 summarises the characteristics of the training and testing database.

### 4.2 Metrics and protocols

Biometric performance in the verification scenario was reported in accordance with the metrics defined by ISO/IEC19795-1:2021 [IS21]. The Equal Error Rate (EER), which represents the operating point at which False Match Rates (FMR) and False Non-Match Rates (FNMR) equalise, is computed. In addition, the FNMR values for several security thresholds, *i.e.* $0 \leq \text{FMR} \leq 40$ are depicted as Detection Error Trade-off (DET) curves. We also evaluate the identification rate for different rank values, *i.e.* Rank-N on a closed-set identification scenario. Note that for the verification scenario, all possible comparisons

| Embedding Size (N) | #Operations (O) | Optical | | Capacitive | |
|---|---|---|---|---|---|
| | | FNMR | EER | FNMR | EER |
| 32 | 63 | 5.36 | 1.19 | 10.60 | 2.42 |
| 64 | 127 | 5.37 | 1.26 | 10.49 | 2.38 |
| 128 | 255 | 3.51 | 1.00 | 7.91 | 2.14 |
| 256 | 511 | 3.04 | 0.94 | 8.19 | 2.31 |
| **512** | **1023** | **1.89** | **0.63** | **5.39** | **1.68** |
| 1024 | 2047 | 2.61 | 0.97 | 6.46 | 2.26 |
| 2048 | 4095 | 3.91 | 1.17 | 9.46 | 2.51 |

Tab. 2: Biometric performance (in %) for the verification scenario using the texture-based branch. FNMR values are reported for a FMR at 0.1%. The best result is highlighted in bold.

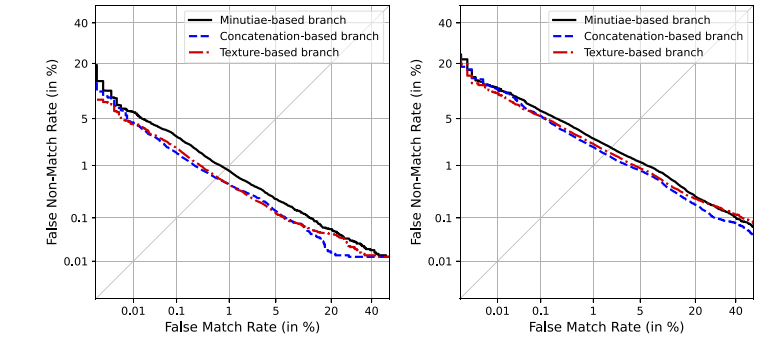
for mated and non-mated comparisons are computed, while 10-fold cross-validation is performed on the closed-set identification protocol. Furthermore, the computational workload of a single comparison in relation to the embedding size (*i.e.* feature dimensions ($N$)) and the number of operations ($O$) according to the cosine comparator are reported. Since the computed embeddings are normalised, the cosine similarity function between two fixed-length representations of size $N$ performs $N$ multiplications followed by $N-1$ additions, resulting in $N+(N-1)$ operations.

## 5 Results

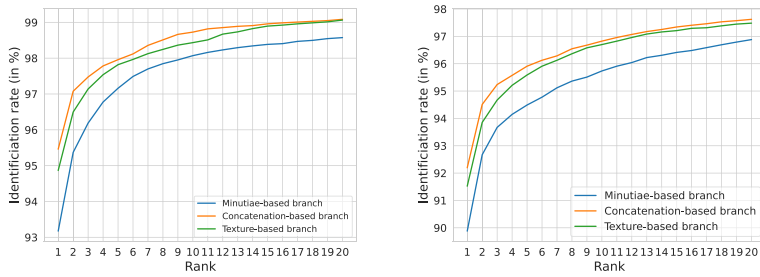
Tab. 2 reports the biometric performance for different embedding sizes ($N$) as well as their respective number of operations ($O$). Following the theoretical behaviour presented in Fig. 3, we can observe that the biometric performance improves with $N$, resulting in the best performance $N = 512$ (*i.e.* EER = 0.63% and EER = 1.68% for optical and capacitive sensors, respectively). Note a slight degradation of performance for $N > 512$, indicating the introduction of unreliable features at larger embeddings. Regarding the comparison between the performance depicted by both sensors, we perceive a significant deterioration in terms of EER and FNMR for the capacitive capture device. In particular, the algorithm for optical sensor images yields a FNMR@FMR = 0.1% of 1.89% for $N = 512$, which is approximately three times lower than the one achieved on capacitive sensor images at the same security threshold (*i.e.* FNMR = 5.39%). These results demonstrate that the feature representation computed by the DeepPrint system is affected by the sensor technology of the fingerprint capture device. Therefore, further research to overcome this deficiency is necessary.

### 5.1 Performance analysis

Fig. 5 depicts performance plots for $N = 512$ for different branches, *i.e.* minutiae, texture and concatenation-based branches, for both verification and closed-set identification scenarios. In this context, each evaluated branch is trained on a fixed-length embedding of



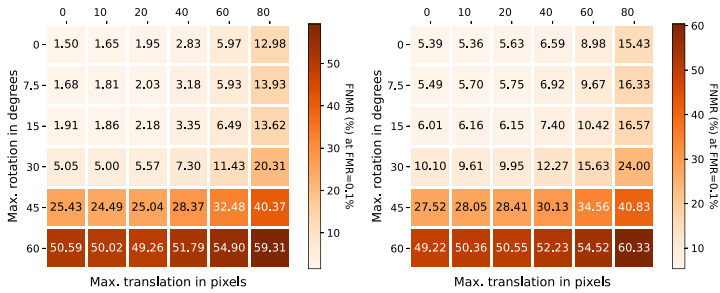

(a) Optical sensor (b) Capacitive sensor




(c) Optical sensor (d) Capacitive sensor

Fig. 5: DET curves for verification (a)-(b) and identification rate for different Rank values (c)-(d) on optical and capacitive sensors.

512 floating points. Note that both texture and the concatenation schemes in Fig. 5a and Fig. 5b achieve similar recognition performance, resulting in a similar FNMR below 2.0% for optical sensor and FNMR = 5.0% for the capacitive sensor at a FMR = 0.1%. These non-significant differences over high-security thresholds (FMR$\leq$0.01%) make the sole use of embeddings computed by the texture-based branch suitable to be combined with other approaches such as BTP and fusion schemes. It also avoids the detection of minutiae points which might lead to undesired recognition performance. On the other hand, we note, for the closed-set identification scenario in Fig. 5c and Fig. 5d, a slight improvement is obtained when the concatenation of minutiae and texture branches is performed. However, the differences between the identification rates reported by the concatenation and the texture-based branch is lower than 1.0% at Rank-1 in the optical and capacitive sensors. We observe that the worst results are obtained with the minutiae-based branch, which results in a decrease of the identification rate down to 93% and 90% at Rank-1 for the optical and capacitive sensors, respectively. We believe that the non-considered alignment step in the optimisation of minutiae maps in the $E_T(x)$ branch leads to this performance deterioration. Despite this negative observation, we confirm that the concatenation of minutiae and texture-based embeddings complements each other to improve both independent branches.




(a) Optical sensor (b) Capacitive sensor

Fig. 6: Verification performance of the 512-dimensional concatenated embeddings for different levels of simulated rotation and translation. The value on the x-Axis is the maximum rotation $r$ where the input images were randomly rotated by a value sampled from the uniform distribution over $[-r,r]$. Respectively, on the y-Axis we have the max. translation $t$, where each fingerprint image is shifted by an amount of pixels sampled from $[-t,t]^2$.

Finally, we note that similar to the results in Tab. 2, the evaluated system reports different recognition performances depending on the sensor employed.

### 5.2 Robustness analysis

As mentioned in Sect. 3 the effect of the alignment of the fingerprint pose is explored in Fig. 6a and Fig. 6b for the optical and capacitive sensors, respectively. To that end, increasing levels of rotation and translation were applied to the testing set and then, fingerprint embeddings were extracted. Here, the performance (FNMR) deteriorates disproportionately as the magnitude of the rotation and translation increases. Interestingly, the rate of deterioration appears to grow faster for rotation ($x \geq 30$, $y = 0$) compared to translation ($x = 0$, $y \geq 40$) for both sensors. These facts confirm the need for the alignment stage. Despite this, we believe that texture information contributes to some extent to reducing these negative effects. Future work should investigate this effect on the independent branches, including for minutiae map representations without texture information.

## 6 Conclusions

In this paper, we evaluated how dimensionality reduction for a state-of-the-art representation of fixed-length fingerprints affects the overall recognition performance. To do so, we analyse the degradation of biometric performance and computational workload in the comparison stage of the DeepPrint approach. Experimental results computed on a publicly available database empirically demonstrated that learned features with a dimension lower or higher than 512 floating points led to a deterioration of biometric performance. Furthermore, the differences in performance between the results obtained for the optical and

capacitive sensor indicated the need for further research in this field. In spite of this drawback, we do confirm that this fixed-length representation enables its use in combination with BTP schemes and multimodal schemes which is subject to our current research.

## Acknowledgements

This work has in part received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 860813 - TReSPAsS-ETN and 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

[Ba22] Bauspieß, P.; Olafsson, J.; Kolberg, J.; Drozdowski, P.; Rathgeb, C.; Busch, C.: Improved Homomorphically Encrypted Biometric Identification Using Coefficient Packing. In: Proc. Intl. Workshop on Biometrics and Forensics (IWBF). 2022.

[Ba23] Bauspieß, P.; Vad, L.; Myrekrok, H.; Costache, A.; Kolberg, J.; Rathgeb, C.; Busch, C.: On the Feasibility of Fully Homomorphic Encryption of Minutiae-Based Fingerprint Representations. In: 9th Intl. Conf. on Information Systems Security and Privacy ICISSP. pp. 462–470, February 2023.

[Bo22a] Boutros, F.; Damer, N.; Kirchbuchner, F.; Kuijper, A.: Elasticface: Elastic margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1578–1587, 2022.

[Bo22b] Boutros, F.; Kaehm, O.; Fang, M.; Kirchbuchner, F.; Damer, N.; Kuijper, A.: Low-resolution Iris Recognition via Knowledge Transfer. In: 2022 Intl. Conf. of the Biometrics Special Interest Group (BIOSIG). IEEE, pp. 1–5, 2022.

[Ca04] Cappelli, R.; Maio, D.; Maltoni, D. et al.: , SFinGe (Synthetic Fingerprint Generator). `http://biolab.csr.unibo.it/sfinge.html`, 2004.

[De19] Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proc. Intl. Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 4690–4699, 2019.

[ECJ19] Engelsma, JJ.; Cao, K.; Jain, AK.: Learning a fixed-length fingerprint representation. IEEE transactions on pattern analysis and machine intelligence, 43(6):1981–1997, 2019.

[GBJ19] Gong, S.; Boddeti, VN.; Jain, AK.: On the intrinsic dimensionality of image representations. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. pp. 3987–3996, 2019.

[GJ22] Grosz, S. A.; Jain, A. K.: AFR-Net: Attention-Driven Fingerprint Recognition Network. arXiv preprint arXiv:2211.13897, 2022.

[Gr22] Grosz, S. A.; Engelsma, J.; Ranjan, R.; Ramakrishnan, N.; Aggarwal, M.; Medioni, G.; Jain, A. K.: Minutiae-Guided Fingerprint Embeddings via Vision Transformers. arXiv preprint arXiv:2210.13994, 2022.

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

[Ka10] Karimimehr, N.; Shirazi, A. A. B. et al.: Fingerprint image enhancement using gabor wavelet transform. In: 2010 18th Iranian conference on electrical engineering. IEEE, pp. 316–320, 2010.

[Li17] Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proc. Intl. Conference on Computer Vision and Pattern Recognition (CVPR). pp. 212–220, 2017.

[Ma04] Maio, D.; Maltoni, D.; Cappelli, R.; Wayman, J.; Jain, A. K.: FVC2004: Third fingerprint verification competition. In: Proc. First Intl. Conf. of Biometric Authentication (ICBA). pp. 1–7, 2004.

[Ma22] Maltoni, D.; Maio, D.; Jain, A-K.; Feng, J.: Fingerprint classification and indexing. In: Handbook of Fingerprint Recognition, pp. 299–338. Springer, 2022.

[Or03] Ortega-Garcia, J.; Fierrez-Aguilar, J.; Simon, D.; Gonzalez, J.; Faundez-Zanuy, M.; Espinosa, V.; Satue, A.; Hernaez, I.; Igarza, JJ.; Vivaracho, C. et al.: MCYT baseline corpus: a bimodal biometric database. IEE Proceedings-Vision, Image and Signal Processing, 150(6):395–401, 2003.

[Sk22] SkyQuest, T.: , Global Fingerprint Sensor Market. `https://www.skyquestt.com/report/fingerprint-sensor-market`, 2022. Last accessed: September 4, 2023.

[Ta20] Takahashi, A.; Koda, Y.; Ito, K.; Aoki, T.: Fingerprint feature extraction by combining texture, minutiae, and frequency spectrum using multi-task CNN. In: 2020 IEEE Intl. Joint Conf. on Biometrics (IJCB). IEEE, pp. 1–8, 2020.

[Wa93] Watson, C.: , NIST Special Database 14, NIST Mated Fingerprint Card Pairs 2 (MFCP2), 1993.

[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: Proc. Intl. Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5265–5274, 2018.

[WD21] Wang, M.; Deng, W.: Deep face recognition: A survey. Neurocomputing, 429:215–244, 2021.

[We16] Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: Computer Vision–ECCV 2016: 14th European Conf., Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14. Springer, pp. 499–515, 2016.

[WW92] Watson, C.; Wilson, C.: NIST special database 4. Fingerprint Database, National Institute of Standards and Technology, 17(77):5, 1992.

[YJ15] Yoon, S.; Jain, AK.: Longitudinal study of fingerprint recognition. Proceedings of the National Academy of Sciences, 112(28):8555–8560, 2015.

A. Brömme, N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb,
A. Sequeira, M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# Fairness and Privacy in Voice Biometrics: A Study of Gender Influences Using wav2vec 2.0

Oubaïda Chouchane[1], Michele Panariello[1], Chiara Galdi[1], Massimiliano Todisco[1], Nicholas Evans[1]

**Abstract:** This study investigates the impact of gender information on utility, privacy, and fairness in voice biometric systems, guided by the General Data Protection Regulation (GDPR) mandates, which underscore the need for minimizing the processing and storage of private and sensitive data, and ensuring fairness in automated decision-making systems. We adopt an approach that involves the fine-tuning of the wav2vec 2.0 model for speaker verification tasks, evaluating potential gender-related privacy vulnerabilities in the process. Gender influences during the fine-tuning process were employed to enhance fairness and privacy in order to emphasise or obscure gender information within the speakers' embeddings. Results from VoxCeleb datasets indicate our adversarial model increases privacy against uninformed attacks, yet slightly diminishes speaker verification performance compared to the non-adversarial model. However, the model's efficacy reduces against informed attacks. Analysis of system performance was conducted to identify potential gender biases, thus highlighting the need for further research to understand and improve the delicate interplay between utility, privacy and equity in voice biometric systems.

**Keywords:** Speaker verification, privacy preservation, fairness, gender concealment, wav2vec 2.0

## 1 Introduction

The voice is an appealing approach to biometric authentication. Its merits include ease of use, contactless and natural interaction, efficiency, and application to authentication at a distance, e.g. over the telephone. However, the voice is a rich source of personal information and recordings of speech can be used to infer far more than just the speaker's identity, e.g. the speaker's gender [Za21], ethnicity [HRC13], and health status [SLRR21]. The safeguarding of such extraneous personal information is nowadays essential; without it, there is no guarantee that recordings of speech will not be used for purposes beyond person authentication [SDAA19].

The General Data Protection Regulation (GDPR)[2] calls for adequate protections for personal data, encompassing both *sensitive* biometric information like voice and *personal* attributes such as gender[3]. In adherence to Art. 4(1) of the GDPR, personal data processing must abide by principles of legality and fairness, managing data in line with reasonable expectations and avoiding unjust harm. Any AI-driven data processing resulting in unfair discrimination violates this principle.

[1] EURECOM, France. {*lastname* [at] *eurecom* [.] *fr*}

[2] https://gdpr-info.eu/

[3] https://www.gdpreu.org/the-regulation/key-concepts/personal-data/

As mandated by GDPR, this study particularly emphasizes privacy and fairness, focusing on gender due to its demonstrated influence on speaker authentication services [HD22] and the observed gender bias in voice assistant responses [Li19]. GDPR aims to protect the rights and freedoms of individuals, including privacy and non-discrimination, with regard to personal data processing. Concealing gender adheres to the principles of data minimization and privacy by design, limiting the risk of misuse or unauthorized data access.

In this research, we grapple with the triple challenge of utility, privacy, and fairness in speaker verification systems. Starting with fine-tuning a pre-trained wav2vec 2.0 for speaker verification tasks, we then evaluate potential vulnerabilities tied to gender privacy and the fairness of Automatic Speaker Verification (ASV) performance across genders. Subsequently, we implement an adversarial technique during the fine-tuning process to conceal gender information in the speaker embeddings, thereby enhancing user privacy. To conclude, we present a comprehensive analysis of the impact of gender information on the utility, privacy, and fairness of the systems we propose.

## 2 Related work

Significant strides have been made in speaker verification, with efforts concentrated on enhancing user privacy. These strategies prioritize the protection of gender-specific data without sacrificing system utility. Noé et al. [No20] suggested an Adversarial Auto-Encoder (AAE) method to separate gender aspects from speaker embeddings while preserving ASV performance. The approach uses an external gender classifier to analyze encoded data. Later, they leveraged a normalizing flow to control gender information in a flexible manner [No22]. In another study, Benaroya et al. [BOR21] developed a novel neural voice conversion framework using multiple AEs to create separate linguistic and extra-linguistic speech representations, allowing adjustments during the voice conversion process. Recently, Chouchane et al. [Ch23] used an adversarial approach to hide gender details in speaker embeddings while ensuring their effectiveness for speaker verification. They incorporated a Laplace mechanism layer, introducing noise to obscure gender information and offering differential privacy during inference.

In terms of fairness, research reveals a distinct disparity in ASV system performance based on gender, exposing gender bias [TD21]. Two primary strategies to mitigate this bias include pre-processing and in-processing. Pre-processing uses balanced datasets for training, as Fenu et al. [Fe20] demonstrated with gender, language, and age-balanced data. In contrast, in-processing infuses fairness directly during training, as seen in Shen et al.'s Group-Adapted Fusion Network (GFN) [Sh22] and Jin et al.'s adversarial re-weighting (ARW) approach [Ji22]. Peri et al. [PSN23] recently proposed adversarial and multi-task learning techniques for bias mitigation, highlighting a potential trade-off between system utility and fairness.

Finally, shifting focus to system utility, a cornerstone in ASV performance, the wav2vec 2.0 [Ba20], a self-supervised framework for speech representation learning, enters the

scene. The wav2vec 2.0 can be effectively adapted for speaker verification tasks [VVL22, Fa20].

## 3 Automatic speaker verification, gender recognition and suppression using wav2vec 2.0

In this section, we outline our use of the wav2vec 2.0 model, a versatile speech feature encoder that is pre-trained through self-supervision and can be adapted to specific tasks. We fine-tuned wav2vec 2.0 for three distinct tasks: speaker recognition, and gender recognition and suppression. Section 3.1 elaborates on the pre-training process, while Section 3.2 details our contributions to fine-tuning. Both procedures are graphically depicted in Fig. 1.

### 3.1 Pre-training

Given a raw audio input signal $x$, wav2vec 2.0 produces a set of $T$ feature vectors $\mathbf{c}_1, \ldots, \mathbf{c}_T$. The model is split into a 1D-convolutional encoder and a Transformer module [Va17] two main parts. First, the encoder maps the raw audio $\mathbf{x}$ to latent feature vectors $\mathbf{z}_1, \ldots, \mathbf{z}_T$. The latent features are then fed into the Transformer module to produce output feature vectors $\mathbf{c}_1, \ldots, \mathbf{c}_T$, and are also used to compute a set of quantised macro-codewords $\mathbf{q}_1, \ldots, \mathbf{q}_T$. Each macro-codeword $\mathbf{q}_t$ is the concatenation of $G$ codewords $\mathbf{q}_{t,1}, \ldots, \mathbf{q}_{t,G}$ selected from $G$ different codebooks $\mathscr{Q}_1, \ldots, \mathscr{Q}_G$, each of size $V$, learned at training time. Each codeword $\mathbf{q}_{t,j}$ is sampled from $\mathscr{Q}_j$ according to a $V$-fold categorical distribution. The distribution is optimized during pre-training and computed as $\mathbf{p}_{t,j} = \mathrm{GS}(\mathbf{z}_t)$, where GS indicates a linear layer projecting $\mathbf{z}_t$ to $V$ dimensions followed by a straight-through Gumbel-softmax estimator [JGP17].

During pre-training, the model attempts to simultaneously minimize a *contrastive* loss $\mathscr{L}_m$ and a *diversity* loss $\mathscr{L}_d$. To compute the former, some of the latent feature vectors $\mathbf{z}_1, \ldots, \mathbf{z}_T$ are randomly masked. Then, for each masked $\mathbf{z}_t$, the Transformer module attempts to compute $\mathbf{c}_t$ so that it is as similar as possible to the corresponding quantised macro-codeword $\mathbf{q}_t$, and as dissimilar as possible from other "distractor" macro-codewords $\tilde{\mathbf{q}}$ randomly sampled from the rest of the batch. The quantised macro-codewords are computed with no masking. The *diversity* loss $\mathscr{L}_d$ encourages the model to make uniform use of all the $V$ codewords in each codebook by maximizing the entropy of the average probability distribution $\bar{\mathbf{p}}_g$ produced by all $\mathbf{z}_t$ in a batch for each codebook $g$. The overall loss is:

$$\mathscr{L} = \underbrace{-\sum_{\substack{\text{masked}\\ \text{steps } t}} \log \frac{\exp\left(s(\mathbf{c}_t, \mathbf{q}_t)/\kappa\right)}{\sum_{\tilde{\mathbf{q}}} \exp\left(s(\mathbf{c}_t, \tilde{\mathbf{q}})/\kappa\right)}}_{\mathscr{L}_m} \underbrace{-\alpha \frac{1}{GV} \sum_{g=1}^{G} H\left(\bar{\mathbf{p}}_g\right)}_{\mathscr{L}_d} \tag{1}$$

Where $\kappa$ is a temperature coefficient, $s$ is the cosine similarity, $\alpha$ is a weight hyperparameter and $H$ indicates entropy.

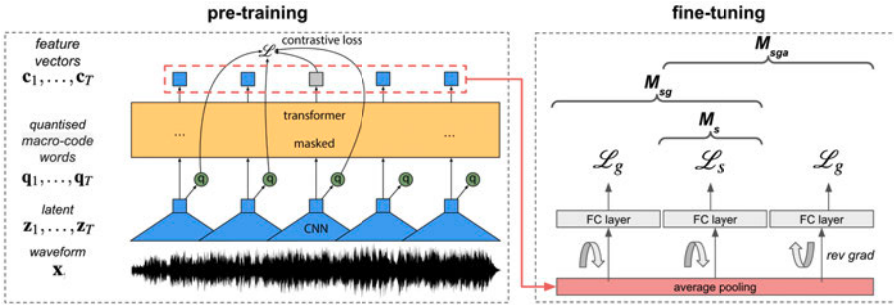



Fig. 1: Graphical depiction of the proposed systems. $M_s$: fine-tuning the speaker identification task. Msg: fine-tuning gender and speaker identification. $M_{sga}$: similar to $M_{sg}$, but the gender identification task is made adversarial.

### 3.2 Fine-tuning for speaker verification and gender recognition

In this paper, we fine-tune wav2vec 2.0 for the downstream tasks of speaker verification and gender recognition. In both cases, for each input utterance $\mathbf{x}$, the output features $\mathbf{c}_1,\ldots,\mathbf{c}_T$ are averaged across time to obtain a 1-dimensional embedding $\mathbf{c}$. In the case of gender recognition, $\mathbf{c}$ is then passed through a linear layer $f_g$ which is trained by optimising the cross-entropy loss $\mathscr{L}_g$ between the predicted logits and the true gender label for each utterance (0 for male, 1 for female). For speaker verification, $\mathbf{c}$ is passed through a different linear layer $f_s$ of $N$ output neurons, where $N$ is the number of speakers in the training dataset. The layer is then optimized to perform speaker identification by minimizing the additive angular margin (AAM) softmax loss $\mathscr{L}_s$ [Xi19]. At test time, the final embedding $\mathbf{c}$ is used as a trial or enrollment vector. Overall, the final loss can be formulated as:

$$\mathscr{L} = \lambda \mathscr{L}_s + (1-\lambda)\mathscr{L}_g \tag{2}$$

where $\lambda$ is a hyper-parameter between 0 and 1 that controls the weight of each loss component. We experimented with three different model configurations: Model 1 ($M_s$) is fine-tuned for speaker verification, i.e. $\lambda = 1$; Model 2 ($M_{sg}$) is fine-tuned for both tasks, i.e. $\lambda = 0.5$; Model 3 ($M_{sga}$) is optimised in a similar manner, though with a gradient reversal layer [Ga16] $g_r$ to suppress gender information.

The optimization process becomes an adversarial game between $f_g$, which attempts to minimize $\mathscr{L}_g$, and the backbone, which attempts to maximize it. Meanwhile, the $\mathscr{L}_s$ component is optimized as usual.

## 4 Experimental setup

Described in this section are the databases used for all experimental work, the metrics used for evaluation, and the fine-tuning procedure.

### 4.1 Databases

We used the VoxCeleb1 and VoxCeleb2 speaker recognition databases [NCZ17, CNZ18]. VoxCeleb1 includes over 100,000 utterances from 1,251 celebrities, while VoxCeleb2 contains over a million utterances from 6,112 speakers. Both datasets, compiled from YouTube videos, are widely used for speaker recognition and voice-related machine-learning tasks. Fine-tuning is performed using the VoxCeleb2 development set which contains data collected from 5994 unique speakers of which 3682 are male and 2312 are female, corresponding to an imbalance in favour of male speakers of 22.9%. To assess the performance of our systems, we used the VoxCeleb1 test set, which consists of 40 unique speakers of which 25 are male and 15 are female.

### 4.2 Metrics

A range of key metrics was selected, many of which are derived from the evaluation of biometric classification systems, e.g. speaker verification and gender classification. The following describes how they are used to jointly assess the utility, privacy, and fairness of the models under scrutiny.

**Utility** is measured by assessing the performance for the task of automatic speaker verification (ASV) in terms of equal error rate (EER). EER is the operating point defined by the detection threshold $\tau$ at which the false acceptance rate (FAR) and the false rejection rate (FRR) are equal.

**Privacy** relates to the difficulty of an adversary to infer sensitive attributes. We use AUC (area under the receiver operating characteristic curve) metric to gauge privacy. In contrast to EER, AUC provides a comprehensive view, which is ideal for evaluating system security across diverse threshold selections.

**Fairness** is aimed at ensuring that a system behaves equally with all subgroups of the target population. Many approaches for measuring fairness have been proposed recently and there is still no agreement on which is the most appropriate. We adopted two different metrics with the aim of giving a more meaningful insight into the fairness of the models.

The first adopted approach aims at ensuring that the error rates for all demographic groups fall within a small margin $\varepsilon$. However, for practical purposes, given a pair of demographic groups $D = d_1, d_2$, we calculate $A(\tau)$ and $B(\tau)$, as:

$$A(\tau) = max\left(\left|FAR^{d_1}(\tau) - FAR^{d_2}(\tau)\right|\right) \tag{3}$$

$$B(\tau) = max\left(\left|FRR^{d_1}(\tau) - FRR^{d_2}(\tau)\right|\right). \tag{4}$$

These represent the maximum absolute differences in FAR and FRR across all groups. In a perfect system, both $A(\tau)$ and $B(\tau)$ would equal 0, reflecting identical error rates across all groups.

The Fairness Discrepancy Rate (FDR) [dFPM21] is defined as:

$$FDR(\tau) = 1 - (\alpha A(\tau) + (1-\alpha)B(\tau)) \tag{5}$$

where the hyper-parameter $\alpha \in [0,1]$ determines the relative importance of false alarms. FDR ranges between 0 and 1 and would equal 1 in the case of a perfectly fair system. However, achieving perfect fairness is often unrealistic, leading to the introduction of $\varepsilon$ which allows for certain discrepancies. Though $\varepsilon$ isn't included in the FDR calculation, it's vital for defining an acceptable level of fairness and interpreting FDR results.
Given the absence of a universal $\varepsilon$ and the complexities of biometrics, absolute fairness often isn't achievable. Thus, FDR and Area Under FDR (auFDR) are used to compare the fairness of different biometric systems. The auFDR is calculated by integrating the FDR over a specific threshold range $\tau$, denoted as $FAR_x$. To fairly compare the auFDR between different systems, the specific range of $\tau$ used must be reported, as the value of the auFDR depends on this range. Like the FDR, the auFDR varies from 0 to 1, with higher values denoting better fairness. In our experiments, we set the range to FARs below 0.1; FARs above this value correspond to a system with little practical interest.

The second metric is the fairness activation discrepancy (FAD), which we use to investigate fairness *within* the network. FAD is inspired by *InsideBias* [Se21], a fairness metric developed originally for the study of face biometrics and which we adapt to our study of voice biometrics. Notably, this adaptation of FAD for voice biometrics is a novel metric in this context.

*InsideBias* is based upon the examination of neuron activations and the comparison of model responses to demographic groups within distinct layers. In [Se21], the authors observed that underrepresented groups corresponded to lower average activations. In the case of voice biometrics, the output of each network layer can be viewed as a bi-dimensional tensor of neurons over temporal frames:

$$A_{ij}^{[l]} = \Psi^{[l]}(\cdot) \tag{6}$$

where $i = 1,...,N$, $j = 1,...,M$, $A_{ij}$ is the activation of the $i^{th}$ neuron for the $j^{th}$ temporal frame, $\Psi^{[l]}$ is the activation function at layer $l$, and $N$ and $M$ are the total number of neurons and frames respectively. For each layer $l$ we calculate the root mean square of $A_{ij}$ over the $j^{th}$ frame which serves to account for large positive or negative activations. Then, we take the maximum along the $i^{th}$ feature dimension:

$$\Lambda^{[l]} = \max_i \sqrt{\left(\frac{1}{M}\sum_j A_{ij}^2\right)} \tag{7}$$

The FAD is defined as the absolute difference between $\Lambda$ for a pair of two distinct groups and is given by $FAD = |\Lambda_{d_1} - \Lambda_{d_2}|$. Near-zero values of FAD indicate better fairness.

### 4.3 Fine-tuning procedure

$M_s$, $M_{sg}$ and $M_{sga}$ models are fine-tuned as described in Section 3.2. An initial warm-up is applied to the linear classification heads for the first $10k$ optimization steps, keeping the wav2vec 2.0 backbone frozen. The entire model is then fine-tuned in an end-to-end fashion for the remaining steps. We use the pre-trained model provided by Baevski et al. [Ot19][4]. Performance for the speaker identification task exceeded 95% accuracy for all three models whereas the adversarial system delivered a gender recognition accuracy of only 47%.

### 4.4 Gender privacy threat models

The ability of the systems to conceal the gender information contained in its embeddings is measured by simulating the presence of a third party (an *attacker*) training a 2-layer fully-connected neural network $\mathcal{N}$ to infer the speaker gender from utterance embeddings. We consider two threat models. In the first one, the attacker is not aware that gender concealment has taken place (*uninformed attack* (uIA)) and therefore trains $\mathcal{N}$ on embeddings that are not gender-protected (in this case, those produced by $M_s$ and $M_{sg}$). In the second one, the attacker is aware that model $M_{sga}$ was used to protect the gender identity (*informed attack* (IA)), has access to that model, and trains $\mathcal{N}$ on embeddings produced by that same model. We expect this to result in a more effective attack.

## 5 Experimental results

We present results for each of the three models $M_s$, $M_{sg}$, and $M_{sga}$. Performance is assessed in terms of utility, privacy, and fairness.

In terms of utility, the performance of model $M_s$ is in line with state-of-the-art automatic speaker verification systems, achieving an EER of 2.36% as shown in Table 1. The performance of model $M_{sg}$ and $M_{sga}$ are slightly worse, 3.23% and 3.89% respectively, showing that gender influence does not improve speaker recognition. Furthermore, an analysis of the EER broken down by gender shows small differences in speaker recognition for the two genders.

Fairness performances are shown at the bottom of the Table 1 in terms of the auFDR for different values of $\alpha$. All auFDR results are close to 1, indicating reasonable fairness for each group. Fig. 2 depicts a plot of the FDR against the threshold for $\alpha = 0.5$. Profiles are shown for all three systems. The FDR is in all cases above 0.9, and the $M_s$ system is always the fairest for each $\tau$. Again, gender influence does not improve fairness.

Privacy performances are presented in Table 2. AUC results for uninformed attacks (uIA) are shown at the top. When training and testing are performed using embeddings generated using the same, unprotected models, the AUC is 97.09% and 98.07% for $M_s$ and

[4] https://github.com/facebookresearch/fairseq/tree/main/examples/

| | | | Models | | |
|---|---|---|---|---|---|
| | | | $M_s$ | $M_{sg}$ | $M_{sga}$ |
| EER(%) | Overall | | 2.36 | 3.23 | 3.89 |
| | Male | | 3.12 | 4.22 | 4.98 |
| | Female | | 3.05 | 4.21 | 5.26 |
| auFDR | $\alpha$ | 0 | 0.98 | 0.97 | 0.96 |
| | | 0.25 | 0.97 | 0.97 | 0.95 |
| | | 0.5 | 0.97 | 0.96 | 0.94 |
| | | 0.75 | 0.96 | 0.95 | 0.92 |
| | | 1 | 0.95 | 0.94 | 0.91 |

Tab. 1: Performance analysis of the three models for utility and fairness, including EER breakdown by gender and auFDR across various $\alpha$ values (refer to eq.5) for $\tau$ ranging from 0.1% to 10%.

| | Data | | Attack |
|---|---|---|---|
| | Training | Test | AUC (%) |
| uIA | $M_s$ | $M_s$ | 97.09 |
| | $M_s$ | $M_{sga}$ | **46.80** |
| | $M_{sg}$ | $M_{sg}$ | 98.07 |
| | $M_{sg}$ | $M_{sga}$ | **40.76** |
| IA | $M_{sga}$ | $M_{sga}$ | 96.27 |

Tab. 2: Assessment of gender concealment effectiveness under different threat scenarios in terms of AUC.

$M_{sg}$ models, respectively, demonstrating a lack of privacy protection. In contrast, when the same uninformed attack is made on the gender-protected model $M_{sga}$, the AUC drops to 46.80% and 40.76% respectively. This significant decrease indicates that the gender classifier predictions become nearly random, successfully concealing the gender information, demonstrating effective protection of privacy.
Performances for the informed attack (IA) are shown in the last row of Table 2. When embeddings are extracted with the $M_{sga}$ model, the AUC is much higher, at 96.27%. This result underlines the difficulty of obfuscating gender information from embeddings. Fig. 3 reveals an explanation. It illustrates a projection by principal component analysis of the embeddings generated by each of the three models. While the $M_{sga}$ model is adversely trained with respect to gender cues, Fig. 3c shows that they persist. We see that, rather than fully obfuscating gender cues, $M_{sga}$ only rotates the principal components hence why, when trained on similarly-treated training data, gender can still be recognised.

Finally, an analysis of internal bias in terms of FAD has been performed at different network layers considering male and female groups. This analysis aims to provide insights into the comparative measures of fairness across three distinct models and how they dynamically propagate through the various layers. By examining the internal bias at each layer, we can better understand the impact of model architecture and training data on fairness outcomes. As illustrated in Fig. 4, 32 layers were selected in total from the wav2vec 2.0 model. These include 8 layers from the 1D-convolutional encoder and 24 intermediate activation layers from the Transformer modules.
Fig. 4 shows the FAD values calculated at different layers. The first layers of the CNNs display similar fairness, likely due to their focus on low-level features. Contrastingly, Transformer layers, which handle high-level features, have wider fairness variations. $M_s$ and $M_{sga}$ show a complementary behavior as when one achieves high FAD, the other has lower FAD, and vice versa. This could be because $M_s$ was fine-tuned for speaker verification, while $M_{sga}$, with its gradient reversal layer, was trying to suppress gender information. As layers progress, all models converge to FAD values, with $M_s$ being the fairest at the end, confirming what is observed in terms of auFDR.

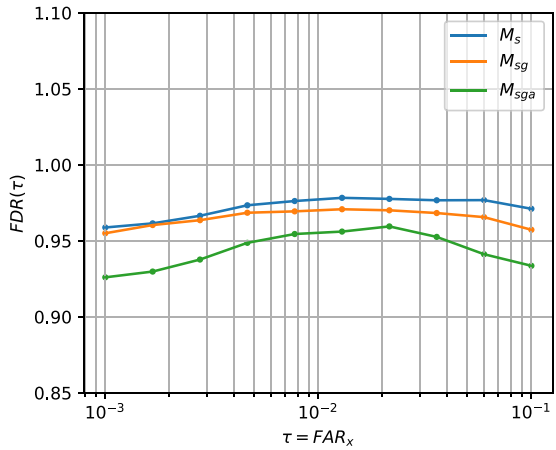



Fig. 2: FDR of different ASV systems for different decision thresholds for $\tau$ from 0.1% to 10%

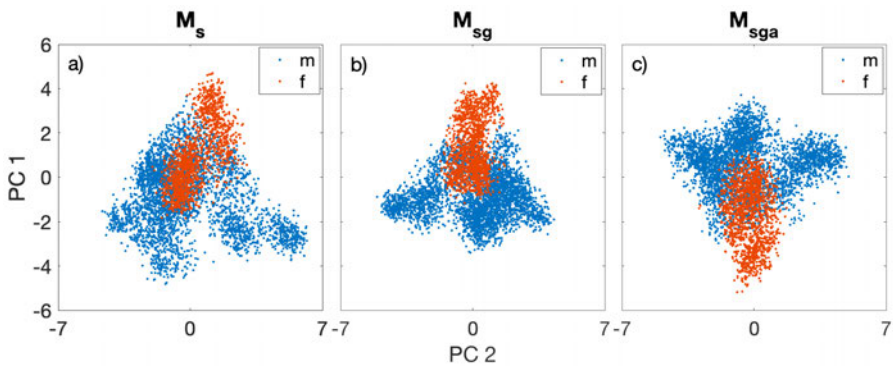



Fig. 3: PCA visualizations of features from three models illustrating gender recognition capabilities. Blue points correspond to males and red to females.

## 6 Conclusions and Future Directions

This research explored the influence of gender information while fine-tuning wav2vec 2.0 for speaker verification. We proposed three models: $M_s$, $M_{sg}$, and $M_{sga}$, each with a different focus: speaker recognition, speaker recognition with gender classification, and speaker recognition with gender obfuscation, respectively. Our experiments revealed that $M_s$ succeeds in speaker verification (EER of 2.36%), while $M_{sga}$, designed to hide gender information, performed much worse (EER of 3.89%). Interestingly, improving gender recognition in the $M_{sg}$ model did not lead to better speaker verification performance (EER of 3.23%). Privacy evaluations showed effective gender obfuscation against uninformed attacks, but informed attackers could still extract gender information. Fairness evaluations,

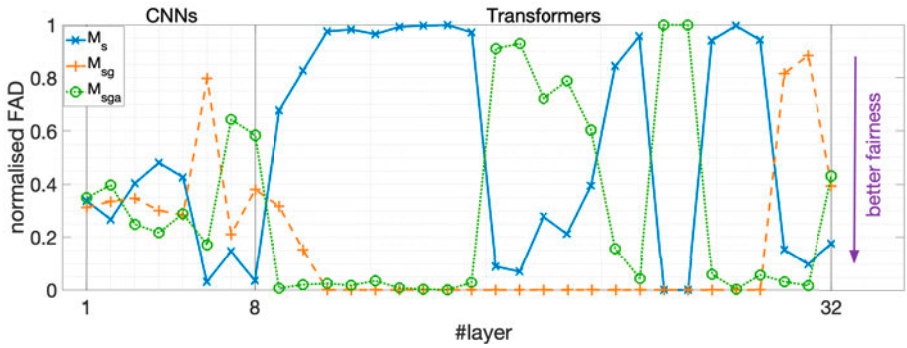



Fig. 4: Normalised Fairness Activation Discrepancy (FAD) of different systems at different wav2vec 2.0 module layers.

based on FDR, revealed that highlighting or hiding gender did not significantly impact the fairness of the systems. Furthermore, an analysis of FAD across model layers showed more disparities within Transformer layers, but all systems eventually converged to FAD values that match the auFDR assessment, with system $M_s$ showing superior fairness. In summary, while we achieved notable results in utility and privacy protection against uninformed attacks, future work includes strengthening gender obfuscation against informed attacks and enhancing fairness across systems.

## 7 Acknowledgements

This work is supported by the TReSPAsS-ETN project funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 860813 and partly supported by the VoicePersonae project funded by the French Agence Nationale de la Recherche (ANR) and the Japan Science and Technology Agency (JST).

## References


[Ba20] Baevski, Alexei; Zhou, Yuhao; Mohamed, Abdelrahman; Auli, Michael: wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In (Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.F.; Lin, H., eds): Advances in Neural Information Processing Systems. volume 33. Curran Associates, Inc., pp. 12449–12460, 2020.

[BOR21] Benaroya, Laurent; Obin, Nicolas; Roebel, Axel: Beyond Voice Identity Conversion: Manipulating Voice Attributes by Adversarial Learning of Structured Disentangled Representations. arXiv preprint arXiv:2107.12346, 2021.

[Ch23] Chouchane, Oubaïda; Panariello, Michele; Zari, Oualid; Kerenciler, Ismet; Chihaoui, Imen; Todisco, Massimiliano; Önen, Melek: Differentially Private Adversarial Auto-Encoder to Protect Gender in Voice Biometrics. In: Proceedings of the 2023 ACM Workshop on Information Hiding and Multimedia Security. pp. 127–132, 2023.

[CNZ18] Chung, Joon Son; Nagrani, Arsha; Zisserman, Andrew: Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622, 2018.

[dFPM21] de Freitas Pereira, Tiago; Marcel, Sébastien: Fairness in biometrics: a figure of merit to assess biometric verification systems. IEEE Transactions on Biometrics, Behavior, and Identity Science, 4(1):19–29, 2021.

[Fa20] Fan, Zhiyun; Li, Meng; Zhou, Shiyu; Xu, Bo: Exploring wav2vec 2.0 on speaker verification and language identification. arXiv preprint arXiv:2012.06185, 2020.

[Fe20] Fenu, Gianni; Medda, Giacomo; Marras, Mirko; Meloni, Giacomo: Improving fairness in speaker recognition. In: Proceedings of the 2020 European Symposium on Software Engineering. pp. 129–136, 2020.

[Ga16] Ganin, Yaroslav; Ustinova, Evgeniya; Ajakan, Hana; Germain, Pascal; Larochelle, Hugo; Laviolette, François; March, Mario; Lempitsky, Victor: Domain-Adversarial Training of Neural Networks. Journal of Machine Learning Research, 17(59):1–35, 2016.

[HD22] Hutiri, Wiebke Toussaint; Ding, Aaron Yi: Bias in automated speaker recognition. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 230–247, 2022.

[HRC13] Hanani, Abualsoud; Russell, Martin J; Carey, Michael J: Human and computer recognition of regional accents and ethnic groups from British English speech. Computer Speech & Language, 27(1):59–74, 2013.

[JGP17] Jang, Eric; Gu, Shixiang; Poole, Ben: Categorical Reparameterization with Gumbel-Softmax. In: International Conference on Learning Representations. 2017.

[Ji22] Jin, Minho; Ju, Chelsea J-T; Chen, Zeya; Liu, Yi-Chieh; Droppo, Jasha; Stolcke, Andreas: Adversarial reweighting for speaker verification fairness. arXiv preprint arXiv:2207.07776, 2022.

[Li19] Lima, Lanna; Furtado, Vasco; Furtado, Elizabeth; Almeida, Virgilio: Empirical analysis of bias in voice-based personal assistants. In: Companion Proceedings of the 2019 World Wide Web Conference. pp. 533–538, 2019.

[NCZ17] Nagrani, Arsha; Chung, Joon Son; Zisserman, Andrew: Voxceleb: a large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612, 2017.

[No20] Noé, Paul-Gauthier; Mohammadamini, Mohammad; Matrouf, Driss; Parcollet, Titouan; Nautsch, Andreas; Bonastre, Jean-François: Adversarial disentanglement of speaker representation for attribute-driven privacy preservation. arXiv preprint arXiv:2012.04454, 2020.

[No22] Noé, Paul-Gauthier; Nautsch, Andreas; Matrouf, Driss; Bousquet, Pierre-Michel; Bonastre, Jean-François: A bridge between features and evidence for binary attribute-driven perfect privacy. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 3094–3098, 2022.

[Ot19] Ott, Myle; Edunov, Sergey; Baevski, Alexei; Fan, Angela; Gross, Sam; Ng, Nathan; Grangier, David; Auli, Michael: fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In: Proceedings of NAACL-HLT 2019: Demonstrations. 2019.

[PSN23] Peri, Raghuveer; Somandepalli, Krishna; Narayanan, Shrikanth: A study of bias mitigation strategies for speaker recognition. Computer Speech & Language, 79:101481, 2023.

[SDAA19] Shaqra, Ftoon Abu; Duwairi, Rehab; Al-Ayyoub, Mahmoud: Recognizing emotion from speech based on age and gender using hierarchical models. Procedia Computer Science, 151:37–44, 2019.

[Se21] Serna, I.; Pena, A.; Morales, A.; Fierrez, J.: InsideBias: Measuring Bias in Deep Networks and Application to Face Gender Biometrics. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE Computer Society, Los Alamitos, CA, USA, pp. 3720–3727, jan 2021.

[Sh22] Shen, Hua; Yang, Yuguang; Sun, Guoli; Langman, Ryan; Han, Eunjung; Droppo, Jasha; Stolcke, Andreas: Improving fairness in speaker verification via group-adapted fusion network. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 7077–7081, 2022.

[SLRR21] Solana-Lavalle, Gabriel; Rosas-Romero, Roberto: Analysis of voice as an assisting tool for detection of Parkinson's disease and its subsequent clinical interpretation. Biomedical Signal Processing and Control, 66:102415, 2021.

[TD21] Toussaint, Wiebke; Ding, Aaron Yi: Sveva fair: A framework for evaluating fairness in speaker verification. arXiv preprint arXiv:2107.12049, 2021.

[Va17] Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N; Kaiser, Ł ukasz; Polosukhin, Illia: Attention is All you Need. In (Guyon, I.; Luxburg, U. Von; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; Garnett, R., eds): Advances in Neural Information Processing Systems. volume 30. Curran Associates, Inc., 2017.

[VVL22] Vaessen, Nik; Van Leeuwen, David A: Fine-tuning wav2vec2 for speaker recognition. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 7967–7971, 2022.

[Xi19] Xiang, Xu; Wang, Shuai; Huang, Houjun; Qian, Yanmin; Yu, Kai: Margin Matters: Towards More Discriminative Deep Neural Network Embeddings for Speaker Recognition. In: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). pp. 1652–1656, 2019.

[Za21] Zaman, Syed Rohit; Sadekeen, Dipan; Alfaz, M Aqib; Shahriyar, Rifat: One Source to Detect them All: Gender, Age, and Emotion Detection from Voice. In: 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC). pp. 338–343, 2021.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# Face Image De-identification Based on Feature Embedding for Privacy Protection

Goki Hanawa, Koichi Ito, Takafumi Aoki [1]

**Abstract:** With the expansion of social networking services, a large number of face images have been disclosed on the Internet. Since face recognition makes it easy to collect face images of specific persons, the collected face images can be used to attack face recognition systems, such as spoofing attacks. Face image de-identification, which makes face recognition difficult without changing the appearance of the face image, is necessary for disclosing face images safely on the Internet. In this paper, we propose a face image de-identification method by embedding facial features of another person into a face image. The proposed method uses a convolutional neural network to generate a face image that can be recognized as that of another person while preserving the appearance of the face image. Through a set of experiments using a public face image dataset, we demonstrate that the proposed method preserves the appearance of face images and has high de-identification performance against unknown face recognition models compared to conventional methods.

**Keywords:** De-identification, Face recognition, Privacy protection

## 1 Introduction

Face Recognition [LJ11] is a technology that identifies individuals using features such as facial texture and the position of facial parts. In face recognition, a face image taken from a distance using a standard RGB camera can be used for recognition. This technology is low-cost since it does not require a dedicated sensor, and it is highly convenient since it is non-contact and non-intrusive. Due to the above advantages, face recognition is used for login authentication in smartphones and personal authentication at immigration control, etc. On the other hand, there is a problem that face images can be easily collected. With the increasing use of Social Networking Services (SNS), a large number of face images have been available on the Internet. Malicious persons can not only collect a large number of face images from the Internet, but also can use face recognition to collect face images of specific persons. The collected face images can be used to attack face recognition systems, such as spoofing attacks [Ma19].

De-identification, which makes face recognition difficult while preserving the appearance of the face image, has been investigated to protect the privacy of SNS users and to allow them to safely disclose their face images. Major methods [Ya20, Ya21, Sh20] for de-identifying face images employ Adversarial Examples (AEs) [GSS15]. AEs are images that have been perturbed to induce misclassification of the classification models. It is known that the perturbation for generating AEs is strongly depending on the classification

[1] Graduate School of Information Sciences, Tohoku University, Japan

model used in training. Conventional methods exhibit high de-identification performance for the face recognition model used to generate the perturbation in training. The problem is that the de-identification performance degrades for unknown face recognition models.

In this paper, we propose a face image de-identification method using Convolutional Neural Network (CNN) to embed facial features of another person into a target face image, making face recognition difficult while preserving the appearance of the face image. The proposed method makes the features extracted from the target face image closer to the other person, and thus enables de-identification independent of face recognition models. By embedding features from non-real face images generated by StyleGAN2 [Ka20], the features extracted from the target face image do not correspond to real persons, thus protecting the privacy of both the target face and the face to be embedded. Through a set of experiments using the Labeled Faces in the Wild (LFW) dataset [Hu07], we compare the image quality and de-identification performance of the de-identified images generated by the proposed method with those of conventional methods, and demonstrate the effectiveness of the proposed method.

## 2 Related Work

It is well known that the traditional methods of face image de-identification are blurring and masking of face images [RAP16]. These methods make face images difficult to identify for both humans and authentication models. The problem is that the appearance of the original face image is not preserved, limiting the applications of the de-identified face images. When a person discloses his or her own face images on the Internet, such as in SNS, it is important that the face images can be recognized by humans. Therefore, it is necessary to develop a de-identification method that makes face recognition difficult while preserving the appearance of the original face images.

Recently, several methods have been proposed to de-identify face images while preserving their appearance using deep learning. Major methods [Ya20, Ya21, Sh20] utilize AEs [GSS15], which are images that are perturbed to induce misclassification in the classification models. Face images can be de-identified by adding perturbations that make face recognition difficult. Larger perturbations enhance the de-identification performance, although they significantly change the appearance of the face images. On the other hand, smaller perturbations preserve the appearance of the face images, although they do not provide sufficient de-identification performance. Therefore, a balance between the appearance of the de-identified image and the de-identification performance is important in face image de-identification. Yang et al. proposed two types of face de-identification methods: Landmark-Guided Cutout (LGC) [Ya20] and the Targeted Identity-Protection Iterative Method (TIP-IM) [Ya21]. LGC [Ya20] adds constraints based on facial landmarks to the Fast Gradient Sign Method (FGSM) [GSS15], which can be applied to a variety of images, and specializes on de-identification of face images. The balance between appearance and de-identification performance can be adjusted by the hyperparameter $\varepsilon$ that controls the magnitude of the perturbation. TIP-IM [Ya21] generates de-identified images based on the Maximum Mean Discrepancy (MMD), which is the difference between the data distri-

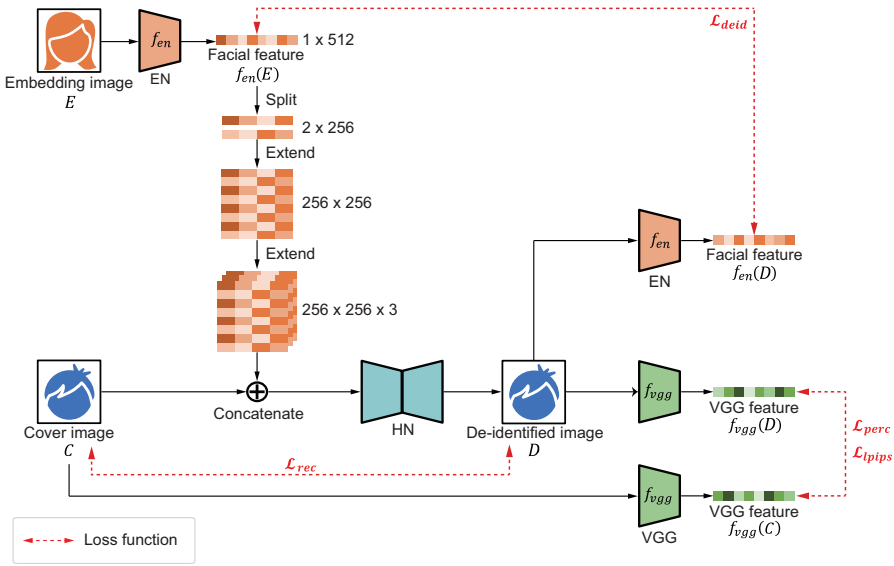



Fig. 1: Overview of the network architecture and loss functions in the proposed method.

bution of the set of original face images and the set of de-identified images. The balance between appearance and de-identification performance can be adjusted by the weight $\gamma$ of the MMD-based loss function in addition to the $\varepsilon$. Shawn et al. propose Fawkes [Sh20], which generates de-identified images by adding a minimal perturbation such that the facial features extracted from the original face image are significantly shifted in the feature space. The balance between appearance and de-identification performance can be adjusted through three modes[3] that control the magnitude of the perturbation.

The perturbations added to AEs are highly depending on the face recognition model on which the perturbation is generated. Although AEs are effective for the face recognition model on which the perturbation is generated, they may not exhibit sufficient de-identification performance for unknown face recognition models. The facial recognition model targeted by the attackers is basically unknown. Therefore, the face de-indentification methods independent of face recognition models are necessary to enhance the practicality of face image de-identification.

## 3 Face Image De-Identification Based on Feature Embedding

This section describes a de-identification method for face images by embedding facial features of other persons into the face images. Face images de-identified by the proposed method have high image quality since the face images are not perturbed like AEs. Fig. 1

[3] `https://github.com/Shawn-Shan/fawkes`

illustrates the overview of the proposed method. In the following, we describe the details of the network architecture of the proposed method and the loss function used in training.

### 3.1 Network Architecture

The proposed method consists of an Extracting Network (EN) that extracts facial features from face images and a Hiding Network (HN) that embeds facial features into a face image. In the following, the face image to be de-identified is denoted as cover image $C$, the original face image of the facial features to be embedded into $C$ is denoted as embedding image $E$, and the image where the features extracted from $E$ are embedded into $C$ is denoted as de-identified image $D$. EN is a trained face recognition model that extracts facial features $f_{en}(E)$ and $f_{en}(D)$ from an embedding image $E$ and a de-identified image $D$, respectively. Note that the same face recognition model $f_{en}$ must be used to extract $f_{en}(E)$ and $f_{en}(D)$ in training, while a different face recognition model can be used for feature extraction in test. HN generates de-identified image $D$ by embedding facial features $f_{en}(E)$ into cover image $C$ to be de-identified. The proposed method employs U-Net [RFB15] as the network architecture of HN. U-Net consists of an encoder and a decoder, and these are connected by skip connections to suppress gradient vanishing. To further suppress gradient vanishing, residual blocks used in ResNet [He16] are used in the encoder. The facial features $f_{en}(E)$ are not directly embedded, but are replicated to the same size as cover image $C$ before embedding. First, a $512 \times 1$ face feature $f_{en}(E)$ is transformed into a 2D matrix of $256 \times 2$ and expanded to $256 \times 256$ by duplicating and merging 128 of them in the height direction. Next, this $256 \times 256$ matrix is replicated and combined in the channel direction to expand it to the same size as cover image $C$ with $256 \times 256 \times 3$. Then, the expanded facial features $f_{en}(E)$ are concatenated with cover image $C$ in the channel direction and input to HN.

### 3.2 Loss Functions

As mentioned above, the two ENs in Fig. 1 are trained face recognition models with fixed weights. Therefore, only the HN need to be trained in the proposed method. We use the following four loss functions to train HN.

**(i) Reconstruction loss** $\mathscr{L}_{rec}$: $\mathscr{L}_{rec}$ is a loss function that reduces the pixel-wise difference between cover image $C$ and de-identified image $D$, and is defined by

$$\mathscr{L}_{rec} = ||C - D||_2. \tag{1}$$

**(ii) Perception loss** $\mathscr{L}_{perc}$ [JAL16]: Cover image $C$ and de-identified image $D$ are input to VGG-19 [SZ15] trained on ImageNet [De09] to obtain the features $f_{vgg}(C)$ and $f_{vgg}(D)$ output from the final layer. $\mathscr{L}_{perc}$ is a loss function that reduces the difference of global features by reducing the difference between $f_{vgg}(C)$ and $f_{vgg}(D)$, and is defined by

$$\mathscr{L}_{perc} = ||f_{vgg}(C) - f_{vgg}(D)||_2. \tag{2}$$

**(iii) Learned Perceptual Image Patch Similarity (LPIPS) loss** $\mathscr{L}_{lpips}$ [Zh18]: Cover image $C$ and de-identified image $D$ are input to VGG-16 [SZ15] trained on ImageNet [De09] to obtain the features $c^l$ and $d^l$ $(l = 1,2,\ldots,L)$ output from each layer, where $c^l = f^l_{vgg}(C)$, $d^l = f^l_{vgg}(D)$, and $L$ is the total number of layers in VGG-16. $\mathscr{L}_{lpips}$ is a loss function that reduces the difference between local and global features by reducing the difference between the weighted sum of $c^l$ and $d^l$, and is defined by

$$\mathscr{L}_{lpips} = \sum_{l=1}^{L} \frac{1}{H^l W^l} \sum_{i=1}^{H^l} \sum_{j=1}^{W^l} ||w^l \odot (c^l_{ij} - d^l_{ij})||_2, \tag{3}$$

where $H^l$ and $W^l$ are the height and width of the feature map output from layer $l$, respectively. $w^l$ indicates the weights of each channel for the features output from layer $l$, and $\odot$ indicates an operator for element-wise product.

**(iv) De-identification loss** $\mathscr{L}_{deid}$: Embedding image $E$ and de-identified image $D$ are input to EN to obtain facial features $f_{en}(E)$ and $f_{en}(D)$, respectively. $\mathscr{L}_{deid}$ is a loss function that makes $f_{en}(D)$ extracted from $D$ similar to $f_{en}(E)$ extracted from $E$ by increasing the cosine similarity between $f_{en}(E)$ and $f_{en}(D)$, and is defined by

$$\mathscr{L}_{deid} = 1 - \cos(f_{en}(E), f_{en}(D)), \tag{4}$$

where $\cos(f_{en}(E), f_{en}(D))$ indicates the cosine similarity between $f_{en}(E)$ and $f_{en}(D)$.

The total loss function $\mathscr{L}$ for training HN is defined by

$$\mathscr{L} = \lambda_{rec}\mathscr{L}_{rec} + \lambda_{perc}\mathscr{L}_{perc} + \lambda_{lpips}\mathscr{L}_{lpips} + \lambda_{deid}\mathscr{L}_{deid}, \tag{5}$$

where $\lambda_{rec}$, $\lambda_{perc}$, $\lambda_{lpips}$, and $\lambda_{deid}$ are weights for $\mathscr{L}_{rec}$, $\mathscr{L}_{perc}$, $\mathscr{L}_{lpips}$, and $\mathscr{L}_{deid}$, respectively. De-identified image $D$ that preserves the appearance of $C$ is generated by using $\mathscr{L}_{rec}$, $\mathscr{L}_{perc}$, and $\mathscr{L}_{lpips}$. De-identified image $D$ that makes the face recognition model misidentify $D$ as the person in $E$ is generated by using $\mathscr{L}_{deid}$.

## 4 Experiments and Discussion

This section describes experiments to evaluate the performance of the proposed method for face image de-identification.

### 4.1 Datasets

In the training of the proposed method, we use CelebFaces Attributes (CelebA)[4], which is a large-scale public face image dataset, and Generated Faces by StyleGAN2 (GFSG2)[5],

[4] `https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html`

[5] `https://github.com/NVlabs/stylegan2` images/100k-generated-images/ffhq-1024x1024/stylegan2-config-f-psi-0.5


Fig. 2: Examples of face images in each dataset used in the experiments.

which is a synthetic face image dataset generated by StyleGAN2 [Ka20]. CelebA [Li15] consists of 202,599 face images of 10,177 people. The randomly selected 199,599 images are used for training, and the remaining 3,000 images are used for validation. GFSG2 consists of 5,000 synthetic face images, and all of them are used for training. We use Labeled Faces in the Wild (LFW)[6] to evaluate the performance of de-identified images. LFW [Hu07] consists of 13,233 face images of 5,749 people. According to the evaluation protocol recommended by LFW, we extract 3,000 pairs of face images of the same person for evaluation. All the face images are resized to $256 \times 256$ pixels. Fig. 2 shows examples of face images in CelebA, GFSG2, and LFW.

### 4.2 Experimental Condition

The proposed method trains HN so that de-identified image $D$ is identified as the person in embedding image $E$, resulting in a risk of privacy violation of the person in embedding image $E$. If the performance of de-identification is not degraded by using a face image of a fake person instead of a real person as embedding image $E$, the privacy can be protected. Therefore, we evaluate the performance of de-identification using face images of real and fake persons as embedding images $E$. We denote the proposed method (R) as using a face image of a real person, and the proposed method (F) as using a face image of a fake person in the experiments. In both methods, a face image in the CelebA dataset is used as cover image $C$ for training HN. In the training of the proposed method (R), a face image of a real person randomly selected at each epoch from the CelebA dataset is used as embedding image $E$. In the training of the proposed method (F), a face image of a fake person, randomly selected at each epoch from the GFSG2 dataset, is used as embedding image $E$. Adam [KB15] is used as the optimizer, and the learning rate is dynamically adjusted from $10^{-5}$ based on the loss on the validation data. Data augmentation that randomly flips cover image $C$ and embedding image $E$ to the left and right, respectively, is introduced during training of the proposed method.

To demonstrate the effectiveness of the proposed method, we compare its performance with that of the de-identification methods using AEs: Landmark-Guided Cutout (LGC) [Ya20], Targeted Identity-Protection Iterative Method (TIP-IM) [Ya21], and Fawkes [Sh20]. LGC uses $\varepsilon = 3.2$ as the hyperparameter that controls the magnitude of the perturbation. TIP-IM uses $\varepsilon = 12$ as the hyperparameter that controls the magnitude of the perturba-

[6] `http://vis-www.cs.umass.edu/lfw/`

Tab. 1: Experimental results of face image de-identification methods, where "Original" indicates the accuracy of original face recognition models.

| Method | PSNR [dB]↑ | SSIM↑ | LPIPS↓ | ASR [%]↑ | | |
|---|---|---|---|---|---|---|
| | | | | FaceNet | CosFace | Softmax |
| Original | — | — | — | 0.73 | 0.73 | 0.90 |
| LGC [Ya20] | 30.64 | 0.943 | 0.0682 | 43.23 | 46.76 | 62.76 |
| TIP-IM [Ya21] | 30.34 | 0.927 | 0.1560 | 56.70 | 56.53 | 62.36 |
| Fawkes [Sh20] | **35.03** | **0.980** | 0.0905 | 51.73 | 45.96 | 63.53 |
| Proposed (R) | 28.13 | 0.957 | 0.0553 | 73.43 | **61.86** | **81.20** |
| Proposed (F) | 28.53 | 0.959 | **0.0547** | **78.13** | 60.96 | 79.66 |

tion, and $\gamma = 0$ as the weights of MMD [Bo06] that is the loss function to control the appearance. Fawkes sets the mode controlling the magnitude of the perturbation to high.

In training, we use the face recognition models trained with ArcFace [DGZ19] as EN. In test, we evaluate the performance of de-identification against unknown face recognition models: FaceNet [SKP15], CosFace [Wa18], and iResNet-50 [DGZ19] with softmax. Note that face recognition model used in test is differ from EN because of black-box.

### 4.3 Evaluation Metrics

In the experiments, we evaluate the image quality of de-identified images and the performance of de-identification in 1-to-1 matching (verification).

The image quality of the de-identified images is evaluated using Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM) [Wa04], and LPIPS [Zh18]. PSNR evaluates image quality based on Mean Squared Error (MSE) between images. SSIM [Wa04] evaluates image quality based on the difference of pixels, contrast, and structure between images. Higher PSNR and SSIM indicates higher image quality. LPIPS [Zh18] evaluates image quality based on the difference of weighted sums of features output from each layer when $C$ and $D$ are input to AlexNet [KSH12]. LPIPS provides an evaluation metric that is closer to human perception than PSNR and SSIM. Lower LPIPS indicates higher image quality.

The performance of de-identification in 1-to-1 matching is evaluated by the Attack Success Ratio (ASR). ASR indicates the ratio of pairs that are verified as impostor pairs among the genuine pairs after de-identification. Higher ASR indicates better de-identification performance. We de-identify the face image of one of the 3,000 genuine pairs selected from LFW. Then, we verify the pairs after de-identification using face recognition models for test, and calculate ASR based on the verification results. The proposed method (R) uses the face image of a real person randomly selected from the face images in LFW as embedding image $E$. The proposed method (F) uses the face image of a fake person in GFSG2, which is not used for training, as embedding image $E$.


Fig. 3: Examples of de-identified images generated by each method.

### 4.4 Comparison of Face Image De-identification Methods

We evaluate the performance of face image de-identification methods. Table 1 shows the quality of the generated de-identified image $D$ and its de-identification performance for unknown face recognition models. Fig. 3 shows examples of the de-identified image $D$ generated by each method. The proposed methods (R) and (F) exhibit higher de-identification performance for all face recognition models than the conventional methods, and can generate de-identified images with higher image quality due to the lower LPIPS. The de-identified images generated by the proposed method are less noisy and look natural, while those generated by the conventional method are noisy and look unnatural. The proposed method de-identifies a face image by locally applying the features of another person's face to the face image, and thus can achieve high de-identification performance against unknown face recognition models, while preserving the appearance of the face image. The de-identified images generated by the proposed methods (R) and (F) exhibit comparable image quality and de-identification performance. By using a face image of a fake person as the embedding image, it is possible to de-identify face images, taking into account the privacy of the person in the embedding image.

## 5 Conclusion

In this paper, we proposed a face image de-identification method to embed facial features of another person into a target face image. The proposed method makes face recognition difficult while preserving the appearance of the face image. Through a set of experiments, we demonstrated the effectiveness of the proposed method compared with the conventional methods using AEs. We presented that the proposed method can protect the privacy of both the target face and the face to be embedded by embedding features from non-real face images.

## References

[Bo06] Borgwardt, K. M.; Gretton, A.; Rasch, M. J.; Kriegel, H.-P.; Scholkopf, B.; Smola, A. J.: Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics, 22:e49–e57, July 2006.

[De09] Deng, J.; Dong, W.; Socher, R.; Li, L. J.; Li, K.; Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 248–255, June 2009.

[DGZ19] Deng, J.; Guo, J.; Zafeiriou, S.: ArcFace: Additive angular margin loss for deep face recognition. Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 4685–4694, June 2019.

[GSS15] Goodfellow, I.; Shlens, J.; Szegedy, C.: Explaining and harnessing adversarial examples. Proc. Int'l Conf. Learning Representations, pp. 1–11, May 2015.

[He16] He, K.; Zhang, X.; Ren, S.; Sun, J.: Deep residual learning for image recognition. Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 770–778, June 2016.

[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, October 2007.

[JAL16] Johnson, J.; Alahi, A.; Li, F.: Perceptual losses for real-time style transfer and super-resolution. Proc. European Conf. Computer Vision, pp. 694–711, March 2016.

[Ka20] Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T.: Analyzing and improving the image quality of StyleGAN. Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 8110–8119, June 2020.

[KB15] Kingma, D.; Ba, J.: Adam: A Method for stochastic optimization. Proc. Int'l Conf. Learning Representations, pp. 1–15, May 2015.

[KSH12] Krizhevsky, A.; Sutskever, I.; Hinton, G. E.: ImageNet classification with deep convolutional neural networks. Proc. Advances Neural Information Processing Systems, 25:1106–1114, December 2012.

[Li15] Liu, Z.; Luo, P.; Wang, X.; Tang, X.: Deep learning face attributes in the wild. Proc. Int'l Conf. Computer Vision, pp. 3730–3738, December 2015.

[LJ11] Li, S.Z.; Jain, A.K.: Handbook of Face Recognition. Springer, 2011.

[Ma19] Marcel, S.; Nixon, M. S.; Fierrez, J.; Evans, N.: Handbook of Biometric Anti-Spoofing. Springer, 2019.

[RAP16] Ribaric, S.; Ariyaeeinia, A.; Pavesic, N.: Deidentification for privacy protection in multimedia content: A survey. Signal Processing: Image Communication, 47:131–151, September 2016.

[RFB15] Ronneberger, O.; Fischer, P.; Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. Proc. Int'l Conf. Medical Image Computing and Computer Assisted Intervention, pp. 234–241, October 2015.

[Sh20] Shan, S.; Wenger, E.; Zhang, J.; Li, H.; Zheng, H.; Zhao, B.: Fawkes: Protecting privacy against unauthorized deep learning models. Proc. the 29th USENIX Security Symposium, pp. 1589–1604, August 2020.

[SKP15] Schroff, F.; Kalenichenko, D.; Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 815–823, June 2015.

[SZ15] Simonyan, K.; Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556:1–14, May 2015.

[Wa04] Wang, Z.; Bovik, A. C.; Sheikh, H. R.; Simoncelli, E. P.: Image quality assessment: From error visibility to structual similarity. IEEE Trans. Image Processing, 13(4):600–612, April 2004.

[Wa18] Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Li, Z.; Gong, D.; Zhou, J.; Liu, W.: Cosface: Large margin cosine loss for deep face recognition. Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 5265–5274, June 2018.

[Ya20] Yang, X.; Yang, D.; Dong, Y.; Su, H.; Yu, W.; Zhu, J.: RobFR: Benchmarking adversarial robustness on face recognition. CoRR, abs/2007.04118:1–28, July 2020.

[Ya21] Yang, X.; Dong, Y.; Pang, T.; Su, H.; Zhu, J.; Chen, Y.; Xue, H.: Towards face encryption by generating adversarial identity masks. Proc. Int'l Conf. Computer Vision, pp. 3897–3907, March 2021.

[Zh18] Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 586–595, June 2018.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# Robust Sclera Segmentation for Skin-tone Agnostic Face Image Quality Assessment

Wassim Kabbani [1], Christoph Busch [2], Kiran Raja [3]

**Abstract:** Face image quality assessment (FIQA) is crucial for obtaining good face recognition performance. FIQA algorithms should be robust and insensitive to demographic factors. The eye sclera has a consistent whitish color in all humans regardless of their age, ethnicity and skin-tone. This work proposes a robust sclera segmentation method that is suitable for face images in the enrolment and the border control face recognition scenarios. It shows how the statistical analysis of the sclera pixels produces features that are invariant to skin-tone, age and ethnicity and thus can be incorporated into FIQA algorithms to make them agnostic to demographic factors.

**Keywords:** FIQA, Face Recognition, Facial Landmarks, Eye Sclera, Skin-tone, Illumination, Natural Color, Color Imbalance

## 1 Introduction

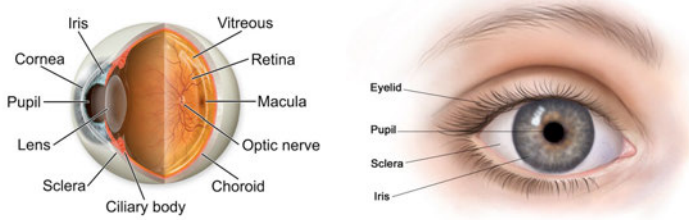
Face image quality assessment refers to the process of evaluating the utility of a face image for face recognition. It involves analyzing various quality factors that may impact the recognition performance. The quality measures produced from analyzing the image can be in the form of individual quality components, such as background uniformity, illumination uniformity, pose, exposure, dynamic range, sharpness, facial expressions, or in the form of a unified quality score.

The ISO/IEC CD on 29794-5 [IS] (Information technology — Biometric sample quality — Part 5: Face image data) specifies that a face image quality assessment algorithm should be insensitive to demographic factors such as age, skin-tone or ethnicity.

The eye sclera refers to the outer layer of the eyeball surrounding the iris. It is the opaque, whitish portion of the eye that surrounds the colored iris and the dark circular opening called the pupil. Figure 1 illustrates the anatomy of the eye including the sclera. This characteristic of being whitish in color regardless of age, ethnicity and skin-tone [Ka23] is what makes it interesting for the task of face image quality assessment.

Analyzing the eye sclera in a face image can help in making the quality assessment algorithms of some of the face image quality components invariant to skin-tone and ethnicity. Not all of the quality components specified in ISO/IEC CD on 29794-5 can make use

[1] IIK, Info. Sec. and Comm. Technology, Gjovik, Norway, wassim.h.kabbani@ntnu.no
[2] IIK, Info. Sec. and Comm. Technology, Gjovik, Norway, christoph.busch@ntnu.no
[3] IIK, Info. Sec. and Comm. Technology, Gjovik, Norway, kiran.raja@ntnu.no




(a) (b)

Fig. 1: Anatomy of the Eye. Image src: perspectiveopticians.co.uk

of this but many of them can. This can be, for example, useful for all the illumination and color related components such as illumination uniformity, no under or over exposure, natural color, dynamic range as well as for eyes visible and eyes open.

## 2 Related Work

The blood vessels structure of the eye sclera is unique to each person, hence it could be used for identification [Zh12]. Therefor, sclera segmentation methods have started to emerge since a high quality segmentation of the sclera from the eye and the iris is required before any further processing and recognition can take place.

Most sclera segmentation methods are deep learning based models trained on large scale datasets with ground truth segmentation masks. Some of them can only segment the sclera, others perform full eye segmentation for the sclera, the iris, and the pupil.

ScleraSegNet [Wa19] is a sclera segmentation method based on an attention assisted U-Net model. It utilizes attention modules in addition to the U-Net to imporve the segmentation performance. In a following improved version of ScleraSegNet [Wa20], the authors suggest to adjust the architecture by placing the attention modules into the central bottelneck part between the contracting path and the expansive path of the U-Net to strengthen the learning capacity of the network and this proves to improve the segmentation performance.

RITnet [Ch19] is a real-time eye segmentation deep neural network model that is trained on the OpenEDS dataset [Ga19]. It is the winning model of OpenEDS Semantic Segmentation Challenge 2019 [4] and achieves state-of-the-art results on the OpenEDS's testset.

Segmentation models are usually trained on specialized datasets collected for training eye segmentation models and gaze tracking models. Among the largest and most recent datasets are OpenEDS [Ga19], and NVGaze [Ki19].

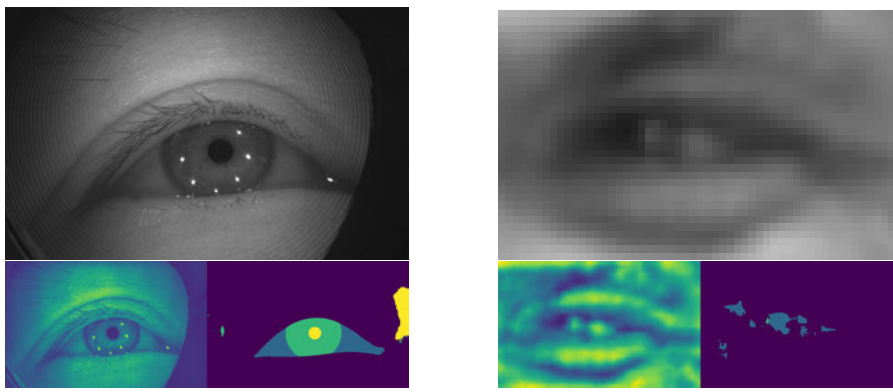
[4] https://research.facebook.com/openeds-challenge/

The OpenEDS:OpenEyeDataset [Ga19], contains eye-images captured using a virtual-reality head mounted display with two eye-facing cameras and under controlled illumination. It contains high quality images of 400x640 pixels of the eye region only.

The NVGaze [Ki19] dataset, from Nvidia, is created to satisfy the criteria for near-eyegaze estimation under infrared illumination. It comprised two types of images, synthetic images of 1280x960 pixels, and real images of 640x480 pixels of the eye region only.

While face images could be of non-uniform illumination, imbalanced color, and low resolution, existing eye segmentation models are trained on datasets of high-resolution images captured under controlled illumination in a specialized setting. This makes them less suited for the task of segmenting the eye region in a face image in order to perform further analysis. Furthermore, the face parsing network [5] that is standardized in ISO/IEC CD 29794-5 [IS] and which segments the face into 19 classes such as hair, eyeglasses, eyes, eyebrows, nose, mouth and ears, does not give a segmentation for the different regions inside the eye but rather for the eye as a whole. Thus, a dedicated sclera segmentation method that is suitable for face images commonly encountered during the face recognition process is needed.

Figure 2 shows eye segmentation results of the RITnet model. In figure 2a it can be seen that the model achieves very good results on a high-quality image from the OpenEDS dataset. However in figure 2b, it can be seen that the segmentation process fails when used on an eye region crop taken from a face image of 224x224 pixels from the LFW dataset [Hu07].



(a) Image from OpenEDS [Ga19]

(b) Crop from a face image from LFW [Hu07]

Fig. 2: Eye segmentation results of RITnet [Ch19]

## 3 Sclera Segmentation

The proposed sclera segmentation method is based on the facial landmarks and it uses MediaPipe [Lu19] as the landmarks extractor. In particular, the landmarks of both eyes as

[5] https://github.com/zllrunning/face-parsing.PyTorch

well as the landmarks of both the left and the right irises are utilized. The process is the same for both eyes so it is explained for one eye only. To decide which pixels belong to the sclera, first a convex hull of the eye's landmarks $ch(eye)$ is computed, this encloses the entire area of the eye including the sclera, the iris and the pupil. Second, the minimum enclosing circle of the iris' landmarks $ec(iris)$ is computed, this encloses the iris and the pupil. Then, all the points in the minimum bounding rectangle of the eye's landmarks $br(eye)$ that test positive for being inside the convex hull $ch(eye)$ and outside the minimum enclosing circle of the iris landmarks $ec(iris)$ (the euclidean distance between the point and the center of the circle is greater than the radius) are considered to belong to the sclera region. Figure 3 illustrates the process, where figure 3a shows the original image, figure 3b shows the convex hull $ch(eye)$ in green and the enclosing circle $ec(iris)$ in yellow, figure 3c shows the bounding rectangle of the eye $br(eye)$ in blue, and figure 3d shows the sclera pixels painted with white.

(a) (b) (c) (d)

Fig. 3: Landmark-based Sclera Segmentation Method. Image from FRLL [DJ21]

The landmark-based sclera segmentation method can successfully segment the sclera regardless of the skin tone, and is also robust to the size of the eyes and the presence of transparent eyeglasses, as shown in figure 4.

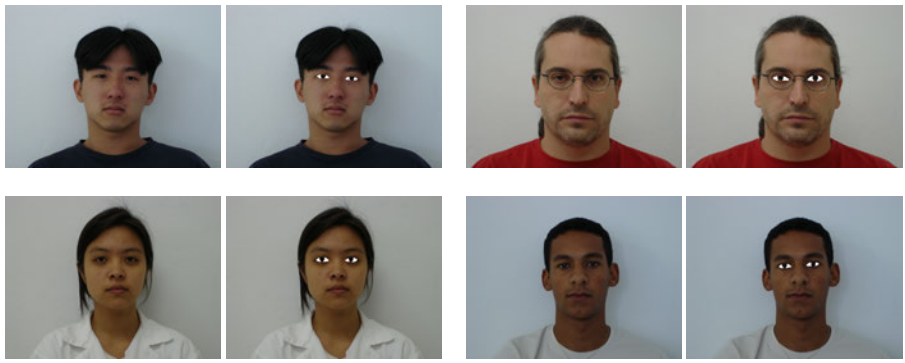


Fig. 4: Sclera segmentation in the enrolment scenario. Images from FEI [Th]

The method works well not only in the enrolment scenario where photos of subjects are taken under controlled environment, but also works well for the border control scenario

where images could be of lower quality. Figure 5 shows segmentation examples of small images of 224x224 pixels from the LFW dataset [Hu07]. The results show that the method is robust to image resolution, skin-tone, the presence of eyeglasses and small pose variations.

Fig. 5: Sclera segmentation on smaller, in the wild images from LFW [Hu07]

## 4 Unnatural Color and Color Balance

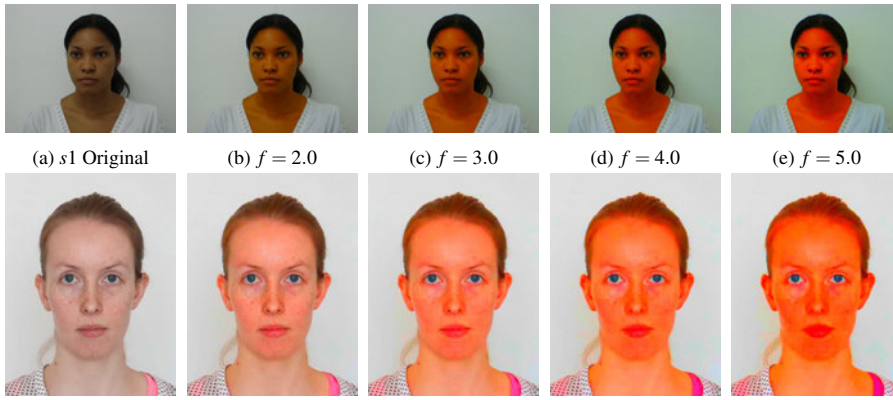
Global adjustments of color intensities that affect the entire image can result in color imbalance. Color imbalance usually takes the form of color casts or extreme color saturation. Color casts can by created synthetically as a post-processing step after taking the photo by manipulating the intensities of individual color channels. They can also result from illuminating the subject with light sources of different color temperatures, while taking the photo, causing digital cameras to render a color cast. Extreme color saturation, on the other hand, is created when the intensities of all colors in the image, not individual channels, are manipulated to take much lower or much higher values than normal resulting in under or over color saturation.

The ”No Unnatural Color” is specified as a quality component measure in ISO/IEC CD 29794-5 [IS] because the skin color is a discriminative personal quality and thus affects the face recognition performance. However, the wide variety of skin tones and the potential presence of factors such as tattoos, moles and other facial anomalies, makes detecting unnatural color in face images reliably a challenging task.

Since the color of the sclera is uniformly whitish across all skin-tones, it should be the case that the pixel values of the sclera region show consistent changes when a face image undergoes global adjustments of color intensities such as over saturation, regardless of the skin-tone of the subject.

In figure 6, sub figures 6a and 6f show the original images of two subjects $s1$ and $s2$ of two different skin tones. The rest of the sub figures show synthetically created images with different saturation factors, four for each of the original images. Tables 1 and 2 show the mean pixel values of the face region as well as the left and right sclera regions in each of the images for subject 1 and subject 2 respectively.

In table 1, which shows the pixel values statistics of subject $s1$ images, it can be seen that the mean pixel value of the face region increases slightly resulting in brighter color as



(a) $s1$ Original (b) $f = 2.0$ (c) $f = 3.0$ (d) $f = 4.0$ (e) $f = 5.0$

(f) $s2$ Original (g) $f = 2.0$ (h) $f = 3.0$ (i) $f = 4.0$ (j) $f = 5.0$

Fig. 6: Images of two subjects of different skin tones with different saturation factors. Images from FEI [Th]

the image gets more saturated. In table 2, on the other hand, which shows the statistics of subject $s2$ images, it can be seen that the mean pixel value of the face region decreases slightly resulting in darker color as the image gets more saturated. However, looking at the mean pixel values of both the left and the right sclera regions, in both tables, it can be seen that both are clearly increasing in value resulting in a brighter color in all images and for both subjects.

| Saturation | Face Oval (s1) | Left Sclera (s1) | Right Sclera (s1) |
|---|---|---|---|
| Original | 60.30 | 67.28 | 71.94 |
| f=2.0 | 56.54 | 77.02 | 80.67 |
| f=3.0 | 60.96 | 83.81 | 86.97 |
| f=4.0 | 66.76 | 94.07 | 103.23 |
| f=5.0 | 72.38 | 97.45 | 106.73 |

Tab. 1: Mean pixel values for $s1$ images

| Saturation | Face Oval (s2) | Left Sclera (s2) | Right Sclera (s2) |
|---|---|---|---|
| Original | 153.60 | 165.41 | 168.16 |
| f=2.0 | 152.26 | 178.93 | 182.13 |
| f=3.0 | 145.95 | 191.13 | 192.40 |
| f=4.0 | 134.54 | 197.93 | 199.21 |
| f=5.0 | 122.48 | 212.34 | 214.46 |

Tab. 2: Mean pixel values for $s2$ image

The purpose here is not to show an unbalanced color detection algorithm, but rather to show that a detection algorithm that relies on analyzing the sclera region, rather than the entire face, has better chances of being more reliable and skin-tone invariant. The consistent behavior of statistical values, even simple ones like the mean, across images of people with different skin colors and even different initial illumination conditions as in images 6a

and 6f, when exposed to various color manipulations, and given that the ground truth color of the sclera region is the same, makes the detection algorithms more robust and more agnostic to demographic factors.

## 5 Illumination

In the same way in which the pixel values of the left and the right sclera regions can be used with algorithms that detect unnatural color in images, they can also be employed to get useful information about the illumination of the face images and to estimate how well the subject in a face image is illuminated and how uniform the illumination is in a way that is completely agnostic to the skin-tone of the subject.

Figure 7 shows examples of face images with varying illumination quality. Figure 7a shows a very dark image where the subject is barely visible. Looking at the histogram of the pixel values of the sclera regions, it can be clearly seen that most pixel values are on the lower end of the value scale and thus have darker colors. Figure 7b shows an image with well illuminated subject. This can also be deduced by looking at the histogram which also confirms that the illumination is symmetric between the left and the right side. Figure 7c shows a poorly lit subject, with a light source from the top causing the eyes region to be dark. The same information can also be deduced by looking at the histogram of the pixel values of the sclera region. Lastly, figure 7d shows a face image with non-uniform illumination, where the right side of the face is well illuminated while the left side is rather dark. This can be confirmed by looking at the histograms which show that the distribution of the pixel values of the left sclera is shifted to the left and has lower pixel values, thus the left side is less illuminated than the right side of the face. All this analysis can be done independently from the subject in the image and without considering what skin-tone they might have.

(a) Bad Illumination (b) Good Illumination (c) Top Illumination (d) Nonuniform

Fig. 7: Sclera segmentation for assessing face image illumination. Images from FEI [Th], LFW [Hu07], and Illinois DOC labeled faces [Fi19]

## 6 Conclusion and Future Work

Face image quality assessment algorithms should be invariant to demographic factors. The eye sclera is one region of the face which has consistent whitish color across demographic

boundaries. This work introduced a novel algorithm for sclera segmentation that is suitable for face images used during the enrolment and the verification scenarios. It then presented how the behavior of the statistical features of the pixel values of the sclera regions is consistent across different demographic boundaries which makes them very useful for creating FIQA algorithms that are more robust and invariant to demographic factors [6].

A follow up work will utilize the sclera segmentation method and incorporate the demonstrated consistent statistical behavior of the pixel values of the sclera regions into the face image quality assessment algorithms of various face image quality components to make them more robust to demographic factors.

## 7 Acknowledgement

This work was supported by the European Union's Horizon 2020 Research and Innovation Program under Grant 883356.

## References

[Ch19] Chaudhary, Aayush K; Kothari, Rakshit; Acharya, Manoj; Dangi, Shusil; Nair, Nitinraj; Bailey, Reynold; Kanan, Christopher; Diaz, Gabriel; Pelz, Jeff B: RITnet: real-time semantic segmentation of the eye for gaze tracking. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). IEEE, pp. 3698–3702, 2019.

[DJ21] DeBruine, Lisa; Jones, Benedict: Face Research Lab London Set. 4 2021.

[Fi19] Illinois DOC labeled faces dataset, https://www.kaggle.com/davidjfisher/illinois-doc-labeled-faces-dataset.

[Ga19] Garbin, Stephan J.; Shen, Yiru; Schuetz, Immo; Cavin, Robert; Hughes, Gregory; Talathi, Sachin S.: OpenEDS: Open Eye Dataset. CoRR, abs/1905.03702, 2019.

[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.

[IS] ISO/IEC WD 29794-5 Information technology — Biometric sample quality — Part 5: Face image data, https://www.iso.org/standard/81005.html.

[Ka23] Kano, Fumihiro: Evolution of the Uniformly White Sclera in Humans: Critical Updates. Trends in Cognitive Sciences, 27(1):10–12, 2023.

[Ki19] Kim, Joohwan; Stengel, Michael; Majercik, Alexander; Mello, Shalini; Dunn, David; Laine, Samuli; McGuire, Morgan; Luebke, David: NVGaze: An Anatomically-Informed Dataset for Low-Latency, Near-Eye Gaze Estimation. pp. 1–12, 05 2019.

[Lu19] Lugaresi, Camillo; Tang, Jiuqiang; Nash, Hadon; McClanahan, Chris; Uboweja, Esha; Hays, Michael; Zhang, Fan; Chang, Chuo-Ling; Yong, Ming; Lee, Juhyun; Chang, Wan-Teh; Hua, Wei; Georg, Manfred; Grundmann, Matthias: MediaPipe: A Framework for Perceiving and Processing Reality. In: Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019. 2019.

[6] https://github.com/wkabbani/sclera-segmentation

[Th] FEI Face Database, https://fei.edu.br/ cet/facedatabase.html.

[Wa19] Wang, Caiyong; He, Yong; Liu, Yunfan; He, Zhaofeng; He, Ran; Sun, Zhenan: ScleraSegNet: an Improved U-Net Model with Attention for Accurate Sclera Segmentation. In: 2019 International Conference on Biometrics (ICB). pp. 1–8, 2019.

[Wa20] Wang, Caiyong; Wang, Yunlong; Liu, Yunfan; He, Zhaofeng; He, Ran; Sun, Zhenan: ScleraSegNet: an Attention Assisted U-Net Model for Accurate Sclera Segmentation. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2(1):40–54, 2020.

[Zh12] Zhou, Zhi; Du, Eliza Yingzi; Thomas, N. Luke; Delp, Edward J.: A New Human Identification Method: Sclera Recognition. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 42(3):571–583, 2012.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# Exploiting Face Recognizability with Early Exit Vision Transformers

Seth Nixon[1] Pietro Ruiu[2] Marinella Cadoni[3] Andrea Lagorio[4] Massimo Tistarelli[5]

**Abstract:** Face recognition with Deep Learning is generally approached as a problem of capacity. The field has seen progressively deeper, more complex models or larger, more highly variant datasets. However, the carbon footprint of machine learning is a concern. A real push is developing to reduce the energy consumption of machine learning as we strive for a more eco-friendly society. Lower energy consumption or compute budget is always desirable, if accuracy is not reduced below a usable level. We present an approach using the state of the art Vision Transformer and Early Exits for reducing compute budget without significantly affecting performance. We develop a system for face recognition and identification with a closed-set gallery and show that with a small reduction in performance, a reasonable reduction in compute cost can be obtained using our method.

**Keywords:** Biometrics, Face Recognition, Vision Transformer, Early Exit

## 1 Introduction

The variability in distinctiveness of humans can be observed when looking at human faces which are exceptionally recognisable, while others can fade into a crowd. “Doddington’s Zoo”[Do98, RRT09] is a well known taxonomy of this phenomenon, where individuals are categorised into different zoomorphic groups based on how easily they can be distinguished or recognized. Principally, a large proportion of individuals can be tagged as highly recognisable, or ”Sheep” in the taxonomy, where only a small amount fall into the ”difficult to recognise” group. However, this small population will disproportionately contribute to the error rate of an automatic recognition system. Additionally, for the best recognisable population it may not be necessary to process their biometric data into complex descriptions, as simpler feature vectors may suffice. In this paper we take inspiration from this phenomenon to propose a novel approach to reduce the computational complexity of an inference model for face recognition.

Transformers are a state of the art machine learning (ML) algorithm, initially presented as a model for natural language processing [Va17] and popularised in mainstream culture with ChatGPT. These models are notorious for their high compute requirement, due to their fundamental operation (Multi-Headed Self Attention) being quadratic, and their exceptional

[1] University of Sassari, Department of Biomedical Sciences, Sassari, Italy, swsnixon@uniss.it
[2] University of Sassari, Department of Biomedical Sciences, Sassari, Italy, pruiu@uniss.it
[3] University of Sassari, Department of Biomedical Sciences, Sassari, Italy, maricadoni@uniss.it
[4] University of Sassari, Department of Biomedical Sciences, Sassari, Italy, lagorio@uniss.it
[5] University of Sassari, Department of Biomedical Sciences, Sassari, Italy, tista@uniss.it

performance in a vast number of applications. They have been extended to vision with the Vision Transformer (ViT) [Do20]. ViT's obviate the usage of convolutions and instead use Multi-Headed Self Attention (MHSA) to encode an image. MHSA has been shown to be an extremely powerful operation, with similar expressive power to convolution, and has some similarities with human vision [Tu21, Ca22]. ViTs generally have shown state of the art performance in a wide range of applications such as image classification[Do20], object detection [Fa21] and text to image generation [Ra21]. Attention within a ViT has also been investigated with respect to human vision, with the aim of reducing compute budget or improving performance.

Early Exit (EE) approaches in machine learning are one solution for implementing time critical applications on resource-constrained devices. In these instances compute budget is something which must be considered. EE systems allow a machine learning process to use only a portion of the full network where using the full network would be either unnecessary or infeasible. They have found applications in computing continuum applications where earlier exits of the model are computed on edge devices and later exits can be offloaded to the cloud [LZC18].

In this paper we present an EE approach using ViTs for face recognition. We explore a novel application of percentile based exit criterion on a closed set gallery, allowing these to be computed ad-hoc, only dependant on the outputs of a ML model. Additionally we present how Early Exits added to a ViT can reduce the compute budget at inference with only a small loss in performance.

## 2 Related works

There are many techniques for reducing the compute budget of a deep network, including network pruning [Re93], distillation [To21] and quantisation [Go14]. Early Exits (EE) are another form in the group of Dynamic Inference (DI) techniques. DI modulates the complexity of an ML model in relation to some constraint (e.g. compute budget, energy consumption) or to reduce complexity where it offers little benefit. The final goal is to provide a result similar to the final result of the neural network while only using a subset of the layers of the full model. Generally, accuracy is reduced but with a worthwhile reduction in computational cost.

In [Ba18] the authors study the effect of adding EE to any ML model, proposing a general framework that systematically "elastifies" an arbitrary network. There are also DI techniques which approach the problem from the opposite direction. [Wa21] allows inputs for the "easier" images to be processed at a lower resolution. The authors offer a cascade of Transformers with varying token input counts, allowing the bulk of input images to be processed less, where the higher representational capacity models are reserved for the more challenging inputs.

The most similar to our work is that of [PT21, Pa20]. They proposed a metric-learning EE methodology for Deep Learning models applied to face verification. Three EEs were placed after the residual blocks of a ResNet-50 [Sa18], trained on the large-scale MS-

Celeb-1M dataset [Gu16]. The approach reduced computational complexity while maintaining similar accuracy to the final output of the network.

With respect to EE one of the drawbacks of CNNs is that the descriptive capability of earlier layers is challenging to express. This is due to the feature maps remaining wide with limited channel dimension. For example a ResNet-18 has feature map dimension 56x56x64 (*H*xWxC) after the first block compared to 7x7x512 at the final block. Obviously, average pooling in the *HW* dimension loses a significantly higher amount of information at the earlier layers compared to the later. ViTs instead have uniform scale throughout processing, and while this may at first glance appear to be a disadvantage in terms of robustness to scale, these models have been shown to be sensitive to scale even without the prior embedded [Do20]. In the context of our work, this means that EEs can be operated uniformly throughout the model.

EEs have been explored in ViT. In [BZI22] seven different architectures for EE branches that can be inserted into ViT backbones are proposed. In [BZI21] the authors propose a novel hybrid approach using Transformer-based EE branches on CNN backbones, terming them single-layer vision transformer (SL-ViT). We instead determine our exit criterion based on the matching scores of a gallery of identities, where the above methods balance computational budget against accuracy loss with the trained classifiers of the model.

## 3 Model

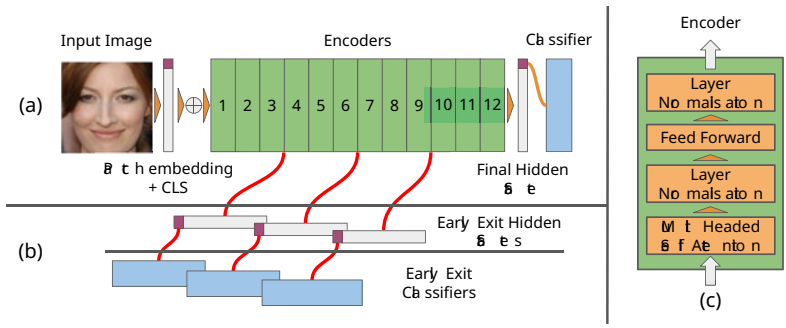



Fig. 1: (a) A ViT. (b) EEs added to the base model. (c) A ViT Encoder

We begin with an off-the-shelf ViT which we term ViT-N [Do20] (vit-base-patch16-224 from *huggingface*[Hu]). A class (CLS) token is appended to the input to allow for image classification. Where most face recognition systems are trained with more complex loss functions, we train with Cross Entropy Loss to maintain proximity to the vast majority of the ViT literature.

For the EE model we take the base ViT and add additional classifiers at layers 3, 6 and 9, the hidden state of the CLS token at these layers is used for classification. To train this model we also use Cross Entropy Loss, however we adapt it to incorporate the additional classifiers. There are many strategies for combining losses in EE models [Pa20, Bo17], we

simply use the average loss of all the classifiers for the backward pass. For a model with $m$ exits the Early Exit Cross Entropy Loss $L_{CE}^{EE}$ is then defined in equation 1:

$$L_{CE}^{EE} = -\frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n} T_i^j log(p_i^j) \tag{1}$$

Where $T$ and $p$ are respectively the Target and predicted probabilities for class $i$, and $m$ is the number of exits. For both ViT-N and ViT-EE we transfer learn from weights pre-trained on Imagenet1k [De09]. Transfer learning from a general purpose, large scale dataset such as Imagenet to other tasks has been shown to generalise as well, if not better, than models trained from scratch on a specific task [KSL19, Me18], as well as significantly reduce the time to train. Additionally, this allows us to utilise a large portion of the approximately 2500 days of TPU processing performed by Google [Do20], further increasing the energy-efficiency of this system. The architecture of ViT-N is shown in Figure 1(a), the additional exits are shown in Figure 1(b).

## 4 Exit criterion

In [RRT09] the taxonomy of the recognizability of humans defined by [Do98] was used to improve the performance of a biometric system. Within the taxonomy there are various zoomorphic categories which define the recognizability characteristics of a person. These categories can be utilised to determine exit criterion in a system with a closed-set gallery.

The simplest and most commonly occurring category of recognizability is the *Sheep*: an identity which exhibits high genuine matching scores and low imposter matching scores. Practically, this is a person who is easy to recognise. Another category we consider is the *Goat* which exhibits low genuine match scores, i.e. these are less likely to be recognised correctly. Finally we consider *Lambs* which have high impostor scores, these are more likely to be recognised incorrectly.

To determine if an identity is a *Goat* we first compute the mean genuine score for that identity $S_{GEN}$. If this identities $S_{GEN}$ is below a threshold then the identity is classified as a *Goat*, otherwise it is either a *Sheep* or a *Lamb*.

To determine if an identity is a *Lamb* we compute the maximum impostor score between the current identity and every other identity $S_{IMP}^i$, $i = 1...n$ where n is the number of identities. We then take the mean of $S_{IMP}^i$ to form the mean impostor score for this identity $S_{IMP}$:

$$S_{IMP} = \frac{1}{n}\sum_{i=1}^{n} S_{IMP}^i \tag{2}$$

If $S_{IMP}$ is above a threshold, then the identity is classified as having high impostor scores, and is a *Lamb*. Should an identity not pass either the *Goat* or the *Lamb* test, then it is

considered a *Sheep*. Should an identity be simultaneously a *Goat* and a *Lamb* it is considered a *Goat* as we compute this first. While in Doddingtons taxonomy there are a much larger number of categories (worms, wolves, doves etc..) we will focus on the three defined above, as they fully cover the spectrum of recognizability we are interested in.

## 5 Experimental setup

We train both ViT-N and ViT-EE on CASIA-WebFace [Yi14], a dataset of 494414 face images of 10575 identities. We train for 120 epochs on 95% of the dataset, leaving 5% for validation. We use the AdamW optimiser with a learning rate of $1e^{-4}$, decayed according to a cosine schedule, a weight decay of $1e^{-2}$ and default betas. We stopped training at 70 epochs for both models as they had converged.

To test we use the *controlled* samples from the FRGC dataset [Ph05], we discard the *uncontrolled* examples to remove as much noise as possible from the testing set, we also discard any identities with less than 10 images. What results is a curated set of 24120 images of 472 identities. We then extract the face from each image at 2 times the width of the eyes, offset vertically by 0.3 times eye width to remove as much background as possible and approximately center the faces.

To classify with a standard ViT, we extract the final hidden state of the CLS token from two images passed through the model, and compare this with Cosine similarity. To classify with EEs, we extract the hidden state of the class token at the end of layers 3, 6, 9 and the final layer. We only compare hidden states extracted from the same exit.

We split our subset of FRGC into a 30/70 (7236/16884) probe/gallery set. This is approximately 122m probe/gallery and 285m gallery/gallery comparisons. To reduce the total volume of comparisons, and through empirical testing we have seen it makes little difference to the results, we sample 10 million gallery/gallery pairs to compute exit identities, and 5 million probe/gallery pairs to present results. In both cases we ensure approximate identity-size-parity between the sampled pairs and the original dataset, and that approximately 10% of pairs are matches.

## 6 Results

First we present the results from both the ViT-N and ViT-EE. Here, each of the four exits of ViT-EE are used to classify all of the probe/gallery pairs of images. We also extract final hidden states from layers 3, 6, 9 and 12 of ViT-N and present how the different training methodologies impact the hidden states extracted at each level.

The first observation is that a standard ViT performs excellently on our test set (Figure 2 ViT-N exit 4). The implication that ViTs are excellent face recognizers is unsurprising, they have offered state of the art performance in nearly all, if not every, application they have been applied to. This being said, experiments on more complicated face datasets would be required to confirm this.

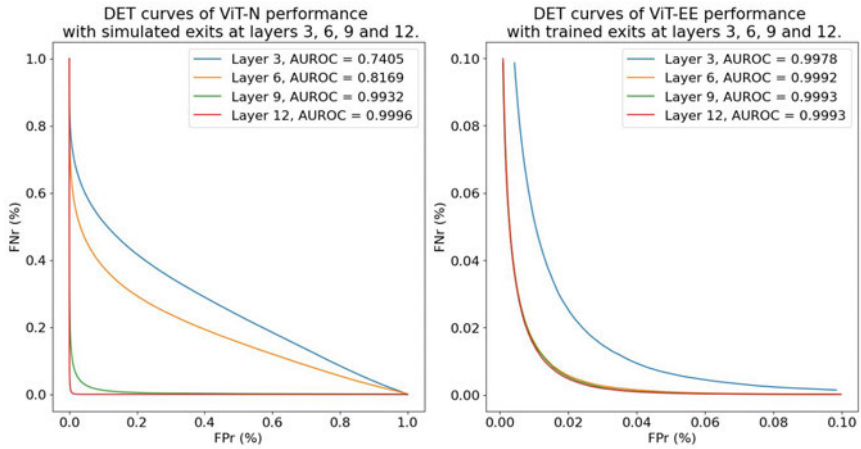



Fig. 2: Detection Error Tradeoff (DET) curves of the recognition performance of hidden states extracted from four different levels of ViT-N and ViT-EE.

Secondly, as shown in Figure 2, by computing the average loss of all 4 classifiers in ViT-EE, we have significantly increased the descriptiveness of the output feature vectors at the earlier layers of the model, with a small reduction at the final layer.

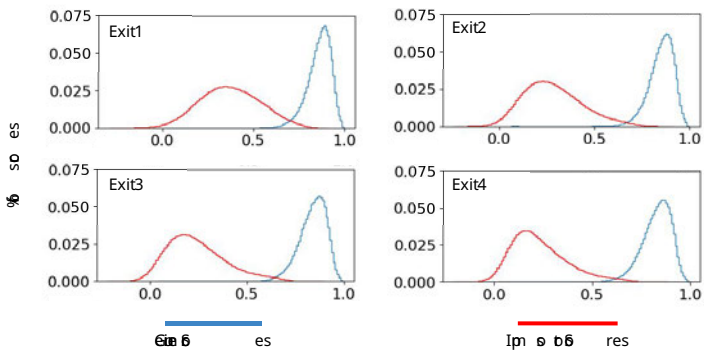



Fig. 3: Score distributions at each exit of ViT-EE.

Figure 3 shows the matching score distribution at each exit of ViT-EE for the full set of Probe/Gallery pairs. We can see that all exits have reasonably good separation. Most importantly, as we progress further through the model, both the impostor scores and the genuine scores tend to reduce, with the impostor scores reducing more thereby increasing the overall performance. Additionally, the variance in the impostor scores generally reduces as well, indicating that the model's representation progressively becomes more robust to the inter-class variation of the dataset, though there does remain a consistent contingent of overlapping measurements.

### 6.1 Early exits

In this Section we present results for the ViT-EE model with exit criterion enforced. This forms a single system where identities are classified at each layer based on their gallery scores. We compute the Goat and Lamb threshold percentiles from 0 to 100 in increments of 10. For a pair of Goat/Lamb thresholds ($T_{GOAT}$/$T_{LAMB}$) if an identities gallery match scores have a mean genuine score ($S_{GEN}$) above $T_{GOAT}$ and a mean impostor score ($S_{IMP}$) below $T_{LAMB}$ that identity is assigned to the current exit, and discarded from computation at the later exits. If not then they have their $S_{GEN}$ and $S_{IMP}$ re-computed at the next exit within the full set of identities which did not pass the test at the current exit. This is repeated until the final exit, where all remaining identities are assigned. Goat and Lamb percentiles remain fixed throughout the model. For an exit, probe samples who's claimed identity is assigned to this exit are scored against only the gallery samples with identities assigned to this exit. We present results as True Match rate at False Match Rate (TMr@FMr) at three FMr, 1%, 0.1% and 0.01%. Figure 4 shows heat maps of TMr@FMr1% and the corresponding Floating Point Operations (FLOPs) of that system.

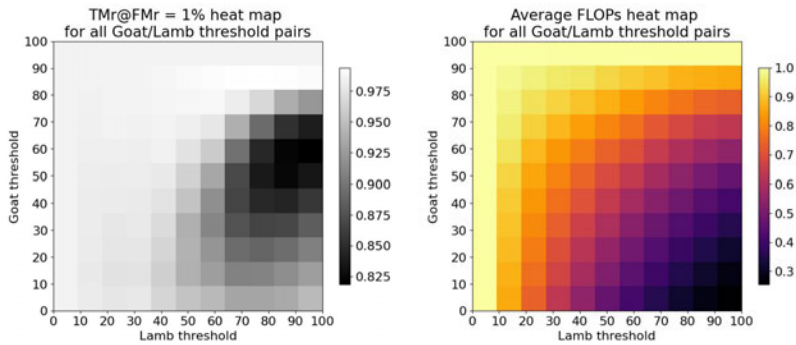



Fig. 4: Heat maps of TMr@FMr1% at each pair of Goat/Lamb thresholds, and the corresponding FLOPs.

If we first examine the FLOPs heat map of Figure 4. We can see that as we decrease the Goat threshold $T_{GOAT}$, and increase the Lamb threshold $T_{LAMB}$, the %FLOPs decreases. In fact, a $T_{LAMB}$ of 0 or a $T_{GOAT}$ of 100 correspond to all identities being assigned to the final exit. Vice-versa a $T_{GOAT}$ of 0 and a $T_{LAMB}$ of 100 correspond to all identities being assigned to the first exit. One would expect the TMr@FMr1% to follow the same pattern, however this is not the case. There is a region with very high lamb thresholds and medium Goat thresholds which offers the absolutely worst performance of the systems. When the system has a high Lamb threshold, i.e. it is less punishing on identities with poor impostor scores, it is the Goat Threshold that dominates. It is guaranteed that at the earlier exits the identities classified have high $S_{GEN}$, however if the Lamb Threshold is over the 50th percentile, then we are naturally going to propagate more error. However, it is likely that the identities which fit into the 90th percentile of genuine scores have excellent separation between their genuine and impostor distributions. As such the performance peaks at this

point. Figure 5 shows a scatter plot of TMr at FMr = 1%, 0.1% and 0.01% for all pairs of thresholds.

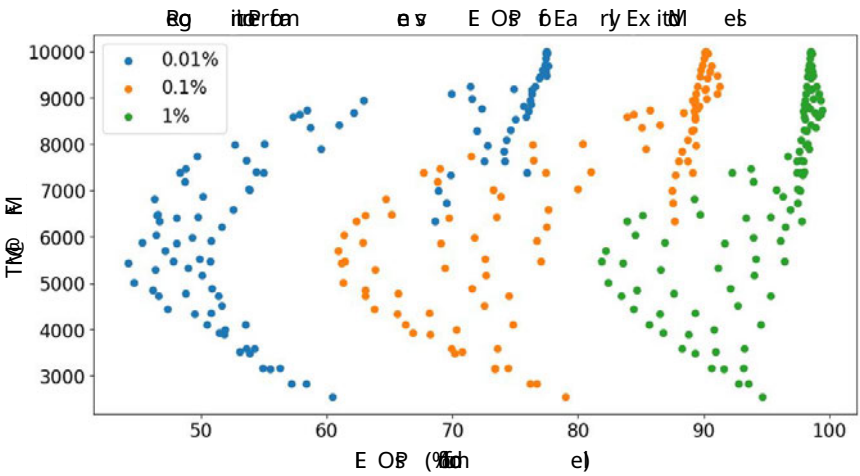



Fig. 5: TMr@FMr for FMr = 1%, 0.1% and 0.01% across the full range of Goat/Lamb threshold pairs. Each point corresponds to ViT-EE with a single pair of Goat/Lamb thresholds.

To compare maximum performance of the models we compare the TMr@FMr of EE systems to ViT-N. ViT-N achieves a TMr@FMr1% of 99.72%, TMr@FMr0.1% of 95.28% and TMr@FMr0.01% of 85.27%. Our best performing system at TMr@FMr1% has 99.4%, a loss of only 0.3%, with a flops reduction of 12.73%, a significant gain. At TMr@FMr0.1% 91.28%, a loss of 4% with a more modest FLOPs reduction of 7.48%. At TMr@FMr0.01% the maximum performance does not reduce the FLOPs.

Within ViT-EE alone, from Figure 5 we can see that there are systems at the medium to high FMr rates which perform better, with a lower performance budget. We can increase TMr@FMr1% by 0.9% (98.51% to 99.4%) and save 12.73% of FLOPs. We can increase TMr@FMr0.1% by 1.15% (90.13 to 91.28%) but with a more modest reduction in FLOPs of 7.47%. We do not see any increases in TMr@FMr0.01%. The performance where all identities are assigned to the first exit is surprisingly high, even more surprising is that it is much higher than many of the other combinations of thresholds, where one would expect that using the more processed feature vectors at the later exits would offer higher discriminative capability. Regardless, with 25.49% FLOPs of the total model we only lose 3.84% TMr@FMr1%, 11.14% TMr@FMr0.1% and 17.02% TMr@FMr0.01% compared to using the final exit of ViT-EE.

Finally we can examine the Attention maps of the models to gain some insight into how the two models differ in terms of their encoding of an image, see Figure 6.

Figure 6 shows the Attention Rollout [AZ20] of the CLS token up to layers 3, 6, 9 and 12 of ViT-N and ViT-EE. The attention of the CLS token gives an indication of which

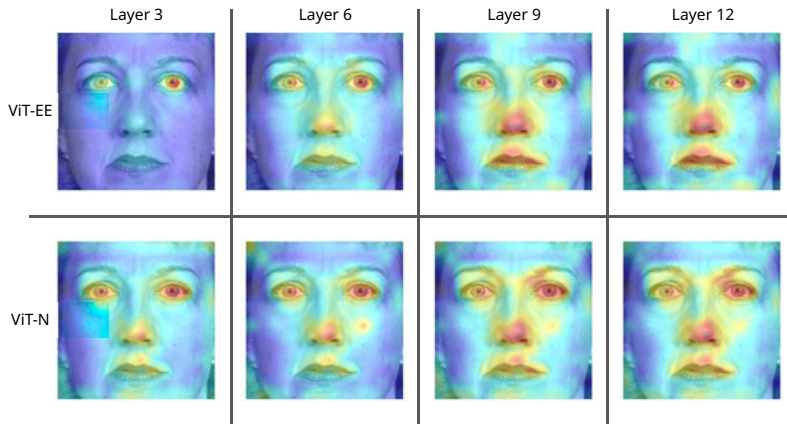



Fig. 6: Attention at the exit layers of ViT-EE, and the corresponding layers of ViT-N.

areas of the image contributed more to the tokens hidden state. Firstly, the attention of layers 9 and 12 in ViT-EE is very similar, this aligns with the DET curve in Figure 2 where both layers 9 and 12 offer similar classification performance when presented with the full set of testing pairs. We can also observe that ViT-EE has much more focused attention in layer 3 around the eyes, with very limited attention across the rest of the face. This suggests that the model, when describing faces at the earlier layers of the model, or alternatively the *easier* identities in the dataset, finds sufficient discriminative information in just a few small features of the face. Conversely as we progress through the model the attention increasingly encapsulates the other salient features of the face, strengthening its reliance on these as we go further. This supports the hypothesis that some identities can be described sufficiently with a much simpler feature vector, where others need more complex descriptions, involving more components of a biometric trait. Interestingly, with reference to Figure 2, ViT-N appears to have relatively uniform attention across the different layers, despite the hidden states at these layers offering vastly different performance capability. As ViT-N appears to be focusing generally on the face at each level, we hypothesize that ViT-N does not effectively encode all salient features of the face with only a portion of the encoders of the model. This makes sense as the model is trained only to classify at the final layer. In this instance, the earlier layers are purely in support of the latter, where the layers in ViT-EE simultaneously support the later layers and offer their own classifications. As ViT-EE is trained with the usage of the hidden states at the earlier layers in mind, the model learns to encode only those features which it can sufficiently describe to give the best classification performance at each exit.

## 7 Conclusions

In the proposed approach, by adding EEs to a ViT, the computational burden of a face recognition model can be considerably reduced without significantly degrading the accu-

racy. As it has been shown, by grouping identities according to their matching scores, some individuals can be still well described with only 25% of a full ViT.

While exploiting the averaged loss throughout the classifiers clearly allows us to reduce the computational burden while only slightly degrading the accuracy, this is a somewhat naive solution and not necessarily the best way of training the model. Ideally it would be desirable to reserve the representational capacity at the later layers of the model for more challenging identities, rather than using a smoothed version of the original representation as in the current approach. Adopting a cascade of models may be a viable solution to improve the representational capability of the model.

The proposed approach focused on one particular configuration of the EEs, but many others could be investigated. For example, an EE configuration considering all layers of the model is also a viable alternative solution.

To properly operate, the proposed system requires a minimum of two gallery images per identity. This requirement obviously impairs the application of the system whenever only one gallery image is available. A slight modification of the model, which is currently under investigation, would allow the system to operate with a single gallery image, however the exit criterion would be only based on Lamb scores.

Finally, an extension of the current method to operate on both closed set and open set gallery data is being considered.

## 8 Acknowledgements

The research activities described in this paper have been conducted within the “Bando Fondazione di Sardegna 2022-2023” grant, the *Secure Passwordless Authentication for Digital Identities* (SPADA) project funded by the Italian Ministry of Education, University and Research (project code: ARS01_00785) and the PNRR M4C2 project Ecosystem of Innovation for Next Generation Sardinia (e.INS) - Spoke 06 ”Digital Transformation”.

## References

[AZ20] Abnar, S.; Zuidema, W.: Quantifying Attention Flow in Transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, pp. 4190–4197, 2020.

[Ba18] Bai, Y.; Bhattacharyya, S.S.; Happonen, A.P.; Huttunen, H.: Elastic neural networks: A scalable framework for embedded computer vision. In: Proceedings of the 26th European Signal Processing Conf. IEEE, pp. 1472–1476, 2018.

[Bo17] Bolukbasi, T.; Wang, J.; Dekel, O.; Saligrama, V.: Adaptive neural networks for efficient inference. In: International Conf. on Machine Learning. PMLR, pp. 527–536, 2017.

[BZI21] Bakhtiarnia, A.; Zhang, Q.; Iosifidis, A.: Multi-exit vision transformer for dynamic inference. arXiv preprint arXiv:2106.15183, 2021.

[BZI22] Bakhtiarnia, A.; Zhang, Q.; Iosifidis, A.: Single-layer vision transformers for more accurate early exits with less overhead. Neural Networks, 153:461–473, 2022.

[Ca22] Cadoni, M.; Nixon, S.; Lagorio, A.; Fadda, M.: Exploring attention on faces: similarities between humans and Transformers. In: 18th IEEE International Conf. on Advanced Video and Signal Based Surveillance. IEEE, pp. 1–8, 2022.

[De09] Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conf. on Computer Vision and Pattern Recognition. IEEE, pp. 248–255, 2009.

[Do98] Doddington, G.; Liggett, W.; Martin, A.; Przybocki, M.; Reynolds, D.: Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. Technical report, NIST, 1998.

[Do20] Dosovitskiy, A. et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[Fa21] Fang, Y.; Liao, B.; Wang, X.; Fang, J.; Qi, J.; Wu, R.; Niu, J.; Liu, W.: You only look at one sequence: Rethinking transformer in vision through object detection. Advances in Neural Information Processing Systems, 34:26183–26197, 2021.

[Go14] Gong, Y.; Liu, L.; Yang, M.; Bourdev, L.: Compressing deep convolutional networks using vector quantization. arXiv preprint arXiv:1412.6115, 2014.

[Gu16] Guo, Y.; Zhang, L.; Hu, Y.; He, X.; Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: Proceedings of the 14th European Conf. on Computer Vision. Springer, pp. 87–102, 2016.

[Hu] Hugging Face - the AI community building the future. https://huggingface.co/. Accessed: 8-05-2023.

[KSL19] Kornblith, S.; Shlens, J.; Le, Q.V.: Do better imagenet models transfer better? In: Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. IEEE, pp. 2661–2671, 2019.

[LZC18] Li, E.; Zhou, Z.; Chen, X.: Edge intelligence: On-demand deep learning model co-inference with device-edge synergy. In: Proceedings of the 2018 Workshop on Mobile Edge Communications. pp. 31–36, 2018.

[Me18] Mehra, R. et al.: Breast cancer histology images classification: Training from scratch or transfer learning? ICT Express, 4(4):247–254, 2018.

[Pa20] Passalis, Nikolaos; Raitoharju, Jenni; Tefas, Anastasios; Gabbouj, Moncef: Efficient adaptive inference for deep convolutional neural networks using hierarchical early exits. Pattern Recognition, 105:107346, 2020.

[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: IEEE computer society Conf. on Computer Vision and Pattern Recognition. volume 1. IEEE, pp. 947–954, 2005.

[PT21] Passalis, N.; Tefas, A.: Adaptive Inference for Face Recognition leveraging Deep Metric Learning-enabled Early Exits. In: 29th European Signal Processing Conf. (EUSIPCO). IEEE, pp. 1346–1350, 2021.

[Ra21] Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I.: Zero-shot text-to-image generation. In: International Conf. on Machine Learning. PMLR, pp. 8821–8831, 2021.

[Re93] Reed, R.: Pruning algorithms-a survey. IEEE transactions on Neural Networks, 4(5):740–747, 1993.

[RRT09] Ross, A.; Rattani, A.; Tistarelli, M.: Exploiting the "doddington zoo" effect in biometric fusion. In: IEEE 3rd International Conf. on Biometrics: Theory, Applications, and Systems. IEEE, pp. 1–7, 2009.

[Sa18] Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: IEEE Conf. on Computer Vision and Pattern Recognition. IEEE, pp. 4510–4520, 2018.

[To21] Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conf. on Machine Learning. PMLR, pp. 10347–10357, 2021.

[Tu21] Tuli, S.; Dasgupta, I.; Grant, E.; Griffiths, T.L.: Are convolutional neural networks or transformers more like human vision? arXiv preprint arXiv:2105.07197, 2021.

[Va17] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems, 30, 2017.

[Wa21] Wang, Y.; Huang, R.; Song, S.; Huang, Z.; Huang, G.: Not all images are worth 16x16 words: Dynamic transformers for efficient image recognition. Advances in Neural Information Processing Systems, 34:11960–11973, 2021.

[Yi14] Yi, D.; Lei, Z.; Liao, S.; Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923, 2014.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# Contactless Palmprint Recognition for Children

Akash Godbole[1], Steven A. Grosz [2], and Anil K. Jain [3]

**Abstract:** Effective distribution of nutritional and healthcare aid for children, particularly infants and toddlers[1], in the world's least developed and most impoverished countries, is a major problem due to lack of reliable identification documents. We present a mobile-based contactless palmprint recognition system, Child Palm-ID, which meets the requirements of usability, cost, and accuracy for child recognition. On a contactless child palmprint database, Child-PalmDB1, with 1,020 unique palms (age range of 6 mos. to 48 mos.), Child Palm-ID achieves a TAR=94.8% at FAR=0.1%. Child Palm-ID is also able to recognize adults, achieving a TAR=99.5% on the CASIA contactless palmprint database and a TAR=100% on the COEP contactless adult palmprint database, both at FAR=0.1%. For child palmprint images captured at an interval of five months with differences in standoff distance, illumination and motion blur, the TAR drops to 80.5% at FAR=0.1%. This indicates that more research opportunities remain in contactless child palmprint recognition.

## 1 Introduction

In 2020, 22% of the world's 680 million children[2], under the age of 5 years, were physically stunted due to malnourishment and lack of adequate medication[3]. A majority of these children live in developing or least developed countries.


Fig. 1: Example face (a) and corresponding contactless palmprint images (b) of subjects in in Child-PalmDB2. The predicted keypoints are also shown (in red) in (b). (c) Corresponding Regions of Interest (ROIs) extracted from the palm images in (b) based on the predicted keypoints. The white polygons in (b) represent the palmar friction ridge area that is captured in the ROIs in (c). Face images are collected only for record keeping and are not used for matching.

[1] Graduate Student, Computer Science, Michigan State University, godbole1@cse.msu.edu
[2] Doctoral Candidate, Computer Science, Michigan State University, groszste@cse.msu.edu
[3] University Distinguished Professor, Computer Science, Michigan State University, jain@cse.msu.edu
[1] https://www.cdc.gov/ncbddd/childdevelopment/ positiveparenting/index.html
[2] https://ourworldindata.org/grapher/under-5-population
[3] https://www.who.int/data/gho/data/themes/topics/joint-child-malnutrition-estimates-unicef-who-wb

To address this problem, international organizations such as the World Health Organization (WHO)[4], Bill and Melinda Gates Foundation (BMGF)[5] and the World Food Programme (WFP)[6] have made substantial efforts to reduce malnourishment as well as improve vaccination coverage among this vulnerable population. However, the lack of reliable identification documents makes it difficult to authenticate the recipients of the services and curtail the occurrence of fraud.

While biometric recognition has been utilized for the identification of children [En21, Ra18, Ka22, RDS22], the available solutions have yet to meet the requirements of low-cost acquisition, high accuracy, robustness to capture environment (e.g. dust, humidity, and temperature), and high throughput for field deployments. It is worth noting that India's Aadhaar program, with an enrolment database of $\sim$1.4 billion, does not enroll anyone under the age of 5 [7].

A biometric trait must meet the *persistence* and *individuality* requirements for the child population under consideration [JRN11]. These requirements make it difficult to justify using an infant or toddler's rapidly changing facial appearance. While a few studies have suggested using footprints [Ko19] and toe prints, their use requires the awkward process of removing socks and shoes and, in cases where the child is barefooted, cleaning their feet. Iris images are difficult to capture if the child is sleeping or crying. While fingerprint recognition has been studied in the context of infant and toddler recognition [En21, Sa19], however, images of infant fingerprints do not contain sufficient friction ridge details for accurate recognition. These limitations, paired with the occurrence of COVID-19 and related concerns about hygiene, has motivated the development of mobile-based contactless biometric systems [8] [9]. We posit contactless palmprint recognition is a cost-effective and feasible solution for child identification. Palmprints provide a large surface area along with well formed principal lines and creases for recognition of infants and toddlers. The proposed Child Palm-ID system does not require custom sensors, as smartphone cameras have sufficient resolution to capture contactless palmprint images of children (a Samsung Galaxy S22 has a 50MP primary camera). Our entire Child Palm-ID system, from image capture to feature extraction and recognition runs on a Samsung Galaxy S22 at 167 ms per comparison.

Prior attempts at palmprint-based recognition for newborns and infants faced a number of challenges in acquisition such as i) motion blur due to hand movements and ii) deformation and low quality introduced by contact-based sensing and oil on the child's palms, etc. [Le11]. To keep the child recognition problem tractable, children in this study are in the age group of 6 to 48 months. Child development studies [St88] report that starting at the age of about 12 months, a child can follow instructions such as opening the fist and holding the palm in front of a mobile phone camera.

Concretely, the contributions of this study are as follows:

[4] https://www.afro.who.int/news/strategic-plan-reduce-malnutrition-africa-adopted-who-member-states
[5] https://www.gatesfoundation.org/our-work/programs/global-growth-and-opportunity/nutrition
[6] https://www.wfp.org/nutrition
[7] https://uidai.gov.in/en/my-aadhaar/about-your-aadhaar/aadhaar-enrolment.html
[8] https://one.amazon.com/
[9] https://www.fujitsu.com/global/services/security/offerings /biometrics/

- An end-to-end mobile-based contactless palmprint recognition system, Child Palm-ID, designed and prototyped for infants and toddlers. Code and contactless databases collected by the authors will be released upon acceptance of this paper.
- Automatic keypoint detection along with homographic alignment for region of interest (ROI) extraction and Thin Plate Spline (TPS) re-alignment to account for non-linear distortion and pose variations in palmprint images.
- State-of-the art recognition accuracies of Child Palm-ID on child as well as adult contactless palmprints; we used a COTS system [Pr] and two publicly available algorithms [Zh17, Ma19] for adult palmprint matching as baselines.
- Evaluation on time-separated contactless child palmprints to demonstrate the need for robust alignment and representation in the presence of differences in standoff distance, illumination and motion blur.

(a) (b) (c)

Fig. 2: Child palmprint collection camp in Dayalbagh, India, January 2023. (a) Parents bringing their children for palmprint collection must sign a consent form and provide the child's name and age along with a mobile number for a possible second round of data collection. (b) and (c) Authors collecting contactless palmprint images using Armatura PalmMobileSDK [Pr]. The palmprint images were collected indoors in a pediatrician's clinic.

## 2 Palmprint Databases

While there are a number of adult contactless palmprint databases[10] available in the public domain [Zh17, Ha08, Iz19, Su05, CO], there are no contactless child palmprint databases available in the public domain. For this reason, we use the available adult palmprint databases to pre-train our recognition model which is then fine-tuned using the self-collected child palmprint databases, to account for the differences in child and adult palmprint images (i.e. size of the hand, level of motion blur and deformation, etc.).

We collected the two child palmprint datasets containing over 40,000 images from 1,824 unique child palms in two different sessions: August 2022 and January 2023 at the Saran Ashram Hospital, Dayalbagh, India (see Fig. 2). The two databases are called Child-PalmDB1 and Child-PalmDB2, respectively [GGJ23]. To enlarge the public adult palmprint databases for pre-training our models, we also collected palmprints of the caregiver-

[10] We were not provided access to PolyU-IITD [Ku18] and IITD Touchless Palmprint Database [Ku08] hence we had to resort to using a private database provided to us under NDA.

Tab. 1: Details of contactless palmprint databases used in this study

| **Training Database*** | **Capture Device** | **Image Size (pixels)** | **# Unique Palms** | **Total # images** |
|---|---|---|---|---|
| Tongji Adult [Zh17] | JAI AD-80 GE | 600x800 | 600 | 12,000 |
| CASIA Multispectral [Ha08] | CCD Camera | 768x576 | 200 | 7,200 |
| Child-PalmDB2[1] | Samsung Galaxy S22 | 1080x1440 | 963 | 18,277 |
| Adult-PalmDB2[1] | Samsung Galaxy S22 | 1080x1440 | 1,227 | 22,548 |
| SMPD [Iz19] † | iPhone 6 | 3264x2448 | 92 | 3,677 |
| Private Database | Redmi Note 9 Pro | 1080x1920 | 1,016 | 28,748 |
| **Testing Database** | **Capture Device** | **Image Size (pixels)** | **# Unique Palms** | **Total # images** |
| CASIA Adult [Su05] | CMOS Camera | 640x480 | 614 | 5,502 |
| COEP Adult [CO] | Undisclosed | 1600x1200 | 168 | 1,344 |
| Child-PalmDB1 | Samsung Galaxy S22 | 1080x1440 | 1,020 | 19,158 |
| Child CrossDB | Samsung Galaxy S22 | 1080x1440 | 318 | 12,720 |

[1] *Collected by authors. Will be released once the paper is accepted for publication.*
* *Training and testing databases are disjoint.*
† *https://www.kaggle.com/datasets/mahdieizadpanah/sapienza-university-mobile-palmprint-databasesmpd*

s/parents who brought the child to the data collection camp. This database is called Adult-PalmDB2 and contains 1,227 unique adult palms. Child-PalmDB1 and Child-PalmDB2 contain 159 common subjects (318 palms) which allows us to evaluate the performance of time-separated verification (in our case, about 5 months); this subset is referred to as Child CrossDB.

The ages of the children in Child-PalmDB1 and Child-PalmDB2 range from 6-48 months. The palmprint images were collected using the Armatura PalmMobile SDK [Pr] installed on a Samsung Galaxy S22 at a size of 1080x1440 pixels. Multiple images/child were collected with intentional variations in roll, pitch, and yaw to capture the full range of pose variations present in child palmprint images.

Table 1 shows statistics on the datasets used in this study. The training and testing sets are disjoint and captured at different times and/or by different research groups (Table 1). The number of children in the various age groups in the child palmprint databases are as follows: i) Child-PalmDB1[11]: 73 (6-12 months); 161 (12-24 months); 230 (24-48 months) and ii) Child-PalmDB2: 105 (6-12 months); 202 (12-24 months); 375 (24-48 months). We collect images from both palms. So, the total number of unique child palms is 2,142, including the 318 common palms between the two child palmprint databases.

Collecting palmprint images of a child is challenging and requires carefully designed protocols. We used the following procedure: i) An operator opens and holds the child's palm to prevent unexpected movements, ii) another operator captures the palm image at a small

[11] Age information in Child-PalmDB1 is available for only 444 subjects out of 515 subjects.

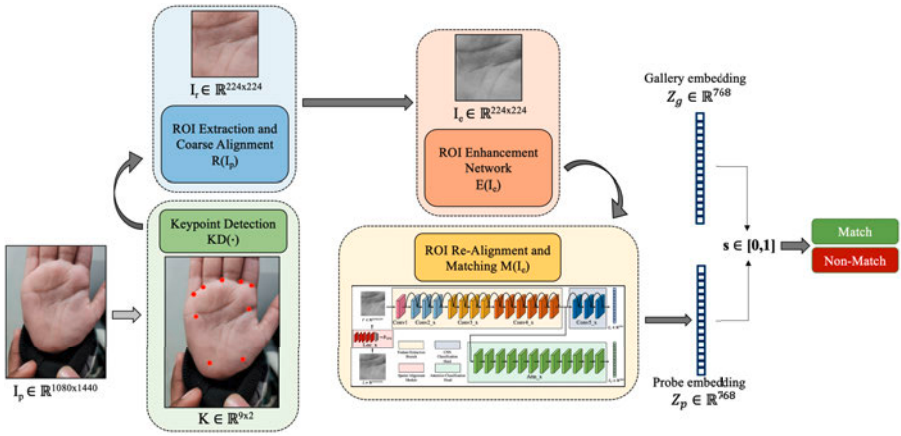



Fig. 3: A schematic diagram of the of proposed Child Palm-ID system. The input image $I_p$ is passed to the keypoint detection network $KD(\cdot)$, followed by ROI extraction. The coarse alignment between the probe and gallery images is based on a homographic transformation, followed by the AFR-Net architecture [GJ22] with a TPS re-alignment module.

standoff to capture the entire palmar surface under uniform lighting to reduce shadows and maintain high contrast in the images.

## 3 Child Palm-ID System

Contactless palmprint recognition consists of three main modules: i) Region of Interest (ROI) extraction and enhancement, ii) ROI alignment, and iii) ROI matching (Fig. 3). The predominant effort in the literature has been in building palmprint recognition systems for adults [DD16, Zh17, LK20, MFK11, Wu14, Le17, LJT13, JF08, DJM02, Ku18, ZKP16] rather than children [RDS22, Ra18]. The two studies that did use palmprints for children used only a small number of unique palms for training and evaluation (100 and 50 subjects, respectively) and lacked a cross-database evaluation. Further, the authors of these two studies did not make the child palmprint database available in the public domain.

### 3.1 ROI Extraction and Enhancement

Due to the potential of large pose variations in contactless palmprint image acquisition, it is important to obtain a consistent region of interest (ROI) across all the captured images [Zh17, DD16, LK20]. We use a deep network to predict a set of nine *keypoints* which localize the ROI via a homographic transformation, an approach commonly used in face recognition [WJ19, Zh14] with large pose variations. This keypoint-based ROI can only provide *coarse alignment*, meaning we may require and additional non-linear re-alignment for an accurate pairwise comparison of extracted ROIs. The re-alignment is particularly helpful in the case of child palmprints due to the unconstrained nature of the data collection.
The keypoint detection module $KD(\cdot)$ uses a ResNet-18 architecture with two fully connected layers inserted at the end to predict 9 keypoints $K \in \mathbb{R}^{9x2}$ in the input image ($I_p$). These 9 keypoints (Fig. 4) were selected to provide a degree of symmetry between the right and left hand while encompassing the palmar boundary containing salient information. As ground-truth for training $KD(\cdot)$, we use the keypoints generated by the Armatura

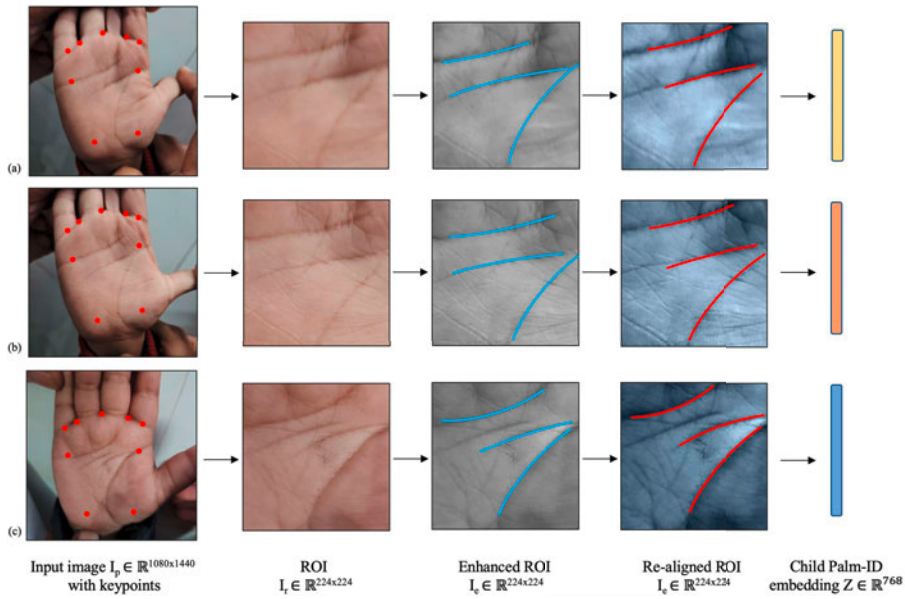



Fig. 4: Intermediate processing steps in Child Palm-ID system to obtain an embedding. (a) and (b) are two different contactless images of the same palm and (c) is an image of a different palm. The re-aligned ROIs include data augmentations that are part of the TPS STN. The principal lines before and after re-alignment are shown in blue and red, respectively. The similarity scores when comparing the ROIs are: i) (a) and (b) = 0.64, ii) (b) and (c) = 0.38, and iii) (a) and (c) = 0.29. The similarity score threshold at FAR=0.1% is 0.46.

PalmMobile SDK COTS system[12]. An MSE objective function is minimized to predict the location of the 9 keypoints.

To place the ROI into a uniform coordinate system, a set of 9 destination points $D$ are manually selected to perform a 9-point homographic transformation $H(\cdot)$ between $K$ and $D$ yielding a perspective transform matrix $\theta_h$. The ROI extraction module $R(\cdot)$ applies $\theta_h$ to $I_p$ to get a warped image $I_p^w$ followed by a 224 x 224 crop $C$ yielding the coarsely aligned ROI, $I_r$ based on the following equations: i) $\theta_h = H(K,D)$; ii) $I_p^w = R(I_p;\theta_h)$; iii) $I_r = C(I_p^w, 224)$. To enhance the extracted ROI, we utilize the latent enhancement network from [GJ23] and adapt it to enhance contactless palmprint images. In particular, we simulated low quality palm images by blurring and down-sampling high quality captured palm images. The enhancement network is then trained via an MSE loss between the high quality palm-print ground-truths and the reconstructed outputs of the enhancement network (Fig. 4). The benefit of enhancement is shown in the ablation study in Table 3.

### 3.2 ROI Alignment and Matching

Adult palmprint recognition systems have utilized the principal lines [St88] for the re-alignment of ROIs [Wu04, Zh03]. Spatial Transformer Networks (STN) have been used to predict alignment parameters that maximize the recognition accuracy [ZC20, ECJ19, GJ22]. Additionally, fine-tuned, non-linear alignment using a Thin Plate Spline (TPS) STN

[12] https://armatura.us/

has shown even higher recognition performance in more unconstrained scenarios such as 3D facial recognition, contact-to-contactless fingerprint matching as well as unconstrained palmprint recognition [Bh17, Gr21, Ma19]. In this paper, we implement a semi-supervised TPS STN module that learns an optimal non-linear distortion field for a coarsely aligned ROI that maximizes the accuracy of Child Palm-ID.

The feature extraction and matching architecture of Child Palm-ID is based on AFR-Net [GJ22], a fingerprint recognition model based on ResNet50 and Vision Transformers (ViT). AFR-Net uses an STN to predict an affine alignment of the input images. We modify the STN to predict a TPS alignment that applies a learned distortion field, $\theta_{TPS}$, to the coarsely aligned palmprint ROIs ($I_r$) producing an aligned ROI, $I'$ based on the following equation: $I' = T(I_r; \theta_{TPS})$. Fig. 4 shows the improved alignment between two ROIs after the re-alignment with $T(I_r)$. Affirming the intuition behind the use of $T(\cdot)$, a significant boost in recognition performance was observed compared to the use of the pre-existing STN in AFR-Net (from TAR = 73.8% to TAR = 88.3%, both at FAR = 0.1%. See Table 3.). Finally, normalized embeddings $Z_p$ and $Z_g$ are obtained for a given probe and gallery image, respectively, and are compared to obtain a similarity score $s \in [0,1]$ based on the following equation: $s = Z_p^T \cdot Z_g, \in [0,1]$.

To further boost the recognition performance, we divide the 224x224 coarsely aligned and enhanced ROIs into 4 quadrants and train an ensemble of models, one per quadrant to complement the model trained on the entire ROI, as has been demonstrated in several facets of deep learning [SR18]. Using the ensemble of these five embeddings, we obtain a final similarity score based on mean score fusion (Table 3).

## 4 Experimental Results

We evaluate the verification performance of Child Palm-ID and compare it to the baseline accuracy of a COTS system [Pr] as well as two open source algorithms [Zh17, Ma19]. We report the accuracies on Child CrossDB as well as separately for the three age groups (6-12 mos., 12-24 mos. and 24-48 mos.) from Child-PalmDB1. Finally, we report results of our ablation study.

### 4.1 Verification Results

We report verification performance on four evaluation databases that were altogether kept separate from the training set (see Table 1). The recognition performance of the proposed Child Palm-ID is competitive with COTS[13]. We also report the longitudinal verification performance on Child CrossDB containing the 159 subjects present in both Child-PalmDB1 and Child-PalmDB2 in Table 2. Child CrossDB is disjoint from the training set. It is instructive to notice the trend in performance of Child Palm-ID on different age groups. Intuitively, a recognition system would perform better on relatively older children since they are likely to be more cooperative during data acquisition. Child Palm-ID shows an accuracy of TAR=92.57% on children between the ages of 6 to 12 mos., TAR=96.41% on children between the ages of 12 to 24 mos. and TAR=98.92% on children in the age

[13] The architecture and training set for the COTS is not known to us. Both the adult databases used for evaluation are in the public domain.

Tab. 2: TAR(%) @ FAR=0.01% (FAR=0.1%) of the proposed Child Palm-ID system (**A**) and the three baselines (**B**, **C**, and **D**.)

| **Database** | Child Palm-ID **(A)** | COTS **(B)** | CR-Comp Code [Zh17]**(C)** | Matkowski et al. [Ma19] **(D)** | Sum Score Fusion of COTS **(B)** with **A** and **D** | |
|---|---|---|---|---|---|---|
| | | | | | **A+B** | **B+D** |
| CPDB1† | 91.48 (94.8) | 90.85 (92.7) | 74.71 (78.7) | 80.91 (83.7) | 92.16 (94.87) | 91.04 (92.7) |
| CPDB1† (6-12 mos.) | 91.11 (92.57) | 87.79 (89.88) | 72.68 (74.25) | 78.91 (80.62) | 90.57 (92.68) | 86.19 (88.65) |
| CPDB1† (12-24 mos.) | 94.97 (96.41) | 91.02 (93.89) | 76.84 (79.82) | 82.8 (85.78) | 95.1 (96.52) | 91.14 (93.9) |
| CPDB1† (24-48 mos.) | 96.97 (98.92) | 94.02 (96.32) | 81.76 (84.39) | 87.51 (89.63) | 97.13 (98.99) | 94.13 (96.36) |
| Child CrossDB | 77.68 (80.5) | 75.93 (78.2) | 61.17 (64.8) | 68.9 (71.7) | 79.86 (82.4) | 77.11 (79.8) |
| CASIA | 98.89 (99.5) | 100 (100) | (96.16) 97.2 | 97.98 (99.2) | 100 (100) | 100 (100) |
| COEP‡ | 100 (100) | 100 (100) | 100 (100) | 100 (100) | 100 (100) | 100 (100) |

† *We abbreviate Child-PalmDB1 as CPDB1 in this table to save space.*
‡ *17 mislabelled identities were removed.*

group of 24-48 mos., all @ FAR=0.1%. Child Palm-ID outperforms the COTS system at FAR = 0.1% in each of the three evaluation age groups.

Lastly, we note an improvement by sum score fusion of Child Palm-ID and COTS, especially in the case of Child CrossDB at FAR=0.1%. This suggests a potential for improvements in our algorithm, given that Child Palm-ID and COTS are complementary.

### 4.2 Ablation Study

In the ablation study shown in Table 3, we examine the effects of the autoencoder enhancement module, TPS alignment module, ensemble of multi-patch embeddings and data augmentations for training. The training datasets were fixed (Table 1) in these ablations. The TPS re-alignment module in row 2 of Table 3, gives the biggest boost in accuracy on all the four evaluation databases. The image enhancement, ensemble of embeddings and data augmentations provide a further boost in accuracy.

### 4.3 Failure Cases

Fig. 5 shows four failure cases of Child Palm-ID system on Child CrossDB. These examples highlight the challenges in cross-dataset comparison when there are significant differences in standoff distance, lighting and rotation between two time-separated acquisitions. Other challenges we noticed include motion blur from movement of the palm, partially closed palms, and large pose variations.

In conjunction, Fig. 6 shows four successful cases of Child Palm-ID on Child CrossDB. It is evident from these examples that a genuine pair of images is correctly matched under relatively similar lighting conditions and overall orientation of the child's palm.

Tab. 3: Ablation Study for Child Palm-ID. Results are reported as TAR (%) @ FAR = 0.1%

| **Modules Used** | | | | | **Evaluation Databases** | | | |
|---|---|---|---|---|---|---|---|---|
| Coarse Align-ment | Re-Align ment | Augmen-tation | Ensem-ble | Enhanc-ement | CASIA Adult Database | COEP Adult Database | Child-PalmDB1 | Child CrossDB |
| ✓ | ✗ | ✗ | ✗ | ✗ | 92.4 | 91.6 | 73.8 | 66.56 |
| ✓ | ✓ | ✗ | ✗ | ✗ | 98.8 | 99.1 | 88.3 | 74.68 |
| ✓ | ✓ | ✓ | ✗ | ✗ | 99.1 | 100 | 92.43 | 76.67 |
| ✓ | ✓ | ✓ | ✓ | ✗ | 99.5 | 100 | 93.41 | 77.4 |
| ✓ | ✓ | ✓ | ✓ | ✓ | 99.5 | 100 | 94.8 | 80.5 |

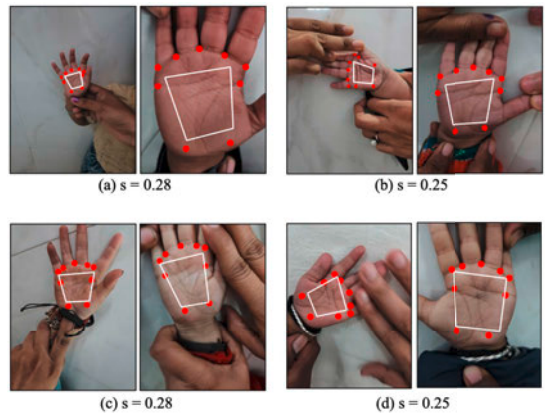



Fig. 5: Example failure cases of Child Palm-ID system in time-separated Child CrossDB. For each genuine pair of images in (a)-(d), the similarity score $s$ is below the threshold of 0.46 at FAR = 0.1%. In both (a)-(d), the left image is from Child-PalmDB1 and the right image is from Child-PalmDB2 with 5 mos. of time-separation.

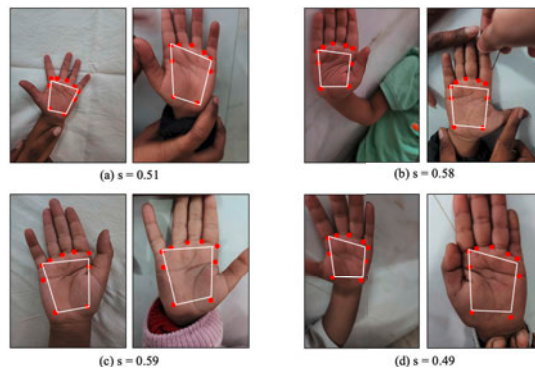



Fig. 6: Example successful cases of Child Palm-ID system in time-separated Child CrossDB. For each genuine pair of images in (a)-(d), the similarity score $s$ is above the threshold of 0.46 at FAR = 0.1%. In (a)-(d), the left image is from Child-PalmDB1 and the right image is from Child-PalmDB2 with 5 mos. of time-separation.

## 5 Conclusion and Future Work

Biometric recognition systems have made great strides over the past 20 years. However, so far, all these systems have been primarily designed to be used by adults. Yet there are numerous social good tasks ranging from eradicating vaccine preventable diseases to child malnutrition where biometric recognition can play a significant role to prevent misery and loss of life.

We have designed and prototyped Child Palm-ID, a contactless mobile-based palmprint recognition system. We have evaluated verification performance of Child Palm-ID on both child as well as adult contactless palmprint databases. We show competitive recognition performance of our system against a SOTA COTS system @ FAR=0.1% and superior performance over two academic algorithms available in the public domain. The main technical contributions of our paper include a re-alignment strategy for palmprint images using a TPS alignment module and an autoencoder-based image enhancement. Both these modules are critical for success in the case of child palmprint recognition. Our ongoing work includes i) Child Palm-ID mobile app displaying the faces of the top $N$ retrievals from a gallery so the operator is able to manually confirm the identity of the child, ii) introduction of a palmprint image quality metric to flag images of poor quality for recapture, iii) synthetic palmprint generation to amplify the amount of data available for training, and iv) representation learning to account for differences in standoff distance, illumination, orientation and motion blur.

## References


[Bh17] Bhagavatula, Chandrasekhar et al.: Faster than Real-time Facial Alignment: A 3D Spatial Transformer Network Approach in Unconstrained Poses. In: Proc. ICCV. 2017.

[CO] COEP: COEP Palm Print Database. COEP Palm Print Database — College of Engineering, Pune. https://www.coep.org.in/resources/coeppalmprintdatabase.

[DD16] Dian, Liu; Dongmei, Sun: Contactless Palmprint Recognition Based on Convolutional Neural Network. In: IEEE 13th ICSP. 2016.

[DJM02] Duta, Nicolae; Jain, Anil K; Mardia, Kanti V: Matching of Palmprints. Pattern Recognition Letters, 2002.

[ECJ19] Engelsma, Joshua J; Cao, Kai; Jain, Anil K: Learning a Fixed-Length Fingerprint Representation. IEEE Trans. PAMI, 43(6):1981–1997, 2019.

[En21] Engelsma, Joshua J; Deb, Debayan; Cao, Kai; Bhatnagar, Anjoo; Sudhish, Prem S; Jain, Anil K: Infant-ID: Fingerprints for Global Good. IEEE Trans. PAMI, 44(7), 2021.

[GGJ23] Godbole, Akash; Grosz, Steven A; Jain, Anil K: Child Palm-ID: Contactless Palmprint Recognition for Children. arXiv preprint arXiv:2305.05161, 2023.

[GJ22] Grosz, Steven A; Jain, Anil K: AFR-Net: Attention-Driven Fingerprint Recognition Network. arXiv preprint arXiv:2211.13897, 2022.

[GJ23] Grosz, Steven A; Jain, Anil K: Latent Fingerprint Recognition: Fusion of Local and Global Embeddings. arXiv preprint arXiv:2304.13800, 2023.

[Gr21] Grosz, Steven A; Engelsma, Joshua J; Liu, Eryun; Jain, Anil K: C2CL: Contact to Contactless Fingerprint Matching. IEEE Trans. IFS, 17, 2021.

[Ha08] Hao, Ying; Sun, Zhenan; Tan, Tieniu; Ren, Chao: Multispectral Palm Image Fusion for Accurate Contact-Free Palmprint Recognition. In: 15th IEEE ICIP. 2008.

[Iz19] Izadpanahkakhk, Mahdieh et al.: Novel Mobile Palmprint Databases for Biometric Authentication. International Journal of Grid and Utility Computing, 10(5):465–474, 2019.

[JF08] Jain, Anil K; Feng, Jianjiang: Latent Palmprint Matching. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(6):1032–1047, 2008.

[JRN11] Jain, Anil K.; Ross, Arun A.; Nandakumar, Karthik: Introduction to Biometrics. Springer Publishing Company, 2011.

[Ka22] Kalisky, Tom et al.: Biometric Recognition of Newborns and Young Children for Vaccinations and Health Care: a Non-randomized Prospective Clinical Trial. Scientific Reports, 2022.

[Ko19] Kotzerke, Johannes et al.: Newborn and Infant Discrimination: Revisiting Footprints. Australian Journal of Forensic Sciences, 2019.

[Ku08] Kumar, Ajay: Incorporating Cohort Information for Reliable Palmprint Authentication. In: Sixth ICVGIP. IEEE, 2008.

[Ku18] Kumar, Ajay: Toward More Accurate Matching of Contactless Palmprint Images Under Less Constrained Environments. IEEE Trans. IFS, 2018.

[Le11] Lemes, Rubisley P; Bellon, Olga RP; Silva, Luciano; Jain, Anil K: Biometric Recognition of Newborns: Identification Using Palmprints. In: (IJCB). IEEE, 2011.

[Le17] Leng, Lu et al.: Dual-source Discrimination Power Analysis for Multi-instance Contactless Palmprint Recognition. Multimedia Tools and Applications, 2017.

[LJT13] Liu, Eryun; Jain, Anil K; Tian, Jie: A Coarse to Fine Minutiae-based Latent Palmprint Matching. IEEE Trans. PAMI, 2013.

[LK20] Liu, Yang; Kumar, Ajay: Contactless Palmprint Identification Using Deeply Learned Residual Features. IEEE Trans. BIOM, 2020.

[Ma19] Matkowski, Wojciech Michal et al.: Palmprint Recognition in Uncontrolled and Uncooperative Environment. IEEE Trans. IFS, 2019.

[MFK11] Morales, Aythami; Ferrer, Miguel A; Kumar, Ajay: Towards Contactless Palmprint Authentication. IET Computer Vision, 5(6):407–416, 2011.

[Pr] Products Powered by Armatura. https://armatura.us/.

[Ra18] Ramachandra, Raghavendra; Raja, Kiran B; Venkatesh, Sushma; Hegde, Sneha; Dandappanavar, Shreedhar D; Busch, Christoph: Verifying the Newborns Without Infection Risks Using Contactless Palmprints. In: IEEE ICB. 2018.

[RDS22] Rajaram, Kanchana; Devi, Arti; Selvakumar, S: PalmNet: A CNN Transfer Learning Approach for Recognition of Young Children Using Contactless Palmprints. In: Machine Learning and Autonomous Systems, pp. 609–622. Springer, 2022.

[Sa19] Saggese, Steven et al.: Biometric Recognition of Newborns and Infants by Non-Contact Fingerprinting: Lessons Learned. Gates Open Research, 2019.

[SR18] Sagi, Omer; Rokach, Lior: Ensemble Learning: A Survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4):e1249, 2018.

[St88] Stevens, Cathy A; Carey, John C; Shah, Madhuri; Bagley, Grant P: Development of Human Palmar and Digital Flexion Creases. The Journal of Pediatrics, 1988.

[Su05] Sun, Z; Tan, T; Wang, Y; Li, SZ: Ordinal Palmprint Representation for Personal Identification. In: Proceedings of the IEEE CVPR. 2005.

[WJ19] Wu, Yue; Ji, Qiang: Facial Landmark Detection: A Literature Survey. International Journal of Computer Vision, 127:115–142, 2019.

[Wu04] Wu, Xiangqian; Zhang, David; Wang, Kuanquan; Huang, Bo: Palmprint Classification Using Principal Lines. Pattern Recognition, 37(10):1987–1998, 2004.

[Wu14] Wu, Xiangqian et al.: A SIFT-based Contactless Palmprint Verification Approach Using Iterative RANSAC and Local Palmprint Descriptors. Pattern Recognition, 2014.

[ZC20] Zhang, Hongxin; Chi, Liying: End-to-End Spatial Transform Face Detection and Recognition. Virtual Reality & Intelligent Hardware, 2(2):119–131, 2020.

[Zh03] Zhang, David; Kong, Wai-Kin; You, Jane; Wong, Michael: Online Palmprint Identification. IEEE Trans. PAMI, 2003.

[Zh14] Zhang, Zhanpeng; Luo, Ping; Loy, Chen Change; Tang, Xiaoou: Facial Landmark Detection by Deep Multi-Task Learning. In: 13th ECCV. pp. 94–108, 2014.

[Zh17] Zhang, Lin; Li, Lida; Yang, Anqi; Shen, Ying; Yang, Meng: Towards Contactless Palmprint Recognition: A Novel Device, a New Benchmark, and a Collaborative Representation Based Identification Approach. Pattern Recognition, 69:199–212, 2017.

[ZKP16] Zheng, Qian; Kumar, Ajay; Pan, Gang: A 3D Feature Descriptor Recovered From a Single 2D Palmprint Image. IEEE Trans. PAMI, 2016.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# Exploring the Untapped Potential of Unsupervised Representation Learning for Training Set Agnostic Finger Vein Recognition

Tugce Arican, [1] Raymond Veldhuis, [2] Luuk Spreeuwers [3]

**Abstract:**

Finger vein patterns are a promising biometric trait because of their higher privacy and security features compared to face and finger prints. Finger vein recognition methods have been researched extensively, especially deep learning based methods such as Convolutional Neural Networks. These methods show promising recognition performance, but their low degree of generalization and adaptability results in much lower and inconsistent recognition performance in cross database scenarios. Despite these drawbacks, much less research has gone into the generalization and adaptability of these deep learning methods.This study addresses these issues and proposes an unsupervised learning approach, namely a patch-based Convolutional Auto-encoder for learning finger vein representations. Our proposed approach outperforms traditional baseline finger recognition methods on the UTFVP, SDUMLA-HMT, and PKU datasets, and achieves state-of-the-art performance on the UTFVP dataset with 0.24% EER. It also indicates a noticeably higher generalization of finger vein features across different datasets compared to a supervised method. The findings of this work offer promising advancements in achieving robust finger vein recognition in real-life scenarios, due to the enhanced generalization and adaptability of our proposed method.

**Keywords:** Finger vein recognition, unsupervised learning, auto-encoders, cross-database

## 1 Introduction

Finger vein patterns are invisible to the naked eye and leave no discernible trace, endowing them with exceptional privacy and security characteristics compared to other biometric modalities like facial features or finger prints. The foundation of finger vein recognition lies in the acquisition and comparison of random vein patterns. Finger vein recognition has gained significant popularity among researchers, engaging their efforts in various aspects, including acquisition and extracting vein patterns, as well as the comparison and analysis of these patterns.

Deep learning methods are extensively employed for finger vein recognition because of their superior generalization abilities in contrast to traditional feature extraction. Researchers propose a variety of supervised and unsupervised learning architectures such as Convolutional Neural Networks (CNNs)[Hu18, Wa19, Ze19, Ku20, Ku20], Convolutional

[1] EEMCS, DMB, University of Twente, Enschede, The Netherlands, t.arican@utwente.nl
[2] EEMCS, DMB, University of Twente, Enschede, The Netherlands, r.n.j.veldhuis@utwente.nl
[3] EEMCS, DMB, University of Twente, Enschede, The Netherlands, l.j.spreeuwers@utwente.nl

Auto-encoders (CAEs)[Ch22, Pa23], and recently Visual Transformers[Hu22]. While supervised learning achieves state-of-the-art recognition performance in finger vein recognition, unsupervised methods demonstrate tremendous potential in learning meaningful representations of finger vein patterns.

While the majority of the existing literature focuses on enhancing the recognition performance of finger vein patterns, only a few researchers[Ta19, Pr22] point out the limitations in the generalization and adaptability of the supervised methods to real-world scenarios, such as cross-database comparisons. In these studies, the performance of cross-database recognition, where the model is trained on a dataset different from the evaluation set, exhibits significantly lower performance compared to the single-database case. These results demonstrate that the state-of-the-art models are not robust against variations in dataset characteristics between train and evaluation sets. This study introduces an unsupervised learning method called patch-based Convolutional Auto-encoder(P-CAE) for learning representations of finger vein patterns. The effectiveness of the P-CAE is compared against a supervised method proposed by Kuzu et.al[KMC21]. The results achieved by the P-CAE demonstrate significant promise for unsupervised methods in terms of generalization and adaptability of learned finger vein features. This holds immense potential to facilitate advancements in cross-device finger vein recognition[4].

## 2 Related Work

Several researchers have proposed various deep learning architectures for finger vein recognition. Tang et. al.[Ta19] propose a Siamese architecture with a Contrastive Loss to reduce intra-class variance and increase inter-class variance. The proposed architecture achieves state-of-the-art results on publicly available finger vein datasets. Ou et.al.[Ou22] utilize a Generative Adversarial Network (GAN) to address intra-class variance issues by artificially increasing the number of finger vein samples. Kuzu et al.[Ku20] highlight the effectiveness of transfer learning over training from scratch and emphasise the importance of the choice of loss function[KMC21]. Both works achieve competitive results on publicly available finger and palm vein datasets. Bros et.al.[BKM21] address contrast issues observed in finger vein images and propose an enhancement approach using a Convolutional Auto-encoder (CAE) which learns a linear combination of finger vein images with their annotated vein patterns. On the other hand, Chen et.al[Ch22] propose a CAE architecture for automatic vein annotation from the finger vein images. Pan et.al[Pa23] highlight the effectiveness of processing texture and shape features of finger vein images separately through a dual-branch CAE architecture.

Despite the extensive literature on finger vein recognition using deep learning methods, only a limited number of researchers have explored the generalization and adaptability of these methods. In their study, Tang et.al[Ta19] present that the recognition performance substantially degrades when the model is evaluated on a finger vein dataset different from the dataset it is trained on. Similarly, Prommegger et.al[Pr22] claim that when the evalu-

[4] Reference and probe images are captured by different devices

ation set possesses distinct characteristics from the training set, the segmentation performance of the compared CNN models undergoes a substantial degradation. Though these studies acknowledge the generalization and adaptability challenges of supervised learning for finger vein recognition, these issues have not been addressed much. Noh et.al[No21] introduce a cycle-consistent adversarial network(CycleGAN) to address heterogeneity issues between datasets. The CycleGAN is trained to generate domain-adapted images between the source and target domains. The generated image is subsequently used as input to a CNN model for recognition purposes. While the authors claim that the CycleGAN can handle unobserved data without retraining the model, the study does not include a comparative analysis on completely unseen data to validate this claim. Chen et.al. [Ch22] propose a vein extraction method as a domain adaptation strategy for improving generalization in finger vein recognition. The authors employ a U-Net model as a domain adaptation network to facilitate the mapping of finger vein images from grey domain to binary domain. The objective of this mapping is to minimise variations between datasets. Though the study demonstrates that the proposed approach can achieve satisfactory recognition performances even when the model is trained on a different dataset than the evaluation set, the authors present the results on a model trained on a particular dataset. It remains unclear how the characteristics of the training data impact the cross-database comparisons. We aim to fill this gap by introducing an unsupervised learning approach that provides finger vein features with better generalization properties. This approach potentially enables a more robust finger vein recognition system, capable of handling operational scenarios where data characteristics may vary.

## 3 Methodology

### 3.1 Patch-based Convolutional Auto-encoder

This study proposes a patch-based Convolutional Auto-encoder(P-CAE) for finger vein representation learning. Finger vein images are primarily composed of the finger background, and the vein patterns have low contrast and are sparsely distributed. A lower dimensional patch input is utilized instead of the entire finger region in order to prioritise extraction of important vein features over less significant finger background information. The P-CAE is trained to reconstruct patches extracted solely from the finger region. Once the training is completed, the comparison of vein patterns is performed on a patch-by-patch basis.

The proposed P-CAE architecture consists of 6 compression and 6 de-compression blocks in the encoder and decoder respectively. Each block involves a convolution (or transposed-convolution) layer, batch normalisation, and the LeakyReLu activation function. The encoder compresses 64 x 64-pixel patches to a latent vector with a size of 32. The P-CAE is trained for 50 epochs with a learning rate of $5x10^{-5}$. Figure 1 summarises the P-CAE architecture. The cosine similarity metric is employed to evaluate the similarity between the embeddings of a pair of patches. Subsequently, the similarity scores of all patch pairs within an image pair are averaged to determine the overall similarity between the pair of images.

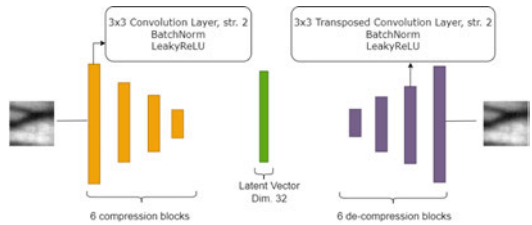



Fig. 1: Patch-based CAE architecture

### 3.2 Alignment

Patch-based methods are susceptible to alignment errors. To attain optimal results with the P-CAE, a number of alignment steps are implemented during the preprocessing step. First, the Iterative Closest Point (ICP)[BM92] aligns a pair of finger vein images using the finger edge information. Then, the vein patterns are horizontally aligned using the method proposed in [Qi16], and out-of-plane rotation is corrected using the approach presented in [Pr19]. Furthermore, the comparison between image pairs is conducted utilizing a sliding window approach. During this step, probe image patches are systematically moved over the reference image, and the location that produces the highest cosine similarity score from a combination of patch pair scores, is utilized as the similarity score for the image pair.

## 4 Dataset and Experiments

The patch-based CAE model is evaluated on three finger vein datasets: UTFVP[TV13], SDUMLA-HMT[YLS11], and PKU[Pe13]. Each dataset exhibits distinct characteristics such as quality, and illumination. The UTFVP dataset consists of high quality finger vein images, while SDUMLA-HMT contains images of lower quality, and exhibits high degree of out-of-plane rotations and translations. Since this study does not focus on addressing rotation and translation issues, a subset of the SDUMLA-HMT dataset with a relatively mild out-of-plane rotation is utilized for the evaluation of this dataset. Further details about the datasets and the divisions used for training and evaluation are presented in Table 1.

The CAE model is compared to two traditional baseline finger vein recognition methods, namely Maximum Curvature(MC)[MNM07] and Repeated Line Tracking(RLT)[MNM04], and a CNN architecture proposed by Kuzu et.al[KMC21]. The CNN utilizes DenseNet-161 as its backbone and showcases outstanding recognition performance on publicly available finger vein datasets. The CNN model is trained with the parameters presented in the original work. To ensure equitable comparison, the CNN is trained and evaluated using identical datasets as the CAE.

| | Number of Subjects | Fingers per Subject | Total number of images | Train Subjects | Evaluation Pairs (M / NM*) |
|---|---|---|---|---|---|
| UTFVP | 60 | 6 | 1440 | 20 | 2880 / 2880 |
| SDUMLA-HMT | 106 | 6 | 3816 | 76 | 5400 / 5400 |
| PKU | 200 | 1 | 1528 | 100 | 5018 / 5018 |

Tab. 1: Finger vein database details

* Mated / Non-mated

The comparison performances of the aforementioned models are evaluated by comparing their Equal Error Rates (EER) and False Non Match Rates(FNMR), where the False Match Rate(FMR) is set to 0.1%(FNMR1000). The EER represents the point at which the FNMR is equal to the FMR, while FNMR1000 indicates the FNMR at a specific FMR value. Detection Error Tradeoff (DET) curves are employed to compare the tradeoff between FMR and FNMR for different models. Furthermore, the behaviour of these two models is investigated by comparing histograms of image pair similarity.

## 5 Results

Table 2 shows within dataset evaluation performances of the recognition methods mentioned in Section 4. The evaluation performance of MC and RLT methods is derived from existing literature, as indicated in Table 2. The CNN and the P-CAE are trained using the same finger vein dataset utilized for evaluation. The results obtained indicate that the P-CAE outperforms the traditional methods on all three datasets, and achieves the state-of-the-art performance on the UTFVP dataset with 0.24% EER. In comparison to other methods, the CNN demonstrates a superior performance on two of the three finger vein datasets.

| | MC | RLT | Results taken from | CNN | P-CAE |
|---|---|---|---|---|---|
| UTFVP | 0.4 | 0.9 | [TV13] | 6.87 | 0.24 |
| SDUMLA-HMT | 3.65 | 5.85 | [Ya19] | 1.56 | 2.67 |
| PKU | 3.14 | 3.7 | [Sy17] | 2.40 | 2.43 |

Tab. 2: Within database evaluation performances of different recognition methods in EER(%)

Table 3 presents cross-database comparison performances in terms of EER(%) for the CNN and the P-CAE models. MC and RLT do not involve a training step, hence, their evaluation performances are fully presented in Table 2. The rows of Table 3 represent the evaluation datasets, while the columns indicate the corresponding training sets. Notably, the P-CAE demonstrates a superior performance compared to the CNN model in cross-database comparisons especially on the UTFVP and SDUMLA-HMT datasets. Moreover, the P-CAE achieves comparable performances despite variations between train and evaluation sets. For example, when evaluating the SDUMLA-HMT dataset using a model trained on the UTFVP dataset, the CNN exhibits a notable inferior performance with 6.63% EER where the performance of the SDUMLA-HMT dataset is presented as 1.56% EER with this

model. On the other hand, under identical conditions, the P-CAE demonstrates a comparable comparison performance. Conversely, the CNN not only exhibits poor performance in cross-database comparisons but also demonstrates a fluctuating behaviour, unlike the P-CAE. For instance, when evaluating the PKU dataset, the CNN achieves an EER of 2.40%. However, when the same dataset is evaluated on a model trained using the SDUMLA-HMT dataset, the comparison performance significantly decreases to 18.8% EER.

| | CNN | | | CAE | | |
|---|---|---|---|---|---|---|
| Evaluation \ Train | UT* | SD** | PKU | UT | SD | PKU |
| UTFVP | 6.87 | 12.2 | 3.68 | 0.24 | 0.38 | 0.24 |
| SDUMLA-HMT | 6.63 | 1.56 | 3.36 | 2.26 | 2.67 | 1.96 |
| PKU | 19.9 | 18.8 | 2.40 | 2.59 | 3.15 | 2.43 |

Tab. 3: Cross-database evaluation performances of CNN and CAE in EER(%)
* UT - UTFVP, SD** - SDUMLA-HMT

Table 4 shows cross-database comparison performances for the CNN and the P-CAE models in terms of FNMR1000 in percentage. The rows in the table corresponds to the evaluation sets, while the columns represents the training sets. Similar to the performances presented in Table 3, particularly in cross-database comparisons, the P-CAE demonstrates notably lower FNMR at an FMR of 0.1%. Furthermore, in comparison to the CNN model, the FNMR is subject to less pronounced impact from the differences in training sets with the P-CAE model. To illustrate, while evaluating PKU dataset, the CNN model achieves 29.27% FNMR.Conversely, the FNMR increases to 95.60% when the training set is utilised as UTFVP. In contrast, the P-CAE exhibits a mere 1.36% disparity under identical conditions.

| | CNN | | | CAE | | |
|---|---|---|---|---|---|---|
| Evaluation \ Train | UT* | SD** | PKU | UT | SD | PKU |
| UTFVP | 55.4 | 80.3 | 30.9 | 0.31 | 0.73 | 0.73 |
| SDUMLA-HMT | 46.0 | 14.1 | 25.5 | 8.94 | 12.6 | 8.94 |
| PKU | 95.6 | 99.4 | 29.2 | 19.2 | 23.8 | 20.57 |

Tab. 4: Cross-database evaluation performances of CNN and CAE in FNMR1000(%)
* UT - UTFVP, SD** - SDUMLA-HMT

Figure 2 presents a comparison of DET curves for both the CNN and the P-CAE models for each training set. Figure 2b illustrates a notable fluctuation in FNMR when the CNN is trained using a different dataset than the evaluation set. In contrast, the P-CAE (Fig. 2a) presents consistent performances despite the differences between the training and the evaluation sets. In particular, when the P-CAE is trained on the SDUMLA-HMT or PKU dataset, the DET curves exhibit a remarkably consistent comparison performance on all three evaluation sets.

Similarity scores of mated and non-mated pairs highlight the differences between the P-CAE and the CNN models. Figure 3 illustrates how the similarity scores vary when the training set changes while evaluating the UTFVP dataset, with both the CNN and the P-CAE models. Notably, the distance scores exhibit significant fluctuations with the change

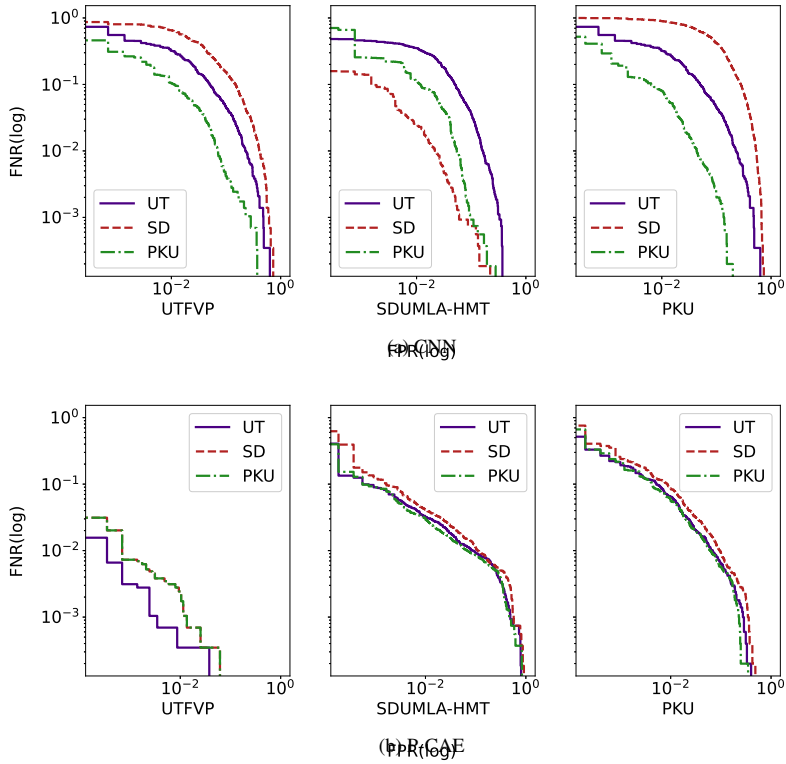
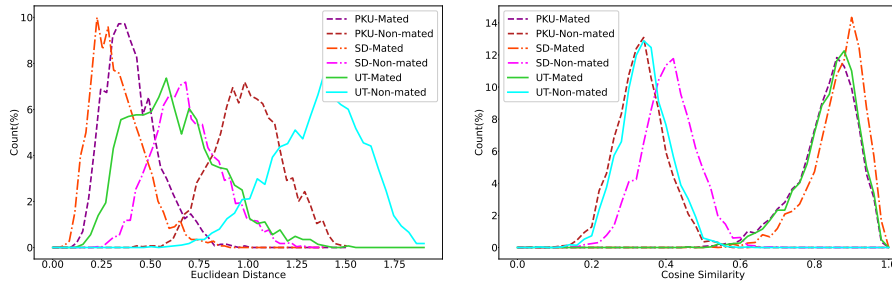



Fig. 2: DET curves of (a) CNN, (b) P-CAE models. Legends indicate the training dataset.

of training set in the CNN model(Fig. 3a). Moreover, when the training set differs from the evaluation set, the distances between all image pairs decrease. This implies that the CNN identifies more similarities between both mated and non-mated pairs in cross-database comparisons. In contrast, Figure 3b demonstrates the notably more stable behaviour of the P-CAE model despite the different characteristics of the training sets. Particularly, the similarities among mated pairs are preserved almost perfectly, as evident from the minimal changes in mated pair similarity histograms with different datasets. On the other hand, a slight increase in similarity scores of non-mated pairs is observed when the SDUMLA-HMT dataset is used for training. This emphasizes the ability of the P-CAE to provide higher generalization and adaptability for finger vein features through the patches and unsupervised learning compared to the CNN model.

Figure 3b shows that training the P-CAE on the SDUMLA-HMT dataset leads to a slight increase in similarity scores in non-mated pairs. Figure 4 demonstrates how the similarity scores for non-mated patch pairs are influenced by the characteristics of the training set. The image pair is taken from the UTFVP dataset. Figure 4b shows that when evaluating




(a) CNN (b) P-CAE

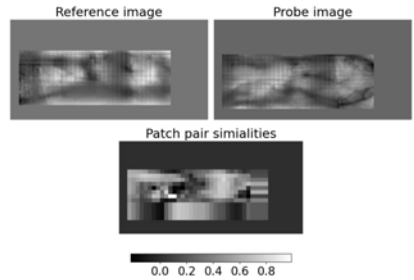
Fig. 3: Similarity histograms of (a) the CNN and (b) the P-CAE models trained on the UTFVP dataset evaluated on the UTFVP (UT), SDUMLA-HMT (SD), and PKU datasets

patch pairs on a model trained with the SDUMLA-HMT dataset, regions lacking distinct vein structures demonstrate significantly higher similarity scores compared to the model trained on the UTFVP dataset(Fig. 4a).

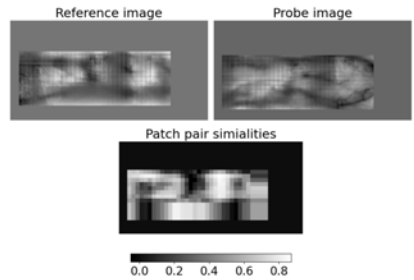
## 6 Discussion and Future Work

Table 3 indicates that the CNN demonstrates a significant fluctuation in cross-database comparison performances, while the P-CAE achieves comparable performances across training sets. This difference is likely due to the influence of background information during the training and the evaluation process. A considerable portion of a finger vein image includes the finger background, which can vary significantly across different acquisition devices. The CNN may inadvertently utilize this information when learning identity information. On the other hand, the P-CAE focuses on a much smaller region of the finger vein image, which limits exposure to the background information and instead emphasises the reconstruction of finger vein information.

Observations indicate that training the CNN model using the PKU dataset leads to noticeably improved comparison performance of the UTFVP dataset when compared to the model trained with the UTFVP dataset. Upon further examination, it comes to light that the images of the same fingers within the UTFVP dataset display notable differences in terms of finger shape and width, and even the image background, in contrast to relatively fewer variations observed in the PKU dataset. It is likely that when the CNN is trained using the UTFVP dataset, it tends to give more importance to extracting feature related to finger shape. Consequently, this leads to higher distances for mated pairs compared to training with the PKU dataset, even though the vein patterns appear to be similar. Although finger shape and width carry identity-related information, the experiments suggest that these features can also perplex the model, resulting in an increase in both false non-match and false match rates.




(a) Trained on UTFVP




(b) Trained on SDUMLA

Fig. 4: Patch pair similarity comparisons of a non-mated pair from the UTFVP dataset scored by the P-CAE trained on the (a) UTFVP, (b) SDUMLA-HMT datasets

In the experiments, the P-CAE is found to be sensitive to the alignment of image pairs. The estimation of translation and out-of-plane rotation parameters relies on the correlations between finger vein image pairs. Therefore, the parameter estimation not only involves the vein patterns but also the finger background. Especially on low quality datasets, such as SDUMLA-HMT, the vein patterns are observed as blending into the finger background. In such cases, the correlation method yields false matches between image pairs, leading to an improper alignment. In future research, it could be beneficial to estimate translation parameters between vein patterns instead of grey images. This approach would minimize the contribution of the background to the estimation of translation parameters, thereby leading to an improved alignment accuracy.

The experiments reveal that when the P-CAE is trained on a low quality dataset, such as SDUMLA-HMT, the model struggles to learn strong discriminative vein representations. Figure 4 demonstrates that in this case, even if the evaluation set exhibits fine vein details,

the P-CAE fails to identify them, which leads to higher similarity scores for non-mated pairs and ultimately results in higher false match rates. As an area for future research, it would be worthwhile to explore reinforcing this lost information through preprocessing techniques or reintroducing it to the model during training. Therefore, the P-CAE would be trained to place more emphasis on capturing fine vein details.

## 7 Conclusion

This work explores the generalization and adaptation abilities of an unsupervised learning method for finger vein recognition. The results indicate that the proposed P-CAE not only outperforms traditional baseline finger vein recognition methods but also achieves state-of-the-art performance on the UTFVP dataset with 0.24% EER.

The cross-database comparisons suggest that the P-CAE exhibits a greater resilience to the variations between the training and the evaluation datasets. Unlike its supervised counterpart, the P-CAE consistently demonstrates stable comparison performance and behaviour despite the differences between training and evaluation dataset characteristics. This observation is crucial in terms of the adaptability of deep learning methods to diverse acquisition conditions, including cross-device recognition.

This study highlights the advantage of the proposed patch-based CAE approach over a supervised counterpart in finger vein recognition, particularly in terms of the generalization and adaptability of the learned features. The outcomes showcased in this research offer promising prospects for the advancement of more robust finger vein recognition techniques in real-life scenarios, including the challenging domain of cross-device finger vein recognition.

## References

[BKM21] Bros, Victor; Kotwal, Ketan; Marcel, Sébastien: Vein enhancement with deep auto-encoders to improve finger vein recognition. In: 2021 International Conference of the Biometrics Special Interest Group (BIOSIG). IEEE, pp. 1–5, 2021.

[BM92] Besl, Paul J; McKay, Neil D: Method for registration of 3-D shapes. Sensor fusion IV: Control paradigms and data structures. Int. Soc. Opt. Photonics, 1611:586–606, 1992.

[Ch22] Chen, Ziyan; Liu, Jiazhen; Cao, Changwen; Jin, Changlong; Kim, Hakil: FV-UPatches: Enhancing Universality in Finger Vein Recognition. arXiv preprint arXiv:2206.01061, 2022.

[Hu18] Hu, Hui; Kang, Wenxiong; Lu, Yuting; Fang, Yuxun; Liu, Hongda; Zhao, Junhong; Deng, Feiqi: FV-Net: learning a finger-vein feature representation based on a CNN. In: 2018 24th international conference on pattern recognition (ICPR). IEEE, pp. 3489–3494, 2018.

[Hu22] Huang, Junduan; Luo, Weijian; Yang, Weili; Zheng, An; Lian, Fengzhao; Kang, Wenxiong: FVT: Finger vein transformer for authentication. IEEE Transactions on Instrumentation and Measurement, 71:1–13, 2022.

[KMC21] Kuzu, Rıdvan Salih; Maiorana, Emanuele; Campisi, Patrizio: Loss functions for CNN-based biometric vein recognition. In: 2020 28th European Signal Processing Conference (EUSIPCO). IEEE, pp. 750–754, 2021.

[Ku20] Kuzu, Ridvan Salih; Piciucco, Emanuela; Maiorana, Emanuele; Campisi, Patrizio: On-the-fly finger-vein-based biometric recognition using deep neural networks. IEEE Transactions on information Forensics and Security, 15:2641–2654, 2020.

[MNM04] Miura, N.; Nagasaka, A.; Miyatake, T.: 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, 2004.

[MNM07] Miura, N.; Nagasaka, A.; Miyatake, T.: Extraction of finger-vein patterns using maximum curvature points in image profiles. IEICE Trans. Inf. Syst., 90(8):1185–1194, 2007.

[No21] Noh, Kyoung Jun; Choi, Jiho; Hong, Jin Seong; Park, Kang Ryoung: Finger-vein recognition using heterogeneous databases by domain adaption based on a cycle-consistent adversarial network. Sensors, 21(2):524, 2021.

[Ou22] Ou, Wei-Feng; Po, Lai-Man; Zhou, Chang; Xian, Peng-Fei; Xiong, Jing-Jing: GAN-Based Inter-Class Sample Generation for Contrastive Learning of Vein Image Representations. IEEE Transactions on Biometrics, Behavior, and Identity Science, 4(2):249–262, 2022.

[Pa23] Pan, Zaiyu; Wang, Jun; Shen, Zhengwen; Han, Shuyu: Disentangled Representation and Enhancement Network for Vein Recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2023.

[Pe13] Peking University PKU Finger Vein Database.

[Pr19] Prommegger, Bernhard; Kauba, Christof; Linortner, Michael; Uhl, Andreas: Longitudinal finger rotation?deformation detection and correction. IEEE Transactions on Biometrics, Behavior, and Identity Science, 1(2):123–138, 2019.

[Pr22] Prommegger, Bernhard; Söllinger, Dominik; Wimmer, Georg; Uhl, Andreas: Cnn based finger region segmentation for finger vein recognition. In: 2022 International Workshop on Biometrics and Forensics (IWBF). IEEE, pp. 1–6, 2022.

[Qi16] Qiu, Shirong; Liu, Yaqin; Zhou, Yujia; Huang, Jing; Nie, Yixiao: Finger-vein recognition based on dual-sliding window localization and pseudo-elliptical transformer. Expert Systems with Applications, 64:618–632, 2016.

[Sy17] Syarif, Munalih Ahmad; Ong, Thian Song; Teoh, Andrew BJ; Tee, Connie: Enhanced maximum curvature descriptors for finger vein verification. Multimedia Tools and Applications, 76:6859–6887, 2017.

[Ta19] Tang, Su; Zhou, Shan; Kang, Wenxiong; Wu, Qiuxia; Deng, Feiqi: Finger vein verification using a Siamese CNN. IET biometrics, 8(5):306–315, 2019.

[TV13] Ton, Bram T; Veldhuis, Raymond NJ: A high quality finger vascular pattern dataset collected using a custom designed capturing device. In: 2013 International conference on biometrics (ICB). IEEE, pp. 1–5, 2013.

[Wa19] Wang, Xian; Wang, Huabin; He, Ying; Ding, Yijun; Tao, Liang: Novel algorithm for finger vein recognition based on inception-resnet module. In: Eleventh international conference on digital image processing (ICDIP 2019). volume 11179. SPIE, pp. 367–375, 2019.

[Ya19] Yang, Wenming; Hui, Changqing; Chen, Zhiquan; Xue, Jing-Hao; Liao, Qingmin: FV-GAN: Finger vein representation using generative adversarial networks. IEEE Transactions on Information Forensics and Security, 14(9):2512–2524, 2019.

[YLS11] Yin, Yilong; Liu, Lili; Sun, Xiwei: SDUMLA-HMT: A multimodal biometric database. In: Biometric Recognition: 6th Chinese Conference, CCBR 2011, Beijing, China, December 3-4, 2011. Proceedings 6. Springer, pp. 260–268, 2011.

[Ze19] Zeng, Junying; Chen, Yao; Qin, Chuanbo; Wang, Fan; Gan, Junying; Zhai, Yikui; Zhu, Boyuan: A novel method for finger vein recognition. In: Biometric Recognition: 14th Chinese Conference, CCBR 2019, Zhuzhou, China, October 12–13, 2019, Proceedings 14. Springer, pp. 46–54, 2019.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# Utility prediction performance of finger image quality assessment software

Olaf Henniger[1]

**Abstract:** A biometric sample is the more utile for biometric recognition the greater the distance between the sample-specific non-mated and mated comparison score distributions. Finger image quality scores turn out to be only weakly correlated with the observed utility. This is worth investigating because finger image quality assessment software is widely used to predict the biometric utility of finger images in many public-sector applications. This paper shows that a weak correlation between predicted and observed utility does not matter if the quality scores are used to decide whether to discard or retain biometric samples for further processing. The important point is that useful samples are not mistakenly discarded or less useful samples are not mistakenly retained. This can be measured by quality-assessment false positive and false negative rates. In cost-benefit analyses, these metrics can be used to chose suitable quality-score thresholds for the use cases at hand.

## 1 Motivation

Several finger image quality assessment algorithms have been developed, e.g., the NIST Fingerprint Image Quality (NFIQ) software version 1 [TWW04], NFIQ version 2 [Ta21] and Minutia Detection Confidence (MiDeCon) [Te21]. For many public-sector applications such as border control, enrolment for biometric identity documents, alien register enrolment, general identification scenarios, NFIQ 2 is required to be used for assessing the quality of finger images [BSI].

NFIQ version 1 [TWW04] distinguished five levels of quality: 1 (excellent), 2 (very good), 3 (good), 4 (fair), and 5 (poor). The improved version 2 of NFIQ [Ta21] provides a higher resolution of quality scores in the range from 0 to 100 (lowest to highest quality, respectively). NFIQ 2 is an open-source reference implementation of ISO/IEC 29794-4 [ISO17].

Either NFIQ version computes several global and local features from a finger image to derive a unified quality score predictive of the sample's utility for automated biometric recognition. For NFIQ 2, a random decision forest was trained for binary classification into two utility classes (high or low utility). The trained random decision forest outputs the probability that an image belongs to the high-utility class multiplied by 100 and rounded to the nearest integer [Ta21, ISO17].

NFIQ 2 was trained on finger images captured using optical sensors based on frustrated total internal reflection and finger images scanned from inked fingerprint cards, all with a

[1] Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany,
olaf.henniger@igd.fraunhofer.de

spatial sampling rate of 500 dpi (i.e., 196,85 pixels per centimetre). Though the application of NFIQ 2 to other kinds of finger images than it was trained for is not recommended, some papers report NFIQ 2 scores for deviate finger images. For finger images from other sensor types or with a different spatial sampling rate, the NFIQ 2 score is hardly predictive of utility. For instance, a recent study confirming the persistence of fingerprints over time [KHB21] noted that the NFIQ 2 scores of 508-dpi finger images from a capacitive sensor are not predictive of the utility of these samples.

The quality score of a biometric sample can be used, e.g., for deciding whether re-acquisition of a biometric sample is necessary. The quality-score threshold for discarding low-quality samples depends on the intended use of the retained samples. For instance, for public-sector applications in Germany, a technical guideline specifies that the NFIQ 2 score threshold for plain left and right index-finger images captured for enrolment purposes is 30 [BSI].

The contribution of this paper is to study metrics summarizing the prediction performance of biometric sample quality assessment algorithms: quality-assessment false positive rate and quality-assessment false negative rate.

## 2 Experimental setup

### 2.1 Finger image data set

When testing biometric sample quality assessment tools, large biometric data sets containing more than one sample per subject and covering wide ranges of potential quality issues are an advantage. Due to quality assurance measures during data collection, finger image data sets including low-quality images are hard to find.

We chose a subset consisting of 2 500 images of 500 different right index fingers, five images per finger instance, taken from the CASIA Fingerprint Image Database Version 5.0 (CASIA-FingerprintV5) [CAS09]. To generate noticeable intra-class variation within a single session, the test subjects rotated their fingers and used various levels of pressure. The plain finger images of size $328 \times 356$ pixels were captured using an optical fingerprint sensor (URU4000) with a spatial sampling rate of 512 dpi (201,57 pixels per centimetre). To enable use of NFIQ 2, we re-sampled all finger images with a spatial sampling rate of 500 dpi (196,85 pixels per centimetre), using the Cognaxon WSQ viewer 4.1.

### 2.2 Biometric comparisons

Minutiae were extracted using the open-source FingerNet framework [Ta17], which deploys a deep neural network for minutia detection. No failures to extract occurred. Minutia positions and angles were converted to MCC (Minutia Cylinder-Code) format for comparison [CFM10]. Then, each finger minutiae template was compared with each mated finger minutiae template and with 1 245 non-mated finger minutiae templates. For fingerprints,

no canonical representation is defined. Hence, any available sample is used as reference. Each comparison yielded a dissimilarity score, which is the lower the more similar the two finger images are.

### 2.3 Utility prediction method

We used NFIQ version 2.2 to assess the finger image quality in the data set re-sampled at 500 dpi. The NFIQ 2 score of a 500 dpi finger image is meant to predict the utility of this image for automated recognition.

### 2.4 Utility assessment method

A biometric sample's actual utility for automated biometric recognition can be assessed by comparing it with mated and non-mated samples from a biometric data set. This paper uses the utility assessment method described in [HFC22]. The utility of a biometric sample $i$ for a comparison algorithm that outputs dissimilarity scores is measured as normalized difference between the mean of $i$'s non-mated dissimilarity scores and the mean of $i$'s mated dissimilarity scores:

$$u_i = \frac{\mu_{\mathrm{n}i} - \mu_{\mathrm{m}i}}{\sqrt{\sigma_{\mathrm{n}}^2 + \sigma_{\mathrm{m}}^2}} \tag{1}$$

where $\mu_{\mathrm{n}i}$ is the arithmetic mean of the dissimilarity scores for $i$ and non-mated references; $\mu_{\mathrm{m}i}$ is the arithmetic mean of the dissimilarity scores for $i$ and mated references; $\sigma_{\mathrm{n}}$ is the standard probe deviation of all non-mated dissimilarity scores; $\sigma_{\mathrm{m}}$ is the standard probe deviation of all mated dissimilarity scores. Comparing with an as large as possible number of other samples ensures that $u_i$ is dominated by the influence of the sample to be assessed and not by individual other samples.

To map $u_i$ (Eq. 1) to the range from 0 to 100 required by [ISO16], the following sigmoid function having an S-shaped curve is used [HFC22]:

$$u_i^* = \frac{100}{1 + 3^{1 - \frac{u_i}{u_A}}} \tag{2}$$

$$u_A = \min(\{u_i \mid i \in A\}) \tag{3}$$

is the lowest of the $u_i$ values of the samples in the set $A$ of unobjectionable samples. A sample belongs into the set $A$ of unobjectionable samples if and only if all its mated dissimilarity scores are less than any sample's non-mated dissimilarity scores. Eq. 2 yields $S(0) = 25$ and $S(u_A) = 50$. $S(0) = 25$ means that clearly deficient samples with $u_i \leq 0$,

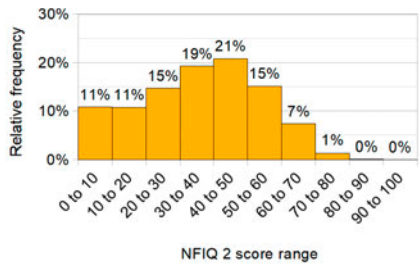
which on average appear more similar to non-mated samples than to mated ones, are assigned utility scores from 0 to 25. $S(u_A) = 50$ means that unobjectionable (i.e., adequate or even excellent) samples are assigned utility scores from 50 to 100.

To fit into the standardized biometric data interchange formats, quality scores must be quantized to integers [ISO16], which is achieved by rounding to the nearest integer.

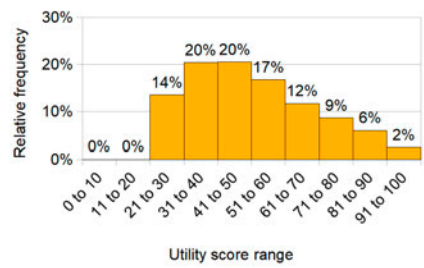
## 3 Experimental results

### 3.1 Predicted vs. observed utility scores

Fig. 1 shows the distributions of predicted and observed utility scores. Fig. 1(a) shows the NFIQ 2 score distribution for the finger images in the chosen data set. We calculated utility scores $u_i^*$ (Eq. 2) using the comparison scores from the finger image data set and comparison algorithm given in Sec. 2. Fig. 1(b) shows the distribution of utility scores of the finger images in the studied data set. Note that a great deal of the finger images is, as for the chosen open-source comparison software, of objectionable quality ($u_i^* < 50$). The presence of low-utility images makes the chosen data set especially suitable for studying finger image quality assessment.




(a) NFIQ 2 scores




(b) Observed utility scores

Fig. 1: Predicted and observed utility score distributions

A common method for characterizing a prediction model's performance is to use the root mean square error (RMSE) between observed and predicted values. The observed utility scores for the data set were calculated according to Eq. 2. The RMSE for the finger images from the data set was 26.8.

Scatterplots allow identifying the type of relationship (if any) between two quantities. Fig. 2 shows a scatterplot between NFIQ 2 scores and utility scores (Eq. 2). Apparently, the scores are rather weakly correlated.

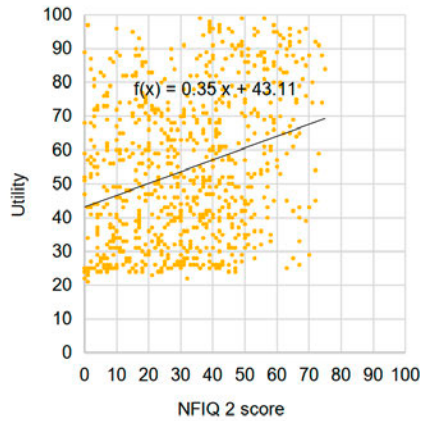
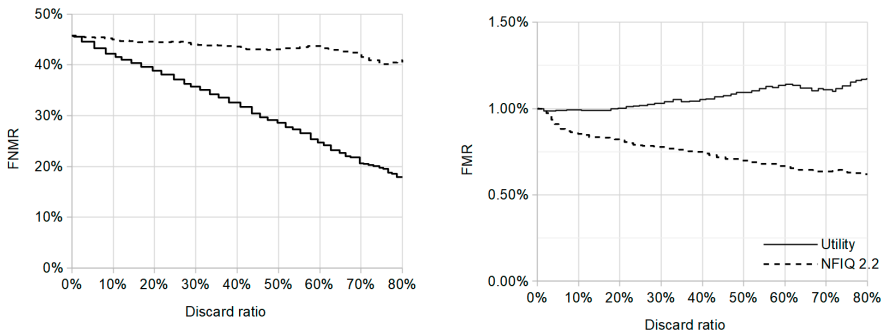



Fig. 2: NFIQ 2 vs. utility score scatterplot

### 3.2 Error vs. discard characteristics

False non-match error vs. discard characteristics (EDC), also known as error vs. reject characteristics [GT07], show the dependence of the false non-match rate (FNMR) at a fixed decision threshold on the percentage of reference and probe images discarded based on lowest quality scores. Fig. 3(a) shows the false non-match EDCs with respect to the a-posteriori utility scores from Eq. 2 and with respect to NFIQ 2 scores. The false non-match EDCs vary for different decision threshold values. The decision threshold was fixed to give an initial FMR value of 1 %. Fig. 3(a) shows that discarding images with low NFIQ 2 scores from the data set lead to a decline in FNMR as desired. As not only false non-match but also false match errors happen, a false non-match EDC always needs to be considered together with the corresponding false match EDC. A decrease in FNMR may




(a) False non-match errors

(b) False match errors

Fig. 3: Error vs. discard characteristics at a fixed decision threshold for an initial FMR of 1 %

inadvertently come along with an increase in FMR. Fig. 3(b) shows the false match EDCs with respect to the utility scores from Eq. 2 and to NFIQ 2 scores.

### 3.3 $d'$ vs. discard characteristics

To summarize the utility-prediction performance in a single plot, we use $d'$ vs. discard characteristics showing the dependence of

$$d' = \frac{\mu_n - \mu_m}{\sqrt{\sigma_n^2 + \sigma_m^2}} \tag{4}$$

on the percentage of reference and probe samples discarded based on lowest quality scores (discard ratio) [HFC22]. The better the utility prediction works, the steeper $d'$ (the distance between the mated and non-mated comparison score distributions) increases with increasing discard ratio.

Fig. 4 shows the $d'$ vs. discard characteristics [HFC22] with respect to the utility scores from Eq. 2 and with respect to NFIQ 2 scores. The $d'$ vs. discard characteristics show that exclusion of images with low quality scores leads to a noticeable improvement in the separability of the mated and non-mated comparison score distributions.

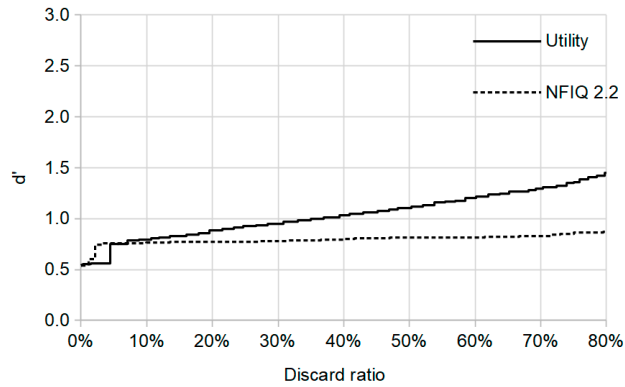



Fig. 4: $d'$ vs. discard characteristics

### 3.4 Quality-assessment error rates

If the quality scores are used to make decisions on whether to discard or to retain biometric samples for further processing, then a weak correlation between predicted and observed utility does not matter. The important point is to avoid mistakenly discarding useful samples or retaining less useful ones. To express this, we define the following two metrics:

- The quality-assessment false negative rate (QFNR) is the proportion of biometric samples the quality scores of which are lower than or equal to a quality-score threshold $u$, but the utility scores of which are greater than 50.
- The quality-assessment false positive rate (QFPR) is the proportion of biometric samples the quality scores of which are greater than a quality-score threshold $u$, but the utility scores of which are lower or equal to 50.

Similar metrics (named incorrect sample rejection rate and incorrect sample acceptance rate) are defined in [Gr22], however, taking only a single mated comparison score into account. QFNR and QFPR both depend on the chosen data set and comparison algorithm and on the quality-score threshold $u$ and should be quoted together at the same quality-score threshold. The higher the quality-score threshold, the more samples are discarded, including samples of unobjectionable quality that had better been retained. The lower the quality-score threshold, the more samples are retained, including samples of objectionable quality that had better been discarded.

Fig. 5 shows QFNR and QFPR for the chosen data set and comparison algorithm over the NFIQ 2 score threshold. At the quality-score threshold of 30 for plain index-finger images captured for enrolment purposes [BSI], about 32 % of the retained finger images are of objectionable (low) quality, and about 14 % of the discarded finger images are of unobjectionable (high) quality. For data sets containing fewer objectionable (low-quality) finger images from the beginning, QFPR would be lower.

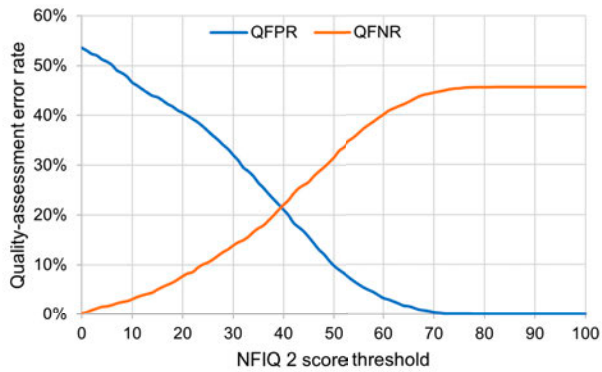



Fig. 5: Quality-assessment error rates over the NFIQ 2 score threshold

The quality-score threshold value of 30 is the result of a cost-benefit analysis that takes into account the costs of the different types of errors, i.e., of mistakenly discarding unobjectionable samples and of mistakenly retaining objectionable samples.

## 4 Conclusions

If the quality assessment algorithm is used to make decisions on whether to discard or to retain a biometric sample for further processing, QFNR (the proportion of unobjectionable samples mistakenly discarded) and QFPR (the proportion of objectionable samples mistakenly retained) are useful metrics. As long as the quality scores of good samples are greater than a chosen quality-score threshold, and the quality scores of bad samples are less than the quality-score threshold, the exact magnitude of deviations between predicted and observed utility scores do not matter.

Note that the utility of a biometric sample depends on the comparison algorithm and on the data set used for comparison. It would be interesting to repeat the experiments with commercial fingerprint comparison algorithms and larger finger image data sets captured on a best effort basis.

## 5 Acknowledgments

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. Portions of the research in this paper use the CASIA Fingerprint Image Database Version 5.0 collected by the Chinese Academy of Sciences' Institute of Automation (CASIA).

## References

[BSI] Biometrics for public sector applications – Part 3: Application profiles, function modules and processes. BSI Technical Guideline TR-03121-3. 1, 2, 7

[CAS09] CASIA Fingerprint Image Database (CASIA-FingerprintV5). Available at http://biometrics.idealtest.org/, 2009. 2

[CFM10] Cappelli, R.; Ferrara, M.; Maltoni, D.: Minutia cylinder-code: A new representation and matching technique for fingerprint recognition. IEEE TPAMI, 32(12), 2010. 2

[Gr22] Grother, P.; Hom, A.; Ngan, M.; Hanaoka, K.: Ongoing Face Recognition Vendor Test (FRVT) – Part 5: Face Image Quality Assessment. Draft NIST Interagency Report, NIST, 2022. Retrieved from https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing. 7

[GT07] Grother, P.; Tabassi, E.: Performance of biometric quality measures. IEEE TPAMI, 29(4), 2007. 5

[HFC22] Henniger, O.; Fu, B.; Chen, C.: Utility-based performance evaluation of biometric sample quality assessment algorithms. In: Proc. of the Int. Conf. of the Biometrics Special Interest Group BIOSIG. 2022. 3, 6

[ISO16] Information technology – Biometric sample quality – Part 1: Framework. International Standard ISO/IEC 29794-1, 2016. 3, 4

[ISO17] Information technology – Biometric sample quality – Part 4: Finger image data. International Standard ISO/IEC 29794-4, 2017. 1

[KHB21] Kessler, R.; Henniger, O.; Busch, C.: Fingerprints, forever young? In: Proc. of the Int. Conf. on Pattern Recognition ICPR. 2021. 2

[Ta17] Tang, Y.; Gao, F.; Feng, J.; Liu, Y.: FingerNet: A unified deep network for fingerprint minutiae extraction. In: Proc. of the Int. Joint Conf. on Biometrics (IJCB). 2017. 2

[Ta21] Tabassi, E.; Olsen, M.; Bausinger, O.; Busch, C.; Figlarz, A.; Fiumara, G.; Henniger, O.; Merkle, J.; Ruhland, T.; Schiel, C.; Schwaiger, M.: NFIQ 2.0 – NIST Fingerprint Image Quality. NIST Interagency Report 8382, NIST, 2021. 1

[Te21] Terhörst, P.; Boller, A.; Damer, N.; Kirchbuchner, F.; Kuijper, A.: MiDeCon: Unsupervised and accurate fingerprint and minutia quality assessment based on minutia detection confidence. In: Proc. of the Int. Joint Conf. on Biometrics (IJCB). 2021. 1

[TWW04] Tabassi, E.; Wilson, C.L.; Watson, C.I.: Fingerprint image quality. NIST Interagency Report 7151, NIST, 2004. 1

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# Statistical Methods for Testing Equity of False Non Match Rates across Multiple Demographic Groups[1]

Michael Schuckers[2] Kaniz Fatima[3] Sandip Purnapatra[4] Joseph Drahos[5] Daqing Hou[6]
Stephanie Schuckers[7]

**Abstract:** Biometric recognition is used for a variety of applications including authentication, identity proofing, and border security. One recent focus of research and development has been methods to ensure fairness across demographic groups and metrics to evaluate fairness. However, there has been little work in this area incorporating statistical variation. This is important because differences among groups can be found by chance when no difference is present or may be due to an actual difference in system performance. We extend previous work to consider when individuals are members of one or more demographics (age, gender, race). Our methodology is meant to be more comprehendable by a non-technical audience and uses a robust bootstrap approach for estimation of variation in false non-match rates. After presenting our methodology, we present a simulation study and we apply our approach to MORPH-II data.

**Keywords:** Fairness, Confidence Intervals, Demographics, Multiple Comparisons

## 1 Introduction

There has been significant attention to face recognition and artificial intelligence as a whole as it relates to equity. For example, the U.S. Federal Trade Commission released guidance on AI fairness, highlighting that "[i]t's essential to test your algorithm [for discrimination] based on race, gender, or other protected classes" [Ji21]. In a review of face recognition literature, demographic factors may have a significant influence on the performance of some biometric recognition algorithms, resulting in a lower biometric performance for demographic groups, such as females, dark-skinned, and/or youngest subjects [Dr20]. Research has shown that results differ depending on the specific algorithms, capture conditions, use cases, and a host of additional factors [HSV19, GZ19, Go21, We22, Yu22, CKG23] .

This paper develops statistical methods for determining if there are statistically distinguishable false non-match rates (FNMR's) simultaneously across multiple demographics each having more than one category. These methods are aimed at non-technical audience, such as policymakers, rather than the complicated analysis of variance and p-value approaches taken for similar circumstances by [Sc10] which can be problematic [WL16]. Building upon the concept of margins of error which are widely known in the public, we derive methods usable for each demographic group or for all demographic groups simultaneously. Specifically, we extend the work of [Sc22] who considered the case where all the

[1] This material is based upon work supported by the Center for Identification Technology Research and the National Science Foundation under Grant 1650503

[2] Math, CS & Stats, St. Lawrence University, Canton, NY, USA, schuckers@stlawu.edu
[3] Computer and Electrical Engineering, Clarkson University, Potsdam, NY, USA, fatimak@clarkson.edu
[4] Computer and Electrical Engineering, Clarkson University, Potsdam, NY, USA, purnaps@clarkson.edu
[5] Computer and Electrical Engineering, Clarkson University, Potsdam, NY, USA, drahosj@clarkson.edu
[6] Computer and Electrical Engineering, Clarkson University, Potsdam, NY, USA, dhou@clarkson.edu
[7] Computer and Electrical Engineering, Clarkson University, Potsdam, NY, USA, sschucke@clarkson.edu

demographic categories were non-overlapping. demographic groups. Here, we consider the case where individuals are members of multiple groups or categories in several demographics and we will refer to demographics as different dimensions while categories will be the values that each of those demographics can take. For example, our methods apply for simultaneously comparing individuals across racial, educational, and age demographics where each individual is classified into one category within each of those groups. In that instance for the demographic Age, an individual might be in the '25 to 40' category. Additionally, for practitioner flexibility, we present methods for both creating a single margin of error for all demographic groups or simultaneously creating intervals for each demographic separately. For the purposes of this paper, we think of fairness as meaning that the FNMR's are not statistically different across one or more demographic categories and we are motivated by an access application.

## 2 Related Work

Metrics for the assessment of fairness have been proposed in the literature. [dFPM22] introduce the Fairness Discrepancy Rate (FDR) which is a summary of system performance accounting for both FNMR and FMR. Their approach uses a "relaxation constant" rather than trying to assess the sampling variation or statistical variation between FNMR's from different demographic groups. Howard et al. present an evaluation of FDR noting its scaling problem. To address this scaling problem, the authors propose a new fairness measure called Gini Aggregation Rate for Biometric Equitability (GARBE) [Ho22]. NIST scientists also propose the Inequality Rate (IR) metric [Gr21]. In addition, the ISO/IEC working draft 19795-10 [IS23] proposes several metrics for demographic performance differentials, including the error rate ratio in case of two groups, and the worst case error rate relative to the geometric mean in case of three or more groups.

While there are several metrics of fairness, there has been little research or use of statistical methods for fairness metrics. The United States National Institute for Standards and Technology (NIST) has performed the most extensive evaluation of biometric recognition as part of a technology evaluation [GNH19]. Results are continually updated at [NI]. Commercial software biometric algorithms are submitted to NIST for testing. Evaluation is performed across a variety of datasets including border, visa application, and mugshot images and for both identification (1:N) and verification (1:1). Performance is reported in terms of FNMR and FMR for verification and FNIR and FPIR for identification. Bootstrapping is provided as a measure of variability and presented throughout their analysis enabling the reader to assess differences, if any, in the context of its statistical variability. Some of the earliest work on the impact of demographics on biometric matching performance was done by [Gi04, Be08, Be09]. More recently, [Co19] look at the impact of demographics on facial recognition.

Bhatt et al. [Bh23] documented and explained the causal understanding of the gender gap problem in the popular deep learning-based facial recognition techniques. The authors claimed the gender gap problem is caused by the imbalance of the test dataset rather than the training set and sorting the images based on hairstyle can reduce the gender gap margin significantly. Other research has also performed extensive evaluations of face recognition

across demographic groups, e.g. [Zh17, Co19, Bu17, GNH19, Kr20, Gr21, Pa22, Te20, Yu22], but have not presented statistical fairness evaluation methods as part of their work.

A definitive methodology for statistical hypothesis testing of the equality of biometric error rates was given in [Sc10]. That approach used resampling methodology to create analysis of variance-like tests for comparing FNMR rates across groups equivalent to a single demographic here. As mentioned above, [Sc22] derived a statistical margin of error via bootstrapping for determining which, if any, FNMR's were different from the rest. However, that paper did not address the practical case when testing across multiple groups simultaneously. In this paper, we generalize their approach to handle the more general and more realistic case when individuals are classified into categories in one or more demographics. One obvious application of this work is the determination of fairness or statistically equal false non-match rates across demographic categories.

## 3 Methodology

The methods proposed here are motivated by an application where biometric devices are tested across multiple demographics and where each individual is classified into categories separately within each demographic. The aim here is to determine if any of the FNMR's from the categories within demographics are statistically different from the overall FNMR assuming a fixed decision threshold for all categories. Below we will provide methods for that determination within a single demographic or across all of the demographics. The techniques here are useful for assessing the equity of performance across demographics.

Our flexible approach is to bootstrap individuals across groups to obtain an understanding of the variation of the error rates in each category and use that variation to build a distribution of the maximal variation for the overall error rate. For our resampling, we follow the bootstrap methodology for FNMR of [Sc10]. Having obtained a reference distribution of the maximal variation, we then create intervals to determine if there are groups that are statistically different. It is important to note that this approach requires no distributional assumptions about the data. Here we present methods for both additive intervals and multiplicative intervals.

Denote the number of demographics by $D$ and let $G_d$ be the number of categories within each demographic $d$ where $d = 1,\ldots,D$ and $k = 1,\ldots,G_d$. Let $\pi$ represent a population FNMR and $\hat{\pi}$ represent the estimated FNMR from our sample. The estimated FNMR for category $k$ within demographic $d$ will be denoted by $\hat{\pi}_{dk}$. This is calculated by the total number of false non-matches divided by the total number of attempts of individuals in that category. The number of false non-matches for individual $i$ will be denoted by $y_i$ for $i = 1,2,\ldots,n$. We allow for a different number of attempts per individual which we denote by $m_i$ for individual $i$. For a multiplicative interval, our equation for the weighted geometric mean FNMR is $\dot{\pi} = (\prod_d \prod_k \hat{\pi}_{dk}^{n_{dk}})^{1/(\sum_d \sum_k n_{dk})}$ where $n_{dk}$ is the number of individuals in category $k$ of demographic $d$.

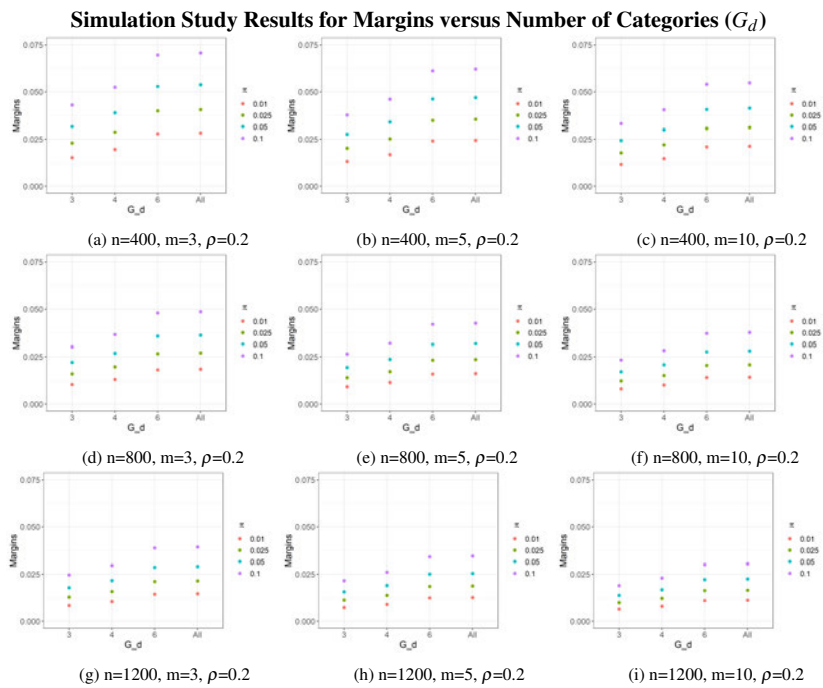
Here we propose two types of inferential intervals: additive and multiplicative. Additive intervals are the most commonly used in practice and involve an estimate plus or minus some margin of error (M). Multiplicative intervals are less common but involve ratios

and an estimate multiplied and divided by a ratio of error (R). We incorporate the latter approach since [IS23] is considering using ratios and geometric means for evaluating the fairness of a biometric device.

Below we present four different approaches to assessing fairness: an additive approach for comparing the FNMR's for all categories with a single interval, an additive approach for comparing FNMR's with each demographic separately, a multiplicative approach for comparing the FNMR's for all categories with a single interval and a multiplicative approach for comparing FNMR's with each demographic separately. The following are the steps for our algorithm.

1. Calculate the error rate, $\hat{\pi}$ and the error rate in each category $k$ within demographic $d$, $\hat{\pi}_{dk}$. Likewise, calculate the weighted geometric mean for the entire test, $\hat{\pi}$, across the various categories $k$ and demographics $d$.
2. Sample with replacement the $n$ individuals. For the analysis below, carry along the corresponding demographic information (to which categories they belong) and the corresponding matching performance information (how many errors from how many attempts) for the selected individuals.
3. Calculate the bootstrapped category error rates. Denote them as $\hat{\pi}^b_{dk}$ for each category $k$ in each demographic $d$.
4. Next calculate and store $\phi = \max_{dk} |\hat{\pi}^b_{dk} - \hat{\pi}_{dk}|$, $\phi_d = \max_k |\hat{\pi}^b_{dk} - \hat{\pi}_{dk}|$, $\psi = \max_{dk} \left(\hat{\pi}^b_{dk}/\hat{\pi}_{dk}, \hat{\pi}_{dk}/\hat{\pi}^b_{dk}\right)$, or $\psi_d = \max_k \left(\hat{\pi}^b_{dk}/\hat{\pi}_{dk}, \hat{\pi}_{dk}/\hat{\pi}^b_{dk}\right)$.
5. Repeat the previous three steps some large number of times, say B times.
6. Let $M$ be the $1-\alpha/2^{th}$ percentile of the distribution of $\phi$, let $M_d$ be the $1-\alpha/2^{th}$ percentile of the distribution of $\phi_d$, let $R$ be the $1-\alpha/2^{th}$ percentile of the distribution of $\psi$, and let $R_d$ be the $1-\alpha/2^{th}$ percentile of the distribution of $\psi_d$.
7. Having obtained values for $M$, $M_d$, $R$ or $R_d$ we can create additive intervals for each $\pi_{dk}$ using $\hat{\pi} \pm M$ and $\hat{\pi}_d \pm M_d$, respectively, as well as multiplicative intervals for $\pi$ and each $\pi_{dk}$ using $\hat{\pi}R^{\pm 1}$ and $\hat{\pi}R_d^{\pm 1}$, respectively.

From the intervals derived in the last step of the above algorithm, we can use them to determine which, if any, groups have error rates that differ from the rest. Outstanding category FNMR's, $\hat{\pi}_{dk}$'s, will lie outside of the intervals calculated via the algorithm above. One use for this approach is to look at the equity of FNMR's across all of the demographics and determining which categories have FNMR's outside of the obtained intervals. *A priori* practitioners should decide if they are interested in differences across all demographic groups (using $M$ or $R$) or in differences within each demographic group (using $M_d$ or $R_d$ for each $d$). Only one approach should be used since it is possible that there may be differences in the outcomes between the approaches and using multiple approaches induces issues with the familywise confidence level of the interval.




(a) n=400, m=3, $\rho$=0.2 (b) n=400, m=5, $\rho$=0.2 (c) n=400, m=10, $\rho$=0.2

(d) n=800, m=3, $\rho$=0.2 (e) n=800, m=5, $\rho$=0.2 (f) n=800, m=10, $\rho$=0.2

(g) n=1200, m=3, $\rho$=0.2 (h) n=1200, m=5, $\rho$=0.2 (i) n=1200, m=10, $\rho$=0.2

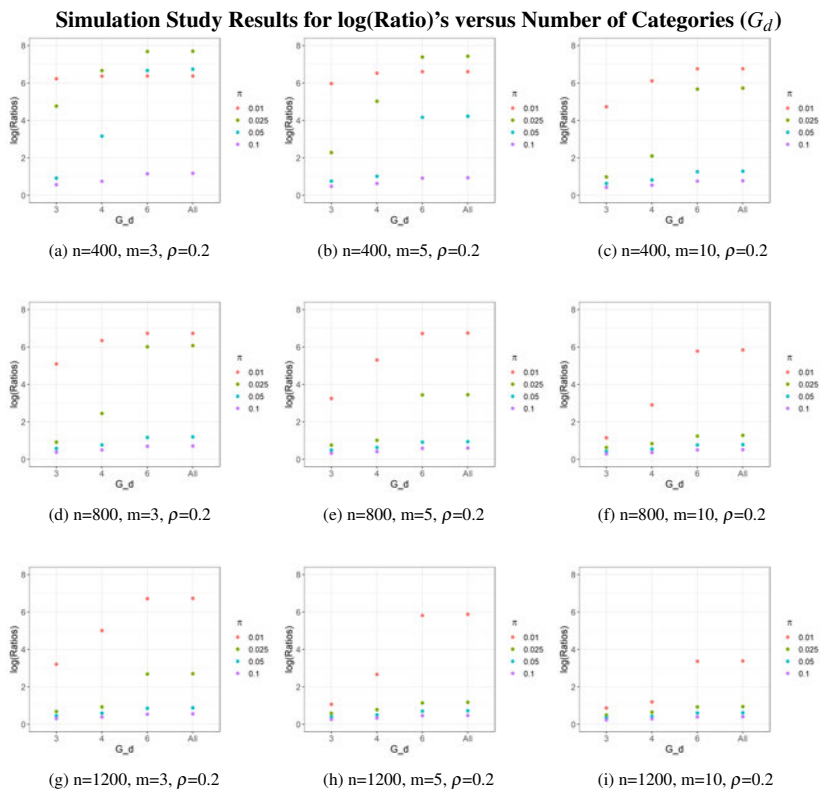
Fig. 1: Results of a simulation study for margins of error as a function of the number of individuals (n), number of attempts (m), the correlation between attempts ($\rho$), and FNMR ($\pi$). Subfigures are organized by columns where $m$ increases from left to right and by rows where $n$ increases from top to bottom. Each figure plots $M$ versus $G$ for fixed $\rho$=0.2 and with different values for $\pi$ denoted by color.

## 4 Simulation Study

To explicate our methodology, we present a simulation study to understand how these performances will differ for different size demographic groups, for different overall error rates and for sample sizes. For a combination of parameters, we generated average values of $M$, $M_d$, $R$, and $R_d$ in order to understand the impact of changes to the parameters on those quantities.

We have the following steps to our simulations having set values for the number of demographics ($D$), the number of categories in demographic $d$ ($G_d$), the False Non-Match Rate ($\pi$), the intra-individual correlation ($\rho$), the number of individuals ($n$), and the number of attempts per individual ($m$).

1. Generate $m$ attempts from $n$ individuals with an FNMR of $\pi$ and an intra-individual correlation of $\rho$.




(a) n=400, m=3, $\rho$=0.2 (b) n=400, m=5, $\rho$=0.2 (c) n=400, m=10, $\rho$=0.2

(d) n=800, m=3, $\rho$=0.2 (e) n=800, m=5, $\rho$=0.2 (f) n=800, m=10, $\rho$=0.2

(g) n=1200, m=3, $\rho$=0.2 (h) n=1200, m=5, $\rho$=0.2 (i) n=1200, m=10, $\rho$=0.2

Fig. 2: Results of simulation study for Ratios as a function of number of individuals (n), number of attempts (m), correlation between attempts ($\rho$), and FNMR ($\pi$). Subfigures are organized by columns where $m$ increases from left to right and by rows where $n$ increases from top to bottom. Each figure plots natural logarithm of $R_d$ and $R$ versus $G_d$ for fixed $\rho$=0.2 and with different values for $\pi$ denoted by color.

2. For each individual $i, i = 1,2,\ldots,n$, and each demographic $d$, $d = 1,2,\ldots,D$, randomly select a category in $\{1,2,\ldots,G_d\}$ for demographic $d$.
3. Bootstrap individuals and their corresponding performance/matching measurements and their demographic categories using the algorithm given in the previous section.
4. Find and store $M$, $M_d$ for each $d$, $R$, and $R_d$ for each $d$.
5. Repeat the previous four steps some larger number of times, say $Z = 1000$.
6. Calculate the mean value for $M$, $M_d$, $R$ and $R_d$.

For our simulation study we used $D = 3$ demographics and $G_1 = 3$, $G_2 = 4$ and $G_3 = 6$. We ran all combinations of the following values for each of the these parameters: $\pi = 0.01$,

| Number of Categories, $G_d$ | Total Subjects, $n$ | Percentiles | | | |
|---|---|---|---|---|---|
| | | 80% | 90% | 95% | 97.5% |
| 3 | 400 | 0.0128 | 0.0153 | 0.0177 | 0.0200 |
| 4 | 400 | 0.0160 | 0.0191 | 0.0221 | 0.0250 |
| 6 | 400 | 0.0221 | 0.0262 | 0.0303 | 0.0345 |
| All | 400 | 0.0228 | 0.0268 | 0.0309 | 0.0350 |
| 3 | 800 | 0.0090 | 0.0107 | 0.0123 | 0.0137 |
| 4 | 800 | 0.0112 | 0.0132 | 0.0151 | 0.0169 |
| 6 | 800 | 0.0153 | 0.0179 | 0.0204 | 0.0229 |
| All | 800 | 0.0158 | 0.0183 | 0.0207 | 0.0232 |
| 3 | 1200 | 0.0073 | 0.0087 | 0.0099 | 0.0111 |
| 4 | 1200 | 0.0091 | 0.0107 | 0.0121 | 0.0135 |
| 6 | 1200 | 0.0123 | 0.0143 | 0.0162 | 0.0180 |
| All | 1200 | 0.0127 | 0.0146 | 0.0164 | 0.0182 |

Tab. 1: Percentiles from the distribution of $M_d$'s and $M$'s with parameters $\rho$=0.2, $m$=10 and $\pi$=0.025

| Number of Categories, $G_d$ | Total Subjects, $n$ | Percentiles | | | |
|---|---|---|---|---|---|
| | | 80% | 90% | 95% | 97.5% |
| 3 | 400 | 2.00 | 2.30 | 2.66 | 3.08 |
| 4 | 400 | 2.59 | 3.21 | 8.20 | 15.22 |
| 6 | 400 | 42.06 | 128.27 | 290.14 | 536.41 |
| All | 400 | 42.46 | 130.29 | 307.77 | 557.58 |
| 3 | 800 | 1.60 | 1.74 | 1.88 | 2.03 |
| 4 | 800 | 1.86 | 2.07 | 2.30 | 2.54 |
| 6 | 800 | 2.48 | 2.93 | 3.47 | 4.11 |
| All | 800 | 2.58 | 3.04 | 3.58 | 4.23 |
| 3 | 1200 | 1.45 | 1.55 | 1.64 | 1.73 |
| 4 | 1200 | 1.63 | 1.76 | 1.90 | 2.04 |
| 6 | 1200 | 2.02 | 2.26 | 2.52 | 2.80 |
| All | 1200 | 2.07 | 2.31 | 2.57 | 2.85 |

Tab. 2: Percentiles from the distribution of $R_d$'s and $R$'s with parameters $\rho$=0.2, $m$=10 and $\pi$=0.025

0.025, 0.05, 0.10, $\rho = 0$,0.05, 0.1, 0.2, $n = 400$, 800, 1200, and $m = 1,3,5,8,10$. To generate to which category of demographic $d$ an individual belonged, we used equal probability though the methodology could easily be extended to consider non-equal probabilities. We generated $Z = 1000$ datasets for each combination of parameters to ensure that our results were statistically robust. Note that the average number of match decisions or attempts per category was $nm/G_d$ and, thus, the average number of errors was $nm\pi/G_d$. Thus, the number of observations and the number of errors per category decreased as $G_d$ increased. In these simulations, if the number of errors in a given category was zero, we used a small value, $\varepsilon = 1.5/n_{dk}$, the midpoint of a Rule of 30 interval [JL97], to ensure a well-defined values for the ratio.

Tab. 3: FNMR Statistical Summaries for MORPH-II Analysis

| | Race | | Gender | | Age | | |
|---|---|---|---|---|---|---|---|
| | Black | White | Female | Male | 17-30 | 31-45 | 45+ |
| $\sum_i m_i$ | 41964 | 9885 | 7927 | 43922 | 23837 | 18781 | 9231 |
| $n_{dk}$ | 10561 | 2599 | 2074 | 11086 | 6163 | 4657 | 2340 |
| $\hat{\pi}_{dk}$ | 0.0241 | 0.0530 | 0.0566 | 0.0247 | 0.0347 | 0.0258 | 0.0242 |

### 4.1 Results for $M$ and $M_d$

We start by considering results for $M$ and $M_d$ from our simulation study described above. Figure 1 shows the $95^{th}$ percentiles of the average error margin across sets of parameters. There the x-axis of the subfigures is the number of categories, $G_d$, except for the last category on the right which is labeled as 'All.' This category represents the values for $\phi$ which is based upon the maximal absolute value of the differences across all categories in all demographics. For the first the values on the x-axis in each subfigure, the quantity plotted is $M_d$. From each subfigure, we can see that the margin of error grows as $G_d$ increases. Moving down subfigure rows, i.e. as $n$ increases we see that $M$ and $M_d$ decrease. Similarly, going from left to right across subfigure columns, i.e. as $m$ increases we see decreases in the margins of error. Within each subfigure, we can see that $M$ becomes smaller as $\pi$ decreases. Similar results with specific values can be found in Table 1 which give specific values for the percentiles of $M_d$'s and $M$'s for value of $n$, when $\rho = 0.2$, $m = 10$ and $\pi = 0.025$.

### 4.2 Results for $R$ and $R_d$

Next, we discuss the results of our simulation study for the distributions of ratios that were generated. Figure 2 has the $95^{th}$ percentiles of the parameter combinations for $R_d$ and $R$ from our simulation study. Because of the large range of values, the y-axis is on a natural logarithmic scale. As we did above in Figure 1, Figure 2 varies $n$ along the subfigure columns and varies $m$ along the subfigure rows. This highlights one of the results of our simulation study which is the ratios generated by our simulation study were sometimes quite large. This was particularly the case when the expected number of errors per number of categories, $nm\pi/G_d$, was small. As above, as either $n$ or $m$ increased these average ratios generally decreased. Increases in $\pi$ tended to result in decreased values for $R_d$ and $R$. This pattern, increases in $\pi$, differs from the trend for additive intervals and is likely a function of the instability of ratios of small values of $\pi$. Table 2 presents the average percentiles for $R_d$'s and $R$'s for three values of $n$, when $\rho = 0.2$, $m = 10$ and $\pi = 0.025$. Here the same pattern of results as in Figure 2 and the impact of small errors on these ratios is clear as $G_d$ increases when $n = 400$.

## 5 Illustration using MORPH-II Data

In this section, we apply our methodology to data from the MORPH-II dataset. The MORPH-II dataset is a longitudinal dataset consisting of mugshots images selected from repeat

offenders, taken over the course of 5 years. For our analysis of the MORPH-II mugshot dataset, we used a Resnet50 face recognition model pre-trained on the VGGFace2 dataset from an open-source code repository [He15, Ca18]. Using this model, we extracted the 512-dimensional embeddings from each sample within the dataset. Then we performed comparisons within each individual and computed FNMR. The comparison score was computed using the cosine similarity between two sample embeddings. We computed every permutation of genuine comparisons for each individual. For this analysis because of sample size considerations, we considered (D=3) three demographics: race, gender, and age. Race had two categories (black and white), gender had two categories (female and male), and age had three categories (young adults [17-30], middle-aged adults [30-45], and old-aged adults [45+]). The data analyzed for this project are from 13160 individuals resulting 51844 intra-individual comparisons. Table 3 has the summary for all categories across the various demographics. The total number of attempts per category, $\sum_i m_i$, is given by the first row. The second and third rows have the number of individuals, $n_{dk}$, and the FNMR, $\hat{\pi}_{dk}$, for demographic $d$ and category $k$, respectively.

For this application, we set the False Match Rate to 0.10 and had an overall FNMR of $\hat{\pi} = 0.0296$ for all individuals and a weighted geometric mean of $\dot{\pi} = 0.0285$. As expected there is variation between the categories in the FNMR's. We applied our methods above to determine if those differences were statistically discernible. For this bootstrap, we did 5000 replications of the data and results for $95^{th}$ and $97.5^{th}$ percentiles can be found in Table 4.

If we want to have a single additive interval for all categories, we should start with the first row, $M$, and an 95% confidence rate would give an range of $\hat{\pi} \pm M = 0.0296 \pm 0.0094 = (0.0202, 0.0390)$. From this we would conclude that any category that fall outside this interval would be statistically different from the overall FNMR. In this case, that would mean that the FNMR's for Whites and Females were statistically larger than the FNMR for all groups. Likewise if we were using a multiplicative interval for all categories, we would find the appropriate interval by taking $\dot{\pi} \cdot R^{\pm 1} = 0.0285(1.252)^{\pm 1} = (0.0228, 0.0357)$. As above, our conclusions would be that the FNMR's for Whites and Females are larger than the overall FNMR.

Tab. 4: Bootstrap Percentiles for FNMR Intervals

| | | $95^{th}$ | $97.5^{th}$ | | $95^{th}$ | $97.5^{th}$ |
|---|---|---|---|---|---|---|
| All | $M$ | 0.0082 | 0.0094 | $R$ | 1.221 | 1.252 |
| Race | $M_1$ | 0.0068 | 0.0079 | $R_1$ | 1.141 | 1.161 |
| Gender | $M_2$ | 0.0078 | 0.0090 | $R_2$ | 1.152 | 1.170 |
| Age | $M_3$ | 0.0049 | 0.0055 | $R_3$ | 1.121 | 1.247 |

It is conceivable that the focus of an analysis will be on one specific demographic rather than across all demographics. In that case, the appropriate tool would be the intervals based upon the appropriate demographic. For example, if for the MORPH-II data we are solely interested in Gender, then we would make an additive interval via $\hat{\pi} \pm M_2$. So that a 90% interval would be $0.0296 \pm 0.0078 = (0.0218, 0.0374)$ and we would conclude that Females were discernibly different from average. A similarly constructed 90% multiplicative

interval for Age, $0.0285(1.121)^{\pm 1} = (0.0254, 0.0319)$ would find that individuals aged 17 to 30 had a detectably higher FNMR.

## 6 Discussion

Equity and fairness in biometrics are important issues. The declaration of differences between demographic groups is a consequential one. Such conclusions about differences between groups need to be statistically sound and recognize the presence of sampling variation. In this paper, we have proposed interpretable methods for the determination of statistical differences in FNMR's in categories across multiple demographics based upon bootstrapping biometric match data. The first approach is an additive bootstrap-based one that extends previous work and deals with the dependence on the FNMR when individuals are classified in categories across multiple demographics. The second approach is similar to the first but uses a multiplicative methodology from ratios in order to generate ranges of values that are statistically similar. Both approaches yield intervals based on the sampling variation in the relevant metrics and can be used for the identification of demographic categories with FNMR's that are statistically discernible. Our resampling-based approach is focused on creating a simple interval that can be explained to a broad audience.

For the application here and the simulation study we have described above, each individual appeared in only one category for each demographic. However, the methodology is flexible enough to support the case where an individual is in (or selects) multiple categories within a demographic.

The simulation study illustrated that ratio-based confidence intervals are less stable than additive confidence intervals when the expected number of errors is small. This instability in the ratios is more pronounced as the overall error rate decreases.

As with any statistical intervals, the choice of $1-\alpha$, the confidence level, is important. Here the intervals chosen are derived to define family-wise error rates and control the effects of multiplicity. While we prefer the use of a single interval across all demographics for ease of interpretation, we have provided methods for the creation of demographic specific intervals.

To illustrate the utility of our methodology, we have applied our approach to MORPH-II data. In this application, we note that percentiles for $M$ and $R$ are larger than values for any of the $M_d$ and $R_d$, respectively. This is expected since the former quantities account for variation across all demographics, rather than across a single set of demographic categories. Additionally, we see that the variation within a demographic depends upon the error rate within each category and upon the group with the smallest number of match decisions. This is because biometric error rates are inherently binary. The application of our bootstrap methodology to the MORPH-II data took less than one minute to complete using the `R` programming language on a standard laptop.

The focus of this paper has been on false non-match rates since we are motivated in fairness in access but it is possible to extend the work here to false match rates though the variance structure of false match rates requires a more complicated bootstrap resampling structure, see Schuckers [Sc10].

## References


[Be08] Beveridge, J. Ross; Givens, Geof H.; Phillips, P. Jonathon; Draper, Bruce A.; Lui, Yui Man: Focus on quality, predicting FRVT 2006 performance. In: 2008 8th IEEE International Conference on Automatic Face Gesture Recognition. pp. 1–8, 2008.

[Be09] Beveridge, J. Ross; Givens, Geof H.; Phillips, P. Jonathon; Draper, Bruce A.: Factors that influence algorithm performance in the Face Recognition Grand Challenge. Computer Vision and Image Understanding, 113(6):750–762, 2009.

[Bh23] Bhatta, Aman; Albiero, Vítor; Bowyer, Kevin W; King, Michael C: The Gender Gap in Face Recognition Accuracy Is a Hairy Problem. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 303–312, 2023.

[Bu17] Buolamwini, Joy Adowaa: , Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers, 2017. MSc Thesis; http://hdl.handle.net/1721.1/114068; Last accessed: July 10, 2022.

[Ca18] Cao, Qiong; Shen, Li; Xie, Weidi; Parkhi, Omkar M.; Zisserman, Andrew: , VGGFace2: A dataset for recognising faces across pose and age, 2018.

[CKG23] Cheong, Jiaee; Kalkan, Sinan; Gunes, Hatice: Causal Structure Learning of Bias for Fair Affect Recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 340–349, 2023.

[Co19] Cook, Cynthia M.; Howard, John J.; Sirotin, Yevgeniy B.; Tipton, Jerry L.; Vemury, Arun R.: Demographic Effects in Facial Recognition and Their Dependence on Image Acquisition: An Evaluation of Eleven Commercial Systems. IEEE Transactions on Biometrics, Behavior, and Identity Science, 1(1):32–41, 2019.

[dFPM22] de Freitas Pereira, Tiago; Marcel, Sébastien: Fairness in Biometrics: A Figure of Merit to Assess Biometric Verification Systems. IEEE Transactions on Biometrics, Behavior, and Identity Science, 4(1):19–29, 2022.

[Dr20] Drozdowski, Pawel; Rathgeb, Christian; Dantcheva, Antitza; Damer, Naser; Busch, Christoph: Demographic bias in biometrics: A survey on an emerging challenge. IEEE Transactions on Technology and Society, 1(2):89–103, 2020.

[Gi04] Givens, Geof H.; Beveridge, J. Ross; Draper, Bruce A.; Bolme, David: Using a Generalized Linear Mixed Model to Study the Configuration Space of a PCA+LDA Human Face Recognition Algorithm. In: Articulated Motion and Deformable Objects. volume 3179 of Lecture Notes in Computer Science, pp. 1–11, 2004.

[GNH19] Grother, P.; Ngan, M.; Hanaoka, K.: Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. Technical report, United States National Institute of Standards and Technology, 2019. NIST.IR 8280, https://doi.org/10.6028/NIST.IR.8280.

[Go21] Gong, Sixue: Face Recognition: Representation, Intrinsic Dimensionality, Capacity, and Demographic Bias. Michigan State University, 2021.

[Gr21] Grother, P: Demographic differentials in face recognition algorithms. Virtual Events Series–Demo-Graphic Fairness in Biometric Systems, 2021.

[GZ19] Guo, Guodong; Zhang, Na: A survey on deep learning based face recognition. Computer vision and image understanding, 189:102805, 2019.

[He15] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian: , Deep Residual Learning for Image Recognition, 2015.

[Ho22] Howard, John J; Laird, Eli J; Sirotin, Yevgeniy B; Rubin, Rebecca E; Tipton, Jerry L; Vemury, Arun R: Evaluating Proposed Fairness Models for Face Recognition Algorithms. arXiv preprint arXiv:2203.05051, 2022.

[HSV19] Howard, John J; Sirotin, Yevgeniy B; Vemury, Arun R: The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance. In: 2019 ieee 10th international conference on biometrics theory, applications and systems (btas). IEEE, pp. 1–8, 2019.

[IS23] ISO/IEC 19795-10: , Information technology – Biometric performance testing and reporting — Part 10: Quantifying biometric system performance variation across demographic groups (draft), 2023.

[Ji21] Jillson, Elisa: , Aiming for truth, fairness, and equity in your company's use of AI, 2021. https://www.ftc.gov/business-guidance/blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai; Last accessed: July 7, 2022.

[JL97] Jovanovic, B. D.; Levy, P. S.: A Look at the Rule of Three. The American Statistician, 51(2):137–139, may 1997.

[Kr20] Krishnapriya, K. S.; Albiero, Vítor; Vangara, Kushal; King, Michael C.; Bowyer, Kevin W.: Issues Related to Face Recognition Accuracy Varying Based on Race and Skin Tone. IEEE Transactions on Technology and Society, 1(1):8–20, 2020.

[NI] NIST: , NIST FRVT Demographics. https://pages.nist.gov/frvt/html/frvt_demographics.html. Accessed: 2023-04-13.

[Pa22] Pahl, Jaspar; Rieger, Ines; Möller, Anna; Wittenberg, Thomas; Schmid, Ute: Female, white, 27? bias evaluation on data and algorithms for affect recognition in faces. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 973–987, 2022.

[Sc10] Schuckers, Michael E.: Computational Methods in Biometric Authentication. Springer, 2010.

[Sc22] Schuckers, Michael; Purnapatra, Sandip; Fatima, Kaniz; Hou, Daqing; Schuckers, Stephanie: Statistical Methods for Assessing Differences in False Non-Match Rates Across Demographic Groups. In: Pattern Recognition. ICPR International Workshops and Challenges. IEEE, Montreal, QC, pp. 207–216, 2022.

[Te20] Terhörst, Philipp; Fährmann, Daniel; Damer, Naser; Kirchbuchner, Florian; Kuijper, Arjan: Beyond identity: What information is stored in biometric face templates? In: 2020 IEEE international joint conference on biometrics (IJCB). IEEE, pp. 1–10, 2020.

[We22] Wehrli, Samuel; Hertweck, Corinna; Amirian, Mohammadreza; Glüge, Stefan; Stadelmann, Thilo: Bias, awareness, and ignorance in deep-learning-based face recognition. AI and Ethics, 2(3):509–522, 2022.

[WL16] Wasserstein, Ronald L.; Lazar, Nicole A.: The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician, 70(2):129–133, 2016.

[Yu22] Yucer, Seyma; Poyser, Matt; Al Moubayed, Noura; Breckon, Toby P: Does lossy image compression affect racial bias within face recognition? In: 2022 IEEE International Joint Conference on Biometrics (IJCB). IEEE, pp. 1–10, 2022.

[Zh17] Zhang, Xiao; Fang, Zhiyuan; Wen, Yandong; Li, Zhifeng; Qiao, Yu: Range loss for deep face recognition with long-tailed training data. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5409–5418, 2017.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# Voice Morphing: Two Identities in One Voice

Sushanta K. Pani[1], Anurag Chowdhury, Morgan Sandler, Arun Ross

**Abstract:** In a biometric system, each biometric sample or template is typically associated with a single identity. However, recent research has demonstrated the possibility of generating "morph" biometric samples that can succesfully match more than a single identity. Morph attacks are now recognized as a potential security threat to biometric systems. However, most morph attacks have been studied on biometric modalities operating in the image domain, such as face, fingerprint, and iris. In this preliminary work, we introduce Voice Identity Morphing (VIM) - a voice-based morph attack that can synthesize speech samples that impersonate the voice characteristics of a pair of individuals. Our experiments evaluate the vulnerabilities of two popular speaker recognition systems, ECAPA-TDNN and x-vector, to VIM, with a success rate (MMPMR) of over 80% at a false match rate of 1% on the Librispeech dataset.

**Keywords:** Identity Morphing, Morph Attack, Speaker Recognition, Speech Synthesis

## 1 Introduction

Biometric systems use physical or behavioral traits to recognize individuals [JFR07]. A biometric system acquires a biometric sample of an individual (e.g., voice) using a sensor (e.g., microphone) and extracts a salient feature set (or template). This template is then used to recognize the individual. Typically, a template is associated with a single identity. However, over the past decade, several adversarial techniques, called *morph attacks*, have been developed to create synthetic biometric samples that can successfully match multiple identities [FFM14].[2] Furthermore, in recent times, DeepFake based synthetic image generators have been used to launch morph attacks on image-based biometric systems, viz., face, fingerprint, and iris, with high success rates [Ve21]. The success of such attacks can potentially lead to compromise of security in sensitive applications where a single biometric ID card could be shared by two or more individuals for nefarious purposes.

Existing literature on morph attacks demonstrates its potency against biometric modalities such as face, fingerprint, and iris [Ve21], [Sc17], [SR21]. For example, landmark-based [MFFM19, RRB16] and deep learning-based [Zh21, Da18] face morph attacks have been shown to be effective against face recognition systems. Similarly, researchers have shown the possibility of launching a morph attack against iris matchers both at the image level [FFM14, SR21] and feature level [FCM16, RB17].

[1] All authors are affiliated with Michigan State University, USA, Corresponding author: Arun Ross (rossarun@cse.msu.edu)

[2] A related vulnerability known as MasterPrint attack [RMR17] or MasterFace attack [Ng22] has also been studied.

The voice modality, on the other hand, has seemingly been spared from morph attacks until now. The use of voice biometrics is especially relevant in some commercial applications, such as digital voice assistants [Ho18] and telephone banking [MSI01]. The voice morphing attack may be particularly harmful in scenarios where verification of a single identity is essential to proceed. For instance, consider an online spoken language test. In this context, the test-taking system might require the candidate to enroll their voice beforehand to ensure that the same individual appears for the test. This step is typically achieved using a speaker recognition system, designed to prevent an accomplice from taking the test on behalf of the candidate. However, with a voice morphing attack method, the candidate could enroll a morphed combination of their voice and that of an accomplice. This blend would match both identities, allowing the accomplice to take the test on the candidate's behalf by successfully matching their voice to the enrolled morphed template. This situation, coupled with the rapid adoption of voice biometric-enabled devices and services, has heightened interest in understanding their vulnerabilities to morphing attacks. Therefore, it is essential to investigate the viability and success rate of such attacks on popular speaker recognition systems.

In this paper, we propose a voice morphing technique called Voice Identity Morphing (VIM)[3] that can synthesize artificial voice samples containing the voice characteristics of a pair of identities. Experimentally we show that the morph voice samples generated from two identities can successfully match target audio samples of both constituent identities using two different popular speaker recognition systems. The proposed method uses the DeepTalk network [CRD21] to extract speaker embeddings from two source identities. Then, it performs a feature-level fusion of the two embeddings producing a new embedding corresponding to the morphed identity. Finally, the morphed embedding is input to a Tacotron 2-based Text-to-Speech synthesizer to generate a morphed audio sample.

The main contributions of this preliminary work are as follows: (a) We propose a voice identity morphing technique capable of generating speech samples that can successfully match two identities within the framework of a speaker recognition system. (b) We evaluate and demonstrate the vulnerability of two popular speaker recognition systems, namely x-vector [Sn18] and ECAPA-TDNN [DTD], to our proposed method. (c) We perform an ablation study to better understand this vulnerability, and we initiate a discussion on potential forensic measures that may counteract it. (d) We propose directions for future study on this topic.

## 2 Proposed Method: Voice Identity Morphing

Voice Identity Morphing (VIM), as shown in Figure 1, has two stages: a) synthetic voice generation and b) morph attack on a speaker recognition system. In the first stage, the proposed method generates synthetic speech samples exhibiting speaker-dependent speech characteristics pertaining to two different speakers, also referred to as the target speaker pair. The synthetic speech sample, called the morphed speech sample, is then compared to

[3] Note that *voice morphing* as defined in this work is different from previous use of this terminology in the speech literature, where it denotes modifying an individual's voice to sound like another individual.

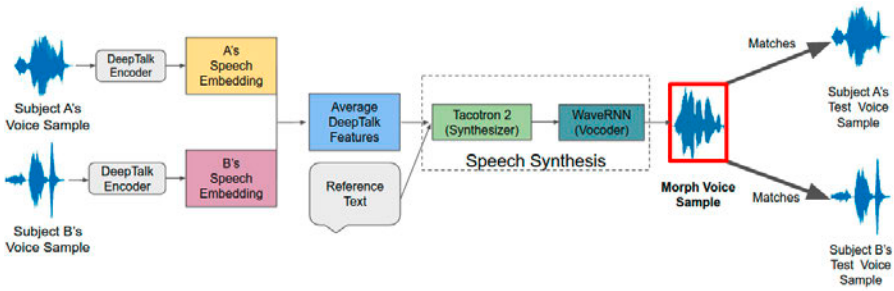



Fig. 1: Illustration of Voice Identity Morphing: Initially, the DeepTalk encoder processes and extracts embeddings that capture the unique speaker characteristics of two distinct individuals. Subsequently, to create a morphed identity, we compute the average of these two embeddings. This averaged embedding then serves as a reference point for our speech synthesis module. Ultimately, employing this reference, the vocoder module generates a spectrogram that merges elements from both contributing speakers.

individual voice samples from the target speaker pair to launch the morph attack. An attack is successful if the morphed speech sample matches both the target speaker pair's speech samples. The morph voice generation architecture has three separate modules: Encoder, Synthesizer, and Vocoder.

We use a pre-trained **DeepTalk encoder** model to generate vocal style-based speaker identity embeddings of voice samples. We choose this encoder for its competitive performance with the x-vector system and its robustness to degraded audio scenarios. This encoder architecture consists of a 1D-CNN based speech filter bank also known as DeepVOX network [CR20] and Global Style Token (GST) [Wa18] based prosody embedding network. The DeepVOX network generates short-term speaker-dependent DeepVOX features (see Table 1 for architecture details). The GST based prosody embedding network generates a fixed dimensional reference embedding from DeepVOX features by using a 2D-CNN followed by a 128-unit GRU. The DeepTalk encoder is pre-trained on the Librispeech, VoxCeleb1 and VoxCeleb2 datasets. The synthesizer module uses these embeddings as an input during the morph sample generation stages (speech synthesis and vocoding). As an initial step of the morph sample generation stage, we average the embeddings ($Emb_a$ and $Emb_b$) of two voice samples from separate speakers to generate a morph embedding $Emb_{morph} = (Emb_a + Emb_b)/2$. We perform this averaging step to incorporate features from both constituent identities. This assumes that there is an underlying geometric relationship between the identities in the learned embedding space from the DeepTalk encoder. We illustrate these relationships using t-SNE in Figure 3.

We use **Tacotron 2 speech synthesizer** [Sh18] to generate a mel-spectrogram for the corresponding text input. We use the Tacotron 2 synthesizer to retain consistency with the original DeepTalk architecture. Tacotron 2 architecture consists of an encoder and a decoder with an attention mechanism. The encoder creates an internal representation of input text, and the decoder uses the internal representation to generate features that encode the audio as a frame-level mel-spectrogram. The attention mechanism helps the decoder

Tab. 1: DeepVOX network setup for learning a 40-dimension feature representation from speech frames. All rows are convolutional layers separated by a SELU activation function.

| In Channels | Out Channels | Kernel | Dilation |
|---|---|---|---|
| 1 | 2 | 5x1 | 2x1 |
| 2 | 4 | 5x1 | 2x1 |
| 4 | 8 | 7x1 | 3x1 |
| 8 | 16 | 9x1 | 4x1 |
| 16 | 32 | 11x1 | 5x1 |
| 32 | 40 | 11x1 | 5x1 |

learn from the internal representation by weighting out potential failure cases where some subsequences of text are repeated or ignored by the decoder.

We use a **WaveRNN-based neural vocoder** [Ka18] pretrained model to generate morph samples by inverting the mel-spectrogram output from the Tacotron 2 synthesizer into audio samples. WaveRNN aims to have an expressive and non-linear transformation of the context and minimize the number of operations each step. An RNN addresses this purpose by combining the context and input within a single transformation.

## 3 Experimental Protocol

### 3.1 Dataset

We conducted experiments using the publicly accessible Librispeech dataset [Pa15], an audiobook corpus derived from Librivox projects. This dataset includes 1000 hours of audio data, in which, for each sample, a speaker reads English text. The dataset is divided into three subsets (100hr, 360hr, 500hr), all sampled at 16kHz. For our experiment, we utilized the 500 hour subset that consists of 1,166 participants (554 female and 612 male). We selected the 500hr subset not only because it is the largest subset, but also because it encompasses 440 speakers, each with more than 30 minutes of speaking time – a factor crucial for the morph generation process.

### 3.2 Baseline Recognition Performance

We assess speaker recognition systems' vulnerability to morph samples using two popular speaker recognition systems: x-vector [Sn18] and ECAPA-TDNN [DTD]. We choose these systems as they are freely available and are used in a wide range of systems.[4] We use the implementation of these systems in Speechbrain [RPO21] toolkit. The x-vector matcher is a TDNN (Time delay neural network) architecture and applies statistical pooling to extract 512-dimensional embedding for variable length utterances. The matcher utilizes categorical cross-entropy loss for training. The ECAPA-TDNN matcher architecture

[4] ECAPA-TDNN amassed 553,704 downloads in one month (June 2023) according to the HuggingFace website [In23]

consists of convolutional layers, residual blocks, and attentive statistical pooling layers. It utilizes Additive Margin SoftMax Loss to generate a 192-dimensional embedding. Both matchers utilize Voxceleb1 [NCZ17] and Voxceleb2 [CNZ18] datasets to train the models. They use cosine distance similarity of speaker embeddings to compare a pair of speaker identities.

Before assessing their vulnerability, we evaluate the baseline recognition performance of these speaker recognition systems on 440 subjects in the 500-hr subset of the Librispeech dataset [Pa15]. Table 2 provides the performance of these speaker recognition systems in terms of True Match Rate (TMR) at 1%, 0.1%, and 0.01% False Match Rate (FMR). TMR is the proportion of genuine samples that were correctly matched, whereas FMR was the proportion of impostor samples that were incorrectly matched. ECAPA-TDNN model performs better than the x-vector model in correctly classifying genuine and impostor pairs.

Tab. 2: Performance of two speaker recognition systems in terms of TMR (%) at 1%, 0.1%, and 0.01% FMR in the Librispeech dataset. The ECAPA-TDNN and x-vector are two popular, high-performing speaker recognition systems available in the Speechbrain toolkit.

| Matcher | TMR (%) | | |
|---|---|---|---|
| | FMR 1% | FMR 0.1% | FMR 0.01% |
| ECAPA-TDNN | 98.91 | 97.50 | 93.25 |
| x-vector | 88.17 | 78.57 | 68.52 |

### 3.3 Morph Generation Setup and Results

To generate morph voice samples that incorporate both identities of two different speakers, we first fine-tune a separate Tacotron 2 synthesizer with speech samples of that speaker pair. A pre-trained Tacotron 2 synthesizer needs approximately 30 minutes of the voice samples for fine-tuning [CRD21]. Therefore, we select 440 speakers (221 female and 219 male) which has 30 minutes or more cumulative duration of voice samples. From 440 speakers, we generate 96,580 speaker pairs ($^{440}C_2$ ). To generate *better quality* morph samples, we consider those speaker pairs which have *high similarity* in their speech. Each instance of Tacotron 2 takes 8-10 hours to fine-tune. Given this, we select the top 100 speaker pairs. We measure the similarity by the cosine distance of their ECAPA-TDNN-extracted speaker embeddings. Through this process, we select the top 100 speaker pairs, out of which only 43 pairs have unique speakers. The trimmed list of speaker pairs has 3 cross gender speaker pairs. Considering these 43 speaker pairs, we fine-tune 43 different Tacotron 2 synthesizers in parallel. For fine-tuning these Tacotron 2 synthesizers, we also provide 256-dimensional speaker embeddings extracted from a pre-trained DeepTalk encoder model [CRD21] as input along with a reference text. The fine-tuned Tacotron 2 synthesizer outputs a morphed mel spectrogram which is then fed as input into the WaveRNN vocoder [Ka18] to generate morphed speech samples. We create 100 such morphed samples from each speaker pair (10 samples per speaker) which results in 4,300 morphed samples. The speech samples used to generate morph samples are different from the ones

used for training the Tacotron 2 synthesizer. We use the remaining voice samples of a speaker for testing. Our experiment has disjoint sets of training (60%), morph (10%) and test (30%) speech samples.

To evaluate the vulnerabilities of the two speaker recognition systems against the generated morph samples (morph attack), we use the Mated Morph Presentation Match Rate (MMPMR) [Sc17] and Morphing Attack Potential (MAP) [Fe22] metrics. MMPMR is a fraction of successful morph attacks out of the total number of morph attacks. A morph attack is considered successful when the morph sample matches with test samples of both speakers. Table 3 provides the performance of morph attacks in terms of MMPMR at different thresholds corresponding to 1%, 0.1%, and 0.01% FMRs. We report the morph attack success rate in two categories: speaker pair level and morph sample level. A successful morph attack at the speaker pair level has at least one morph sample that matches the samples of both speakers. However, morph sample-level MMPMR reports the success of all morph samples irrespective of the speaker. The proposed morphing technique VIM can create morph samples attacking ECAPA-TDNN and x-vector speaker recognition systems with 95.34% and 86.04% respective success rates at 0.1% FMR, for speaker pair level. The results show that the ECAPA-TDNN speaker recognition system is more susceptible to morph attacks compared to the x-vector recognition system. The considerable success rate of morph attacks could likely be related to the morph pair selection process or the effective capturing of subject information by the DeepTalk encoding method. This infers that prior knowledge of the speaker recognition system would generate stronger morph attacks. Also, we hypothesize that state-of-the-art speaker recognition systems are likely to detect vocal features of both the parent speakers in a composite audio. This may make them vulnerable to such morphing attacks as well. We find that the fusion of speech synthesis embeddings generates effective morph audio samples for use in attacks on speaker recognition systems.

Tab. 3: Vulnerability assessment of two speaker recognition systems to voice identity morph attack in terms of MMPMR (%) at different threshold corresponding to 1%, 0.1%, and 0.01% FMR on the Librispeech dataset.

| Matcher | Speaker pair MMPMR (%) | | | Morph sample MMPMR (%) | | |
|---|---|---|---|---|---|---|
| | FMR 1% | FMR 0.1% | FMR 0.01% | FMR 1% | FMR 0.1% | FMR 0.01% |
| ECAPA-TDNN | 100.00 | 95.34 | 81.39 | 91.23 | 62.11 | 21.58 |
| x-vector | 93.02 | 86.04 | 9.30 | 82.13 | 38.95 | 4.32 |

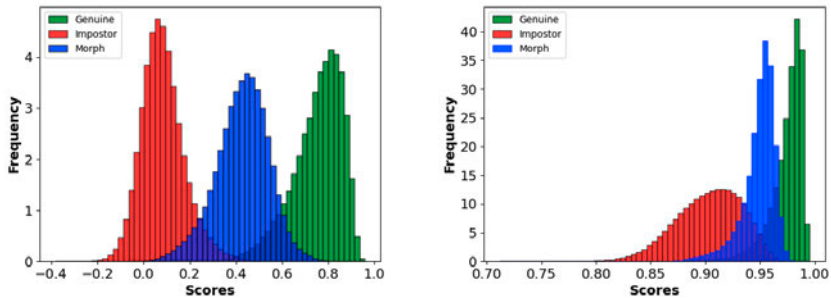
### 3.4 Result Analysis

We further analyze our morph attack performance using: 1) histogram plots, 2) t-SNE plots, and 3) morphing attack potential (MAP). Figure 2 shows the histogram plots of match scores corresponding to genuine pairs (green), impostor pairs (red), and pairs which include at least one morphed sample (blue) for both speaker recognition systems. In both systems, we find that the morphed pairs match score distribution lies between genuine and impostor score distributions. Morph samples are classified as genuine matches in the

Tab. 4: Morphing Attack Potential (MAP) [Fe22]: This metric represents the success rate (%) of a morphed sample matching at least a specified number of probe voice samples (denoted as # of attempts) within the Librispeech dataset, using one or both of the speaker recognition systems (SRS), namely ECAPA-TDNN and x-vector. The success rate is evaluated at three false match rate (FMR) thresholds: 1%, 0.1%, and 0.01%.

| # of Attempts | FMR 1% | | FMR 0.1% | | FMR 0.01% | |
|---|---|---|---|---|---|---|
| | 1 SRS | 2 SRS | 1 SRS | 2 SRS | 1 SRS | 2 SRS |
| 1 | 92.0% | 52.7% | 60.4% | 7.6% | 20.2% | 2.3% |
| 2 | 90.2% | 46.3% | 54.4% | 5.7% | 16.5% | 1.7% |
| 3 | 88.9% | 41.6% | 50.8% | 5.0% | 14.3% | 1.0% |
| 4 | 87.9% | 38.1% | 47.9% | 4.6% | 13.0% | 0.6% |
| 5 | 87.0% | 35.7% | 45.8% | 4.2% | 11.4% | 0.3% |

ECAPA-TDNN and x-vector systems with recognition thresholds of 0.46 and 0.96 respectively at 0.1% FMR.




ECAPA-TDNN x-vector

Fig. 2: Speaker recognition match score distributions of non-morph versus non-morph genuine (Green), non-morph versus non-morph impostor (Red), and morph versus non-morph genuine morph scores (Blue) using ECAPA-TDNN and x-vector embeddings.

The second analysis we perform is based on the t-SNE dimensionality reduction technique. The t-SNE [VdMH08] method helps visualize high-dimensional embeddings in a two-dimensional space by reducing the dimension. Figure 3 shows the t-SNE plot of morph sample embeddings from two speaker pairs (AB and CD) along with non-morph samples of four constituent speakers (A, B, C, and D). The embeddings are extracted by the ECAPA-TDNN recognition system. Here, embeddings of morph samples of one speaker pair (AB) are closer to embeddings of A and B speakers. Similarly, embeddings of morph samples of another speaker pair (CD) are closer to embeddings of C and D speakers. The analysis again validates the effectiveness of the proposed morphing technique and the potential threat of morph attacks.

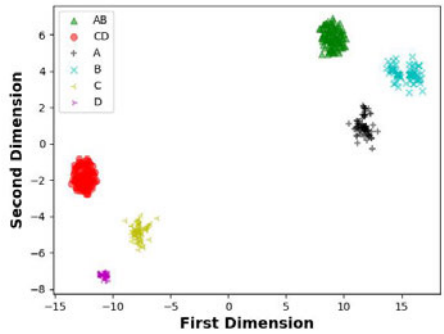



Fig. 3: t-SNE plot which illustrates high-dimensional ECAPA-TDNN embeddings of morph speech samples from two separate speaker pairs (AB and CD) and non-morph speech samples of individual speakers (A, B, C, and D). Morph embeddings of each pair are closer to the non-morph embeddings of their constituent speakers.

The Morphing Attack Potential (MAP) [Fe22] constitutes the third analysis. This metric takes into account multiple Speaker Recognition Systems (SRS) to ensure generality, and a variable number of verified probe samples for robustness. The result is a matrix (Table 4) in which one axis represents the number of probe samples (referred to as the number of attempts), and the other axis represents the number of SRS. The entries in each row represent the success rate (in percentage) of a morphed sample matching at least a specified number of probe voice samples (referred to as the number of attempts) using either or both of the SRS, viz., ECAPA-TDNN and x-vector. We report the success rates over three FMR thresholds of 1%, 0.1%, and 0.01%. The results imply that VIM is effective at a fairly competitive FMR of 1%, but suggest there is still room for improvement in performance at very low FMR thresholds. This may be attributed to the morph selection process or perhaps to the pre-trained models used in the encoder and speech synthesis steps.

## 4 Summary and Future Work

To the best of our knowledge, this preliminary work is the first to demonstrate the vulnerability of speaker recognition systems to morph attacks. In this regard, we propose a voice morphing technique called VIM to generate speech samples corresponding to the identities of two subjects. Using these morph samples, we demonstrate a morph attack success rate of over 80% on two popular speaker recognition systems (ECAPA-TDNN and x-vector). As future work, we propose to select high-similarity pairs for a morphing attack using x-vector to investigate whether the selection process plays a vital role in the performance of such an attack. Additionally, evaluating newer speaker recognition systems such as TitaNet [KPG22] and MFA-Conformer [Zh22] would provide more insight into the generalizability of VIM. Comparing other speech synthesis systems in the speech synthesis step would shed light on the role this step plays in the VIM attack. Furthermore, we aim to develop a

system for detecting morphed speech samples, possibly through the identification of their constituent identities. It may also be interesting to explore the maximum number of identities that can be combined into a single audio sample using VIM.

## 5 Reproducibility

The code for generating VIM samples can be found online at our Github link. [5]

## References

[CNZ18] Chung, J. S.; Nagrani, A.; Zisserman, A.: VoxCeleb2: Deep Speaker Recognition. In: INTERSPEECH. 2018.

[CR20] Chowdhury, Anurag; Ross, Arun: DeepVOX: Discovering Features From Raw Audio For Speaker Recognition in Non-Ideal Audio Signals. arXiv preprint arXiv:2008.11668, 2020.

[CRD21] Chowdhury, Anurag; Ross, Arun; David, Prabu: DeepTalk: Vocal Style Encoding for Speaker Recognition and Speech Synthesis. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6189–6193, 2021.

[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: IEEE 9th international conference on biometrics theory, applications and systems (BTAS). pp. 1–10, 2018.

[DTD] Desplanques, Brecht; Thienpondt, Jenthe; Demuynck, Kris: ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In: INTERSPEECH 2020. pp. 3830–3834.

[FCM16] Ferrara, Matteo; Cappelli, Raffaele; Maltoni, Davide: On The Feasibility of Creating Double-Identity Fingerprints. IEEE Transactions on Information Forensics and Security, 12(4):892–900, 2016.

[Fe22] Ferrara, Matteo; Franco, Annalisa; Maltoni, Davide; Busch, Christoph: Morphing Attack Potential. In: International Workshop on Biometrics and Forensics (IWBF). IEEE, pp. 1–6, 2022.

[FFM14] Ferrara, Matteo; Franco, Annalisa; Maltoni, Davide: The Magic Passport. In: IEEE International Joint Conference on Biometrics. pp. 1–7, 2014.

[Ho18] Hoy, Matthew B: Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. Medical Reference Services Quarterly, 37(1):81–88, 2018.

[In23] Inc., Hugging Face: , Hugging Face: The AI community building the future, 2023. Accessed: 2023-07-08.

[JFR07] Jain, Anil K; Flynn, Patrick; Ross, Arun A: Handbook of Biometrics. Springer Science & Business Media, 2007.

[5] https://github.com/morganlee123/VIM

[Ka18] Kalchbrenner, Nal; Elsen, Erich; Simonyan, Karen; Others: Efficient Neural Audio Synthesis. In: International Conference on Machine Learning. PMLR, pp. 2410–2419, 2018.

[KPG22] Koluguri, Nithin Rao; Park, Taejin; Ginsburg, Boris: TitaNet: Neural Model for Speaker Representation with 1D Depth-Wise Separable Convolutions and Global Context. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 8102–8106, 2022.

[MFFM19] Matteo Ferrara, Matteo; Franco, Annalisa; Maltoni, Davide: Decoupling Texture Blending and Shape Warping in Face Morphing. In: Proceedings of the 18th International Conference of the Biometrics Special Interest Group (BIOSIG). Gesellschaft für Informatik eV, 2019.

[MSI01] Melin, Håkan; Sandell, Anna; Ihse, Magnus: CTT-Bank: A Speech Controlled Telephone Banking System-An Initial Evaluation. TMH-QPSR, 1:1–27, 2001.

[NCZ17] Nagrani, A.; Chung, J. S.; Zisserman, A.: VoxCeleb: A Large-Scale Speaker Identification Dataset. In: INTERSPEECH. 2017.

[Ng22] Nguyen, Huy H.; Marcel, Sebastien; Yamagishi, Junichi; Echizen, Isao: Master Face Attacks on Face Recognition Systems. IEEE Transactions on Biometrics, Behavior, and Identity Science, 4(3):398–411, 2022.

[Pa15] Panayotov, Vassil; Chen, Guoguo; Povey, Daniel; Khudanpur, Sanjeev: Librispeech: An ASR Corpus Based on Public Domain Audio Books. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5206–5210, 2015.

[RB17] Rathgeb, Christian; Busch, Christoph: On The Feasibility of Creating Morphed IrisCodes. In: IEEE International Joint Conference on Biometrics (IJCB). pp. 152–157, 2017.

[RMR17] Roy, Aditi; Memon, Nasir; Ross, Arun: MasterPrint: Exploring the Vulnerability of Partial Fingerprint-Based Authentication Systems. IEEE Transactions on Information Forensics and Security, 12(9):2013–2025, 2017.

[RPO21] Ravanelli, Mirco; Parcollet, Titouan; Others: , SpeechBrain: A General-Purpose Speech Toolkit, 2021. arXiv:2106.04624.

[RRB16] Ramachandra, R; Raja, KB; Busch, C: Detecting Morphed Face Images. In: 8th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS. pp. 1–7, 2016.

[Sc17] Scherhag, Ulrich; Nautsch, Andreas; Rathgeb, Christian; Others: Biometric Systems Under Morphing Attacks: Assessment of Morphing Techniques and Vulnerability Reporting. In: International Conference of the Biometrics Special Interest Group (BIOSIG). IEEE, pp. 1–7, 2017.

[Sh18] Shen, Jonathan; Pang, Ruoming; Weiss, Ron J; Others: Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4779–4783, 2018.

[Sn18] Snyder, David; Garcia-Romero, Daniel; Sell, Gregory; Povey, Daniel; Khudanpur, Sanjeev: X-Vectors: Robust DNN Embeddings for Speaker Recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5329–5333, 2018.

[SR21] Sharma, Renu; Ross, Arun: Image-Level Iris Morph Attack. In: IEEE International Conference on Image Processing (ICIP). pp. 3013–3017, 2021.

[VdMH08] Van der Maaten, Laurens; Hinton, Geoffrey: Visualizing Data Using t-SNE. Journal of Machine Learning Research, 9(11), 2008.

[Ve21] Venkatesh, Sushma; Ramachandra, Raghavendra; Raja, Kiran; Busch, Christoph: Face Morphing Attack Generation and Detection: A Comprehensive Survey. IEEE Transactions on Technology and Society, 2(3):128–145, 2021.

[Wa18] Wang, Yuxuan; Stanton, Daisy; Zhang, Yu; Others: Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. In: International Conference on Machine Learning (ICML). PMLR, pp. 5180–5189, 2018.

[Zh21] Zhang, Haoyu; Venkatesh, Sushma; Ramachandra, Raghavendra; Others: MIPGAN—Generating Strong and High Quality Morphing Attacks Using Identity Prior Driven GAN. IEEE Transactions on Biometrics, Behavior, and Identity Science, 3(3):365–383, 2021.

[Zh22] Zhang, Yang; Lv, Zhiqiang; Wu, Haibin; Others: MFA-Conformer: Multi-scale Feature Aggregation Conformer for Automatic Speaker Verification. In: INTERSPEECH. 2022.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# Compressed Models Decompress Race Biases: What Quantized Models Forget for Fair Face Recognition

Pedro C. Neto[1,2], Eduarda Caldeira[1,2], Jaime S. Cardoso[1,2] and Ana F. Sequeira [1]

**Abstract:** With the ever-growing complexity of deep learning models for face recognition, it becomes hard to deploy these systems in real life. Researchers have two options: 1) use smaller models; 2) compress their current models. Since the usage of smaller models might lead to concerning biases, compression gains relevance. However, compressing might be also responsible for an increase in the bias of the final model. We investigate the overall performance, the performance on each ethnicity subgroup and the racial bias of a State-of-the-Art quantization approach when used with synthetic and real data. This analysis provides a few more details on potential benefits of performing quantization with synthetic data, for instance, the reduction of biases on the majority of test scenarios. We tested five distinct architectures and three different training datasets. The models were evaluated on a fourth dataset which was collected to infer and compare the performance of face recognition models on different ethnicity.

**Keywords:** Racial Bias, Face Recognition, Deep Learning, Compression, Quantization, Synthetic Data.

## 1 Introduction

Face recognition methods have made significant progress over the previous years [Bo22]. Current systems are capable of rivalling with humans under certain conditions and are quickly reducing the gap on the remaining test scenarios [Ph18]. The urge to keep the current rate of improvement on these deep learning-based approaches led to an era of complex and obscure models. As such, despite their extraordinary performance, there are two pressing concerns. First, there are hardware limitations that affect the complexity of the models that can be deployed and used in real scenarios. These limitations affect both storage, memory and processing time. The second concern is that the behaviour of a deep neural network is not easily understood [Ne22]. As such, besides the valuable information, also irrelevant or even harmful correlations can be learnt by these models, and hidden within their obscure nature.

Addressing these two concerns is of utter importance. To tackle them individually, one must be careful to avoid a potential trade-off between their mitigation. For instance, considering the possibility of an existing bias on the models, further reducing the model size can lead to an increased bias. Moreover, unless the model is reduced to an interpretable version of itself, these growing biases remain hidden within the black-box model.

[1] INESC TEC, Porto, Portugal, pedro.d.carneiro@inesctec.pt
[2] Faculdade de Engenharia da Universidade do Porto, Porto, Portugal

Instead of using a smaller model, current work is investigating different model compression approaches. Quantization, knowledge distillation and pruning are the most common. In this work, we aim to study the impact of quantization on the mitigation or amplification of existing biases. Hooker *et al.* [Ho20] presented a set of experiments that indicates a potential increase of the previous biases and tried to identify the elements forgotten by the deep neural network [Ho19]. To further extend this research, we framed our problem within the context of racial biases in face recognition systems. Stoychev *et al.* [SG22] presented mixed results on a face-related task and the effects on the biases were dependent on the training dataset. For this reason, our work, starting from Boutros *et al.* [BDK22] quantization approach, further includes the usage of real and synthetic data and the usage of distinct datasets for training. This study also aims to understand the current trade-offs between small models and hidden biases.

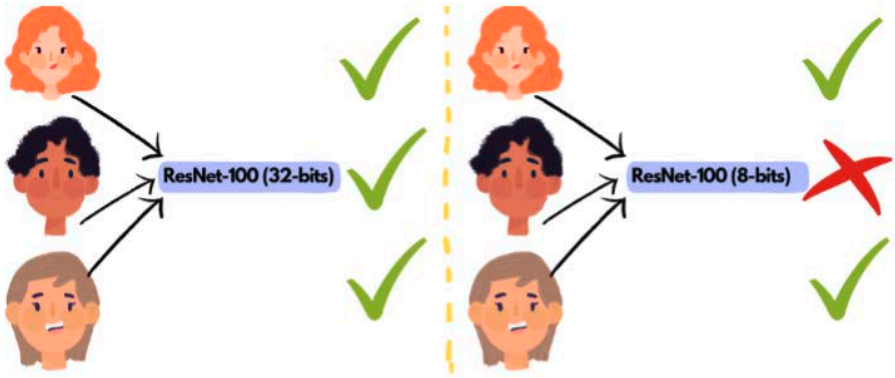



Fig. 1: Representation on a potential source of biases created after quantizing a deep neural network. Less represented classes or classes prone to suffer from discrimination might be easier to forget, leading to errors focused on those classes.

Within the context of this work, we aim to answer three research questions: 1) *Are smaller models more biased?* 2) *Are quantized models more biased? (As represented in Figure 1*) 3) *What is the impact of using synthetic data to quantize these models*? We individually address each of these questions and provide a potential explanation for the behaviour displayed by the models. Furthermore, the usage of synthetic data is motivated by the possibility of further growing our current datasets without compromising ethical concerns or privacy. Besides, with synthetic data, it is possible to create a higher degree of variability. Given these research questions, we present the following contributions:

- A study on racial bias on four differently sized models trained on MS1MV2 [De19] and of two models trained on BUPT-Balancedface and BUPT-Globalface [WZD21];
- Using QuantFace [BDK22] and real data we study the racial bias of the quantized version of all the models previously evaluated;
- We repeat the previous analysis with synthetic data;

- While confirming some of the Hooker *et al.* [Ho20] findings, we have also discovered that quantization with synthetic data mitigates the racial bias of the resulting model.

While we conducted our studies on a face recognition task, we theorise that the results seen are plausible for similar verification or classification problems, and other biometric modalities. In other words, smaller models and quantisation might impact the class with higher risk of discrimination in the same way that we describe in this study.

The following sections are divided into four major sections and a conclusion. Section 2 describes the five distinct datasets utilised in this study. Afterwards, in Section 3 and 4 the methodology and experimental setup are described in detail. Finally, the results are shown and discussed in Section 5.

## 2 Datasets

This study utilized five distinct datasets. MS1MV2 [De19], BUPT-Balancedface and BUPT-Globalface [WZD21] have been used for training the base models and quantization, while the synthetic data [BDK22] was only used for quantization and RFW [Wa19] just for evaluation of the models.

### 2.1 MS1MV2

MS1MV2 is widely used in the literature to train and compare several deep face recognition models [Bo22, Ne21]. It is a refined version of the MS-Celeb-1M dataset [Gu16], which further improved the training of these systems. The dataset contains 85k different identities and almost six million images and it is not balanced with respect to the race.

### 2.2 BUPT-Balancedface and BUPT-Globalface

Wang *et al.* [WZD21] introduced two distinct datasets to train deep face recognition systems. These datasets have been created to mitigate race bias on face recognition through skin tone labelling as African, Asian, Caucasian and Indian. BUPT-Globalface contains two million images from 38k different identities, and the distribution of races follows their distribution in the world. On the other hand, BUPT-Balancedface contains 1.3 million images from 28k identities which are divided into 7k identities per race. As such, this second dataset is race balanced.

### 2.3 Synthetic data

This dataset, introduced in [BDK22] contains approximately 500k unlabelled synthetic images. These images have been generated by a generative adversarial network [Go14,

Ka20]. The noise used as input to generate the images was sampled from a Gaussian distribution and fed to a pretrained generator (official open source implementation [3] of StyleGAN2-AD). The usage of synthetic data is often seen to result in sub-optimal performances [Qi21] which might be caused by a domain gap between real and synthetic data [Xu20, Sa18, Le20]. In this work, the goal is not to use the synthetic data to learn the representations from scratch, and we further argue that there might exist advantages of this domain gap.

### 2.4 RFW

Racial Faces in-the-wild (RFW) [Wa19], was proposed by the same authors of BUPT-Balancedface, and was designed as a benchmarking dataset for fair face verification. Similarly, it includes labels for ethnicity, which allows for a fair assessment of potential biases. It contains 3000 individuals with 6000 image pairs for face verification.

## 3 Methods

The methodology was designed in two different processes. First, it is necessary to understand if there is a bias problem on quantized models, and for this we have used the publicly available QuantFace models. If the problem is identified, it is necessary to understand if it is visible on models trained on other datasets (balanced and non-balanced).

There are four different architectures for QuantFace available: MobileFaceNet [Ch18], ResNet-18 [He16], ResNet-50 and ResNet-100. Each of these architectures is available in five distinct shapes: the original full-precision model, the 8-bit model quantized with real data, the 8-bit model quantized with synthetic data, the 6-bit model quantized with real data and finally the 6-bit model quantized with synthetic data. This part is essential to understand if the behaviour of the quantized model changes with the selected precision and the network architecture. Hence, the dataset for this is fixed as the MS1MV2, so we can ignore the data as a factor of variability.

For the second part of the study, a ResNet-34 was trained on BUPT-Balancedface and a second ResNet-34 was trained on BUPT-Globalface. The first is available in the four different quantized versions described above, whereas the second was only studied in its 8-bit version with real data quantization from two different sources and synthetic quantization. The network architecture is fixed so that the variability factors are limited to the data used for training and quantization. Moreover, the usage of a different real dataset for quantization than the one used for training attempts to further improve the understanding of the reasons behind the performance of a model. For instance, performance changes might be caused by the usage of data from a different distribution than the training data.

In order to gain additional insights regarding the reasons behind the impact of the synthetic data on the bias resulting from the quantization, we further trained an ethnicity classifier

[3] `https://github.com/NVlabs/stylegan2-ada`

on BUPT-Balancedface to estimate the ethnicity distribution in the synthetic data. This classifier comprises a fully-connect layer on top of a pretrained Elastic-Arc model [Bo22] model and achieves accuracies above 95%.

## 4 Experimental Setup

For the quantization process we have utilized the open-source implementation of QuantFace[4] with a batch size of 128 on a Nvidia Tesla v100 32GB. We have utilized the same configuration as proposed by Boutros *et al.* [BDK22]. For training the face recognition models on BUPT-Balancedface and BUPT-Globalface, we have utilized the same protocol of Deng *et al.* [De19] to preprocess the images, reduce the learning rate and stop the training. The ethnicity classifier utilised the Elastic-Arc model [Bo22] with all its layers frozen. An additional classifier layer was added and trained.

### 4.1 Evaluation Metrics

The performance of the evaluated models was measured in terms of accuracy. For the fairness evaluation of these models we have utilised two metrics: the standard deviation between the different accuracies (STD), and the skewed error ratio (SER) seen in Equation 1.

$$SER = \frac{100 - min(acc)}{100 - max(acc)} \tag{1}$$

The STD aims to evaluate the variance between the different accuracy values. The usage of STD in a set with just four different samples might be questioned, however, this has been the approach used in the literature [WZD21]. For the sake of reproducible research and compliance with the literature, we have chosen to retain the metrics previously used. On the other hand, SER measures or much larger is the worst error when compared with the better error. This is important to understand the relative differences between the different accuracy values. As a relative evaluation metric, SER is highly sensitive when the accuracy is above 99%. This happens because as the errors get below 1% their relative difference also change accordingly. For instance, a SER computed for a maximum accuracy of 90% and a minimum accuracy of 80% is the same if these accuracy values were 99.9% and 99.8%. STD is highly sensitive to absolute differences, and grows large on sets with lower accuracy values.

## 5 Results

A careful analysis of the performance of the different sized models at full precision (Table 1) shows that smaller models tend to have higher biases and lower performance in terms

[4] `https://github.com/fdbtrs/QuantFace/`

of average accuracy. ResNet-100 is an exception and this difference might be related to the fact that SER becomes highly sensitive when the errors are below 1%.

Tab. 1: Table comprising the results, evaluated on RFW, from the different models trained on MS1MV2 and their respective quantized versions for different bits and quantization strategies (real or synthetic data). The versions of the models quantized with synthetic data seem to display better fairness metrics at a comparable average performance.

| Model | Bits | Quant. | Caucasian | Indian | Asian | African | Avg. | STD | SER |
|---|---|---|---|---|---|---|---|---|---|
| MobileFaceNets | 32 | - | 95.18% | 92.00% | 89.93% | 90.22% | **91.83%** | 2.41 | 2.09 |
| | 8 | Real | 95.32% | 91.60% | 89.27% | 90.08% | 91.57% | 2.68 | 2.29 |
| | 8 | Synth. | 94.18% | 91.83% | 88.85% | 89.72% | 91.15% | **2.38** | 1.92 |
| | 6 | Real | 90.05% | 86.52% | 82.88% | 83.18% | 85.66% | 3.36 | 1.72 |
| | 6 | Synth. | 89.97% | 86.95% | 83.13% | 84.40% | 86.11% | 3.02 | **1.68** |
| ResNet-18 | 32 | - | 97.48% | 95.38% | 93.72% | 94.27% | **95.21%** | 1.66 | 2.49 |
| | 8 | Real | 97.42% | 95.33% | 93.55% | 94.20% | 95.13% | 1.70 | 2.50 |
| | 8 | Synth. | 96.95% | 95.07% | 93.30% | 93.87% | 94.80% | **1.61** | **2.20** |
| | 6 | Real | 96.93% | 94.65% | 92.52% | 93.22% | 94.33% | 1.95 | 2.44 |
| | 6 | Synth. | 96.80% | 94.78% | 92.35% | 93.28% | 94.30% | 1.94 | 2.39 |
| ResNet-50 | 32 | - | 99.00% | 98.15% | 97.62% | 98.32% | 98.27% | **0.57** | **2.38** |
| | 8 | Real | 99.07% | 98.07% | 97.65% | 98.40% | **98.30%** | 0.60 | 2.53 |
| | 8 | Synth. | 99.02% | 97.72% | 97.33% | 97.88% | 97.99% | 0.73 | 2.72 |
| | 6 | Real | 98.32% | 96.27% | 94.55% | 95.87% | 96.25% | 1.56 | 3.24 |
| | 6 | Synth. | 97.95% | 96.63% | 94.97% | 96.20% | 96.44% | 1.23 | 2.45 |
| ResNet-100 | 32 | - | 99.65% | 98.88% | 98.50% | 99.00% | **99.01%** | **0.48** | 4.29 |
| | 8 | Real | 99.57% | 98.87% | 98.15% | 98.77% | 98.84% | 0.58 | 4.30 |
| | 8 | Synth. | 99.37% | 98.72% | 98.13% | 98.78% | 98.75% | 0.51 | 2.97 |
| | 6 | Real | 95.27% | 93.15% | 90.32% | 91.70% | 92.61% | 2.12 | 2.05 |
| | 6 | Synth. | 95.93% | 93.40% | 91.92% | 92.60% | 93.46% | 1.75 | **1.99** |

The quantized version of these models seems to retain the performance and bias advantages when compared to simpler models. As theorised, the quantization has a negative impact on the bias, and in most cases on the performance too. The lower the number of bit, the higher the bias. However, the usage of synthetic data has shown, for all the different precisions, a capability to reduce the bias while retaining the performance. From this data, it is not clear if the improvement is due to a specific characteristic of the synthetic data.

We have used the ethnicity classifier to get an estimation of the racial balance of the synthetic data. We obtained 365889 Caucasians, 81568 Asians, 81568 Indians and 61966 Africans. Since the data is not balanced, it is not possible to associate the effects of this data to its balance.

Further training two ResNet-34 on BUPT-Balancedface and BUPT-Globalface shows, at full precision, that despite a higher performance of the latter, the balance of the former is essential to ensure better bias metrics. On the model trained with the BUPT-Balancedface the versions quantized with synthetic data has not only kept the same tendency of the previous table, but it has also surpassed by a large margin the version of the method quantized with the balanced data. This might be caused by the lower variability of the BUPT-Balancedface data with respect to the number of identities.

Tab. 2: Table comprising the results, evaluated on RFW, from two ResNet-34 models trained on BUPT-Balancedface (BL) and BUPT-Globalface (GL) and their respective quantized versions for different bits and quantization strategies (BL, GL or synthetic data). The versions of the models quantized with synthetic data seem to perform outstandingly well.

| Train Data | Bits | Quant. | Caucasian | Indian | Asian | African | Avg. | STD | SER |
|---|---|---|---|---|---|---|---|---|---|
| BL | 32 | - | 96.60% | 94.50% | 94.03% | 93.37% | **94.63%** | **1.40** | 1.95 |
| | 8 | BL | 94.98% | 93.60% | 92.77% | 90.95% | 93.08% | 1.68 | 1.80 |
| | 8 | Synth. | 96.03% | 94.40% | 93.97% | 92.50% | 94.23% | 1.45 | 1.89 |
| | 6 | BL | 89.22% | 87.87% | 86.25% | 82.80% | 86.54% | 2.77 | **1.60** |
| | 6 | Synth. | 94.58% | 92.88% | 91.45% | 91.13% | 92.51% | 1.58 | 1.64 |
| GL | 32 | - | 97.67% | 95.52% | 94.15% | 93.87% | **95.30%** | 1.74 | 2.63 |
| | 8 | BL | 95.42% | 92.75% | 91.83% | 89.88% | 92.47% | 2.30 | 2.21 |
| | 8 | GL. | 94.70% | 92.15% | 90.23% | 88.75% | 91.46% | 2.57 | **2.12** |
| | 8 | Synth. | 97.33% | 95.15% | 94.17% | 93.55% | 95.05% | **1.66** | 2.42 |

The ResNet-34 trained on the BUPT-Globalface performs better if quantized with the data from the BUPT-Balancedface instead of using the data from training. Once again, it might be possible that introducing variability and unseen data for the quantization increases the capability of the model to be robust for all ethnicities. This is further validated by the version quantized with the synthetic data, which leads to a performance similar to the full precision model.

## 6 Conclusion

In this document, we tackled three research questions, and we have provided answers to all of them. 1) and 2) It was possible to infer that models quantized with real data and smaller models are indeed more biased; 3) it was also verifiable that using synthetic data for quantization positively impacts the fairness metrics. We have extended previous literature on the assessment of the information that is lost by quantized models and further introduced a novel topic regarding the usage of synthetic data for bias mitigation.

Despite the interesting results shown by our experiments, there are several gaps in the literature that should be tackled in future work. For instance, it is not known if this behaviour is the same for gender biases, or if synthetic data harms gender biases while helping to mitigate race biases. A more comprehensive study is required. The usage of the combined real data that has and has not been seen, and synthetic data should be also analysed to understand how can we, just by changing the training data, mitigate these biases while retaining the original performance. Furthermore, we still do not know if these findings hold for different traits and tasks, and further studies are required to confirm the generalization of these findings.

While the results shown are still preliminary, they introduce a few research directions that might be relevant for the future of biometrics in an era of increasing concern with these biases.

## Acknowledgments

This work is co-financed by Component 5 - Capitalization and Business Innovation, integrated in the Resilience Dimension of the Recovery and Resilience Plan within the scope of the Recovery and Resilience Mechanism (MRR) of the European Union (EU), framed in the Next Generation EU, for the period 2021 - 2026, within project NewSpacePortugal, with reference 11. It was also financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within the PhD grant "2021.06872.BD".

## References

[BDK22] Boutros, Fadi; Damer, Naser; Kuijper, Arjan: Quantface: Towards lightweight face recognition by synthetic data low-bit quantization. In: 2022 26th International Conference on Pattern Recognition (ICPR). IEEE, pp. 855–862, 2022.

[Bo22] Boutros, Fadi; Damer, Naser; Kirchbuchner, Florian; Kuijper, Arjan: Elasticface: Elastic margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1578–1587, 2022.

[Ch18] Chen, Sheng; Liu, Yang; Gao, Xiang; Han, Zhen: MobileFaceNets: Efficient CNNs for Accurate Real-Time Face Verification on Mobile Devices. In: CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings. volume 10996 of Lecture Notes in Computer Science. 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/CVF conference on computer vision and pattern recognition. pp. 4690–4699, 2019.

[Go14] Goodfellow, Ian J.; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil; Courville, Aaron C.; Bengio, Yoshua: Generative Adversarial Nets. In: Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada. pp. 2672–2680, 2014.

[Gu16] Guo, Yandong; Zhang, Lei; Hu, Yuxiao; He, Xiaodong; Gao, Jianfeng: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14. Springer, pp. 87–102, 2016.

[He16] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 770–778, 2016.

[Ho19] Hooker, Sara; Courville, Aaron; Clark, Gregory; Dauphin, Yann; Frome, Andrea: What do compressed deep neural networks forget? arXiv preprint arXiv:1911.05248, 2019.

[Ho20] Hooker, Sara; Moorosi, Nyalleng; Clark, Gregory; Bengio, Samy; Denton, Emily: Characterising bias in compressed models. arXiv preprint arXiv:2010.03058, 2020.

[Ka20] Karras, Tero; Aittala, Miika; Hellsten, Janne; Laine, Samuli; Lehtinen, Jaakko; Aila, Timo: Training Generative Adversarial Networks with Limited Data. In: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual. 2020.

[Le20] Lee, Sangrok; Park, Eunsoo; Yi, Hongsuk; Lee, Sang Hun: StRDAN: Synthetic-to-Real Domain Adaptation Network for Vehicle Re-Identification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020. Computer Vision Foundation / IEEE, pp. 2590–2597, 2020.

[Ne21] Neto, Pedro C; Boutros, Fadi; Pinto, João Ribeiro; Damer, Naser; Sequeira, Ana F; Cardoso, Jaime S: Focusface: Multi-task contrastive learning for masked face recognition. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). IEEE, pp. 01–08, 2021.

[Ne22] Neto, Pedro C; Gonçalves, Tiago; Pinto, João Ribeiro; Silva, Wilson; Sequeira, Ana F; Ross, Arun; Cardoso, Jaime S: Explainable biometrics in the age of deep learning. arXiv preprint arXiv:2208.09500, 2022.

[Ph18] Phillips, P Jonathon; Yates, Amy N; Hu, Ying; Hahn, Carina A; Noyes, Eilidh; Jackson, Kelsey; Cavazos, Jacqueline G; Jeckeln, Géraldine; Ranjan, Rajeev; Sankaranarayanan, Swami et al.: Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. Proceedings of the National Academy of Sciences, 115(24):6171–6176, 2018.

[Qi21] Qiu, Haibo; Yu, Baosheng; Gong, Dihong; Li, Zhifeng; Liu, Wei; Tao, Dacheng: SynFace: Face Recognition With Synthetic Data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10880–10890, October 2021.

[Sa18] Sankaranarayanan, Swami; Balaji, Yogesh; Jain, Arpit; Lim, Ser-Nam; Chellappa, Rama: Learning From Synthetic Data: Addressing Domain Shift for Semantic Segmentation. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. Computer Vision Foundation / IEEE Computer Society, pp. 3752–3761, 2018.

[SG22] Stoychev, Samuil; Gunes, Hatice: The effect of model compression on fairness in facial expression recognition. arXiv preprint arXiv:2201.01709, 2022.

[Wa19] Wang, Mei; Deng, Weihong; Hu, Jiani; Tao, Xunqiang; Huang, Yaohai: Racial Faces in the Wild: Reducing Racial Bias by Information Maximization Adaptation Network. In: The IEEE International Conference on Computer Vision (ICCV). October 2019.

[WZD21] Wang, Mei; Zhang, Yaobin; Deng, Weihong: Meta Balanced Network for Fair Face Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

[Xu20] Xu, Minghao; Zhang, Jian; Ni, Bingbing; Li, Teng; Wang, Chengjie; Tian, Qi; Zhang, Wenjun: Adversarial Domain Adaptation with Domain Mixup. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. AAAI Press, pp. 6502–6509, 2020.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# Fuzzy Signature with Biometric-Independent Verification

Haruna Higo[1] Toshiyuki Isshiki[1] Saki Otsuki[1] Kenji Yasunaga[2]

**Abstract:** Just as biometric authentication has replaced conventional authentication methods to improve convenience and security, fuzzy signatures (FS), a technique that replaces the signing key of a digital signature with biometric information, has been studied. This paper proposes biometric-independent fuzzy signature (BIFS), a novel FS primitive that enables biometric-independent verification, which is inherently unsolvable with conventional FS. To circumvent that problem, BIFS generates signatures through a two-party protocol using biometric information and helper keys. We also propose a construction of BIFS that can easily replace existing signature services because the signature and verification keys are in the same form as those of digital signatures.

**Keywords:** Digital signature, secure sketch, fuzzy signature.

## 1 Introduction

Biometric authentication compensates for the disadvantages of authentication by memory and possession. Biometric characteristics have strong bindings with the owner, so there is less risk of being forgotten or guessed, unlike passwords, and less risk of being lost or stolen, unlike IC cards. Recent technological advances have successfully overcome the difficulties caused by the nature of biometric information, and biometric authentication is now used in large-scale applications such as national IDs.

A digital signature (DS) on an electronic document serves an equivalent purpose as a signature on a paper document. That is, with DS, the signer guarantees that he or she has endorsed some document. Typical applications are e-mail protections (S/MIME), electronic contracts, and blockchain applications such as cryptocurrency transfers and proof of attributes. The basis for the guarantee is the binding between the signing keys and the owners. If a signing key is stolen, attackers can impersonate the owner; if lost, the owner can no longer give guarantees. This feature is similar to that of authentication. Therefore, there are lines of research on fuzzy cryptographic primitives, such as fuzzy extractors [DRS04], to study the use of biometric information as a substitute for signing keys.

Fuzzy signature (FS), proposed by Takahashi et al. [Ta15], enables one to sign a message only with a biometric feature, which is extendedly researched in [Ma16, Ta19, Ka21]. The feature of FS is that the signer can generate signatures without remembering or holding anything. However, the feature of using only biometric features as a signing key has the

[1] NEC Corporation, 1753 Shimonumabe, Nakahara-ku, Kawasaki, Kanagawa 211-8666, Japan, {h-higo-aj, toshiyuki-isshiki, saki-otsuki}@nec.com
[2] Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8550 Japan, yasunaga@c.titech.ac.jp

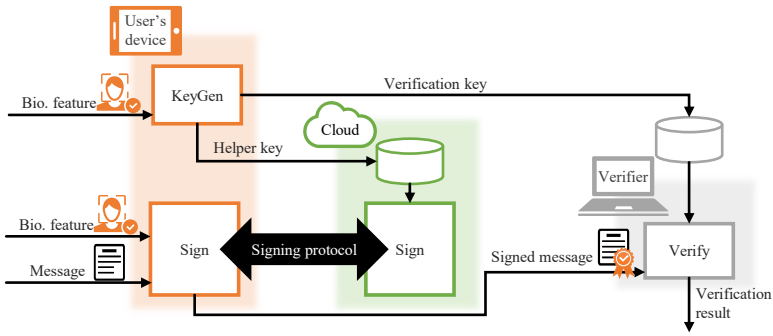



Fig. 1: Algorithms of BIFS

In this paper, we propose a novel FS primitive as illustrated in Figure 1, called BIFS, that enables biometric-independent verification. The core of BIFS is a two-party signing protocol. In the signature protocol, the first signer uses a biometric feature, and the second signer uses a helper key which is generated as a pair with a verification key on a biometric feature in the key generation phase. The requirement to have access to the helper keys is a usability disadvantage of BIFS compared to FS. However, the two fundamental problems of FS are solved with good use of the helper keys. First, multiple keys can be associated with a single biometric characteristic by using different helper keys. Secondly, verification keys and signatures can be constructed as biometric-independent data since the fuzziness can be absorbed in the signing protocol. As a security requirement, we define unforgeability of the signatures and confidentiality of the biometric features from the helper keys, verification keys, and signatures.

| | DS | FS | BIFS |
|---|---|---|---|
| Binding between signing key and owner | Weak | *Strong* | *Strong* |
| Storage for signing | Signing key | *No* | Helper key |
| Biometric-independent verification | *Yes* | No | *Yes* |

Tab. 1: Comparison of the signature methods.

We propose a generic construction for BIFS. Similar to the schemes in [Ta19], the scheme is based on a linear sketch scheme and a DS scheme with key-homomorphism [DS19]. Signatures and the verification keys in the proposed scheme are in the same forms as those of the underlying DS scheme. Therefore, not only can signatures and verification keys be made public, but also the proposed scheme is easy to replace existing signature services without changing the verification applications. This paper further discusses the feasibility of the proposed scheme by comparing it to DS and FS schemes, as summarized in Table 1.

## 2 Preliminaries

$\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{R}$ denote the sets of all natural numbers, integers, and real numbers, respectively. $a \leftarrow A$ denotes that $a$ is the output from the algorithm $A$ or chosen according to the distribution $A$. $b \xleftarrow{U} B$ denotes that $b$ is chosen uniformly at random from the set $B$. $\kappa$ denotes a security parameter, and PPT means probabilistic polynomial time on $\kappa$. $negl(\kappa)$ denotes a negligible function on $\kappa$ that satisfies $negl(\kappa) < 1/p(k)$ for all positive polynomials $p(\cdot)$ and sufficiently large $k$. We omit public parameters among the inputs if it is implicit from the contexts.

We use a model of biometric features as $\mathscr{F} = ((\mathsf{dis}, W), t, \mathscr{W})$ in which $(\mathsf{dis}, W)$ is a metric space where $W$ is a space of biometric features that constitutes an abelian group, and $\mathsf{dis} : W \times W \mapsto \mathbb{R}$ is the distance function, $t \in \mathbb{R}$ is the threshold for comparison where it is assumed that $\mathsf{dis}(x, x') < t$ holds if and only if two biometric features $x$ and $x'$ are extracted from the identical biometric characteristic, and $\mathscr{W}$ is a distribution of biometric features over $W$.

**Linear Sketch** To tolerate the differences between two readings of biometric features, we employ the linear sketch (LS) [Ma16] in the proposed scheme. The concept of LS is designed according to secure sketch [DRS04] that enables precise reconstruction of a noisy input. A sketch is a composition of a key and a biometric feature in both primitives. With linearity, LS enables to reconstruct the difference of the keys if the two features are similar enough.

Formally, an LS scheme for biometric features $\mathscr{F} = ((\mathsf{dis}, W), t, \mathscr{W})$ and an abelian group $(\mathscr{K}, +)$ consists of three PPT algorithms $\mathsf{Setup}_{\mathsf{LS}}$, $\mathsf{Gen}_{\mathsf{LS}}$, and $\mathsf{Rec}_{\mathsf{LS}}$, described as follows: The setup algorithm $\mathsf{Setup}_{\mathsf{LS}}$ takes a description of a model of biometric features and an abelian group $(\mathscr{K}, +)$ as input and outputs a public parameter $pp_{\mathsf{LS}}$. The sketch generation algorithm $\mathsf{Gen}_{\mathsf{LS}}$ takes a public parameter $pp_{\mathsf{LS}}$, an element $x \in \mathscr{K}$, and a biometric feature $w \in W$ as input and outputs a sketch $s$. The difference reconstruction algorithm $\mathsf{Rec}_{\mathsf{LS}}$ takes a public parameter $pp_{\mathsf{LS}}$ and two sketches $s$ and $s'$ as input and outputs a difference $\Delta$.

As correctness, for every $pp_{\mathsf{LS}}$ from $\mathsf{Setup}_{\mathsf{LS}}$, $x, x' \in \mathscr{K}$, and $w, w' \in W$ that satisfy $\mathsf{dis}(w, w') < t$, it is required to hold that $\mathsf{Rec}_{\mathsf{LS}}(pp_{\mathsf{LS}}, \mathsf{Gen}_{\mathsf{LS}}(pp_{\mathsf{LS}}, x, w), \mathsf{Gen}_{\mathsf{LS}}(pp_{\mathsf{LS}}, x', w')) = x - x'$.

Two security requirements are defined for LS schemes. The first one, *linearity* requires existence of a PPT algorithm $\mathsf{Lin}_{\mathsf{LS}}$ that for every $pp_{\mathsf{LS}}$ from $\mathsf{Setup}_{\mathsf{LS}}$, $x, \Delta \in \mathscr{K}$, and $w, e \in W$, the two distributions $\{s \leftarrow \mathsf{Gen}_{\mathsf{LS}}(pp_{\mathsf{LS}}, x, w);\ s' \leftarrow \mathsf{Gen}_{\mathsf{LS}}(pp_{\mathsf{LS}}, x + \Delta, w + e) \colon (s, s')\}$ and $\{s \leftarrow \mathsf{Gen}_{\mathsf{LS}}(pp_{\mathsf{LS}}, x, w);\ s' \leftarrow \mathsf{Lin}_{\mathsf{LS}}(pp_{\mathsf{LS}}, s, \Delta, e) \colon (s, s')\}$ are statistically indistinguishable. The second one, *weak simulatability* requires existence of a PPT algorithm $\mathsf{Sim}_{\mathsf{LS}}$ that for every PPT algorithm $\mathscr{A}$, there exists a polynomial $u$ such that $\Pr[\mathscr{A}(\mathsf{D}_{\mathsf{real}}) = 1] - u \Pr[\mathscr{A}(\mathsf{D}_{\mathsf{sim}}) = 1] \leq negl(\kappa)$ holds for $\mathsf{D}_{\mathsf{real}} := \{pp_{\mathsf{LS}} \leftarrow \mathsf{Setup}_{\mathsf{LS}}(\mathscr{F}, (\mathscr{K}, +));\ x \xleftarrow{U} \mathscr{K};\ w \leftarrow \mathscr{W};\ s \leftarrow \mathsf{Gen}_{\mathsf{LS}}(pp_{\mathsf{LS}}, x, w) \colon (pp_{\mathsf{LS}}, x, s)\}$ and $\mathsf{D}_{\mathsf{sim}} := \{pp_{\mathsf{LS}} \leftarrow \mathsf{Setup}_{\mathsf{LS}}(\mathscr{F}, (\mathscr{K}, +));\ x \xleftarrow{U} \mathscr{K};\ s \leftarrow \mathsf{Sim}_{\mathsf{LS}}(pp_{\mathsf{LS}}) \colon (pp_{\mathsf{LS}}, x, s)\}$.

**Digital Signatures** We also utilize digital signature (DS) schemes that have key-homomorphic properties. We first review the standard definitions for DS schemes. A DS scheme consists of four PPT algorithms $\mathsf{Setup}_{\mathsf{SIG}}$, $\mathsf{KeyGen}_{\mathsf{SIG}}$, $\mathsf{Sign}_{\mathsf{SIG}}$, and $\mathsf{Verify}_{\mathsf{SIG}}$, described as follows: The setup algorithm $\mathsf{Setup}_{\mathsf{SIG}}$ takes a security parameter $\kappa$ as input and outputs a public parameter $pp_{\mathsf{SIG}}$ that determines a signing key space $\mathscr{K}_{pp_{\mathsf{SIG}}}$. The key generation algorithm $\mathsf{KeyGen}_{\mathsf{SIG}}$ takes a public parameter $pp_{\mathsf{SIG}}$ as input and outputs a pair of a signing key $sk_{\mathsf{SIG}}$ and a verification key $vk_{\mathsf{SIG}}$. The signing algorithm $\mathsf{Sign}_{\mathsf{SIG}}$ takes a public parameter $pp_{\mathsf{SIG}}$, a signing key $sk_{\mathsf{SIG}}$, a message $\mu$ as input and outputs a signature $\sigma$. The verification algorithms $\mathsf{Verify}_{\mathsf{SIG}}$ takes a public parameter $pp_{\mathsf{SIG}}$, a verification key $vk_{\mathsf{SIG}}$, a message $\mu$, and a signature $\sigma$ as input and outputs a signal that represents acceptance ($\top$) or rejection ($\bot$) of the message-signature pair.

As correctness, for every $pp_{\mathsf{SIG}}$ from $\mathsf{Setup}_{\mathsf{SIG}}$, $(sk_{\mathsf{SIG}}, vk_{\mathsf{SIG}}) \leftarrow \mathsf{KeyGen}_{\mathsf{SIG}}(pp_{\mathsf{SIG}})$, and message $\mu$, it is required to hold that $\top \leftarrow \mathsf{Verify}_{\mathsf{SIG}}(pp_{\mathsf{SIG}}, vk_{\mathsf{SIG}}, \mu, \mathsf{Sign}_{\mathsf{SIG}}(pp_{\mathsf{SIG}}, sk_{\mathsf{SIG}}, \mu))$.

We require the existential unforgeability against chosen message attacks (EUF-CMA) for secure DS schemes. That is, a DS scheme satisfies *EUF-CMA security* if for every PPT algorithm $\mathscr{A}$, it holds that

$$\Pr\left[\begin{array}{l} pp_{\mathsf{SIG}} \leftarrow \mathsf{Setup}_{\mathsf{SIG}}(1^{\kappa}); \\ (sk_{\mathsf{SIG}}, vk_{\mathsf{SIG}}) \leftarrow \mathsf{KeyGen}_{\mathsf{SIG}}(pp_{\mathsf{SIG}}); \\ (\mu^*, \sigma^*) \leftarrow \mathscr{A}^{\mathsf{Sign}_{\mathsf{SIG}}(pp_{\mathsf{SIG}}, sk_{\mathsf{SIG}}, \cdot)}(pp_{\mathsf{SIG}}, vk_{\mathsf{SIG}}) : \\ \top \leftarrow \mathsf{Verify}_{\mathsf{SIG}}(pp_{\mathsf{SIG}}, vk_{\mathsf{SIG}}, \mu^*, \sigma^*) \end{array}\right] \leq negl(\kappa),$$

where $\mu^*$ must not be queried to the signing oracle.

Simply, the key-homomorphism is a property that a signature can be inverted to a signature of a different key for the same message. Formally, a key-homomorphic DS scheme satisfies the following four properties. First, the signing key space $\mathscr{K}_{pp_{\mathsf{SIG}}}$ is required to constitute an abelian group. Second is *a simple key generation process property* that the key generation algorithm can be divided into two subprocesses: the first randomly chooses the signing key, and the second determines the corresponding verification key. Third is a property of inverting verification keys that a PPT algorithm $\mathsf{VKShift}_{\mathsf{SIG}}$ exists for every $pp_{\mathsf{SIG}}$ generated by $\mathsf{Setup}_{\mathsf{SIG}}$ and $sk_{\mathsf{SIG}}, \Delta \in \mathscr{K}_{pp_{\mathsf{SIG}}}$, it holds that $\mathsf{KeyGen}_{\mathsf{SIG}}{}'(pp_{\mathsf{SIG}}, sk_{\mathsf{SIG}} + \Delta) = \mathsf{VKShift}_{\mathsf{SIG}}(pp_{\mathsf{SIG}}, \mathsf{KeyGen}_{\mathsf{SIG}}{}'(pp_{\mathsf{SIG}}, sk_{\mathsf{SIG}}), \Delta)$. The last is a property of inverting signatures that a PPT algorithm $\mathsf{SignShift}_{\mathsf{SIG}}$ exists for every $pp_{\mathsf{SIG}}$, $sk_{\mathsf{SIG}}, \Delta \in \mathscr{K}_{pp_{\mathsf{SIG}}}$, and message $\mu$, it holds that the two distributions $\{\sigma' \leftarrow \mathsf{Sign}_{\mathsf{SIG}}(pp_{\mathsf{SIG}}, sk_{\mathsf{SIG}} + \Delta, \mu) : \sigma'\}$ and $\{\sigma \leftarrow \mathsf{Sign}_{\mathsf{SIG}}(pp_{\mathsf{SIG}}, sk_{\mathsf{SIG}}, \mu);\ \sigma' \leftarrow \mathsf{SignShift}_{\mathsf{SIG}}(pp_{\mathsf{SIG}}, \sigma, \Delta)\}$ are identical and $\mathsf{Verify}_{\mathsf{SIG}}(pp_{\mathsf{SIG}}, \mathsf{KeyGen}_{\mathsf{SIG}}{}'(pp_{\mathsf{SIG}}, sk_{\mathsf{SIG}} + \Delta), \mu, \mathsf{SignShift}_{\mathsf{SIG}}(pp_{\mathsf{SIG}}, \sigma, \Delta)) = \top$.

## 3 Distributed Biometric Signing Protocol

We propose a naive concept of fuzzy signature that enables biometric-independent verification as BIFS. BIFS is designed to solve the principle problem of FS while using a biometric feature as a part of a signing key. That is, BIFS can bind multiple keys with a single

biometric characteristic, and verification keys and signatures are biometric-independent data. For that purpose, BIFS combines the ideas of key distribution and FS and enables signing by two signers where one of the signers' input is a biometric feature.

Therefore, the key generation algorithm generates a pair of a helper key and a verification key from a biometric feature. A signature is generated with a two-party signature protocol with $\mathsf{Sign}^{\mathsf{B}}_{\mathsf{BIFS}}$ on a biometric feature and $\mathsf{Sign}^{\mathsf{K}}_{\mathsf{BIFS}}$ on a helper key without revealing inputs each other. The verification algorithm works similarly to DS's, where the output $\top$ means that the two biometric features used in the key generation of the verification key and the signing of the message are extracted from an identical biometric characteristic. Formally, a BIFS scheme for a model of biometric features $\mathscr{F}$ is a tuple of five PPT algorithms $\mathsf{Setup}_{\mathsf{BIFS}}$, $\mathsf{KeyGen}_{\mathsf{BIFS}}$, $\mathsf{Sign}^{\mathsf{B}}_{\mathsf{BIFS}}$, $\mathsf{Sign}^{\mathsf{K}}_{\mathsf{BIFS}}$, and $\mathsf{Verify}_{\mathsf{BIFS}}$, described below:

- The setup algorithm $\mathsf{Setup}_{\mathsf{BIFS}}$ takes a security parameter $\kappa$ and a model of biometric features $\mathscr{F}$ as input and outputs a public parameter $pp_{\mathsf{BIFS}}$.
- The key generation algorithm $\mathsf{KeyGen}_{\mathsf{BIFS}}$ takes a public parameter and a biometric feature $w$ as input and outputs a pair of a helper key $hk_{\mathsf{BIFS}}$ and a veriification key $vk_{\mathsf{BIFS}}$.
- The signing protocol is a two party protocol with $\mathsf{Sign}^{\mathsf{B}}_{\mathsf{BIFS}}$ and $\mathsf{Sign}^{\mathsf{K}}_{\mathsf{BIFS}}$, where $\mathsf{Sign}^{\mathsf{B}}_{\mathsf{BIFS}}$ takes a public parameter $pp_{\mathsf{BIFS}}$, a biometric feature $w'$, and a message $\mu$ as input, and $\mathsf{Sign}^{\mathsf{K}}_{\mathsf{BIFS}}$ a public parameter $pp_{\mathsf{BIFS}}$ and a helper key $hk_{\mathsf{BIFS}}$, and the output is a signature $\sigma$.
- The verification algorithms $\mathsf{Verify}_{\mathsf{BIFS}}$ takes a public parameter $pp_{\mathsf{BIFS}}$, a verification key $vk_{\mathsf{BIFS}}$, a message $\mu$, and a signature $\sigma$ as input and outputs a signal that represents acceptance ($\top$) or rejection ($\bot$) of the message/signature pair.

If a BIFS scheme is correct, it holds that $\top \leftarrow \mathsf{Verify}_{\mathsf{BIFS}}(pp_{\mathsf{BIFS}}, vk_{\mathsf{BIFS}}, \mu, \sigma)$ for every $pp_{\mathsf{BIFS}}$ from $\mathsf{Setup}_{\mathsf{BIFS}}(1^{\kappa}, \mathscr{F})$, $w$ and $w'$ that satisfies $\mathsf{dis}(w, w') \leq t$, $\mu$, $(hk_{\mathsf{BIFS}}, vk_{\mathsf{BIFS}}) \leftarrow \mathsf{KeyGen}_{\mathsf{BIFS}}(pp_{\mathsf{BIFS}}, w)$, and $\sigma \leftarrow \langle \mathsf{Sign}^{\mathsf{B}}_{\mathsf{BIFS}}(pp_{\mathsf{BIFS}}, w', \mu), \mathsf{Sign}^{\mathsf{K}}_{\mathsf{BIFS}}(pp_{\mathsf{BIFS}}, hk_{\mathsf{BIFS}}) \rangle$.

As the security of BIFS, we require unforgeability of the signatures and confidentiality of the biometric features. Due to the space limitation, we sketch the ideas of the definitions. The unforgeability of BIFS is defined through security games in a similar way to that of the two-party signature scheme in [Li17] except that BIFS uses biometric features. The game is defined with an experiment in which an adversary who controls one of the parties $\mathsf{BIFS.Sign}^i$ asks stateful oracles for key generation and signing with the instructions of the other party $\mathsf{BIFS.Sign}^j$ where $j \neq i \in \{\mathsf{B}, \mathsf{K}\}$.

We require different levels of confidentiality for the helper keys compared to the verification keys and signatures, since the second signer keeps the former secret while the latter two may be made public. According to the definition of the secure sketch [DRS04], the helper keys are allowed to decrease a limited amount of the entropy of the underlying biometric features. On the other hand, we require the verification keys and signatures to be independent of the underlying biometric features.

## 4 Proposed scheme

**Construction** We propose a generic construction for BIFS. The proposed construction is designed in a similar idea as the schemes in [Ta19] and based on a linear sketch scheme and a key-homomorphic DS scheme. The key generation algorithm first generates a key pair $(sk_{\mathsf{SIG}}, vk_{\mathsf{SIG}})$ for the DS. $sk_{\mathsf{SIG}}$ is virtually used in the signing protocol. The helper key $hk_{\mathsf{BIFS}}$ is a sketch of $sk_{\mathsf{SIG}}$ with a biometric feature, which hides the $sk_{\mathsf{SIG}}$. Signatures are generated through a two-message protocol. On receiving a sketch generated from a newly and randomly selected differential key $\Delta$ by the first signer $\mathsf{Sign}^{\mathsf{B}}_{\mathsf{BIFS}}$, the second signer $\mathsf{Sign}^{\mathsf{K}}_{\mathsf{BIFS}}$ derives a masked signing key $sk'_{\mathsf{SIG}} = sk_{\mathsf{SIG}} - \Delta$ and signs the message. A signature of $sk'_{\mathsf{SIG}}$ can be transformed into one of $sk_{\mathsf{SIG}}$ through the signature inversion operation by the first signer who knows $\Delta$. This way, the signers can generate a signature without revealing their secret inputs and the virtual key. The verification keys and valid signatures of the proposed scheme are in the same form as those of the underlying DS scheme; thus, the verification algorithm is the same. A formal description is given below:

**Setup:** On input $1^\kappa$ and a model of biometric features $\mathscr{F}$, runs $pp_{\mathsf{SIG}} \leftarrow \mathsf{Setup}_{\mathsf{SIG}}(1^\kappa)$ and $pp_{\mathsf{LS}} \leftarrow \mathsf{Setup}_{\mathsf{LS}}(\mathscr{F}, (\mathscr{K}_{pp_{\mathsf{SIG}}}, +))$ and outputs $pp_{\mathsf{BIFS}} := (pp_{\mathsf{SIG}}, pp_{\mathsf{LS}})$.

**Key generation:** On input $pp_{\mathsf{BIFS}}$ and $w$, runs $(sk_{\mathsf{SIG}}, vk_{\mathsf{SIG}}) \leftarrow \mathsf{KeyGen}_{\mathsf{SIG}}$, computes a sketch $s := \mathsf{Gen}_{\mathsf{LS}}(sk_{\mathsf{SIG}}, w) + w$, and outputs $hk_{\mathsf{BIFS}} := s$ and $vk_{\mathsf{BIFS}} := vk_{\mathsf{SIG}}$.

**Signing:** On input $(w', \mu, vk_{\mathsf{BIFS}} = vk_{\mathsf{SIG}})$ for $\mathsf{Sign}^{\mathsf{B}}_{\mathsf{BIFS}}$ and $hk_{\mathsf{BIFS}} = s$ for $\mathsf{Sign}^{\mathsf{K}}_{\mathsf{BIFS}}$, firstly $\mathsf{Sign}^{\mathsf{B}}_{\mathsf{BIFS}}$ chooses a differential key $\Delta \xleftarrow{U} \mathscr{K}_{pp_{\mathsf{SIG}}}$, computes a sketch $s' := \mathsf{Gen}_{\mathsf{LS}}(\Delta, w')$, and sends $s'$ and $\mu$ to $\mathsf{Sign}^{\mathsf{K}}_{\mathsf{BIFS}}$. Then, $\mathsf{Sign}^{\mathsf{K}}_{\mathsf{BIFS}}$ computes a masked key as $sk'_{\mathsf{SIG}} := \mathsf{Rec}_{\mathsf{LS}}(s, s')$ and returns a temporary signature $\sigma' \leftarrow \mathsf{Sign}_{\mathsf{SIG}}(sk'_{\mathsf{SIG}}, \mu)$. Finally, $\mathsf{Sign}^{\mathsf{B}}_{\mathsf{BIFS}}$ generates a final signature $\sigma \leftarrow \mathsf{SignShift}_{\mathsf{SIG}}(\Delta, \sigma')$ and outputs the final signature if $\top \leftarrow \mathsf{Verify}_{\mathsf{SIG}}(vk_{\mathsf{SIG}}, \mu, \sigma)$ and $\bot$ otherwise.

**Verification:** On input $(vk_{\mathsf{BIFS}} = vk_{\mathsf{SIG}}, \mu, \sigma)$, outputs the result of $\mathsf{Verify}_{\mathsf{SIG}}(vk_{\mathsf{SIG}}, \mu, \sigma)$.

If the distance of the two biometric features $w$ and $w'$ is within the threshold, the masked key reconstructed from $s$ and $s'$ satisfies $sk'_{\mathsf{SIG}} = sk_{\mathsf{SIG}} - \Delta$. Therefore, the temporary signature $\sigma'$ is transformed into a valid signature of the virtual key $sk_{\mathsf{SIG}}$ through the key-homomorphic operation, and thus the correctness of the scheme holds.

**Security** Here we briefly describe the security properties of the proposed scheme. The unforgeability is satisfied based on the EUF-CMA security and key-homomorphism of the underlying DS scheme. The proof can be given through a sequence of games where the last game is reduced to the EUF-CMA security game of the underlying DS scheme. The confidentiality for the helper keys is satisfied since a helper key is a sketch. Also, since the verification keys and signatures are in the same form as those of the underlying DS scheme, they are obviously independent of the biometric features.

**Efficiency analysis** Here we compare the data sizes and computational costs of the proposed BIFS scheme with the DS and FS schemes.

| ([byte]) | Signing key | Helper key | Verification key | Signature |
|---|---|---|---|---|
| DS | $sk$ (32) | - | $vk$ (48) | $\sigma$ (96) |
| FS | - | - | $sketch + vk$ (2096) | $sketch + vk + \sigma$ (2192) |
| BIFS | - | $sketch$ (2048) | $vk$ (48) | $\sigma$ (96) |

Tab. 2: Comparison of data sizes.

The BIFS scheme replaces the signing key of the underlying DS scheme with a biometric feature and a helper key which is a sketch of the signing key, while the verification key and signature are in identical form. On the other hand, although the generic construction of FS in [Ta15] replaces the signing key by only a biometric feature, the verification key and signature additionally contain a sketch to that of the underlying DS scheme. Table 2 compares the data sizes of the DS, FS [Ta15], and BIFS schemes where $sk$, $vk$, and $\sigma$ means a signing key, verification key, and signature of a DS scheme, respectively, and $sketch$ means a sketch of the signing key.

In considering the actual sizes, we assume adopting biometric features of ArcFace [De19] of which features are floating point vectors with 512 dimensions and BLS signature [BLS01] with curve BLS12-381 [Bo17] with the same setting as adopted in Ethereum 2.0. A sketch is a floating point vector of which size is the same as the features and is 2048 bytes. A signing key, verification key, and signature are 32, 48, and 96 bytes, respectively. Therefore, the actual sizes of each data are as shown in brackets in Table 2. The sketches dominate the sizes; thus, the size of the helper key is larger than the signing key of the DS scheme. On the other hand, the signature size, which is expected to be generated more than once for each key pair, is the same as the DS scheme's and significantly smaller than the FS scheme's that contains a sketch in a signature.

| ([$\mu$s]) | Key generation | Sign with key | Sign with feature | Verification |
|---|---|---|---|---|
| DS | $\mathsf{KeyGen}_{\mathsf{SIG}}$ (56) | $\mathsf{Sign}_{\mathsf{SIG}}$ (373) | - - | $\mathsf{Verify}_{\mathsf{SIG}}$ (925) |
| FS | $\mathsf{KeyGen}_{\mathsf{SIG}}$ (56) | - - | $\underline{\mathsf{KeyGen}_{\mathsf{SIG}}} + \mathsf{Sign}_{\mathsf{SIG}}$ (429) | $\mathsf{Verify}_{\mathsf{SIG}} + \underline{\mathsf{VKShift}_{\mathsf{SIG}}}$ (980) |
| BIFS | $\mathsf{KeyGen}_{\mathsf{SIG}}$ (56) | $\mathsf{Sign}_{\mathsf{SIG}}$ (373) | $\underline{\mathsf{SignShift}_{\mathsf{SIG}}}$ (373) | $\mathsf{Verify}_{\mathsf{SIG}}$ (925) |

Tab. 3: Comparison of computational costs.

On the computational costs, we compare the required operations of the underlying DS scheme as in Table 3. Since the computational costs of the DS schemes with group operations are significantly larger than that of the LS schemes described in [Ta19] with real number operations, here we consider the former only. Underlines represent the additional costs compared to that of the DS scheme.

As shown in Table 3, the cost for the signing in the BIFS scheme is increased by the cost for $\mathsf{SignShift}_{\mathsf{SIG}}$. In detail, $\mathsf{SignShift}_{\mathsf{SIG}}$ is executed by $\mathsf{Sign}^{\mathsf{B}}_{\mathsf{BIFS}}$ and $\mathsf{Sign}_{\mathsf{SIG}}$ by $\mathsf{Sign}^{\mathsf{K}}_{\mathsf{BIFS}}$.

The verification costs are the same for the BIFS and DS schemes. On the other hand, the FS scheme requires additional costs for both signing and verification.

We added $\mathsf{SignShift}_{\mathsf{SIG}}$ and $\mathsf{VKShift}_{\mathsf{SIG}}$ to the BLS signature implementation published in [Mi] and executed on an Ubuntu 18.04 machine with Intel Core i7-8700 3.2 GHz CPU and 16 GB DDR RAM. The average running times for Ethereum 2.0 parameters of $\mathsf{KeyGen}_{\mathsf{SIG}}$, $\mathsf{Sign}_{\mathsf{SIG}}$, $\mathsf{Verify}_{\mathsf{SIG}}$, $\mathsf{VKShift}_{\mathsf{SIG}}$, and $\mathsf{SignShift}_{\mathsf{SIG}}$ are 56, 373, 925, 55 and 373 $\mu$s, respectively. Table 3 shows the running times with brackets. While singing for the BIFS scheme takes about twice the time compared to the normal BLS signature and the FS scheme, the time is less than 1 ms. The verification time for the BIFS scheme is the same as the BLS signature, which is slightly faster than the FS scheme.

## 5 Discussion

One of the challenges of DS is the difficulty of managing signing keys. Signing keys can be lost or stolen, since they do not have essential bindings with their owners. As a countermeasure, the eIDAS Regulations and others recognize remote signature services, in which a service provider takes custody of the owner's keys and generates signatures according to the owners' instructions, as having the same effect as the services where the owners manage keys. Similarly, for applications such as cryptocurrencies, many custodial wallet services in which service providers take custody of owners' keys are being deployed. While custodial services make it harder to lose signing keys, they do not solve the problem of being stolen; the issue has just shifted to user authentication in the service.

Thanks to the characteristics of biometric information, FS solves the binding problems. However, FS has two fundamental problems: A single biometric characteristic cannot be used as multiple keys since only biometric features play roles of signing keys, and verification keys and signatures are biometric-dependent since the fuzziness can be absorbed only in the verification phase. Especially, the verification keys and signatures of the FS schemes [Ta19] contains sketches which leak a limited amount of information of the biometric features. For example, cryptocurrencies and other applications on the public chain allow individuals to use multiple keys, and signatures are publicly verified by a third party. Thus, it might be challenging to utilize FS as a replacement for existing signature services.

With good use of the helper keys, BIFS solves the two fundamental problems of FS. On the other hand, the requirement to have access to the helper keys can be a usability disadvantage of BIFS compared to FS. One of the possible ways for helper key management is for the service provider to manage them in the same way as signing keys in custodial services. In this case, although an attacker could steal the helper key, signatures cannot be generated from the helper key without the biometric information of the owner. Another possible method is for the user to manage helper keys by storing them on a user's device such as an IC card or token. If multiple helper keys are managed in one device, the user's burden is insignificant. Since we can implement the proposed method to have the same signature and verification key format as the BLS signature used in Ethereum 2.0 and others, it may be possible to replace the existing service without changing the verification algorithm.

## References

[BLS01] Boneh, Dan; Lynn, Ben; Shacham, Hovav: Short Signatures from the Weil Pairing. In (Boyd, Colin, ed.): Advances in Cryptology - ASIACRYPT 2001, 7th International Conference on the Theory and Application of Cryptology and Information Security, Gold Coast, Australia, December 9-13, 2001, Proceedings. volume 2248 of Lecture Notes in Computer Science. Springer, pp. 514–532, 2001.

[Bo17] Bowe, Sean: BLS12-381: New zk-SNARK elliptic curve construction. https://electriccoin.co/blog/new-snark-curve/, March 2017.

[De19] Deng, Jiankang; Guo, Jia; Xue, Niannan; Zafeiriou, Stefanos: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. Computer Vision Foundation / IEEE, pp. 4690–4699, 2019.

[DRS04] Dodis, Yevgeniy; Reyzin, Leonid; Smith, Adam D.: Fuzzy Extractors: How to Generate Strong Keys from Biometrics and Other Noisy Data. In (Cachin, Christian; Camenisch, Jan, eds): Advances in Cryptology - EUROCRYPT 2004, International Conference on the Theory and Applications of Cryptographic Techniques, Interlaken, Switzerland, May 2-6, 2004, Proceedings. volume 3027 of Lecture Notes in Computer Science. Springer, pp. 523–540, 2004.

[DS19] Derler, David; Slamanig, Daniel: Key-homomorphic signatures: definitions and applications to multiparty signatures and non-interactive zero-knowledge. Des. Codes Cryptogr., 87(6):1373–1413, 2019.

[Ka21] Katsumata, Shuichi; Matsuda, Takahiro; Nakamura, Wataru; Ohara, Kazuma; Takahashi, Kenta: Revisiting Fuzzy Signatures: Towards a More Risk-Free Cryptographic Authentication System based on Biometrics. In (Kim, Yongdae; Kim, Jong; Vigna, Giovanni; Shi, Elaine, eds): CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021. ACM, pp. 2046–2065, 2021.

[Li17] Lindell, Yehuda: Fast Secure Two-Party ECDSA Signing. In (Katz, Jonathan; Shacham, Hovav, eds): Advances in Cryptology - CRYPTO 2017 - 37th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 20-24, 2017, Proceedings, Part II. volume 10402 of Lecture Notes in Computer Science. Springer, pp. 613–644, 2017.

[Ma16] Matsuda, Takahiro; Takahashi, Kenta; Murakami, Takao; Hanaoka, Goichiro: Fuzzy Signatures: Relaxing Requirements and a New Construction. In (Manulis, Mark; Sadeghi, Ahmad-Reza; Schneider, Steve A., eds): Applied Cryptography and Network Security - 14th International Conference, ACNS 2016, Guildford, UK, June 19-22, 2016. Proceedings. volume 9696 of Lecture Notes in Computer Science. Springer, pp. 97–116, 2016.

[Mi] Mitsunari, Shigeo: BLS threshold signature. https://github.com/herumi/bls.

[Ta15] Takahashi, Kenta; Matsuda, Takahiro; Murakami, Takao; Hanaoka, Goichiro; Nishigaki, Masakatsu: A Signature Scheme with a Fuzzy Private Key. In (Malkin, Tal; Kolesnikov, Vladimir; Lewko, Allison Bishop; Polychronakis, Michalis, eds): Applied Cryptography and Network Security - 13th International Conference, ACNS 2015, New York, NY, USA, June 2-5, 2015, Revised Selected Papers. volume 9092 of Lecture Notes in Computer Science. Springer, pp. 105–126, 2015.

[Ta19] Takahashi, Kenta; Matsuda, Takahiro; Murakami, Takao; Hanaoka, Goichiro; Nishigaki, Masakatsu: Signature schemes with a fuzzy private key. Int. J. Inf. Sec., 18(5):581–617, 2019.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# Face verification explainability heatmap generation using a vision transformer[1]

Ricardo Correia, Fernando Pereira and Paulo L. Correia[2]

**Abstract:** Explainable Face Recognition (XFR) is a critical technology to support the large deployment of learning-based face recognition solutions. This paper aims at contributing to the more transparent usage of Vision Transformers (ViTs) for face verification (FV) tasks, by proposing a novel approach for generating FV explainability heatmaps, for both positive and negative decisions. The proposed solution leverages on the attention maps generated by a ViT and employs masking techniques to create masks based on the highlighted regions in the attention maps. These masks are applied to the pair of faces, and the masking technique with most impact on the decision is selected to be used to generate heatmaps for the probe-gallery pair of faces. These heatmaps offer valuable insights into the decision-making process, shedding light on the most important face regions for the verification outcome. The key novelty of this paper lies in the proposed approach for generating explainability heatmaps tailored for verification pairs in the context of ViT models, which combines the ViT attention maps regions of the probe-gallery pair to create masks that allow evaluating those region´s impact on the verification decision for both positive and negative decisions.

**Keywords:** Explainable face recognition, vision transformer, face verification heatmaps

## 1 Introduction

The increasing adoption of artificial intelligence (AI) tools, and deep learning (DL) models in particular, for multiple computer vision tasks, impacts users and their lives, thus accentuating the need for explainable artificial intelligence (XAI). In the context of face recognition (FR), this type of technology is known as explainable face recognition (XFR), acknowledging concerns about the lack of transparency of many FR models. In this context, understanding how a model works, especially when it fails, is crucial for improving and developing more effective FR solutions, and increase its societal acceptance. For FV, it is vital to know why impostors are wrongly validated or legitimate users are denied access.

A post-hoc XFR tool is applied after the FR model has made its decision, not changing

[1] This work has been partially supported by the European CHIST-ERA program via the *French National Research Agency (ANR) within the XAIface project (grant agreement CHIST-ERA-19-XAI-011)* and by FCT/MEC under the project UID/50008/2020.

[2] Instituto de Telecomunicações; Instituto Superior Técnico – Universidade de Lisboa; Av. Rovisco Pais, 1, 1049-001 Lisboa, Portugal, ricardo.r.nobre.correia@tecnico.ulisboa.pt, fp@lx.it.pt and plc@lx.it.pt

the model, while aiming at providing insights on how the model arrived at its decision. Various post-hoc XFR tools have been developed to enhance the explainability of FV decisions [Me22], [PDS18], [MM22], which can be categorized based on how they extract information from the FR model into propagation-based or perturbation-based tools. Propagation-based XFR tools leverage specific properties of the model being explained by considering the internal structure of the model. Perturbation-based tools make changes to the input face, e.g., masking, or altering specific input features, to assess their impact on the FV model's decision, not considering the inner working of the model.

ViTs [Va17] have recently emerged as a promising tool also for FR purposes [ZD21a]. The ViT self-attention mechanism can contribute to enhance FR explainability since propagation-based tools can explore it to provide insights about the decision made, and generate attention maps expressing the importance of different input image patches for the ViT created embeddings. Even if a ViT is often trained with a classification logic, it can still be used for FV tasks to compute embeddings for both the probe and gallery images and compare those embeddings to determine their similarity. While the ViT attention maps provide insights about the face salient regions with more influence in obtaining the desired output class, they are not necessarily appropriate to explain a FV decision, where the similarity of two faces is compared. As such, these attention maps provide important information, but they cannot be directly taken as FV explainability heatmaps. On the contrary, perturbation based XFR tools, such as Average Removal/Aggregation (AVG) [MM22] and MinPlus [Me22], do not consider the model internal structure but rather focus on the task decision, and therefore they directly output FV explainability heatmaps.

In this context, this paper proposes a novel XFR post-hoc tool for creating FV explainability heatmaps, leveraging the advantages of ViT propagation-based tools and their attention maps. Positive and negative FV decisions are treated differently since for the first case the goal is to highlight those facial regions contributing most to the similarity between probe and gallery images, while for the latter case the regions contributing to differentiate individuals should be highlighted. The key novelty of the proposed solution lies on the way attention maps regions for the probe-gallery pair are combined to create masks that allow evaluating those region's impact on the ViT FV decision, and then to generate effective FV explainability heatmaps.

This paper is structured as follows: Section 2 provides a brief overview of the state-of-the-art on explainable FV. Section 3 presents the proposed FV explainability XFR post-hoc tool, exploiting the ViT attention maps to derive ViT FV explainability heatmaps. Section 4 reports and discusses the results and findings obtained with the proposed XFR tool, comparing with state-of-the-art methods and highlighting the advantage of the proposed tool to also create explainability heatmaps for negative FV decisions. Section 5 presents final remarks and outlines future research directions.

## 2 Brief review on face verification explainability

Explainability tools play a crucial role to provide insights on the inner working of FV models. While a few works propose XFR ante-hoc tools, i.e. intrinsically interpretable models that inherently provide transparency in the decision-making process [WBT22][JZ21], the main literature focus has been on XFR post-hoc tools. As discussed in the Introduction, post-hoc FV explainability tools (and FR tools) can be categorized as perturbation-based and propagation-based tools. Examples of the former include Local Interpretable Model-agnostic Explanations (LIME) [RSG16], Randomized Input Sampling for Explanation (RISE) [PDS18], AVG [MM22] and MinPlus [Me22]. These tools can be applied to any black-box model without changing the model architecture. LIME works by randomly selecting super-pixels and training a weighted model to determine their importance. RISE, MinPlus and AVG perturb different face regions to measure their effect on the model´s decision, thus creating a FV explainability heatmap.

Propagation-based post-hoc tools based on the ViT offer a more direct way to understand the internal functioning of a ViT model, by leveraging its attention mechanism [CGW20]. Examples of such tools applied to ViT include Rollout [AZ20], Gradient-weighted Class Activation Mapping (Grad-CAM) [Se17], Layer-wise Relevance Propagation (LRP) [Bi16], and a ViT-LRP tool [CGW20] which provide insights into the probe regions that are essential to its embedding representation. Rollout uses the attention matrices computed for the various attention layers to generate an attention map. These attention matrices represent the learned attention weights that capture the relationships between the different patches within the probe face image and are used to derive an attention map, visually representing the importance of each patch. Grad-CAM uses the attention matrices and combines them based on their gradients with respect to the output class to generate an attention map. ViT-LRP adapts LRP to the ViT architecture, propagating the output class relevance scores to the attention matrices of the various attention layers, enabling the identification of the most important probe patches for the obtained embedding representation. Unlike perturbation-based tools, propagation-based ViT tools do not directly provide FV explainability heatmaps, but rather attention maps that highlight important regions in the context of image classification.

## 3 Proposed face verification explainability tool

This section proposes a novel FV explainability heatmap generation tool that provides insights into the decision-making process by highlighting the regions that contribute the most to a positive or a negative FV decision. The architecture for the proposed FV explainability tool is depicted in Fig. 1, and its key modules to explain both types of FV decisions are detailed in the following.

### 3.1 Architecture and walkthrough

The proposed explainability tool complements a *Face Verification Pipeline* using a *ViT model* to perform feature extraction and produce embeddings describing the input probe and gallery images, see overall architecture in Fig. 1. The *Attention Map Generation* module creates attention maps for the probe ($P$) and gallery ($G$) face images using a ViT propagation-based post-hoc tool. These classification-focused attention maps capture the regions that most influence the embedding representation of each input image, probe and gallery. Following the methodology outlined in [ZD21a], the *Face Verification Decision* is based on the distance, $d$, between the embeddings computed for the probe and gallery images using the square Frobenius norm, $d = \| e_P - e_G \|_F^2$, where $\{e_P, e_G\} \in \mathbb{R}^n$ are the probe and gallery embeddings, respectively. If this distance exceeds the decision threshold, a negative FV decision is taken. To find the optimal threshold, cross-validation is performed on the FV dataset, following the same methodology used by [ZD21a].

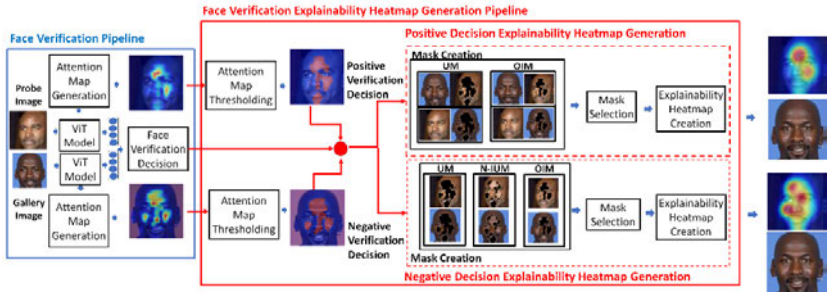



Fig. 1 Proposed face verification explainability architecture

The proposed *Face Verification Heatmap Generation Pipeline* starts with an *Attention Map Thresholding* module, applying Otsu thresholding [Ot79] to obtain a binary thresholded version of the ViT attention map, highlighting the most salient areas. Then, based on the FV decision, a different strategy for *Mask Creation* is adopted, as detailed in Section 3.2. The goal is to determine which input face areas are more relevant for the FV decision. While several masking techniques are discussed, the *Mask Selection* module selects the mask considered as most impactful for the FV decision. This involves feeding new pairs of images, obtained by applying the created masks to the probe and gallery face images, to the *Face Verification Pipeline*, and checking how the distance between the original and new embeddings changes. A selection metric is proposed to assess the impact of each candidate masking technique and guide the selection of the best mask creation technique. Finally, the *Explainability Heatmap Creation* module creates the FV explainability heatmap by considering the selected masking technique and the original probe and gallery attention maps. To get smoother heatmaps, they are filtered, applying an 8×8 dilation followed by a Gaussian filter of size 56×56 and 6.6 variance. Finally, for visualization purposes, the heatmap (with the warmer colours representing the more important regions) is overlaid with the probe face image luminance.

### 3.2 Positive decision explainability heatmap generation

The proposed FV explainability heatmap generation tool for positive verification decisions takes as input the probe and gallery face images and performs four main steps: (i) attention map thresholding; (ii) mask creation, where several alternative solutions are considered; (iii) mask selection; and (iv) explainability heatmap creation. The *Attention Map Thresholding* module converts the attention maps created by the ViT propagation-based tool into a binary mask as described above. *Mask creation* considers alternative masking techniques for creating two image pairs, PP – positive probe, and PG – positive gallery, to be submitted to the FV pipeline as part of the *Mask Selection* module. The PP image pair includes the probe face image and a masked version of the gallery face image while the PG pair includes the gallery face image and a masked version of the probe face image, see Fig. 1. By excluding the regions identified as important in the corresponding attention maps, it is expected that the similarity of the new PP and PG pairs will decrease regarding the original probe and gallery face images. Two masking techniques are proposed for the positive FV case:

- **Only Intersection Masking (OIM)** – The thresholded attention maps computed in the *Attention Map Thresholding* module for both input images are combined to create a single mask, corresponding to the intersection of both the probe and gallery thresholded attention masks, thus including only the regions highlighted as important in both the probe and gallery thresholded attention maps.
- **Unified Masking (UM)** – The thresholded attention maps computed within the *Attention Map Thresholding* module for both input images are combined to create a single mask, corresponding to the union of both the probe and gallery thresholded attention masks, thus including the important regions from both face images.

For each masking technique, the generated mask is applied to both the probe and gallery images and the percentage of removed area is denoted as *RA*. The *Mask selection* module takes the PP and PG pairs generated with each mask creation technique and feeds them to the *Face Verification Pipeline*. The distances between the corresponding embeddings, $d_{PG}$ and $d_{PP}$ , against the original FV embeddings distance, $d$, are then computed; since the areas identified as important in the attention maps were excluded, the distance between embeddings is expected to increase. The selection metric used to evaluate the effectiveness of the mask creation techniques considers: (i) the variation of the distance between the embeddings prior and after the masking, which is related to the relevance of the removed areas; and (ii) the size of the removed areas, to ensure that techniques masking out larger image areas do not receive an unfair advantage, as larger masked areas tend to result in a larger increase in the embeddings distance. The proposed selection metric for positive FV decision cases (SM+) is computed as:

$$\text{SM+} = ((d_{PG} - d)/d) / 2RA + ((d_{PP} - d)/d) / 2RA , \quad (1)$$

where $d_{PG} = \| e_G - e_{MP} \|_F^2$ and $d_{PP} = \| e_P - e_{MG} \|_F^2$ correspond to the distance between

embeddings for the PG and PP image pairs, and $e_{MP}$ and $e_{MG}$ correspond to the embeddings resulting from the *ViT model* module after masking the probe (*P*) and gallery (*G*) images, respectively. The best mask creation strategy is the one leading to a larger SM+ value, as this metric captures the contribution of the masked areas to the FV decision. Finally, the *Explainability Heatmap Creation* module considers the areas of the selected mask as those contributing the most to explain the FV decision. The FV explainability heatmap values are obtained from the original attention maps for the selected area. The OIM heatmap considers only the regions common to both attention maps and, as such, it is generated by computing the average values of the attention maps for these shared regions. The UM heatmap follows the same approach, except when only one attention map contributes to a part of the UM mask, in which case its value is directly used. Finally, the filtering process described in Section 3.1 is applied.

### 3.3 Negative decision explainability heatmap generation

The proposed explainability heatmap generation tool for negative FV decisions differs from the positive case on the mask creation process and the selection metric. *Mask creation* aims to create one image pair, NPG – negative probe and gallery, to be fed to the *Face Verification Pipeline* as part of the *Mask Selection* module, consisting of masked versions of both the probe and gallery face images. The masking techniques aim at removing the regions that contribute to the differentiation between individuals and, by doing so, it is expected that the similarity of the new NPG pairs will increase regarding the original probe-gallery similarity. In addition to the masking techniques proposed in Section 3.2 (OIM and UM), the *Non-Intersecting Unified Masking (N-IUM)* technique combines the thresholded attention maps for the probe and gallery face images to form a single mask, including the union minus the intersection of both masks, to keep regions that are considered important in only one of the attention maps.

The *Mask Selection* module feeds the NPG image pair to the *Face Verification Pipeline* and computes the $d_{NGP}$ distance between the new embeddings, comparing it against the original FV embeddings distance, $d$. By masking out areas identified as important in the attention maps, a smaller distance between embeddings is expected. For the negative FV decisions, the proposed mask technique selection metric (*SM-*) is:

$$\text{SM-} = ((d - d_{NGP})/d)/RA, \tag{2}$$

where $d_{NGP} = \| e_{MG} - e_{MP} \|_F^2$ is the squared Frobenius norm distance between $e_{MP}$ and $e_{MG}$, which correspond to the embeddings resulting from the *ViT model* module after masking the probe (*P*) and gallery (*G*) images, respectively. The best mask creation technique produces the larger *SM-* value, corresponding to a larger impact on the FV decision with a larger reduction of the $d_{NGP}$ value weighted by *RA*. The *Explainability Heatmap Creation* process is completed with the same filtering described in Section 3.1.

# 4 Results and discussion

The ViT code available from [CGW20] was used in this paper for training and testing the ViT model as specified in [ZD21a]. Training used the large-scale database MS-Celeb-1M [Gu16], which contains 5.3 million images of 93,431 celebrities, with the CosFace loss function [Wa18]. For evaluation purposes, several FR datasets were used, notably LFW [Hu07], Similar-looking LFW (SLLFW) [De16], Cross-Age LFW (CALFW) [ZDH17], Cross-Pose LFW (CPLFW) [ZD18] and Transferable Adversarial LFW (TALFW) [ZD21b]. The LFW dataset is split into 10 subsets of image pairs, each with 300 positive and 300 negative pairs; the LFW variants allow testing in more challenging scenarios.

**Masking technique selection**

The performance of the various proposed masking techniques when integrated in the proposed FV explainability heatmap generation tool is evaluated by calculating the average value of *SM+* for each masking technique (OIM, UM) across the positive decision probe-gallery pairs for each dataset, and the average value of *SM-* for each masking technique (OIM, UM, N-IUM) for the negative decision probe-gallery pairs for each dataset. A summary of the results obtained is included in Tab. 1. ViT-LRP [CGW20] is used to generate the attention maps, and all FV pairs of each dataset are considered.

Tab. 1 Masking techniques evaluation results for positive and negative face verification decisions.

| FV decision | Masking technique | LFW | TALFW | CALFW | SLLFW | CPLFW |
|---|---|---|---|---|---|---|
| Positive | OIM | **10.00** | **5.93** | **4.97** | **9.72** | **3.89** |
| | UM | 4.80 | 2.82 | 2.59 | 5.1 | 1.49 |
| Negative | OIM | **1.84** | **2.01** | **2.40** | **2.84** | **2.62** |
| | UM | 1.56 | 1.52 | 1.69 | 1.95 | 1.39 |
| | N-IUM | 1.70 | 1.61 | 1.82 | 2.13 | 1.48 |

The results in Tab. 1 show that, the OIM masks perform the best in capturing the important features for explaining both positive and negative decisions. For the alternative ViT attention map generation tools, notably No-Grad ViT-LRP [CGW20] (a variant of ViT-LRP not integrating gradient information from the attention matrices) and Rollout [AZ20], they both provide consistent results with the ViT-LRP results reported in Table 1 in terms of the best mask creation techniques. In this context, the *Mask Selection* module in the pipeline may not be necessary, as the results clearly indicate that OIM is the most effective masking technique for both types of FV decision. Therefore, to provide an explanation for a FV decision, No-Grad ViT-LRP OIM, ViT-LRP OIM or Rollout OIM should be used.

**Evaluation of face verification explainability heatmaps**

To evaluate the FV explainability heatmaps generated by the proposed explainability tool against the state-of-the-art, a set of perturbation tests were conducted by progressively

masking the probe pixels in descending relevance order, in line with the approach from [CGW20]. Tests were applied to 1000 LFW true positive pairs, as the state-of-the-art explainability tools only explain positive decisions. Using No-Grad ViT-LRP, ViT-LRP and Rollout in the *Attention Map Generation* module and the OIM masking technique, these are compared against the MinPlus, LIME and RISE benchmarks, using Recall as the evaluation metric - see Fig. 2 (left). The tool leading to a faster decrease in Recall is considered the most effective. For negative decisions, only the proposed explainability tool is evaluated as MinPlus, LIME and RISE focus only on similar regions to explain positive decisions. Tests were performed for 1000 LFW true negative pairs, with both the probe and gallery images being progressively masked based on their respective FV explainability heatmaps; here, the True Negative Rate (TNR) metric was used for evaluation. By gradually masking regions that contribute to distinguishing individuals, the expectation is that TNR decreases and the tool leading to a faster decrease will be the most effective, see Fig. 2 (right).

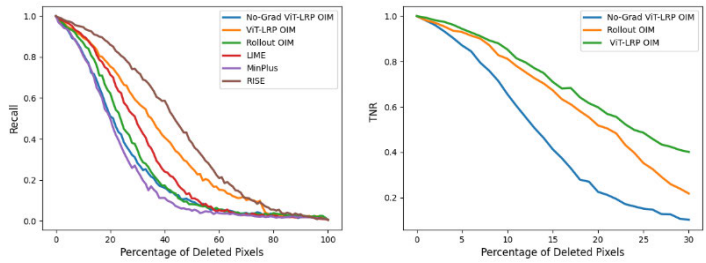



Fig. 2 Post-hoc tools evaluation: left) Recall evaluation; right) TNR evaluation

The Recall results in Fig. 2 show that the proposed No-Grad ViT-LRP OIM tool achieves the best explainability performance for the positive FV decisions along with MinPlus. For negative decisions, the TNR results show that the proposed No-Grad ViT-LRP OIM tool also performs the best. In summary, the proposed No-Grad ViT-LRP OIM tool allows generating effective explainability heatmaps for both positive and negative decisions, a major advantage over perturbation-based tools. Fig. 3 includes a few examples of FV explainability heatmaps for both types of FV decisions.

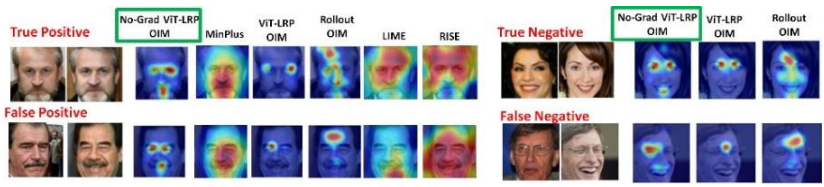



Fig. 3 Face verification explainability heatmap examples

## 5 Conclusions and future work

This paper proposes a novel ViT post-hoc FV explainability tool based on heatmaps, achieving comparable performance to perturbation-based tools for positive decisions. The best solution uses the original attention maps generated by No-Grad ViT-LRP [CGW20] and the OIM masking technique. A key advantage of the proposed propagation-based approach regarding the perturbation-based explainability tools is its ability to effectively explain both positive and negative FV decisions. As future work, the goal is to further leverage the ViT attention mechanisms to create a FV ante-hoc tool.

## References

[AZ20] Abnar, S.; Zuidema W.: Quantifying Attention-flow in Transformers. Annual Meeting of the Association for Computational Linguistics, CoRR, 2020.

[Bi16] Binder, A. *et. al*.: Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers. International Conference on Artificial Neural Networks, 63-71, 2016.

[CGW20] Chefer, H.; Gur S.; Wolf, L.: Transformer Interpretability Beyond Attention Visualization. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Los Alamitos, CA, USA, 782-791, 2021.

[De16] Deng W. *et. al.*: Fine-grained Face Verification: FGLFW Database, Baselines, and Human-DCMN Partnership. Pattern Recognition, 2016.

[Gu16] Guo, Y. *et. al*.: Ms-Celeb-1M: A Dataset and Benchmark for Large-scale Face Recognition. European Conference on Computer Vision, 2016.

[Hu07] Huang G. *et. al.*: Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, University of Massachusetts, Amherst, 2007.

[JZ21] Jiang, H.; Zeng D.: Explainable Face Recognition Based on Accurate Facial Composition. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Montreal, BC, Canada, 1503-1512, 2021.

[Me22] Mery, D.:True Black-Box Explanation in Facial Analysis. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 1595-1604, 2022.

[MM22] Mery, D.; Morris, B.: On Black-Box Explanation for Face Verification. 2022 IEEE/CVF Winter Conference on Application of Computer Vision (WACV), Waikoloa, Hi, USA, 1194-1203, 2022.

[Ot79] Otsu, N.; A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man, and Cybernetics, 62–66, 1979.

[PDS18] Petsiuk, V.; Das A.; Saenko, K.: Randomized Input Sampling for Explanation of Black-box Models. CoRR, 2018.

[RSG16] Ribeiro M.; Singh S.; Guestrin C.: Why Should I Trust You?: Explaining the Predictions of Any Classifier. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics, San Diego, California, USA, 97-101, 2016.

[Se17] Selvaraju, R. *et. al.*: Grad-cam Visual Explanation from Deep Networks via Grandient-based Localization. 2017 IEEE International Conference on Computer Vision (ICCV)., Venice, Italy, 618-626, 2017.

[Va17] Vaswani, A. *et. al.*: Attention Is All You Need. In (I. Guyon and U. Von Luxburg and S. Bengio and H. Wallach and R. Fergus and S. Vishwanathan and R. Garnett): Advances in Neural Information Processing Systems. Curran Associates, Inc., 2017.

[Wa18] Wang H. *et. al.*: Cosface: Large Margin Cosine Loss for Deep Face Recognition. 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 5236-5274, 2018.

[WBT22] Winter, M.; Bailer W.; Thallinger G.: Demystifying Face-Recognition with Locally Interpretable Boosted Features (LIBF). 2022 10th European Workshop on Visual Information Processing (EUVIP), Lisbon, Portugal, 1-6, 2022.

[ZD21a] Zhong, Y.; Deng W.: Face Transformer for Recognition., CoRR, 2021.

[ZD21b] Zhong Y.; Deng W.: Towards Transferable Adversarial Attack Against Deep Face Recognition. IEEE Transactions of Information Forensics and Security, 1452-1466. 2021.

[ZDH17] Zheng T.; Deng W.; Hu J.: Cross-age LFW: A Database for Studying Cross-Age Face Recognition in Unconstrained Environments. CoRR, 2017.

[ZD18] Zheng T.; Deng W.: Cross-Pose LFW: A Database for studying Cross-Pose Face Recognition in Unconstrained Environments, CoRR, 2018.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# Comparison of two architectures for text-independent verification after character-unaware text segmentation

Maria De Marsico [1] Mohammadreza Shabani[2]

**Abstract:** This paper compares the performance of two popular CNN architectures, ResNet-50 and MobileNetV2, fine-tuned for text-independent writer verification. The used benchmark is IAM dataset. The further contributions are an easy and fast sub-region cropping for robust model training, and a biometrics-oriented performance evaluation. The preliminary results are encouraging.

**Keywords:** Text-independent writer verification, ResNet-50, MobileNetV2

## 1 Introduction

Automatic handwriting recognition and writer verification/identification have attracted great interest due to their practical applications. The former supports antique document digital transcription, preservation, and literary analysis. The latter is rather suited for authentication, security, and forensic analysis. The possibly overlapping used features achieve different goals, as in speech vs speaker recognition. Handwriting style is a challenging behavioral biometric trait due to the high variability in individual writing, especially across time. Real-time online recognition also exploits dynamic features. Offline recognition only exploits the graphical sign once the text is complete, through the extraction of visual-spatial features. Writer recognition can be either text-dependent or text-independent. The former exploits stored prototypes of text written by the subjects to recognize. A prototype is represented in many cases by the signature and must be repeated and compared with the reference template during recognition. We tackle offline text-independent verification. Compared to signature verification, it is more challenging since it encompasses a broader scope and covers identity verification through all forms of handwritten text.
Traditional approaches to handwritten text and writer recognition have relied on manual feature extraction [WTB14, Ni15, Ra08]. The advent of deep learning algorithms also influences this field. This work compares the performance of fine-tuning of two pre-trained architectures, namely ResNet-50 and MobileNetV2, for writer recognition. The contributions of this work can be summarized as follows:
- most literature works propose ad-hoc architectures; this work explores the use of pre-trained CNNs with some fine-tuning after data augmentation;
- pre-processing and sample segmentation are generally based on character- or line-based

[1] Sapienza University of Rome, Department of Computer Science, Via Salaria 113, 00198 - Rome, Italy, demarsico@di.uniroma1.it
[2] Sapienza University of Rome, Department of Computer Science, Via Salaria 113, 00198 - Rome, Italy, shabani.1966731@studenti.uniroma1.it

constraints; in this work segmentation constraints are completely relaxed; the clear advantage of the paper is that the analysis is not restricted to a specific line, but can be any part of the written sheet of paper, including multiple lines;
- the machine learning-oriented performance measure via overall accuracy in a closed setting using a softmax layer is well suited for generic recognition/classification tasks and comparable with Rank-1 identification rate; this work adopts standard biometric evaluation metrics for verification; three related aspects are especially relevant: 1) the use of data from different sessions for training and testing, to evaluate the system in a real-world scenario where the collection of reference data is separated in time from probe submission; 2) the preliminary evaluation of the performance when handling unseen data, meaning that, in the testing phase, the templates to compare, both in the system gallery and in the probe set, are represented by the model-computed embeddings, which can be obtained also for subjects not included in the training set; 3) the ability to perform a thorough analysis of operation thresholds to better support both authentication tasks and forensic verification;
- experiments exploited different distance/similarity measures to further assess the generalizability of the approach and the robustness of obtained embeddings.

## 2 Related work

Several available surveys deal with off-line text-independent writer recognition, e.g., [XLW17]; languages with different writing characteristics, i.e., Arabic, Chinese, and English are considered in [TSR17]; [Di19] is a review on signature verification; a more general survey can be found, e.g., in [RNR19]. We only mention here some recent works dealing with text-independent handwriting that lend themselves to some comparison with our approach.
The paper [Ch19] compares approaches based on hand-crafted features derived from Local Binary Patterns (LBP), Local Ternary Patterns (LTP), and Local Phase Quantization (LPQ); these are applied to connected components in a sample document used for writer identification. A simple nearest neighbor classifier (1-NN) with Hamming distance measure is trained to identify the writer according to the similarity of written documents. The writer of the probe sample is recognized via the most similar sample in the gallery. In the experiments on IAM dataset (the same used here), a maximum of 14 randomly-selected text line images is used per writer. A random selection of 60% of the text line images makes the training set while 40% is used as the testing set. It is not clear how often the text lines for training and testing come from the same samples, being acquired in the same lapse of time. This overlooks the fact that handwriting is a behavioral trait affected by even small time differences. The identification seems closed set, with no dissimilarity threshold for the recognition. Therefore, the unseen data could not be correctly classified.
The CNN used in [Ng19] is first trained to extract local features, from both the whole square regions containing generally isolated character images and their sub-regions. Randomly sampled tuples of images are used to train to aggregate the extracted local features of tuple images to form global features. The strategy is applied to both Japanese Kanji handwriting and to text lines in IAM dataset. Since IAM dataset contains a strongly unbalanced set of pages per user, the authors divide the pages for the writers with only one page into two halves (the first for training/validation and the second for testing). Regarding

the writers with two or more pages, only one page is used for training/validation and one page for testing. Hence, 350 writers have a half page and 300 writers have one page for training. Some imbalance remains, with the already mentioned additional issue that more than half of subjects' training and testing samples come from the same document, i.e., they are acquired in the same time. A final softmax layer returns the identity prediction.
The paper [JJ20] adopts an extended version of ResNet-50 joining deep residual networks and a traditional handwriting descriptor. The authors exploit text cropping, but, differently from our proposal (see Section 3.2), it is based on text lines, i.e., the extracted patches are almost equal or less than a word. In the approach proposed in our paper, cropping is based on rectangular areas possibly spanning more lines, therefore taking also into account both horizontal and vertical co-articulation of the written sign, without any further constraint. Further differences are related to the splitting in training and testing sets of IAM dataset. The authors of [JJ20] divide all text lines into two subsets and then extract 1500 training patches from the first subset and 300 test patches from the second subset per writer. This seems to cause a number of patches per writer in the training and testing sets to possibly come from the same document. For this reason, it is possible to draw similar considerations as for [Ng19]. Performance is measured in terms of accuracy using a softmax layer.

## 3 Experimental setup

We re-trained MobileNetV2 [Sa18] and ResNet-50 [He16]. They represent two different classes of CNNs. They support different uses with different accuracy. ResNet-50 obtains the best performance, but the lighter MobileNetV2 lends itself to mobile applications. We used the pre-trained models from Keras API [3]. A novel unconstrained segmentation of written forms is applied to IAM dataset (Section 3.1) to obtain the exploited samples for fine-tuning and testing (Section 3.2), and three different data augmentation strategies were investigated for more robust models (Section 3.3). For each round of experiments, each architecture includes a softmax layer to readily compare the classification with the ground truth. After building the model, differently from the works mentioned in Section 2, we drop the softmax layer to extract the embeddings, and then use them as templates in all possible biometric mated and non-mated comparison trials. The compared samples never come from training forms. The performance achieved with different similarity/distance measures assesses the robustness of the extracted embeddings. A following round of experiments (Section 3.2) adopts a train/test partition where some users only appear in the test set, to test the generalizability of results useful for real-world applications.

### 3.1 The dataset

This work exploited the IAM Handwriting Database [4], which contains forms of unconstrained handwritten English text from 657 writers. The forms were scanned at a resolution of 300dpi and saved as PNG images with 256 gray levels. Figure 1 shows a complete

[3] https://keras.io/api/applications/
[4] https://fki.tic.heia-fr.ch/databases/iam-handwriting-database

form with the specific writer's ID and Name (the latter has been deleted for privacy issues). The forms provided by each writer ranged from 1 page (350 writers) to 59 pages (1 writer). This imbalance has been variously addressed in literature, as can be observed from Section 2. Our strategy is discussed and compared in Section 3.2.

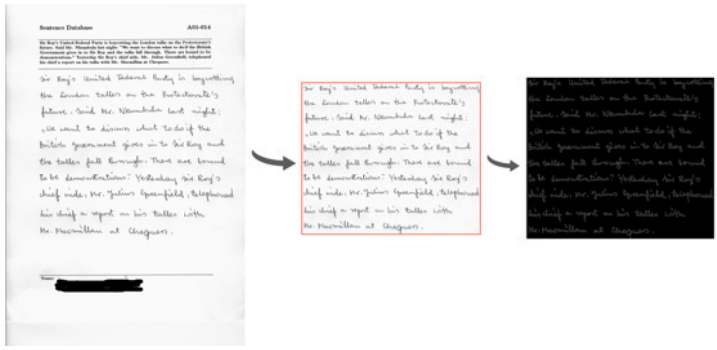


Fig. 1: A form from IAM handwriting database, its handwritten text, and the extracted edges.

The English text and the writer ID on top of the form are removed during pre-processing (Figure1 - left) and the ID is stored as ground truth together with the pre-processed sample. All forms are provided as PNG files together with corresponding XML form label files. These include segmentation information and several estimated parameters that allow extracting only the handwritten portion of the form (Figure 1 - center). The images are captured in a different setting by each writer, so pre-processing also removes the noise and variations in background illumination using a Gaussian blur filter. Finally, Canny Edge Detector extracts only the edges from the handwriting (Figure 1 - right).

### 3.2 Splitting Data and Cropping Data

As a behavioral trait, handwriting can be affected by factors like hurry, mood, etc. Even small time differences can be reflected in variations of some features in the written text. Therefore, we deem that using separate data but extracted from the same sample for both training and testing, or for enrollment and probe, would achieve unrealistic results. We preferred to maintain the set of training and testing forms strictly separated. Unfortunately, this caused leaving out the 350 subjects with a single form, because there would be a complete lack of the time-related variations observed in normal writers. For the same reason, it is not even feasible to use these subjects during testing by splitting the single form into reference/gallery sample and probe sample. The results would be probably better but would not generalize to a real application. The experiments presented here adopted two different strategies to split the data into training and testing sets, still aiming to preserve the balance in the training data for different users. We will indicate them as strategy A and strategy B. In the first strategy A, the training set and the test set contain the same 31 subjects, selected as all those having at least 9 forms split in training and testing. The compared templates are the embeddings computed by the trained model. It is worth underlining the difference with

the strategies of works discussed in Section 2. The test set contains four forms per writer. The others are used for training. To prevent inter-subjects unbalance, only the minimum possible number of forms for training is considered, which is five forms. Differently from other works in literature, we do not segment the text neither into single characters nor into words or lines. Rather, the cropped-edged images are further split into several images with sizes of $500 \times 500$. To increase the size of the training set, a little overlap is allowed between the cropped training images, using a specific threshold: for each form in the training set, the edged image is cropped into eight different smaller images. In contrast, overlap is not allowed for test forms, that are cropped into just three different smaller images (Figure 2). The obtained smaller images are used as either training or testing samples. The adopted segmentation entails faster processing and further allows exploiting both horizontal and vertical co-articulation characteristics of the written sign. In real applications it is possible to classify any part of a written sheet of paper, including multiple lines. Overall,

Fig. 2: Samples of Cropped Images

the training set contains 1240 samples of 31 different writers, while the testing set contains 372 samples (Table 1 - about 77% samples for training and the rest for testing). More samples come from the random on-the-fly augmentation process (Section 3.3).
Experiments with strategy B, all else being equal, aimed to a preliminary test of the generalizability of the obtained models: the training set contains samples from 25 subjects, while all 31 subjects appear in the testing set. Also in this case the compared templates are the embeddings computed by the trained model, to which we also submit the samples of the 6 subjects not previously seen during the training. These 6 subjects can only be included in the probe set as impostors, or also be included in the system gallery as legitimate users. Therefore strategy B does neither simply nor necessarily cause an increase of impostors but is rather useful to evaluate the ability to classify possibly unseen yet legitimate data without re-training the system. On the other hand, subject splitting between training and testing is hardly feasible, due to the huge inbalance in available data and the possible overfitting effect. To the best of our knowledge, no previous work has explored this issue.

### 3.3 Data Augmentation

Data augmentation plays a vital role in preventing overfitting in neural networks by ensuring that the model does not rely too heavily on any specific features and has more data to learn from. Ultimately, the primary objective of data augmentation is to increase the generalizability of the model, enhancing its ability to perform well on new, unseen data.

| | Dataset | #ID | Forms per ID | Cropped Samples | Total |
|---|---|---|---|---|---|
| A | Train | 31 | 5 | 8 (overlap) | 1240 |
| | Test | 31 | 4 | 3 (no overlap) | 372 |
| B | Train | 25 | 5 | 8 (overlap) | 1000 |
| | Test | 31 | 4 | 3 (no overlap) | 372 |

Tab. 1: Summary of used data with the two data splitting strategies

Data augmentation was applied in turn according to three different approaches using Keras Image Generator [5]: 1) no augmentation; 2) augmentation by rotation; 3) augmentation by rotation, shift, brightness, zooming, and re-scaling. Of course, being the samples made of pieces of text, not all kinds of augmentation are suitable. In addition, each transformation can be specified with a range and a degree of randomness. A comparison among the three approaches demonstrated that the models trained with the randomly transformed batch of images after the richer set of augmentation operations provide the best performance. Therefore, for the sake of space and clarity, we report only the related results.

## 4 Experimental results

In biometric verification, the prevailing direction of errors (either type-I or type-II) is critical. The accuracy obtained by a softmax layer does not account for this aspect. We rather compare the embeddings returned by the trained models. This also allows testing verification on unseen data. The experiments further compute the performance of different distance/similarity measures: Euclidean, Mahalanobis, and Manhattan distances, and correlation and cosine similarity. For the sake of space, the reported results only refer to correlation and cosine similarity that provided the best results. However, the performance differences are almost generally negligible, thus testifying to the robustness of the obtained embeddings across different comparison measures. The used performance measures are:
1) the mean and standard deviation (std) of the genuine score distribution (GMEAN and GSTD) account for its quality; the lower the std, the better concentrated the distribution;
2) the same values for the impostor score distribution (IMEAN and ISTD) suggest the separation from the previous one; the lower the std and the farther the mean values, the better; 3) the Area Under Curve (AUC) for the ROC; 4) the equal error rate (EER) is the equilibrium point between False Match Rate (FMR) and False Non Match Rate (FNMR).
5) ZeroFMR, FMR1000, and FMR100, i.e., FNMR when FMR=0, FMR=0.001, and FMR=0.01.

**Results with MobileNetV2 Architecture** Table 2 summarizes the results for MobileNetV2. The correlation appears to be the best comparison method. Genuine and impostor score distributions are quite well separated also passing from training strategy A to B (see upper row of Figure 3). This seems to testify that fine-tuned MobileNetV2, though with lower

[5] https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/image/ImageDataGenerator

performance than ResNet-50 (see below), generalizes quite well on unseen data. ROC comparison confirms the stable behavior of the fine-tuned model.

| | Similarity | GMEAN | GSTD | IMEAN | ISTD | AUC | EER | ZeroFMR | FMR1000 | FMR100 |
|---|---|---|---|---|---|---|---|---|---|---|
| A | correl. | 0.93 | 0.04 | 0.59 | 0.20 | 0.98 | 0.07 | 0.98 | 0.75 | 0.39 |
| | cosine | 0.95 | 0.02 | 0.74 | 0.13 | 0.98 | 0.07 | 0.98 | 0.75 | 0.40 |
| B | correl. | 0.93 | 0.03 | 0.62 | 0.19 | 0.98 | 0.068 | 0.96 | 0.77 | 0.40 |
| | cosine | 0.96 | 0.02 | 0.75 | 0.13 | 0.98 | 0.07 | 0.97 | 0.78 | 0.40 |

Tab. 2: Results by MobileNetV2: training strategy A or B with correlation or cosine similarity

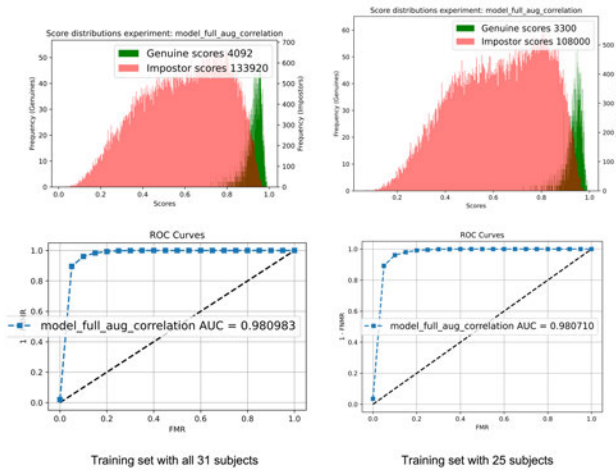



Fig. 3: Upper row: Score distributions for MobileNetV2 using training strategy A (left) or B (right). Bottom row: ROC curve for MobileNetV2 using training strategy A (left) or B (right)

**Results with ResNet-50 Architecture** The best results with ResNet-50 are summarized in Table 3. Further details can be discussed regarding correlation, which is the best compar-

| | Similarity | GMEAN | GSTD | IMEAN | ISTD | AUC | EER | ZeroFMR | FMR1000 | FMR100 |
|---|---|---|---|---|---|---|---|---|---|---|
| A | correl. | 0.92 | 0.05 | 0.46 | 0.18 | 0.99 | 0.03 | 0.71 | 0.33 | 0.09 |
| | cosine | 0.95 | 0.03 | 0.63 | 0.13 | 0.99 | 0.03 | 0.71 | 0.34 | 0.09 |
| B | correl. | 0.92 | 0.04 | 0.48 | 0.20 | 0.99 | 0.04 | 0.90 | 0.61 | 0.20 |
| | cosine | 0.95 | 0.03 | 0.66 | 0.13 | 0.99 | 0.04 | 0.89 | 0.61 | 0.20 |

Tab. 3: Results by ResNet-50: training strategy A or B with correlation or cosine similarity

ison strategy. It is interesting to notice (Figure 4 - top) that passing from training strategy A to B the genuine and impostor score distributions remain quite well separated, with the genuine one being narrow and concentrated on the highest similarity values. This seems to

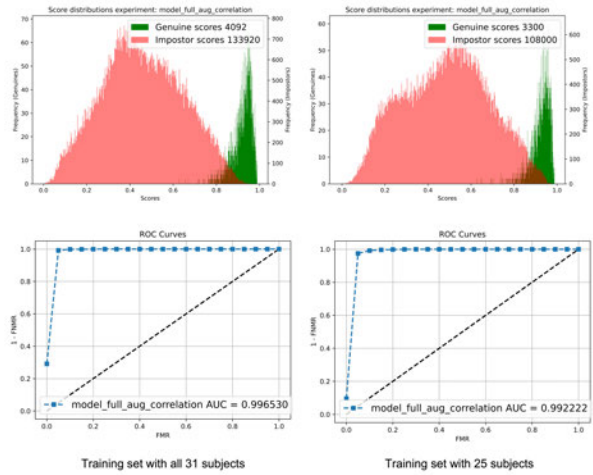



Fig. 4: Upper row: Score distributions for ResNet-50 using training strategy A (left) or B (right). Bottom row: ROC curve for ResNet-50 using training strategy A (left) or B (right)

testify also for this model the quite good generalizability on unseen data. The same considerations stem from ROC curves (Figure 4 - bottom).

**Final observations**. The different evaluation protocols make the comparison with the mentioned papers impossible. However, it is worth reporting the results obtained with softmax with our partition of the suitable testing data. In contrast with those reported above, MobileNetV2 achieves 86.83% accuracy, while ResNet-50 achieves a much lower 61.02%. This seems to confirm that biometric tasks call for biometric performance measures.

## 5 Conclusions

This preliminary study aimed at evaluating the possibility of fine-tuning pre-trained CNNs for offline text-independent writer verification on IAM dataset. The experiments relied on MobileNetV2, producing less accurate embeddings but best suited for mobile applications, and ResNet-50. More distance/similarity measures allowed for checking the robustness of the extracted embeddings, which were also applied in a preliminary study involving some unseen data. The results seem to testify that it is worth continuing along this line, despite the different outcomes of accuracy measure vs. biometric evaluation. Future work includes testing the approach on different languages (e.g., Arabic and Kanji) and more unseen data.

## References

[Ch19] Chahi, Abderrazak; Ruichek, Yassine; Touahni, Raja et al.: An effective and conceptually simple feature representation for off-line text-independent writer identification. Expert Systems with Applications, 123:357–376, 2019.

[Di19] Diaz, Moises; Ferrer, Miguel A; Impedovo, Donato; Malik, Muhammad Imran; Pirlo, Giuseppe; Plamondon, Réjean: A perspective analysis of handwritten signature technology. Acm Computing Surveys (Csur), 51(6):1–39, 2019.

[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.

[JJ20] Javidi, Malihe; Jampour, Mahdi: A deep learning framework for text-independent writer identification. Engineering Applications of Artificial Intelligence, 95:103912, 2020.

[Ng19] Nguyen, Hung Tuan; Nguyen, Cuong Tuan; Ino, Takeya; Indurkhya, Bipin; Nakagawa, Masaki: Text-independent writer identification using convolutional neural network. Pattern Recognition Letters, 121:104–112, 2019.

[Ni15] Nicolaou, Anguelos; Bagdanov, Andrew D; Liwicki, Marcus; Karatzas, Dimosthenis: Sparse radial sampling LBP for writer identification. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). IEEE, pp. 716–720, 2015.

[Ra08] Raju, G: Wavelet transform and projection profiles in handwritten character recognition-A performance analysis. In: 2008 16th International Conference on Advanced Computing and Communications. IEEE, pp. 309–314, 2008.

[RNR19] Rehman, Arshia; Naz, Saeeda; Razzak, Muhammad Imran: Writer identification using machine learning approaches: a comprehensive review. Multimedia Tools and Applications, 78:10889–10931, 2019.

[Sa18] Sandler, Mark; Howard, Andrew; Zhu, Menglong; Zhmoginov, Andrey; Chen, Liang-Chieh: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520, 2018.

[TSR17] Tan, Gloria Jennis; Sulong, Ghazali; Rahim, Mohd Shafry Mohd: Writer identification: a comparative study across three world major languages. Forensic science international, 279:41–52, 2017.

[WTB14] Wu, Xiangqian; Tang, Youbao; Bu, Wei: Offline text-independent writer identification based on scale invariant feature transform. IEEE Transactions on Information Forensics and Security, 9(3):526–536, 2014.

[XLW17] Xiong, Yu-Jie; Lu, Yue; Wang, Patrick SP: Off-line text-independent writer recognition: A survey. International Journal of Pattern Recognition and Artificial Intelligence, 31(05):1756008, 2017.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# Impact of Image Context for Single Deep Learning Face Morphing Attack Detection

Joana Alves Pimenta[1], Iurii Medvedev[2], Nuno Gonçalves[3]

**Abstract:** The increase in security concerns due to technological advancements has led to the popularity of biometric approaches that utilize physiological or behavioral characteristics for enhanced recognition. Face recognition systems (FRSs) have become prevalent, but they are still vulnerable to image manipulation techniques such as face morphing attacks. This study investigates the impact of the alignment settings of input images on deep learning face morphing detection performance. We analyze the interconnections between the face contour and image context and suggest optimal alignment conditions for face morphing detection.

**Keywords:** Face morphing detection; face recognition, deep learning; convolutional neural networks; classification.

## 1 Introduction

The expansion of technological advancements in modern society has led to an increase in security concerns. Traditional identification methods have become less reliable due to their vulnerability to forgetfulness, loss, replication, or theft, thereby compromising their intended security function. As a solution to this issue, biometric approaches are gaining popularity as they utilize physiological or behavioral characteristics to enhance the recognition process. Face image modality took one of the most important roles in modern biometric applications due to the simplicity of face image acquisition and recent advances in computer vision techniques. This led to the widespread use of Face Recognition Systems (FRSs) which utilize facial traits for the purpose of identification or verification [Li20]. Despite the fact that FRSs are currently used in various applications, they are still highly vulnerable to attacks due to the extensive range of image manipulation techniques that can be used to deceive the system.

One of the most important types of threats to FRSs is the face morphing attack. In this attack, facial features from two or more images are merged to create a synthetic image that incorporates features from both faces. The resulting image is similar to the images that gave rise to it, which allows one person to impersonate another, thereby violating the principle of self-ownership. That is why face morphing detection is a critical task in the era of digital manipulation and deep learning techniques. However, the performance of face morphing detection may depend on various factors, such as the alignment and preprocessing of input images. Specifically, the face image alignment setting can impact the amount

[1] University of Coimbra, Institute of Systems and Robotics - Coimbra, Portugal, joana.pimenta@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

of context included in the input image, which in turn can hypothetically affect the performance of the detection algorithm. We conduct our research to define optimal alignment settings for face morphing detection, exploring the possibility of using interconnections between the face contour and image context to improve the performance of the detection algorithm.

Essentially, our purpose is to investigate the relationship between image context and MAD, with the aim of identifying the most effective context properties for detection. Throughout this paper, the term "image context" refers to the background and surrounding elements in the image, i.e., the part of the image that does not contain the face.

As an additional contribution, we combined a dataset that adheres to the International Civil Aviation Organization (ICAO) guidelines for detecting face morphing.

## 2 Related Work

**Face Recognition.** Current advances in face recognition methods use deep learning techniques that employ deep neural networks, allowing the learning of deep facial features, which have high discriminative power.

Face recognition deep networks are commonly trained using classification-based tasks, employing softmax loss or its margin-based alternatives like ArcFace [De19]. The addition of a margin to the softmax loss is crucial because it significantly improves the discriminative power of the learned features. More recently, there has been a focus on incorporating adaptiveness into the margin based on the quality of the input image. For instance, MagFace [Me21] optimizes the feature embedding using an adaptive margin and regularization based on its magnitude. Another approach is AdaFace [KJL22], which proposes adapting the margin function based on the norm of the feature embedding.

**Face Morphing Generation.** Face morphing can be performed using landmark-based or deep learning-based approaches. Landmark-based methods employ a set of fiducial facial points, which are detected on all contributing face images, to generate a morph image by warping and bending procedures [FFM14].

Deep learning-based methods may employ encoder-decoder architectures, such as Generative Adversarial Networks (GANs) [Go14]. For example, the MorGAN [Da18] approach aims to make the generated images look similar to the real images while also encouraging the generators to produce diverse images that differ from each other. Karras et al. [KLA19] proposed the StyleGAN approach, which can be used to generate high-quality morphs.

The MIPGAN [H.21] approach revisits the StyleGAN by introducing an end-to-end optimization approach with a novel loss function that emphasizes preserving the identity of the generated morphed face images by incorporating identity priors. MorDIFF [Da23] proposes the use of diffusion autoencoders to generate high-fidelity and smooth face morphing attacks, which are highly vulnerable to state-of-the-art face recognition models. ReGenMorph [Da21] approach proposes to eliminate blending artifacts by combining image-level morphing and GAN-based generation, resulting in visibly realistic morphed images with high appearance quality.

**Face Morphing Detection.** Morphing attack detection (MAD) methods can be classified into two types, depending on the security application scenario: Single Morphing Attack Detection (S-MAD) and Differential Morphing Attack Detection (D-MAD).

S-MAD refers to techniques that can detect a morphed image without comparing it to an authentic reference image (*non-reference*). They are therefore based on the analysis of visual artifacts or inconsistencies in the morphed image itself. Many approaches rely on the analysis of handcrafted features like Binarized Statistical image features (BSIF) [RRB16], Local Binary Pattern (LBP) [OPH96], Local Phase Quantization (LPQ) [OH08] image descriptors, and Photo Response Non-Uniformity (PRNU) known as sensor noise [Sc19].

Recent works intensively uses deep learning for face morphing detection. OrthoMAD approach [Ne] proposes to use a regularization term for the creation of two orthogonal latent vectors that disentangle identity information from morphing attacks. MorDeephy method [MSG23] introduced fused classification to generalize morphing detection to unseen attacks. The formulation will be followed in this work. Tapia et al. [TB21] proposed a framework using few-shot learning with siamese networks and domain generalization. The framework includes a triplet-semi-hard loss function and clustering to assign classes to image samples. In this work, we focus only on the S-MAD case to perform the analysis of image alignment settings.

## 3 Methodology

**Source Data Curating**. An initial challenge encountered in this research was the lack of a suitably extensive dataset that conformed to ICAO compliance requirements. To address this issue, we combined multiple datasets comprising compliant images, including both publicly available and privately obtained data. When selecting the datasets, we prioritized those that provided a larger number of images per identity and included the following ones: FRGC [Ph05], XM2VTS [Me00], ND Twins [Ph11], FERET [Ph00, Ph98], AR [MB98], PICS [Un99], FEI [Th06], IMMF [FS05] and GTDB [A.99]. Several selected components were filtered to remove non-compliant images, i.e., non-frontal images or other images not suitable for morphing. In the specific case of the ND twins dataset, only one twin from the pair was included due to their striking resemblance, which will be confusing for the methodology of this research. Our result dataset, which we call the ICMD dataset, comprises over 50k images of more than 2.5k individuals.

**Morph Image Generation**. To accompany our training data with face morph samples, we employed landmark-based and deep learning-based (specifically GAN-based) face morphing approaches. These samples are generated using the originals from the ICMD dataset, where pairing is performed following the [MSG23], to ensure unambiguous class labeling in the fused classification task. Namely, the identity list of the dataset is randomly split into two disjoint subsets attributed to the First and Second networks, and the pairing is made between those subsets. In the end, we ensured a consistent classification classification of morphed combinations by the networks. To generalize the detection performance and reduce overfitting for artifact detection, we have included *selfmorphs* for both LDM and StyleGAN approaches. *Selfmorphs* are generated using images of the same individual, resulting in morphed images that continue to represent that same individual but contain

merging artifacts of a different kind. As a result, considering *selfmorphs* as *bona fide* samples we can prioritize morphing detection based on the behavior of deep facial features.

**Alignment settings** Our search for the optimal amount of image context for morphing detection is based on selecting several different alignment settings and running identical experiments for each setting. The face alignment in academia is usually performed by a rigid transformation, which minimizes the coordinate distance between the five facial landmarks (detected by MTCNN [XZ17]) ({left eye}, {right eye}, {nose}, {left mouth corner}, {right mouth corner}) and the definite target list of coordinates (for the resulting image size of $112 \times 112$ - {{38.2, 41.7}, {73.5, 41.5}, {56.0, 61.7}, {41.5, 82.4}, {70.7, 82.2}}) [De19]. The particular list of settings that we used is based on the scaling of this target set of coordinates. The Table 1 presents a schematic correspondence of each alignment with the scale factor utilized, along with its respective indicative ratio of the face's occupancy area in the image. We estimate this face's occupancy as the ratio of face area (limited by a face contour detected using 68 landmarks [Ki09]) to the full image area.

Tab. 1: Summary table of all alignment conditions with their respective scale factors and ratios.

| **Alignments** | **a** | **b** | **c** | **d** | **e** | **f** | **g** | **h** | **i** | **j** | **k** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Scale Factor** | 1.65 | 1.40 | 1.10 | 1.00 | 0.90 | 0.85 | 0.80 | 0.75 | 0.70 | 0.65 | 0.60 |
| **Ratio** | 0.15 | 0.21 | 0.34 | 0.42 | 0.51 | 0.56 | 0.62 | 0.70 | 0.77 | 0.86 | 0.94 |

**S-MAD - Fused Classification.** In our work, we approach *no-reference* face morphing detection in several ways. First, we follow the fused classification approach, where two parallel networks were trained simultaneously. These networks were specifically designed to acquire high-level features by performing classification tasks in order to generalize the performance to unseen attacks [MSG23].

The overall pipeline schematic is presented in Fig.1. Each sample is assigned two class labels: *morphs* inherit them from source identities; *bona fides* have a duplicated original label. The classification task is made differently for each of the networks. First Network labels them by the original identity from the first source image, and the Second Network by the second original label. The main motivation is learning high-level identity discriminative features, which can indicate the presence of face morphing. Such classification is regulated by the explicit binary classification of a dot product of those resulting high-level features.

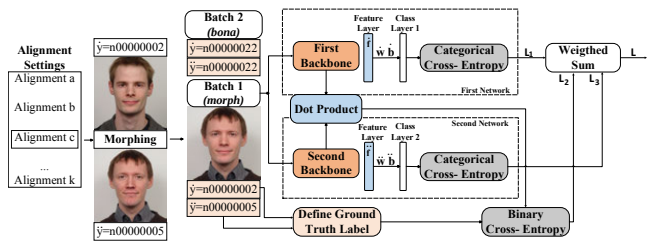



Fig. 1: S-MAD fused approach schematics. In order to simplify the visualization, a single image is shown per batch.

Mathematically, such a schematic is formulated as the weighted sum: $L = \alpha_1 L_1 + \alpha_2 L_2 + \beta L_3$, where $L_1$ and $L_2$ are face recognition components, and $L_3$ is a morphing detection component. Based on the common softmax formulation, each network is regularized by the respective losses:

$$L_1 = -\frac{1}{N}\sum_{i}^{N}\log\left(\frac{e^{\dot{W}_{\dot{y}_i}^T \dot{f}_i + \dot{b}_{\dot{y}_i}}}{\sum_j^C e^{\dot{f}_{\dot{y}_j}}}\right), \qquad L_2 = -\frac{1}{N}\sum_{i}^{N}\log\left(\frac{e^{\ddot{W}_{\ddot{y}_i}^T \ddot{f}_i + \ddot{b}_{\ddot{y}_i}}}{\sum_j^C e^{\ddot{f}_{\ddot{y}_j}}}\right), \tag{1}$$

where $f_i$ are deep features of the $i^{th}$ sample, $y_i$ represents the class index of the $i^{th}$ sample, and $W$ and $b$ denote the weights and biases of the last fully connected layer, respectively. $N$ represents the batch size, while $C$ represents the total number of classes.

Finally, in order to determine the similarity metric based on the ground truth authenticity label of the image, the morphing detection score is obtained by computing the dot product of the backbone outputs ($\dot{f} \cdot \ddot{f}$). This score is then passed through the *sigmoid* function and used to define the binary cross-entropy loss. As a final result, the corresponding loss is defined by:

$$L_3 = -\frac{1}{N}\sum_{i}^{N} t \log\frac{1}{1+e^{-\ddot{f}\cdot\dot{f}}} + (1-t)\log\left(1 - \frac{1}{1+e^{-\ddot{f}\cdot\dot{f}}}\right) \tag{2}$$

The optimization process involves combining the resulting losses as a weighted sum, resulting in $L$, with the goal of minimizing it. This is done to learn facial features that are discriminative and specifically regularized for the task of detecting morphing.

**S-MAD - Binary Classification.** Another approach for face morphing detection is indeed similar to the straightforward binary classification (morph/non-morph). To implement it, we removed the identity classification part from the fused approach and retained only a single deep network in the entire pipeline. The model schema is presented in Fig.2.

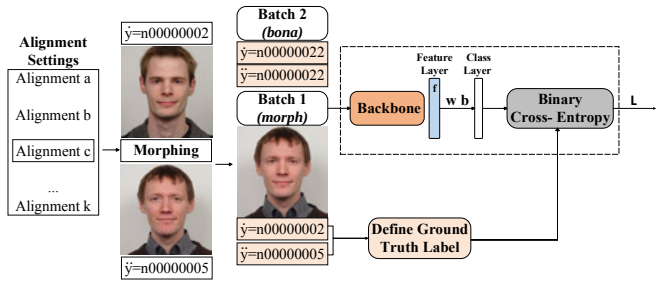



Fig. 2: S-MAD approach model schema for a single network. In order to simplify the visualization, a single image is shown per batch.

**Benchmarking.** For performance estimation, we employ the open-source morphing benchmarking utilities [4] and adopt them into our work. We replace the *bona fide* subset with the images from FRLL-Set [DJ17], Utrecht [Un99], MIT-CBCL [He01] and EFIEP [Av19] (since the default suggested protocols share images with our training data). All protocols

[4] https://github.com/iurii-m/MorDeephy.git

share the same list of *bona fide* images and are only different in the content of morphs, which are taken from the FRLL-Morphs dataset [Sa22] (protocol names correspond to the FRLL-Morph subset names): *protocol-asml* with $\sim$ 2k morphs, *protocol-opencv* with $\sim$ 1.3k morphs, *protocol-facemorpher* with $\sim$ 2k morphs, *protocol-webmorph* with $\sim$ 1k morphs and *protocol-stylegan* with $\sim$ 2k morphs.

**Heatmap Computation.** We analyze the image context impact using the Gradient-Weighted Class Activation Mapping (Grad-CAM) technique and generate a heatmap that highlights the regions of the input image that have the most significant influence on the ground truth binary prediction.

## 4 Experiments and Results

**Training Settings.** As a baseline model in our work, we use EfficientNetB3 [TL19], which is pretrained on the ImageNet dataset. We trained our models for five epochs using a stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and a learning rate linearly decaying from 0.075 to 1e-5 . The batch included 28 images. Separate training experiments are performed for each alignment case on concatenated datasets: original, LDM, StyleGAN morphs, and *selfmorphs*. Face morphs are generated with LDM and StyleGAN approaches. The parameters for the fused approach, which determine the appropriate balance between the different components of the loss function, are taken from the original work [MSG23]: $\alpha = \alpha_1 = \alpha_2$ and $\alpha/\beta$=0.2.

**Binary Classification.** Based on the results presented in Table 2, the alignment range with optimal performance is observed between *e* to *g*, with *e* being the possible optimal case. Based on heatmaps, the face is the principal region for the detection decision, and the regions, which are prompt to contain morphing artifacts, are mainly activated (see Fig. 3).

Tab. 2: BPCER@APCER = (0.1, 0.01) of our S-MAD binary approach for various alignment settings

| Alignments | BPCER@APCER=$\delta$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Protocol-asml | | Protocol-facemorpher | | Protocol-opencv | | Protocol-stylegan | | Protocol-webmorph | |
| | $\delta$=0.1 | $\delta$=0.01 | $\delta$=0.1 | $\delta$=0.01 | $\delta$=0.1 | $\delta$=0.01 | $\delta$ =0.1 | $\delta$=0.01 | $\delta$=0.1 | $\delta$=0.01 |
| a | 0.199 | 0.622 | 0.125 | 0.558 | 0.199 | 0.663 | 0.663 | 0.663 | 0.523 | 0.663 |
| b | 0.143 | 0.380 | 0.131 | 0.387 | 0.144 | 0.440 | 0.586 | 0.586 | 0.340 | 0.586 |
| c | 0.365 | 0.630 | 0.331 | 0.675 | 0.320 | 0.676 | 0.676 | 0.676 | 0.489 | 0.676 |
| d | 0.236 | 0.511 | 0.161 | 0.549 | 0.161 | 0.489 | 0.623 | 0.623 | 0.436 | 0.623 |
| **e** | **0.141** | **0.348** | **0.102** | 0.532 | **0.080** | **0.424** | 0.710 | 0.710 | **0.321** | 0.641 |
| f | 0.199 | 0.455 | 0.127 | 0.551 | 0.125 | 0.533 | 0.675 | 0.675 | 0.328 | 0.579 |
| g | 0.158 | 0.373 | 0.106 | 0.532 | 0.209 | 0.532 | 0.586 | **0.586** | 0.348 | 0.586 |
| h | 0.330 | 0.580 | 0.138 | 0.682 | 0.093 | 0.486 | 0.724 | 0.724 | 0.486 | 0.724 |
| i | 0.214 | 0.408 | 0.174 | 0.476 | 0.149 | 0.442 | **0.573** | 0.573 | 0.396 | **0.573** |
| j | 0.221 | 0.465 | 0.187 | 0.596 | 0.141 | 0.457 | 0.776 | 0.776 | 0.475 | 0.682 |
| k | 0.243 | 0.498 | 0.194 | 0.557 | 0.146 | 0.513 | 0.794 | 0.794 | 0.467 | 0.707 |

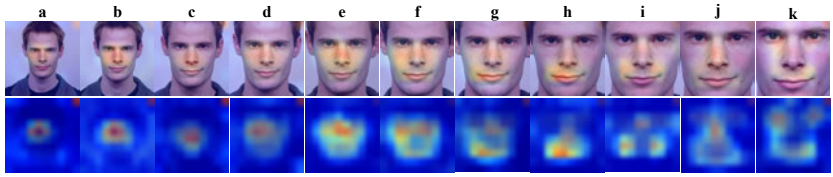



Fig. 3: Grad-CAM morph heatmaps for the S-MAD binary approach under different alignment conditions (Recall that *bona fide* sets are equal across all the protocols).

**Fused Classification.** For this approach, the optimal range is observed at alignment settings from *d* to *i*, with *g* being possibly the optimal case. At the same time, this methodology allows for superior results in comparison to the binary classification case, which may be related to the regularization imposed by the face recognition task. Based on the

Tab. 3: BPCER@APCER = (0.1, 0.01) of S-MAD fused approach for various alignment settings

| Alignments | BPCER@APCER=$\delta$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Protocol-asml | | Protocol-facemorpher | | Protocol-opencv | | Protocol-stylegan | | Protocol-webmorph | |
| | $\delta$=0.1 | $\delta$=0.01 | $\delta$=0.1 | $\delta$=0.01 | $\delta$=0.1 | $\delta$=0.01 | $\delta$ =0.1 | $\delta$=0.01 | $\delta$=0.1 | $\delta$=0.01 |
| a | 0.159 | 0.689 | 0.187 | 0.517 | 0.239 | 0.599 | 0.842 | 0.946 | 0.606 | 0.885 |
| b | 0.063 | 0.495 | 0.072 | 0.646 | 0.099 | 0.658 | 0.671 | 0.946 | 0.702 | 0.964 |
| c | 0.125 | 0.467 | 0.215 | 0.588 | 0.240 | 0.566 | 0.694 | 0.884 | 0.541 | 0.859 |
| d | 0.040 | 0.374 | 0.102 | 0.558 | 0.103 | 0.568 | 0.574 | 0.835 | 0.305 | 0.781 |
| e | 0.162 | 0.580 | 0.149 | 0.582 | 0.177 | 0.602 | 0.566 | 0.767 | 0.605 | 0.870 |
| f | 0.184 | 0.530 | 0.180 | 0.488 | 0.175 | 0.451 | 0.582 | 0.788 | 0.517 | 0.785 |
| **g** | **0.034** | **0.233** | **0.025** | 0.701 | **0.037** | 0.701 | 0.487 | 0.875 | **0.216** | 0.788 |
| h | 0.168 | 0.642 | 0.168 | 0.535 | 0.165 | 0.599 | 0.536 | 0.850 | 0.542 | 0.854 |
| i | 0.046 | 0.255 | 0.036 | **0.365** | 0.044 | **0.390** | **0.305** | **0.583** | 0.246 | **0.554** |
| j | 0.287 | 0.630 | 0.268 | 0.585 | 0.262 | 0.564 | 0.844 | 0.959 | 0.697 | 0.907 |
| k | 0.193 | 0.652 | 0.253 | 0.745 | 0.262 | 0.792 | 0.825 | 0.953 | 0.674 | 0.915 |

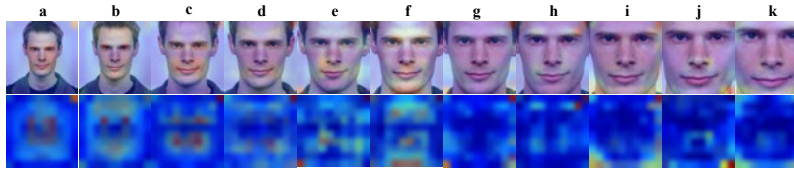



Fig. 4: Grad-CAM morph heatmaps for the S-MAD fused approach under different alignment conditions (Recall that *bona fide* sets are equal across all the protocols).

heatmaps, the detection is mainly focused on the face region and, in many cases, on the regions of intersection between the foreground and background (see Fig. 4).

**NIST FRVT MORPH Results.** We compare the results of our best model (**visteamicao-000**) for fused case with several state-of-the-art (SOTA) MAD approaches, tested on the FRVT NIST MORPH Benchmark [FR]. Each dataset from the benchmark has images generated through different protocols, with distinctions made in tiers such as Tier 2 - Automated Morph Analysis and Tier 3 - High-Quality Morph Analysis.

Tab. 4: Comparison with the SOTA S-MAD approaches using APCER@BPCER = (0.1, 0.01).

| Algorithm | Visa-Border (Tier 2) | | Twente (Tier 2) | | Manual (Tier 3) | |
|---|---|---|---|---|---|---|
| | $\delta$=0.1 | $\delta$=0.01 | $\delta$=0.1 | $\delta$=0.01 | $\delta$=0.1 | $\delta$=0.01 |
| Our | 0.089 | **0.291** | 0.032 | 0.128 | 0.802 | 0.975 |
| Aghdaie et al. [Ag21] | 0.037 | 0.542 | 0.002 | **0.060** | 0.879 | 0.975 |
| Medvedev et al. [MSG23] | 0.232 | 0.555 | 0.174 | 0.493 | 0.641 | **0.926** |
| Ferrara et al. [FFM21] | 0.477 | 0.999 | 0.002 | 0.183 | 0.938 | 0.985 |
| Ramachandra et al. [Ra19] | 0.375 | 0.990 | 0.304 | 0.998 | 0.938 | 0.985 |

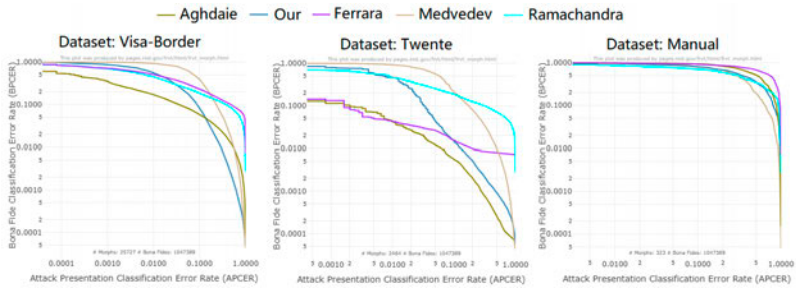



Fig. 5: Detection Error Trade-off curves for different SOTA approaches in different datasets (Visa-Border, Twente and Manual dataset).

Regarding the Visa-Border dataset, our approach outperforms all other SOTA approaches, with a *morph miss rate* of 0.29 at a *false detection* rate of 0.01. In the Twente dataset, when comparing with other approaches, the results demonstrate a highly favorable outcome as well, with a *morph error rate* of 0.128 at a *false detection rate* of 0.01 (See table 4). Although not represented in the table, comparable results were achieved for other datasets, such as the UNIBO Automatic Morphed Face Generation Tool v1.0 and even MIPGAN-II with less dominant but still competitive performances. It is important to take into consideration the influence of the dataset used, and this Tier 2 typology is generally less challenging. When faced with more realistic datasets (Manual dataset), it becomes apparent that overall SOTA approaches show poor generalization across various unseen morphing techniques. Even so, our model results achieved competitive results when compared to those approaches.

## 5 Conclusions

In this work, we aim to identify the context properties that are most effective for S-MAD. The extensive experiments allowed us to determine the alignment range where S-MAD is more effective. Moreover, in this range, there seems to be a certain correspondence between both fused and binary approaches, which translates into a similar area of face occupancy in the image. Despite that, our results also show that face is the most dominant activation region across all the alignment settings, and the impact of context on face morphing detection is limited. Our method achieved state-of-the-art comparable performances on some of the NIST FRVT MORPH benchmark protocols. Our future work will be directed toward investigating similar properties in the differential scenario.

## 6 Acknowledgements

The authors would like to thank the Portuguese Mint and Official Printing Office (INCM) and the Institute of Systems and Robotics - the University of Coimbra for the support of the project FACING2. This work has been supported by Fundação para a Ciência e a Tecnologia (FCT) under the project UIDB/00048/2020.

## References


[A.99] A. V. Nefian: , Georgia Tech face database. http://www.anefian.com/research/face_reco.htm, 1999.

[Ag21] Aghdaie, P.; Chaudhary, B.; Soleymani, S.; Dawson, J.; Nasrabadi, N.: Attention aware wavelet-based detection of morphed face images. CoRR, abs/2106.15686, 2021.

[Av19] Avilés, J.; Toapanta, H.; Morillo, P.; Vallejo-Huanga, D.: Dataset of Ethnic Facial Images of Ecuadorian People. 2019.

[Da18] Damer, N.; Saladié, A. M.; Braun, A.; Kuijper, A.: MorGAN: Recognition Vulnerability and Attack Detectability of Face Morphing Attacks Created by GAN. In: 2018 IEEE 9th International Conference on BTAS). pp. 1–10, Oct 2018.

[Da21] Damer, N.; Raja, K. B.; Sussmilch, M.; Venkatesh, S. K.; Boutros, F.; Fang, M.; Kirchbuchner, F.; Ramachandra, R.; Kuijper, A.: ReGenMorph: Visibly Realistic GAN Generated Face Morphing Attacks by Attack Re-generation. In: ISVC. 2021.

[Da23] Damer, N.; Fang, M.; Siebke, P.; Kolf, J. N.; Huber, M.; Boutros, F.: , MorDIFF: Recognition Vulnerability and Attack Detectability of Face Morphing Attacks Created by Diffusion Autoencoders, 2023.

[De19] Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S.: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: 2019 IEEE/CVF CVPR. pp. 4685–4694, 2019.

[DJ17] DeBruine, L.; Jones, B.: , Face Research Lab London Set, May 2017.

[FFM14] Ferrara, M.; Franco, A.; Maltoni, D.: The magic passport. In: IEEE IJCB. pp. 1–7, 2014.

[FFM21] Ferrara, Matteo; Franco, Annalisa; Maltoni, Davide: Face morphing detection in the presence of printing/scanning and heterogeneous image sources. IET Biometrics, 10, 02 2021.

[FR] FRVT MORPH. https://pages.nist.gov/frvt/html/frvt_morph.html.

[FS05] Fagertun, Jens; Stegmann, Mikkel Bille: , The IMM Frontal Face Database. http://www2.imm.dtu.dk/ aam/datasets/imm_frontal_face_db_high_res.zip, 2005.

[Go14] Goodfellow, I. J.; P.Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.l; Courville, A.; Bengio, Y.: Generative Adversarial Nets. In: NeurIPS. Curran Associates, Inc., 2014.

[H.21] H.Zhang, and Venkatesh, S.; Ramachandra, R.and Raja, Kiran; Damer, N.; Busch, C.: MIPGAN—Generating Strong and High Quality Morphing Attacks Using Identity Prior Driven GAN. IEEE T-BIOM, 3(3):365–383, 2021.

[He01] Heisele, B.; Serre, T.; Pontil, M.; Vetter, T.: , MIT-CBCL FR Database. http://cbcl.mit.edu/software-datasets/heisele/facerecognition-database.html, 2001.

[Ki09] King, Davis E.: Dlib-ml: A Machine Learning Toolkit. Journal of Machine Learning Research, 10:1755–1758, 2009.

[KJL22] Kim, M.; Jain, A. K; Liu, X.: AdaFace: Quality Adaptive Margin for Face Recognition. In: Proceedings of the IEEE/CVF CVPR. 2022.

[KLA19] Karras, T.; Laine, S.; Aila, T.: A Style-Based Generator Architecture for Generative Adversarial Networks. In: 2019 IEEE/CVF CVPR. pp. 4396–4405, 2019.

[Li20] Li, L.; Mu, X.; Li, S.; Peng, H.: A Review of Face Recognition Technology. IEEE Access, pp. 139110–139120, 2020.

[MB98] Martinez, A.; Benavente, R.: , The AR Face Database: CVC Technical Report, 24. https://www2.ece.ohio-state.edu/ aleix/ARdatabase.html, 1998.

[Me00] Messer, K.; Matas, J.; Kittler, J.; Jonsson, K.; Luettin, J.; Maître, G.: Xm2vtsdb: The extended m2vts database. Proc. of Audio- and Video-Based Person Authentication, 04 2000.

[Me21] Meng, Q.; Zhao, S.; Huang, Z.; Zhou, F.: MagFace: A universal representation for face recognition and quality assessment. In: CVPR. 2021.

[MSG23] Medvedev, I.; Shadmand, F.; Gonçalves, N.: MorDeephy: Face Morphing Detection via Fused Classification. In: Proceedings of the 12th ICPRAM. SciTePress, pp. 193–204, 2023.

[Ne] Neto, P. C.; Gonçalves, T.; Huber, M.; Damer, N.; Sequeira, A. F.; Cardoso, J. S.: OrthoMAD: Morphing Attack Detection Through Orthogonal Identity Disentanglement. In: BIOSIG 2022. pp. 1–5.

[OH08] Ojansivu, V.; Heikkilä, J.: Blur Insensitive Texture Classification Using Local Phase Quantization. In: Springer-Verlag. ICISP '08, Berlin, Heidelberg, p. 236–243, 2008.

[OPH96] Ojala, T.; Pietikäinen, M.; Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern recognition, 29(1):51–59, 1996.

[Ph98] Phillips, P.; Wechsler, H.; Huang, J.; Rauss, P.: The FERET database and evaluation procedure for face-recognition algorithms. Image and Vision Computing, 16(5):295–306, 1998.

[Ph00] Phillips, P.J.; Moon, H.; Rizvi, S.A.; Rauss, P.J.: The FERET evaluation methodology for face-recognition algorithms. IEEE TPAMI, 22(10):1090–1104, 2000.

[Ph05] Phillips, P.; Flynn, P.; Scruggs, W.; Bowyer, K.; Chang, J.; Hoffman, K.; Marques, J.; Min, J.; Worek, W.: Overview of the Face Recognition Grand Challenge. In: IEEE CVPR. 2005.

[Ph11] Phillips, P.; Flynn, P.; Bowyer, K.; Vorder, R.; Grother, P.; Quinn, G.; Pruitt, M.: Distinguishing Identical Twins by Face Recognition. In: 9th IEEE FG 2011, Santa Barbara, CA. 2011-03-21 2011.

[Ra19] Ramachandra, R.; Venkatesh, S.; Raja, K.; Busch, C.: Towards making morphing attack detection robust using hybrid scale-space colour texture features. In: 2019 IEEE 5th ISBA. pp. 1–8, 2019.

[RRB16] Raghavendra, R.; Raja, K. B.; Busch, C.: Detecting morphed face images. In: 2016 IEEE 8th International Conference on BTAS. pp. 1–7, Sep. 2016.

[Sa22] Sarkar, E.; Korshunov, P.; Colbois, L.; Marcel, S.: Are GAN-based morphs threatening face recognition? In: ICASSP 2022. pp. 2959–2963, 2022.

[Sc19] Scherhag, U.; Debiasi, L.; Rathgeb, C.; Busch, C.; Uhl, A.: Detection of Face Morphing Attacks Based on PRNU Analysis. IEEE T-BIOM, 1(4):302–317, 2019.

[TB21] Tapia, J. E.; Busch, C.: Single Morphing Attack Detection Using Feature Selection and Visualization Based on Mutual Information. IEEE Access, 9:167628–167641, 2021.

[Th06] Thomaz, C. E.: , FEI- Face Database. https://fei.edu.br/ cet/facedatabase.html, 2006.

[TL19] Tan, Mingxing; Le, Quoc V.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv, 2019.

[Un99] University of Stirling: , Psychological Image Collection at Stirling (PICS). http://pics.stir.ac.uk/, 1999.

[XZ17] Xiang, J.; Zhu, G.: Joint Face Detection and Facial Expression Recognition with MTCNN. In: 2017 4th ICISCE. pp. 424–427, 2017.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# Contactless Fingerprints: Differential Performance for Fingers of Varying Size and Ridge Density

Carson King[1], Evan Garrett[2], Aeddon Berti[3], Nasser Nasrabadi[4] and Jeremy Dawson[5]

**Abstract:** The match performance of contactless fingerprint probes compared to contact-based galleries has increased accuracy. This performance, along with convenience of use, is encouraging the utilization of contactless fingerprint collection methods. However, issues with differential performance for different demographics may still exist. Past works focused mainly on the interoperability of contactless prints with smartphone applications and kiosk devices. This paper focuses on the differential performance of genuine match scores based on the demographic of finger size, ridge density, and total ridge count. Distribution of genuine match scores shows a correlation between an increase in genuine match scores and these variables in contactless smartphone collection methods with the largest correlation appearing in finger size.

**Keywords:** Fingerprint, Interoperability, Contact, Contactless, Finger Size, Ridge Density

## 1 Introduction

The advancement of camera capture quality for mobile devices has sparked interest in the use of these devices as contactless fingerprint capture tools. Smartphones allow for a portable and quick collection that is more accessible and convenient than traditional standalone sensors. Along with this newfound interest comes the set of challenges that are linked to the optimization and accuracy of contactless fingerprints compared to their contact counterparts. These contactless fingerprint tools typically generate a contact equivalent fingerprint, obtained from the fingerphoto, for subsequent matching attempts. Contactless fingerprint imaging systems have been found to have distortion and loss of information, image clarity, and greyscale variations, which tends to be an issue caused by the difference in lighting based on the collection location. Due to the limited amount of datasets available for contactless fingerprints, research has focused on the interoperability of contactless and ink or livescan contact-based fingerprints is limited [MP17]. Other works have reported challenges arising from low ridge/valley contrast, non-uniform illumination, perspective distortions from non-uniform collection distances, differences in the finger orientation, and lack of cross-compatibility when matching

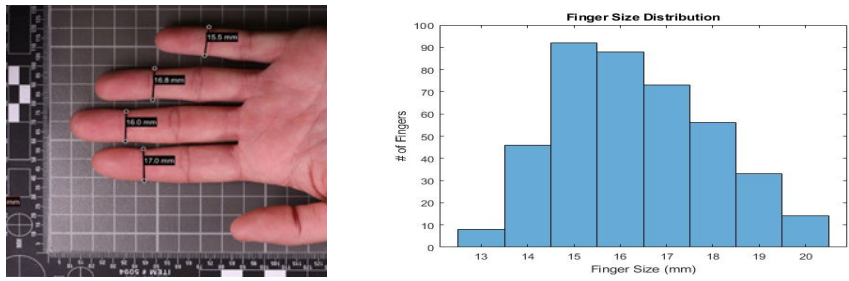
[1] Carson King, LCSEE, West Virginia University, Morgantown WV, US, crk0021@mix.wvu.edu
[2] Evan Garrett, LCSEE, West Virginia University, Morgantown WV, US, erg0015@mix.wvu.edu
[3] Aeddon Berti, LCSEE, West Virginia University, Morgantown WV, US, adb0068@mix.wvu.edu
[4] Nasser Nasrabadi, LCSEE, West Virginia University, Morgantown WV, US, nasser.nasrabadi@mail.wvu.edu
[5] Jeremy Dawson, LCSEE, West Virginia University, Morgantown WV, US, jeremy.dawson@mail.wvu.edu

against legacy datasets [Gr22]. The purpose of the study presented here is to evaluate the effects caused by finger size and ridge density in the interoperability of contact and contactless-based collection methods. Contributions in this paper are 1) an analysis of the correlation between finger size and ridge density in contactless fingerprint datasets, and 2) an analysis of the impact of finger size and ridge density on genuine match scores when matched against a contact dataset. These results will lend critical insight into how finger scaling in collection apps can impact the performance of contactless fingerprints.

There have been two main areas of research in the field of contactless fingerprint technology; differential performance and interoperability between contact and contactless images. The research field of contactless fingerprints has mainly been focused on interoperability since legacy contact-based datasets requiring this functionality. Beyond matching contactless probes to legacy contact galleries, demographic factors have also been explored to determine which variables can influence contactless match scores [Gr22, BND22]. These demographic factors were skin color, skin texture, keratin levels, pigmentation, temperature, elasticity, and finger minutiae. To date, no linear relationship between any of these demographics and match performance has been observed. However, there was a strong correlation between the image quality and match scores observed in [HE16]. Enhancement of fingerprint images has been a major area of study for both contact and contactless fingerprints because distortion generally causes high FNMR [MS16]. Enhancement techniques can be simple, such as removing noise from slap fingerprints to allow for accurate segmentation [RM11], to complex, such as using deep learning to unwarp contactless fingerprint images [Da19]. Finger size has been investigated with differing results. Previous research that investigated the influence of finger size on the interoperability of contactless fingerprints with the acquired match scores against contact-based devices found that there was a correlation between finger size and match scores with one of the matchers evaluated [Wi21]. However, the finger sizes were only separated into two distinct ranges, large or small, and it is difficult to determine the actual effect of finger size with only two subjective and qualitative size variables. This concept was examined in another study that compared fingerprints from smartphones to legacy slaps and found a TAR of 95.79% and a FAR of 0.1% while the baseline using contact-based methods was a TAR of 98.55% with an equal FAR of 0.1% [De18]. An issue that could cause variation in match performance is finger orientation. One study observed variations of match scores based on finger orientation, with results indicating that pose correction caused a decrease in EER and a 9.93%, 10.20%, and 74.97% improvement in rank 1 accuracy from three respective databases [TK20]. Ridge density is the spacing of individual ridges in a fingerprint and is a unique trait that is commonly used for its uniqueness in anti-spoofing liveness detection [AS06]. Contactless fingerprints present a challenge when considering ridge density because of the curvature of the finger when taking the image leads to perspective distortion of the ridges on the periphery of the finger. A resolution of 500 ppi is the minimum sampling rate required, but this causes under-sampling of the edges, so the US National Institute of Standards and Technology (NIST) recommends a 700 ppi sampling rate to accurately capture the edges [Li18].

### 1.1 Dataset and Measurements

The dataset that was utilized for this research[6] consists of contact fingerprints, contactless fingerprints, and hand images. The devices used to collect this data were the Guardian and Kojak for contact fingerprints, Gemalto and Morpho Wave for contactless kiosk capture images, and two third-party apps, on the Android Galaxy S20 and Android Galaxy S21 for contactless fingerprints. A commercial digital camera was also used to capture hand geometry images. The largest demographics for these collections consisted of 20–29-year-old Caucasians. To ensure the uniformity of the measurements, a custom interface was created to measure the width of the first joint closest to each fingertip in the hand geometry images as a baseline finger size measurement (Fig. 1(a)). Finger size distribution is provided in Fig. 1(b), with most finger sizes being between 15 mm and 17 mm in width. Finger size distributions are shown in Fig. 1.




(a) (b)

Fig 1. Finger Size Distribution Chart (a) and Finger Size Program Result in mm (b)

Ridge density was determined in MATLAB by gathering datapoints using images that were collected with the contact-based Kojak device, since they were the baseline for generating match scores. The *regionplots* command was used with the *centroids* parameter to find the center of mass of each image. Multiple centroids were found and averaged to find the horizontal and vertical center of rows of the fingerprint image. Then, the edge detection was done using a Sobel filter with the *edge* function, an example of a centroid image and an edge-detected image are shown in Fig. 2(a) and Fig. 2(b), respectively. Finally, the pixel values at the rows and columns were stored in arrays that were iterated to count the number of ridges in the image that were detected. Fifteen pixels were counted in both the positive and negative direction in both rows and columns to count the ridges. Once the ridges were accounted for, the individual arrays were divided by 2 since the edge detection method counted both the start of the ridge and where it ended. These values were averaged to get a ridge value for each fingerprint. For the ridge density calculation performed in this study, only horizontal ridges (with respect to the orientation of the fingers in the hand photos) were utilized, because the finger measurements were only the width of the finger. The flatter the participant's finger was

[6] Dataset is available upon request.

on the capture platform resulted in more data being collected, increasing the ridge count number. This adds a third variable to be considered: ridge count in each individual capture. For this purpose, total ridges, regardless of direction and finger size, were counted to see the impact of total ridges on genuine match scores. The matcher utilized was the Innovatrics fingerprint matcher version 7.6.0.627, which is a consumer off the shelf system optimized for matching contactless fingerprints.

(a) (b)

Fig. 2: Uncropped Input Image with Centroids (a) & Sobel Edge Detected Fingerprint (b)

## 2 Results

Understanding the relation between fingerphoto ridge density and finger size is imperative to understanding the impact that they influence the genuine match scores. Ridge density was found for each finger by dividing the number of horizontal ridges by the size of the finger, with the resulting values ranging from 0.19 to 1.45 ridges/mm. Fingers were separated into 10 bins based on their ridge density value where each bin is 0.05 ridges/mm in width. In this dataset, there does appear to be a correlation in the relationship between the two variables of finger size and ridge density. This relationship is a positive linear function, as the finger size increases the ridge density increases, with the difference in the average for the smallest finger size bin and the largest finger size bin being over 3 units of ridges per millimeter. The Kojak device fingerprints were used as the gallery and matched against the other devices to produce the match scores to associate with finger size and ridge density. The middle range of sizes do not appear to have a correlation between the finger size and match score, but there is a noticeable difference at both the lower end (13mm and 14mm) and upper end (19mm and 20mm) shown in Fig. 3.

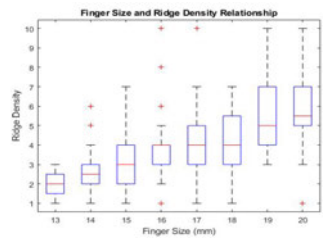
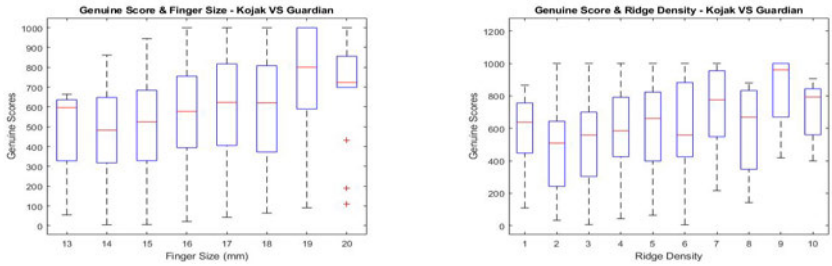



Fig. 3: Finger Size and Ridge Density Relationship

No statistically significant correlation between the three variables investigated and genuine match scores was observed in the baseline Guardian vs. Kojak matching experiment. The smallest finger size had a maximum genuine score below 700, but every size above this had genuine scores over 900, with similar results for all three variables, as shown in Fig. 4 and Fig. 5. This result is to be expected due to the maturity of contact-based fingerprint collection and matching.




(a) (b)

Fig. 4: Genuine Score Against Finger Size (a) & Ridge Density (b) Contact-Based Guardian

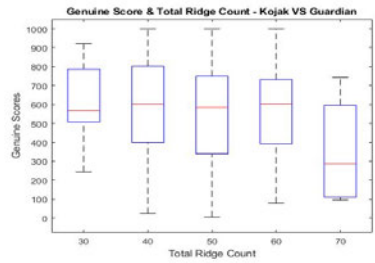
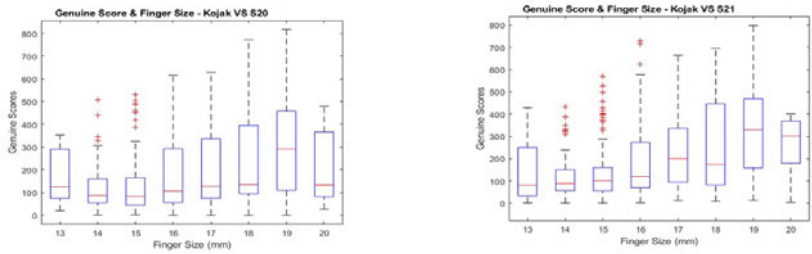



Fig. 5: Genuine Score Against Total Ridge Count Contact-Based Guardian

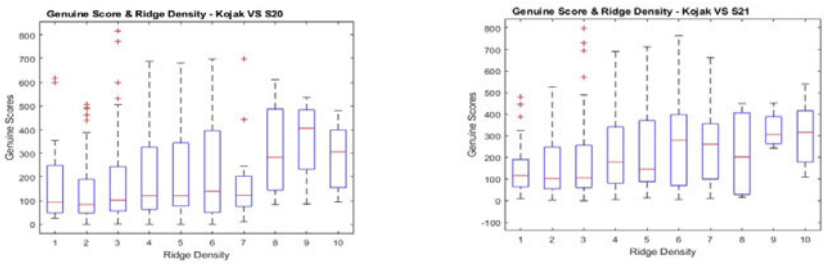
For the contactless kiosk fingerprint images, there were varying results between the two devices. The median values were consistent for all three variables for the Gemalto device, but they had different maximum match scores, while the Morpho device had similar results to the contact-based method with no correlations observed between the genuine scores and any of the three variables. The fingerprints captured using smartphone apps produced results that displayed a correlation between finger size, ridge density, and genuine match scores. The results had variation based on which of the two applications were used. However, between the two models of cellular devices, there was little variation. Application A results were similar to the contactless Gemalto results for the finger size variable. As finger size increased, the median stayed consistent but the maximum score increased. The ridge density plots displayed no correlation between finger size and ridge density for the fingerprints that were captured with application A.

The ridge count plots had a variation in the smallest bin of ridge count between the two devices, which could be caused by a capture issue, such as finger orientation. Application B displayed the highest correlation between genuine scores, finger size, and ridge density. The results between each model of the cellular device had little variation, as observed in application A. Application B finger size and genuine score plots showed the most apparent correlation between finger size and genuine match scores. Genuine scores for fingers sizes between 13 mm and 16 mm had exceptionally low genuine match scores, but for 17 mm and up, the genuine match scores started to drastically increase with finger size. There seemed to be a correlation between ridge density and genuine match score. As ridge density increases there is a slight increase in genuine match scores averages. There was a significant quantity of outliers for the smaller ridge density bins. For ridge count, it appears that lower ridge counts were correlated with a higher average match score, but the middle range of ridge counts had many outliers that were above the average value. The results for each smartphone are displayed side by side for each variable and application in Fig. 6 through Fig. 11.



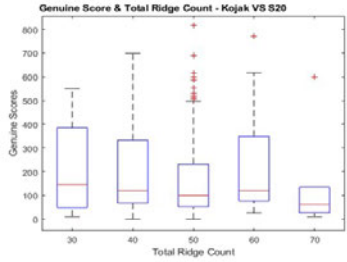

(a) (b)

Fig. 6: Genuine Score Against Finger Size App A S20 (a) & S21 (b)

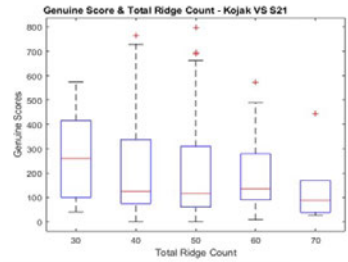



(a) (b)

Fig. 7: Genuine Score Against Ridge Density App A S20 (a) & S21 (b)

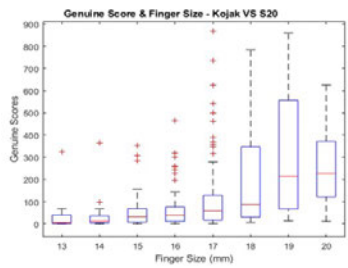



(a)

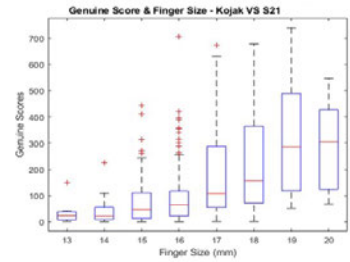



(b)

Fig. 8: Genuine Score Against Total Ridge Count App A S20 (a) & S21 (b)

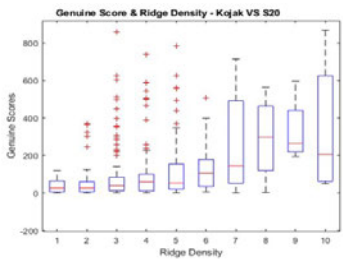



(a)

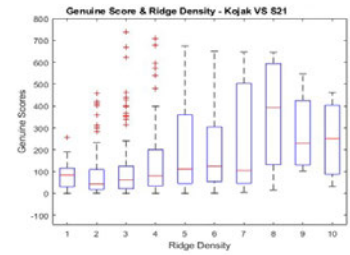



(b)

Fig. 9: Genuine Score Against Finger Size App B S20 (a) & S21 (b)

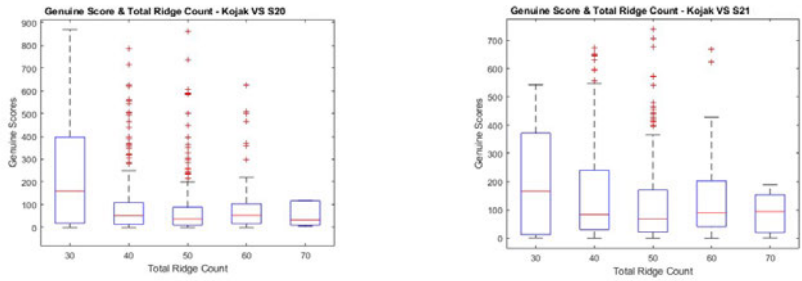



(a)




(b)

Fig. 10: Genuine Score Against Ridge Density App B S20 (a) & S21 (b)




(a) (b)

Fig. 11: Genuine Score Against Total Ridge Count App B S20 (a) & S21 (b)

## 3 Conclusion

Ultimately, these results provide evidence that there is a correlation between finger size and match scores, specifically in contactless fingerprints compared to contact-based prints. It is difficult to accurately determine the total effect that they have on the match score due to the low number of data available at the smallest and largest bins present in the contactless dataset used for this study. The smallest finger size bins typically displayed comparatively low genuine scores, while the scores increased and stayed relatively consistent at sizes of 15 mm and up. The smallest finger sizes did have the most variation between the different devices, and further investigation is needed to determine the cause of this. In a similar fashion, ridge density showed the same trend as the finger size result with little to no correlation in each case except the cellular device fingerprint images. The total ridge count appeared to have little to no correlation across any device. The observation of fingerprints captured using smartphone apps resulted in the highest variability in results expected because this is the newest modality and has had little time for optimization and refinement. These results have major implications on how contactless fingerprint app developers scale finger images prior to image processing to produce a contact-equivalent image. To further this research, these experiments need to be performed on a larger dataset consisting of more variability in finger size, specifically containing exceptionally large and small fingers. The distribution of finger sizes will most likely retain the same distribution observed in this study based on the average finger sizes, but it is desirable to have more data to have a higher sample of the outliers in finger size.

## References


[AS06] Abhyankar, A.; Schuckers, S.: Fingerprint Liveness Detection Using Local Ridge Frequencies and Multiresolution Texture Analysis Techniques, *2006 International Conference on Image Processing*, pp 321-324, 2006.

[BND22] Berti, A.; Nasrabadi, N.; Dawson, J.: Investigating the Impact of Demographic Factors on Contactless Fingerprint Interoperability, *Lecture Notes in Informatics (LNI)*, Gesellschaft fur Informatik, Bonn 2022.

[De18] Deb, D.; Chugh, T.; Engelsma, J.; et.al.: Matching Fingerphotos to Slap Fingerprint Images, arXiv:1804.08122, 2018.

[Da19] Dabouei, A.; Soleymani, S.; Dawson, J.; Nasrabadi, N.: Deep Contactless Fingerprint Unwarping, *2019 International Conference on Biometrics*, Crete, Greece, 2019.

[Gr22] Grosz, S.; Engelsma, J.; Liu, E.; Jain, A.: C2CL: Contact to Contactless Fingerprint Matching, *IEEE Transactions on Information Forensics and Security*, Vol. 17, pp. 196-210, 2022.

[HE16] Hancock, R.; Elliot, S.: Evidence of correlation between fingerprint quality and skin attributes, *2016 IEEE International Carnahan Conference on Security Technology (ICCST)*, 2016.

[Li18] Libert, J.; Grantham, J.; Bandini, B. et.al.: Guidance for Evaluating Contactless Fingerprint Acquistion Devices, *National Institute of Standards and Technology, Gaithersburg*, 2018.

[MP17] Mil'shtein, S.; Pillai, A.: Perspectives and limitations of touchless fingerprints, *IEEE International Symposium on Technologies for Homeland Security (HST)*, pp. 1–6, 2017.

[MS16] Madhavi, K.; Sreenath, B.: Rectification of distortion in a single rolled fingerprint, *2016 International Conference on Circuits, Controls, Communications and Computing (I4C)*, pp. 1-4, 2016.

[RM11] Ramaiah, N.; Mohan, C.: De-noising Slap Fingerprint Images for Accurate Slap Fingerprint Segmentation, *2011 10th International Conference on Machine Learning and Applications and Workshops*, pp. 208-211, 2011.

[TK20] Tan, H.; Kumar, A.: Towards More Accurate Contactless Fingerprint Minutiae Extraction and Pose-Invariant Matching, *IEEE Transactions on Information Forensics and Security*, Vol. 15, pp. 3924-3937, 2020.

[Wi21] Williams. B,; McCauley, J.; Dando, J.; Nasrabadi, N.; Dawson, J.: Interoperability of Contact and Contactless Fingerprints Across Multiple Fingerprint Sensors, *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2021.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# Impact of Data Breadth and Depth on Performance of Siamese Neural Network Model: Experiments with Two Behavioral Biometric Datasets

Ahmed Anu Wahab[1], Daqing Hou[2]

**Abstract:** Deep learning models, such as the Siamese Neural Networks (SNN), have shown great potential in capturing the intricate patterns in behavioral data. However, the impact of dataset breadth (i.e., the number of subjects) and depth (i.e., the amount of data per subject) on the performance of these models remain unexplored. To this end, we have conducted extensive experiments using two publicly available large datasets (Aalto and BrainRun), varying both the number of training subjects and the number of samples per subject. Our results show that dataset depth plays a crucial role in capturing more intricate variations specific to individual subjects, thereby positively influencing the performance of the SNN models. On the other hand, increasing the dataset breadth enables the model to effectively capture more inter-subject variability, which proved to be a more significant factor in improving the overall model performance. Specifically, once a certain threshold for the number of training subjects is surpassed, breadth starts to dominate performance and the impact of dataset depth diminishes and disappears. These findings shed light on the importance of dataset breadth and depth in training deep learning models for behavioral biometrics and provide valuable insights for designing more effective authentication systems.

**Keywords:** Deep learning, Behavioral Biometrics, Siamese Neural Network, Dataset Breadth and Depth.

## 1 Introduction

Behavioral biometrics is an emerging solution that leverages a user's unique behavioral patterns for identity verification. Key advantages of behavioral biometrics include passiveness, unobtrusiveness, and cost-effectiveness (requiring no additional hardware), making it an attractive solution [WHS23]. Moreover, unlike other methods, behavioral biometrics can continuously monitor a user's behavioral patterns for anomalies and prevent account takeovers beyond the login point, which is also known as continuous authentication.

Recent years has seen an increasing trend towards using deep learning models in behavioral biometrics [HWD10, AJT20, DZ13]. However, binary classifiers are commonly used in these work, where a model is required for each subject, making it difficult to scale. Other issues include the need for a substantial volume of data per subject to train adequately, as well as the necessity for retraining when new data is added to a subject.

The Siamese Neural Network (SNN) has been used to overcome these limitations for keystroke dynamics [Ac21]. SNN, originally implemented for image classification and person re-identification [Br93], is specifically designed for measuring the similarity between two or more inputs. It trains two or more identical sub-networks, each of which pro-

[1] ECE, Clarkson University, Potsdam, NY, USA, wahabaa@clarkson.edu
[2] ECE, Clarkson University, Potsdam, NY, USA, dhou@clarkson.edu

| Keystroke Dataset | #Subjects | Data per subject |
|---|---|---|
| CMU [KM09] | 51 | '.tie5Roanl' 400 times |
| GreyC (A) [GEAR09] | 133 | 'greyc laboratory' 51 times |
| GreyC (B) [GEAR12] | 83 | 132 samples |
| Clarkson I [Vu14] | 39 | 43,066 keystrokes |
| Clarkson II [Mu17] | 103 | 125,000 keystrokes |
| Buffalo [SCU16] | 148 | 17,000 keystrokes |
| Account Recovery [Wa21] | 44 | 11,218 keystrokes |
| Multi-Keyboard [Wa22a] | 60 | 14,000 keystrokes |
| CU Multi-modality [Ra23] | 88 | 8,782 keystrokes |
| Aalto Mobile [Pa19a] | 37,370 | 15 sentences |
| Aalto Desktop [Dh18] | 168,000 | 15 sentences |

Tab. 1: Public keystroke datasets, with numbers of subject and amount of data per subject.

duces a vector known as embeddings for an input vector - a lower-dimensional representation of the input. Embeddings are then compared to produce a similarity score between the inputs. This network can be used to compute similarity scores for new subjects never seen during training, making it an effective and scalable model in behavioral biometrics.

While it is generally known that deep learning models such as SNNs require large data to train, there are no specific guidelines available on exactly how much data is needed to train them. For example, as shown in Table 1, with the exception of the Aalto desktop and mobile datasets [Dh18, Pa19a], public keystroke datasets typically consist of only a low number of subjects, ranging from a few tens to about a hundred. Will these datasets still be relevant for training deep learning models? Moreover, it is unclear whether the breadth (i.e., the number of subjects) or depth (i.e., the amount of data per subject) of a dataset is more important. This lack of a clear understanding of the nature of data needed for training effective Siamese networks hinders the development of behavioral biometrics.

To this end, we have conducted experiments using a large publicly available keystroke dataset known as the Aalto dataset [Dh18]. We trained with different subsets of the Aalto dataset (breadth-wise and depth-wise) to determine the optimal training size, amount of data per subject, and number of subjects, for achieving high performance with SNN. To further generalize our findings beyond keystroke dynamics, we also experimented with BrainRun [Pa19b], a fairly large mobile dataset with the gesture and motion modalities.

Our experiments provided valuable insights into the roles of dataset breadth and depth in determining the performance of Siamese networks-based behavioral biometric. Specifically, our findings showed that while larger datasets generally resulted in better performance, the breadth of a dataset, as measured by the number of subjects, had a more significant impact on performance than depth. Furthermore, the results provide insights into the critical aspect of determining the levels of performance that can be expected of Siamese networks based on the dataset's breadth and depth. They also provide guidance on which aspect to improve first, and the level of performance improvement that can be achieved by adding more data to the dataset.

## 2 Related Work

Keystroke dynamics have used simple distance classifier and outlier detection methods. These approaches typically involve capturing features such as monographs and digraphs. Commonly used distance classifiers and outlier detection methods include the Manhattan distance, euclidean distance, Mahanalobis distance, k-nearest neighbours, k-means clustering, and their variants [KM09, Ay20, Wa22b, ZDJ12, ZD15].

However, the distance and outlier detection techniques rely heavily on the extracted features and require strong domain knowledge for manual feature engineering. With the advent of deep learning, the trend has shifted towards utilizing deep neural networks for keystroke biometrics, e.g., [DZ13, HWD10, AJT20]. Deep learning offers several advantages, including the ability to automatically extract relevant features from raw keystroke data without the need for explicit or extensive feature engineering.

Although these early work on the application of deep learning for keystroke dynamics achieved better performance compared to the traditional distance-based or outlier detection methods, they were trained and tested on small datasets. As shown in Table 1, most publicly available keystroke datasets are very limited in size, typically with a few tens up to a hundred subjects. These small datasets have limited the full exploration of deep learning in keystroke dynamics until 2018 when the Aalto dataset [Dh18] was released, which has collected 136 million keystrokes data from 168,000 subjects.

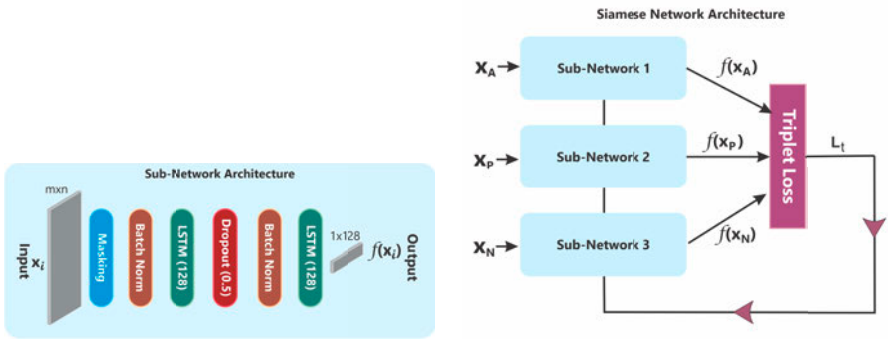
Acien et al. [Ac21] designed an SNN architecture called TypeNet, which is based on Long Short-Term Memory (LSTM) networks. Using the Aalto dataset, the model was trained with 68,000 subjects and tested with 1,000 subjects, achieving a 1.2% EER. Their work showed a significant improvement over past work as it leveraged a large amount of data and was tested on a large number of subjects unseen during training, making it more realistic.

We propose to characterize datasets by breadth and depth, where the breadth is based on the number of subjects in the dataset, and the depth is the number of data per subject. Despite the availability of the large keystroke dataset [Dh18] and the work done by Acien et al. [Ac21], no previous studies have explored the impact of dataset breadth and depth on deep learning performance in keystroke dynamics or any other behavioral biometric modality. Our work is therefore novel in that it addresses this gap in the literature.

## 3 The Siamese Neural Network Architecture

SNN is used to find the similarity between inputs by comparing the output vectors (embeddings) of the sub-networks. As shown in Figure 1a, the Siamese sub-network includes several layers: a masking layer that helps prevent the model from training on zero-padded rows, batch normalization layers that normalize the input data and improve the training speed and stability, two LSTM layers which capture the temporal dependencies in the sequential data, and a dropout layer as regularization to prevent overfitting.

The SNN architecture in Figure 1b consists of three (triplet) sub-networks that share weights and are trained together to learn meaningful representations of input data. Each




(a) Sub-network Architecture (b) Siamese Network Architecture

Fig. 1: (a) The Siamese sub-network, taking a time series input ($\mathbf{x}_i$) of shape $m \times n$ and returning an output vector (embeddings) of shape $1 \times 128$. (b) The Siamese network, consisting of three (3) sub-networks. Loss is calculated from the three output vectors and are back-propagated into the network.

sub-network takes in a single input ($\mathbf{x}_i$) and produces an output vector ($\mathbf{f}(\mathbf{x}_i)$). The first sub-network takes in an anchor sample ($\mathbf{x}_A$), the second a positive sample ($\mathbf{x}_P$), and the third a negative sample ($\mathbf{x}_N$). The anchor and positive samples are drawn from genuine user's data, whereas the negative samples from an impostor. The three output vectors are then passed through the triplet loss function to update the weights of the entire Siamese network. As shown in Equation 1 the triplet loss function minimizes the distance between anchor and positive samples while maximizing the distance between anchor and negative samples, where $\alpha$ is a hyperparameter that controls the degree of separation between the anchor and negative samples in the embedding space.

$$L_t = max\{0, ||\mathbf{f}(\mathbf{x}_A^i) - \mathbf{f}(\mathbf{x}_P^i)||^2 - ||\mathbf{f}(\mathbf{x}_A^i) - \mathbf{f}(\mathbf{x}_N^j)||^2 + \alpha\} \tag{1}$$

## 4 Datasets

The primary dataset used for our experiments is Aalto, a large publicly available keystroke dataset. To generalize our findings, we also utilize BrainRun, a mobile motion dataset.

**The Aalto Keystroke Dataset** The Aalto University desktop [Dh18] dataset is a large-scale controlled free-text dataset collected using an online typing test on desktop computers. The dataset has 136 million keystrokes collected from 168,000 subjects and for a duration of three months, each subject transcribing 15 English sentences which were randomly drawn from a set of 1,525 examples consisting of at least 3 words, and a maximum of 70 characters per sentence. The characters typed can exceed 70 as subjects are allowed to make typing errors, correct them or add new characters when typing. For each character, we extracted 4 time-features (monographs and digraphs) as well as its ASCII value. We filtered out potential outliers by removing rows containing digraphs that exceed 5 seconds.

**The BrainRun Mobile Motion Dataset** The BrainRun dataset [Pa19b] collected raw behavioral data through the BrainRun mobile app, an educational game available on app

stores. This app has five different game types, namely "Focus", "Mathisis", "Memoria", "Reacton", and "Speedy". The data were collected from over 2,000 users using their own devices as they participated in the mini-games. The dataset consists of two behavioral modalities, *gesture* and *motion*. The *gesture* data captures information such as taps and swipes as users plays BrainRun. The *motion* data were collected with a sampling rate of 100 ms, directly from the built-in sensors (accelerometer, gyroscope, magnetometer) in the smartphone, as well as from DeviceMotion, a library in the React Native framework. For our experiments, we utilized the acceleration (along the x, y, and z axis) and rotation data (characterized by alpha, gamma, and beta) that were collected from the React Native framework (DeviceMotion). Although over 2,000 subjects participated in the data acquisition, only 1,132 subjects contributed motion data. The data was originally stored in a json file per user session. To improve data quality, we removed noisy data and applied min-max normalization on each column. Lastly, we split each user's motion data into samples of 200 rows each. To ensure consistency in our experiments, we only considered subjects with at least 15 samples. After the data preprocessing, only 496 out of the initial 1,132 subjects were considered usable for our experiments.

## 5 Experimental Procedures and Results

We conducted multiple experiments to investigate the effect of data size on Siamese networks and determine the most important dimension (breadth or depth) of a dataset for achieving optimal performance. The SNN architecture was implemented using the Tensorflow library on a 24 GB Nvidia GeForce RTX 3090. All experiments were trained with 150 epochs, 150 steps per epoch, 512 sequences per batch, Adam optimizer, and a margin ($\alpha$) of 1.5. Experimental results were reported using the Equal Error Rate (EER) metric. The primary dataset used in these experiments is the keyboard dataset (Aalto), while the mobile motion dataset (BrainRun) was used to validate the generalizability of our findings. To ensure the reliability of our results, we repeated each experiment 10 times, each time with random subject selection. This approach minimizes any potential variation in the results and obtains a more accurate estimate of the model performance.

### 5.1 Experimental Procedures

**Breadth-wise Experiments:** We randomly selected 1,000 out of the 168,000 subjects in Aalto dataset for testing. To ensure that our experiments cover a diverse range of subjects, we randomly selected 10 groups of subjects from the remaining subjects, with replacement, for training. The number of subjects in each group are 125, 250, 500, 1,000, 2,000, 4,000, 8,500, 17,000, 34,000, and 68,000. For each group, we created a data generator for generating the required input triplets for the Siamese network. Since generating all possible triplets from the dataset will be expensive in both time and storage, we randomly generated only 7.6 million triplets, out of the total possible triplets, for training the Siamese network. This number was empirically selected to ensure that the training process was not excessively computationally expensive, while still providing enough triplets to effectively train the model. Furthermore, with this selection, each group of subjects has more than enough triplets data required to train the model. For samples in the Aalto dataset, we employed a sequence length ($m$) of 70, indicating the maximum number of rows of data in each

sample. Any sample data that exceeds this limit would be truncated, while those below it would be zero-padded.

We also conducted an additional experiment in which, instead of generating 7.6 million triplets as described above, we generated a smaller number (120,000 triplets), but from a larger pool of 68,000 subjects. This experiment is meant to investigate how a relatively small amount of data but from a large number of subjects performs, compared with when using a larger data from a smaller pool of subjects.

To further validate our findings, we conducted similar experiments using the BrainRun dataset. The motion data in the BrainRun dataset has more intra-variance and is a less effective modality compared to keystroke data. As a result, we set the sequence length ($m$) to 200, longer than the sequence length used in the Aalto dataset. We set aside 60 subjects out of the 496 usable subjects in the BrainRun dataset for testing purposes. We then randomly created three groups of subjects from the remaining subjects, with replacement, where each group consisted of 125, 250, and 436 training subjects. We generate only 7.6 million triplets for training. With these experiments using the BrainRun dataset, our goal was to show the generalizability of our findings beyond the Aalto dataset.

**Depth-wise Experiments:** The depth-wise experiments aimed to evaluate the impact of the amount of data or samples available per subject on the performance of the Siamese network. To achieve this, we conducted several experiments using the same group of subjects as in the *Breadth-wise* experiments, but with different numbers of samples per user. Specifically, we ran experiments using 5 and 10 samples per user, instead of using only the original 15 samples in the keystroke dataset. By reducing the amount of data available for each subject, we aimed to investigate how well the Siamese network can generalize in a more limited depth-wise data scenario, and whether the network's performance would be significantly affected by this. These experiments provided insights into how much data or samples are required per subject to achieve optimal performance in Siamese networks.

**Model Evaluation:** The evaluation of all models trained on the Aalto keystroke dataset, including both the breadth-wise and depth-wise experiments, was conducted using a dedicated test dataset comprising 1,000 subjects that were excluded from the training process. That is, no overlap exists between the training and test subjects. To ensure fairness and unbiased evaluation across all experiments, the same test users were used for all experiments. Similarly, for the models trained on the BrainRun motion dataset, we conducted testing using a set of 60 users who were not part of the training process.

To evaluate the performance of the models, we followed a standardized testing procedure. For each subject in the $T$ test subjects, we randomly selected one sample from each of the remaining $T-1$ subjects as impostor samples. These impostor samples, along with the $g$ genuine samples, make up the subject's test samples. We computed the pairwise Euclidean distance between the gallery embeddings $(G)$ and the genuine query embeddings $(Q_{\mathrm{g}})$, resulting in genuine similarity scores. Likewise, we computed the pairwise Euclidean distance between the gallery embeddings $(G)$ and the impostor embeddings $(Q_{\mathrm{i}})$, which provided impostor similarity scores. The EER was then computed based on these scores.

| Samples | Average EER (%) for Varying Number of Subjects with 7.6M Triplets | | | | | | | | | | EER - 120K Triplets |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 125 | 250 | 500 | 1K | 2K | 4K | 8.5K | 17K | 34K | 68K | 68K subjects |
| 15 | 7.94 | 4.99 | 2.91 | 1.82 | 1.37 | 1.21 | 1.12 | 1.12 | 1.11 | 1.09 | 2.17 |
| 10 | 8.90 | 5.96 | 3.72 | 2.17 | 1.49 | 1.34 | 1.16 | 1.16 | 1.16 | 1.10 | 2.21 |
| 5 | 10.19 | 8.15 | 5.64 | 3.76 | 2.40 | 1.55 | 1.35 | 1.21 | 1.16 | 1.11 | 2.29 |

Tab. 2: Aalto Dataset: Average EERs for both the breadth-wise experiments (as seen horizontally with varying number of subjects), and the depth-wise experiments (as seen vertically with varying number of samples per subject) with 7.6 million triplets and 120K triplets for training.

### 5.2 Results for Aalto Dataset

**Breadth-wise:** Table 2 shows the average EERs of the SNN models with varying numbers of training subjects from the keystroke dataset. Horizontally the EERs reveal a clear pattern of exponential decay, indicating that the performance of the SNN model improves significantly as the number of training subjects increases. For instance, training the model with 15 samples (7.6 million triplets) obtained from 125 subjects resulted in an average EER of 7.94%. However, when the same number of triplets were obtained from 8,500 subjects with the same 15 samples each, the performance improved to an EER of 1.12%. This observation highlights the importance of increasing the breadth of the training dataset for achieving better performance. Additionally, we observed a diminishing point of improvement, where further increasing the number of subjects had little to no impact on the performance. This trend is evident when comparing the results from 8,500 subjects to 68,000 subjects. Hence, we determined that, for this specific dataset, 8,500 subjects represents an optimal number for effectively training the SNN. These findings underscore the significance of a broader training dataset in achieving significant performance improvements.

Furthermore, we observed a noteworthy finding when training the model with a relatively small number of triplets (120K) generated from a large pool of subjects (68,000) with 15 samples per subject. This model achieved an impressive EER of 2.17%, surpassing the performance of models trained with 7.6 million triplets (15 samples) obtained from 125, 250, and 500 subjects, as shown in Table 2. This result provides further compelling evidence for the significance of dataset breadth in training the SNN models.

**Depth-wise:** The results for the depth-wise experiments with the Aalto dataset are shown vertically in Table 2, highlighting the impact of reducing the number of samples per subject on the performance of SNN. As we decreased the number of samples from 15 to 10 and further to 5, we observed a notable degradation in performance. This finding proves the significant role of dataset depth in effectively training an SNN model. However, it is worth noting that while these performance degradation are more pronounced for models trained with triplets from a smaller pool of subjects (such as 125, 250 or 500 subjects), their significance diminishes and disappears as the number of subjects increases sufficiently. This suggests that the influence of dataset depth becomes less substantial after the number of subjects surpasses a certain threshold, which can be observed at 4K or 8.5K.

| | Avearge EER (%) with 7.6M Triplets | | |
|---|---|---|---|
| **Samples** | **125** | **250** | **436** |
| 15 | 17.56 | 14.01 | 10.31 |
| 10 | 22.61 | 16.18 | 11.53 |
| 5 | 26.80 | 19.35 | 13.78 |

Tab. 3: BrainRun Dataset: Average EERs for both the breadth-wise experiments (as seen horizontally with varying number of subjects), and the depth-wise experiments (as seen vertically with varying number of samples per subject) with 7.6 million triplets for training.

### 5.3 Results for BrainRun Dataset

**Breadth-wise:** We extended our investigation to the BrainRun dataset, which represents an entirely different modality compared to keystroke dataset. Despite maintaining the same training data size, we observed that an SNN model trained with triplets obtained from 436 subjects outperformed those trained with triplets obtained from 125 and 250 subjects as seen from the horizontal values in Table 3.

**Depth-wise:** A vertical look at the results in Table 3, we observed a similar trend showing a performance degradation as the number of samples per user is reduced. While the results of the depth-wise experiment on the Aalto dataset suggest that the influence of dataset depth becomes less significant after a certain threshold of subjects, the limited number of subjects in the BrainRun dataset prevented us from observing the same diminishing effect. However, as the number of subjects increased in the BrainRun dataset, we noticed the EERs of the different sample sizes (15, 10, and 5) converging, leading us to hypothesize that, given a larger number of subjects in the training dataset, the degradation caused by decreasing the dataset depth would also reduce.

## 6 Conclusion

This study investigated the impact of dataset breadth and depth on the performance of a deep learning Siamese network model in the context of behavioral biometrics. The study experimented with two datasets, the Aalto and the BrainRun. The results revealed that both dataset breadth and depth play crucial roles in the model's performance. Increasing the number of subjects involved in the training dataset had a significant positive impact on the model's performance, demonstrating the importance of capturing a wide range of behavioral patterns and accounting for inter-subject variability. On the other hand, the depth of data per subject also influenced performance, but its effect was less pronounced, particularly when a sufficiently large number of subjects were used during training. In conclusion, although dataset depth still holds significance in capturing intricate variations specific to individual subjects, increasing dataset breadth emerged as a more substantial factor in improving the performance of the Siamese network model. The findings highlight the option to increase dataset breadth by including a larger number of subjects, to enhance model generalization and achieve superior performance. In situations where this is not feasible, increasing the dataset depth should be prioritized.

## References

[Ac21] Acien, Alejandro; Morales, Aythami; Monaco, John V; Vera-Rodriguez, Ruben; Fierrez, Julian: TypeNet: Deep learning keystroke biometrics. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2021.

[AJT20] Andrean, Alvin; Jayabalan, Manoj; Thiruchelvam, Vinesh: Keystroke dynamics based user authentication using deep multilayer perceptron. International Journal of Machine Learning and Computing, 10(1):134–139, 2020.

[Ay20] Ayotte, Blaine; Banavar, Mahesh; Hou, Daqing; Schuckers, Stephanie: Fast free-text authentication via instance-based keystroke dynamics. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2(4):377–387, 2020.

[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, 1993.

[Dh18] Dhakal, Vivek; Feit, Anna Maria; Kristensson, Per Ola; Oulasvirta, Antti: Observations on typing from 136 million keystrokes. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. pp. 1–12, 2018.

[DZ13] Deng, Yunbin; Zhong, Yu: Keystroke dynamics user authentication based on gaussian mixture model and deep belief nets. International Scholarly Research Notices, 2013, 2013.

[GEAR09] Giot, Romain; El-Abed, Mohamad; Rosenberger, Christophe: Greyc keystroke: a benchmark for keystroke dynamics biometric systems. In: 2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems. IEEE, pp. 1–6, 2009.

[GEAR12] Giot, Romain; El-Abed, Mohamad; Rosenberger, Christophe: Web-based benchmark for keystroke dynamics biometric systems: A statistical analysis. In: 2012 Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing. IEEE, pp. 11–15, 2012.

[HWD10] Harun, N; Woo, Wai Lok; Dlay, SS: Performance of keystroke biometrics authentication system using artificial neural network (ANN) and distance classifier method. In: International Conference on Computer and Communication Engineering (ICCCE'10). IEEE, pp. 1–6, 2010.

[KM09] Killourhy, Kevin S; Maxion, Roy A: Comparing anomaly-detection algorithms for keystroke dynamics. In: 2009 IEEE/IFIP International Conference on Dependable Systems & Networks. IEEE, pp. 125–134, 2009.

[Mu17] Murphy, Christopher; Huang, Jiaju; Hou, Daqing; Schuckers, Stephanie: Shared dataset on natural human-computer interaction to support continuous authentication research. In: 2017 IEEE International Joint Conference on Biometrics (IJCB). IEEE, pp. 525–530, 2017.

[Pa19a] Palin, Kseniia; Feit, Anna Maria; Kim, Sunjun; Kristensson, Per Ola; Oulasvirta, Antti: How do people type on mobile devices? Observations from a study with 37,000 volunteers. In: Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services. pp. 1–12, 2019.

[Pa19b] Papamichail, Michail D; Chatzidimitriou, Kyriakos C; Karanikiotis, Thomas; Oikonomou, Napoleon-Christos I; Symeonidis, Andreas L; Saripalle, Sashi K: Brainrun: A behavioral biometrics dataset towards continuous implicit authentication. Data, 4(2):60, 2019.

[Ra23] Ray-Dowling, Aratrika; Wahab, Ahmed Anu; Hou, Daqing; Schuckers, Stephanie: Multi-Modality Mobile Datasets for Behavioral Biometrics Research: Data/Toolset paper. In: Proceedings of the Thirteenth ACM Conference on Data and Application Security and Privacy. pp. 73–78, 2023.

[SCU16] Sun, Yan; Ceker, Hayreddin; Upadhyaya, Shambhu: Shared keystroke dataset for continuous authentication. In: 2016 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, pp. 1–6, 2016.

[Vu14] Vural, Esra; Huang, Jiaju; Hou, Daqing; Schuckers, Stephanie: Shared research dataset to support development of keystroke authentication. In: IEEE International joint conference on biometrics. IEEE, pp. 1–8, 2014.

[Wa21] Wahab, Ahmed Anu; Hou, Daqing; Schuckers, Stephanie; Barbir, Abbie: Utilizing Keystroke Dynamics as Additional Security Measure to Protect Account Recovery Mechanism. In: ICISSP. pp. 33–42, 2021.

[Wa22a] Wahab, Ahmed Anu; Hou, Daqing; Banavar, Mahesh; Schuckers, Stephanie; Eaton, Kenneth; Baldwin, Jacob; Wright, Robert: Shared multi-keyboard and bilingual datasets to support keystroke dynamics research. In: Proceedings of the Twelfth ACM Conference on Data and Application Security and Privacy. pp. 236–241, 2022.

[Wa22b] Wahab, Ahmed Anu; Hou, Daqing; Schuckers, Stephanie; Barbir, Abbie: Securing account recovery mechanism on desktop computers and mobile phones with keystroke dynamics. SN Computer Science, 3(5):360, 2022.

[WHS23] Wahab, Ahmed Anu; Hou, Daqing; Schuckers, Stephanie: A User Study of Keystroke Dynamics as Second Factor in Web MFA. In: Proceedings of the Thirteenth ACM Conference on Data and Application Security and Privacy. pp. 61–72, 2023.

[ZD15] Zhong, Yu; Deng, Yunbin: A survey on keystroke dynamics biometrics: approaches, advances, and evaluations. Recent Advances in User Authentication Using Keystroke Dynamics Biometrics, (1):1–22, 2015.

[ZDJ12] Zhong, Yu; Deng, Yunbin; Jain, Anil K: Keystroke dynamics for user authentication. In: 2012 IEEE computer society conference on computer vision and pattern recognition workshops. IEEE, pp. 117–123, 2012.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# Cyclist Recognition from a Silhouette Set

Eijiro Makishima[1], Fumito Shinmura[2], Daigo Muramatsu[3]

**Abstract:** Person recognition from surveillance cameras can be useful for criminal investigations. Currently, gait recognition technology can identify walking individuals, but recognition of people riding bicycles has not been actively investigated, despite cycling being a popular mode of transportation. In this paper, we propose a method to recognize individuals riding bicycles (cyclists) using a silhouette set. We captured two types of cyclist data, normal and rush modes, from five different views, and generated silhouette image sequences from this data. We evaluated accuracy of the proposed method on the silhouette images in identification and verification tasks. The evaluation results demonstrate the effectiveness of our proposed method.

**Keywords:** Cyclist recognition, Bicycle riding person, Silhouette set, Behavioural biometrics, Criminal investigation

## 1 Introduction

Recently, many surveillance cameras have been installed throughout the city, capturing images of stationary or moving individuals. These captured images can provide valuable information for criminal investigations, particularly in identifying the persons depicted. To achieve this, it is essential to develop person recognition methods capable of operating at a distance from surveillance camera images.

One popular approach is gait recognition [NTC05, WWP18], which focuses on recognizing individuals based on their walking patterns. Numerous gait recognition methods have been proposed to achieve robustness against variations in observation views, belongings, clothes [Ch21, Fa20, Hu21], and public databases such as CASIA-B [YTT06] and OUMV-LP [Ta18] are used for accuracy evaluation. Speed-invariant gait recognition methods are also investigated [Xu19]. These methods enable the recognition of walking individuals, even when the resolution of the target person is low. Gait recognition techniques have expanded the capabilities of surveillance cameras in criminal investigations and forensics [Bo11, Iw13, LL14, ICH19].

However, not all individuals in surveillance camera images are walking. Some may be running, while others may be riding bicycles. While there is existing research on recognizing running individuals such as [Xu19], methods for recognizing cyclists have not been adequately discussed to the best of the authors' knowledge. Recently, point clouds gait

[1] Graduate School of Science and Technology, Seikei University, Musashino, Tokyo, dm236210@cc.seikei.ac.jp
[2] Faculty of Science and Technology, Seikei University, shinmuraf@st.seikei.ac.jp
[3] Faculty of Science and Technology, Seikei University, muramatsu@st.seikei.ac.jp

dataset is released [Sh23], and bicycle riding individuals are included in the dataset according to the web page [3], but methods for individuals riding bicycle are not discussed. Therefore, this paper aims to focus on the recognition task of individuals riding bicycles, whom we refer to as ”cyclists” throughout this paper. Examples of the target images for cyclist recognition are shown in Figure 1.


Fig. 1: Silhouette sequences from three cyclists. All individuals are riding bicycles, but regions of bicycles are not included in the target silhouettes.

By comparing gait recognition with cyclist recognition, several common properties can be identified:

- Both walking and pedaling a bicycle involve periodic motion.
- The person's image can be influenced by their shape, pose, and movement.
- Covariates such as observation views, belongings, and clothes can impact the person's images.

On the other hand, there is a significant difference between walking and bicycle riding action. A cyclist can continue moving even without pedaling temporarily. This means that cyclists do not need to pedal continuously to move, and if they stop pedaling, they can still keep moving due to the law of inertia. As an extreme example, cyclist can even move forward while pedaling in the opposite direction. In the case of walking, a person cannot move if they stop moving their feet.

To realize the cyclist recognition, this paper focuses on the following:

**Construction of multi-view cyclist dataset with different modes**
We capture cyclist image data using five cameras. Data associated with two types of speed modes: comfortable and hurry are collected. Together with the RGB images, gender and age information are also collected.

**Implementation of cyclist recognition method from silhouette set**
We propose cyclist recognition pipeline. From the captured image sequences, silhouette

[3] https://lidargait.github.io/ [Confirmed on 10 Jul. 2023]

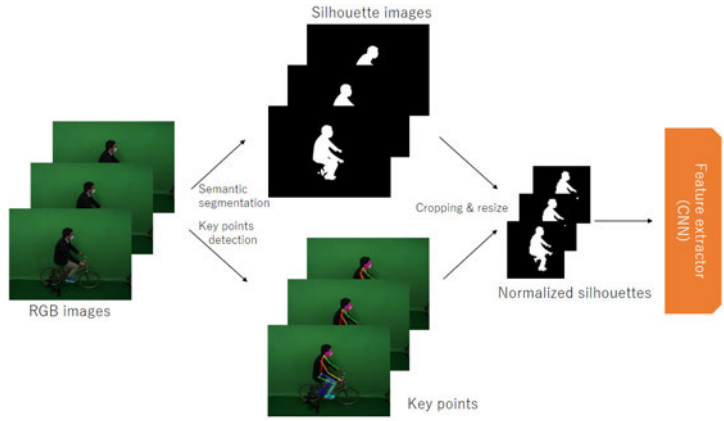



Fig. 2: Overall algorithm of proposed method. From an image, cyclist area is extracted as a silhouette. And using key point information, the silhouette is normalized for feature extraction.

image sequences are generated using semantic segmentation method [Ba22]. These generated silhouettes are normalized using key points information obtained using OpenPose [Ca17]. Considering the properties of bicycle riding actions and inspired by GaitSet [Ch21], we extract cyclist feature from a set of silhouettes, and realize cyclist recognition.

## 2 Cyclist Recognition from Silhouette Set

### 2.1 Overall

Figure 2 shows the overall algorithm of the proposed method. Using the semantic segmentation method [Ba22], each input image is segmented, and segmented part with label ”person” is used as a silhouette of the target person of the image. Then, the silhouette is transformed into size-normalized silhouette. These normalized silhouettes are input to feature extractor and a cyclist feature vector is obtained as the output. This output is used for person recognition.

### 2.2 Pre-processing

Silhouettes of the target person are extracted from RGB images and utilized for recognition. For this extraction, Beit [Ba22] is employed as a semantic segmentation method.

Next, the extracted silhouettes are cropped and resized. While bounding box (BB) information of the target person is commonly used for cropping, it is not suitable for cyclist recognition since the heights of foot points change by pedaling action. A stable reference point is usually waist, as a person sits on the saddle, and the height of the saddle is usually

stable against the ground. To locate the waist point, we employ OpenPose[Ca17] and detect key points of the cyclist. Let $p_n = (x_n, y_n)$, and $p_w = (x_w, y_w)$ be the detected key point of the neck and the waist, respectively. Then we compute the L-2 norm $l_{wn} = ||p_w - p_n||_2$ between two key points of the waist and neck, and set a region of interest (ROI) by two points, the upper left point and the lower right point. The upper left point of the ROI is set by $(x_w - l_{wn}, (y_w + y_n)/2 - c_1 l_{wn})$ and lower right point of $(x_w + l_{wn}, (y_w + y_n)/2 + c_2 l_{wn})$. Here, $(c_1, c_2, c_3)$ is a parameter set for normalization and is set by $(1.5, 2.7, 2.1)$. The cropped data is then resized to $64 \times 64$ [pixels].

### 2.3 Feature Extraction

A silhouette sequence of the target cyclist is input to the feature extractor. Usually, pedaling is a periodic movement, and hence, order information and/or temporal information of the silhouettes can be useful for recognition. However, unlike walking and/or running, a cyclist can keep moving without pedaling a bicycle temporally or with pedaling in reverse direction temporally. This means that usage of temporal information may lead to an erroneous decision. We therefore use silhouette images not as a silhouette sequence, but as a silhouette set. This set means that the order of the silhouette is not used for this method.

In order to realize cyclist recognition from a silhouette set, we focus on a method named GaitSet [Ch21]. GaitSet is proposed for cross-view gait recognition, but we think this method is applicable to cyclist recognition because GaitSet uses a gait silhouette set for input for person recognition and achieves reasonable accuracy.

Let $\mathbf{x} = x_1, x_2, \cdots, x_n$ be a size-normalized silhouette set of cyclist composed of $n$ pieces of silhouette. And let $f(;\theta)$ be a network for feature extraction with parameter $\theta$. In the proposed method, feature $v$ is computed by

$$v = f(\mathbf{x} : \theta). \tag{1}$$

And the parameter $\theta$ is trained by minimizing batch all triplet loss $L$:

$$L = ReLU(\xi + D_{pos} - D_{neg}), \tag{2}$$

where $\xi$ is the margin between intra-class (same person) distance $D_{pos}$ calculated using features pairs originating from the same person, and inter-class (different person) distance $D_{neg}$ calculated using features pairs originating from the different person. And $ReLU()$ is the rectified linear unit function. Please see [Ch21] for detail.

### 2.4 Recognition

Let $\mathbf{v}_i^G$ be the gallery features enrolled in the system associated with the $i$-th person. And let $\mathbf{v}^p$ be a probe feature extracted from an input image sequence. For the recognition, we compute a L1 norm $||\mathbf{v}_i^G - \mathbf{v}^p||_1$, and use the norm for recognition.

## 3 Data construction

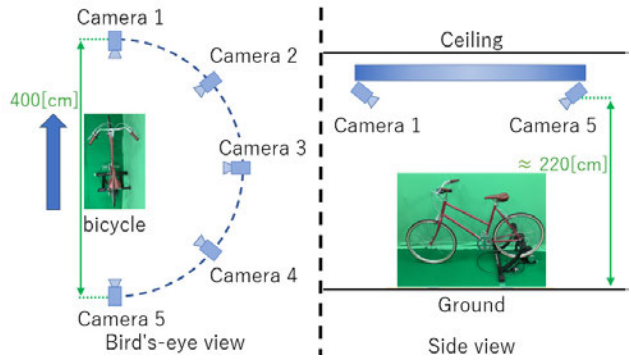



Fig. 3: Environment for data collection

Data collection associated with this research is conducted with the approval of Seikei University Ethics committees. Figure 3 shows the data collection environment. A bicycle is set on the bike trainer, and five cameras are set around the bicycle at a height of about 220cm. For data collection, 38 subjects participated. All subjects pedal a bicycle for less than 1 minute in each session, and they are requested to pedal a bicycle in three sessions. In the first and second sessions, subjects are instructed to pedal a bicycle in a comfortable manner (we call "normal"); on the other hand, subjects are instructed to pedal a bicycle in a hurried manner (we call "rush"). The middle parts of the pedaling are captured for 20 seconds at 30fps and saved for the research. By this collection, cyclist image sequences of three sessions from five views (azimuth angle of 0 (front), 45, 90 (side), 135, and 180 (back) degrees) are collected. Captured images are transformed into silhouette images as explained in preprocessing. Figure 4 shows examples of cyclists' silhouettes from multiple views before size-normalization. Captured data information is summarized in Table 1.

Tab. 1: Captured data information

| | | |
|---|---|---|
| Subjects | 38 | (Male: 31, Female: 7, 20s: 36, 30s: 1, 40s: 1) |
| Sessions | 3 | (Normal, Normal, Rush) |
| Images/session | 600 | (30 [fps] $\times$ 20 [sec]) |
| Views | 5 | (0, 45, 90, 135, 180 [degree]) |

## 4 Experiment

### 4.1 Evaluation protocols

For accuracy evaluation, the subjects are divided into two groups consisting of 19 subjects with no subjects overlapping, and two-fold cross validations are conducted. In order to eliminate the influence of the grouping, the two-fold cross validations were conducted five

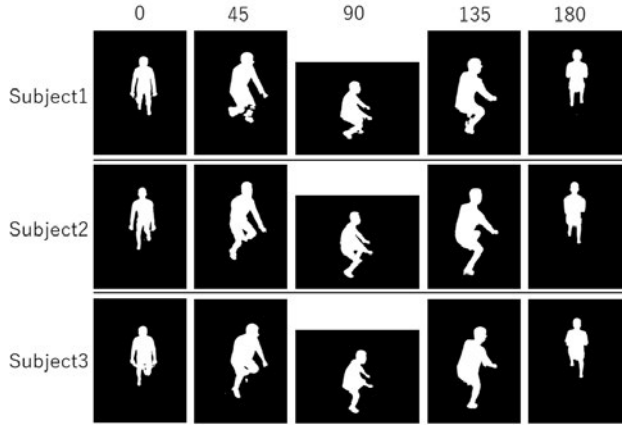



Fig. 4: Silhouette images of three cyclists from five views (azimuth of 0, 45, 90, 135, and 180 [degree])

times with different groupings.
In each session, 600 images of each subject are available for each view. Therefore, we divide the data into 20 sets, so that each set contains 30 images captured in one second. And we use each set as an independent data set.

Accuracy is evaluated through two tasks: identification and verification. Twenty sets from the first session associated with 19 evaluation subjects of one view are enrolled as the gallery ($20 \times 19 = 380$), and the 20 sets from the second and third sessions associated with the 19 evaluation subjects from a view are used for the probe ($20 \times 2 \times 19$). Gallery data are associated with normal pedaling, while the probe data are both of normal and rush pedaling. Because data from five views are available, accuracy associated with $5 \times 5 = 25$ view-settings is evaluated.

For identification, we count the number of probes that have the same identity as the nearest gallery for rank-1 identification rates. For verification, distances associated with all combinations of the gallery and the probe are computed and compared with threshold values; then, false accept rates (FARs), false reject rates (FRRs), and equal error rates (EERs) are measured for accuracy evaluation.

Since data from five views are available in our dataset, we consider 5 times 5 view settings for evaluation. By this evaluation, we can evaluate accuracy of same view settings and cross-view settings.

### 4.2 Experimental results

We summarize the rank-1 identification rates in Table 2 and EERs in Table 3. In these tables, results associated with the probe of normal and the probe of rush are reported

separately; upper parts report the accuracy of normal vs. normal, and lower parts report the accuracy of normal vs. rush. In each part, each diagonal element shows the accuracy with the same view setting, and the others show the accuracy with the cross-view setting.

In the same view settings with normal vs. normal, we can achieve the rank-1s of 99.29, 99.47, 99.42, 99.87, and 96.03% and EERs of 2.37, 2.07, 3.17, 1.93, and 4.51% for views of 0, 45, 90, 135, and 180, respectively. These results show that we can achieve a reasonable recognition accuracy in the case where the views and pedaling modes of the gallery and the probe are the same.

Moreover, in the setting of cross-mode for rush vs. normal, the rank-1s of 86.47, 87.74, 88.40, 88.11, and 83.68% and EERs of 9.21, 9.28, 9.74, 8.30, and 12.00% for views of 0, 45, 90, 135, and 180 degree, respectively. The pose of the upper body in normal pedaling and rush pedaling is significantly different in some cyclists. Figure 5 shows the silhouettes of two cyclists in normal and rush modes. From these silhouettes, we can see that cross-mode cyclist recognition is a challenging task. Considering this fact, achieved recognition accuracy is reasonable.

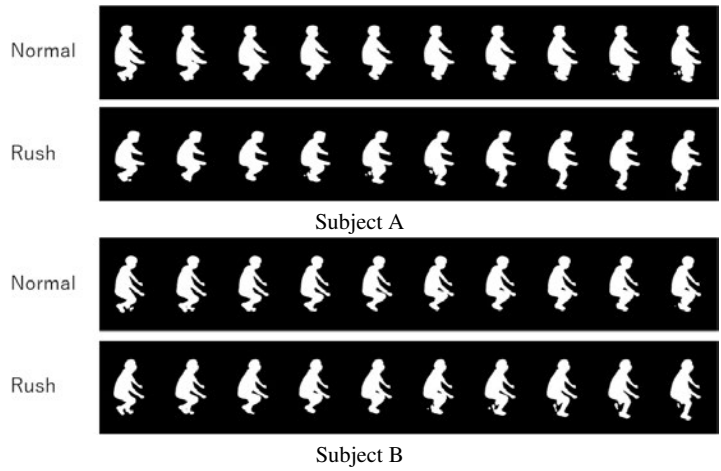



Fig. 5: Difference between normal pedaling and rush pedaling

In the cross-view settings, accuracy greatly deteriorate. The best rank-1 accuracy is 68.74%, and this accuracy is achieved when the view of the probe and the gallery is 135 and 90 degree, respectively. This accuracy should be improved, but it is also promising result because 45 degree is large view difference.

## 5 Conclusion

We propose a cyclist recognition method using a cyclist silhouette set, in this paper. From the image sequences of cyclist, we extract person regions and generate cyclist silhouettes.

Tab. 2: Rank-1 in each view setting

| Probe | | Gallery view [degree], mode=Normal | | | | | |
|---|---|---|---|---|---|---|---|
| Mode | View | 0 | 45 | 90 | 135 | 180 | Average |
| Normal | 0 | 99.29 | 59.50 | 36.82 | 45.42 | 37.66 | 55.74 |
| | 45 | 59.24 | 99.47 | 67.55 | 61.87 | 25.66 | 62.76 |
| | 90 | 37.16 | 67.66 | 99.42 | 68.42 | 24.29 | 59.39 |
| | 135 | 43.05 | 59.15 | 68.76 | 99.87 | 41.45 | 62.46 |
| | 180 | 33.92 | 27.95 | 29.10 | 46.71 | 96.03 | 46.74 |
| Rush | 0 | 86.47 | 53.97 | 33.34 | 41.84 | 30.79 | 49.29 |
| | 45 | 48.74 | 87.74 | 59.87 | 48.31 | 22.03 | 53.34 |
| | 90 | 33.32 | 55.92 | 88.40 | 60.80 | 22.82 | 52.25 |
| | 135 | 41.68 | 54.63 | 57.34 | 88.11 | 35.26 | 55.41 |
| | 180 | 27.95 | 20.61 | 20.11 | 36.34 | 83.68 | 37.74 |

Tab. 3: EER's[%] in each view setting

| Probe | | Gallery view [degree], mode=Normal | | | | | |
|---|---|---|---|---|---|---|---|
| Mode | View | 0 | 45 | 90 | 135 | 180 | All |
| Normal | 0 | 2.37 | 19.63 | 24.39 | 22.60 | 27.37 | 22.10 |
| | 45 | 18.00 | 2.07 | 13.74 | 15.92 | 32.16 | 18.50 |
| | 90 | 24.77 | 13.44 | 3.17 | 13.92 | 32.01 | 20.02 |
| | 135 | 23.44 | 17.25 | 14.20 | 1.93 | 25.16 | 19.17 |
| | 180 | 27.12 | 31.06 | 30.48 | 23.79 | 4.51 | 26.11 |
| Rush | 0 | 9.21 | 21.01 | 26.43 | 25.87 | 32.72 | 24.72 |
| | 45 | 21.86 | 9.28 | 17.84 | 21.07 | 36.47 | 22.49 |
| | 90 | 27.34 | 17.57 | 9.74 | 18.37 | 33.75 | 22.52 |
| | 135 | 23.82 | 16.71 | 17.39 | 8.30 | 27.76 | 20.38 |
| | 180 | 28.86 | 34.10 | 31.88 | 25.77 | 12.00 | 28.43 |

GaitSet [Ch21]-based feature extractor is used for feature extraction from a cyclist silhouette set, and the features are used for identification and verification tasks. For cyclist recognition, we collected cyclist recognition data set from 38 subjects. This is the multi-view data set captured from 5 views, and two types of pedaling, normal and rush, are collected. Proposed method is evaluated on the collected dataset, under the settings of cross-view, and cross-mode. The evaluation results show the potential of cyclist recognition. Now the size of the dataset is small and collected in an indoor environment, we will collect data in outdoor environments from a larger number of subjects. Moreover, in this paper, only one type of bicycle is considered. Different types of bicycles will be considered.

## Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 20H04188.

## References


[Ba22] Bao, Hangbo; Dong, Li; Piao, Songhao; Wei, Furu: BEiT: BERT Pre-Training of Image Transformers. In: International Conference on Learning Representations. 2022.

[Bo11] Bouchrika, I.; Goffredo, M.; Carter, J.; Nixon, M.: On Using Gait in Forensic Biometrics. Journal of Forensic Sciences, 56(4):882–889, 2011.

[Ca17] Cao, Zhe; Simon, Tomas; Wei, Shih-En; Sheikh, Yaser: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In: CVPR. 2017.

[Ch21] Chao, Hanqing; Wang, Kun; He, Yiwei; Zhang, Junping; Feng, Jianfeng: GaitSet: Cross-view Gait Recognition through Utilizing Gait as a Deep Set. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2021.

[Fa20] Fan, Chao; Peng, Yunjie; Cao, Chunshui; Liu, Xu; Hou, Saihui; Chi, Jiannan; Huang, Yongzhen; Li, Qing; He, Zhiqiang: GaitPart: Temporal Part-Based Model for Gait Recognition. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14213–14221, 2020.

[Hu21] Huang, Zhen; Xue, Dixiu; Shen, Xu; Tian, Xinmei; Li, Houqiang; Huang, Jianqiang; Hua, Xian-Sheng: 3D Local Convolutional Neural Networks for Gait Recognition. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14900–14909, 2021.

[ICH19] I, Macoveciuc; CJ, Rando; H, Borrion: Forensic Gait Analysis and Recognition: Standards of Evidence Admissibility. Journal of Forensic Science, 64(5):1294–1303, Sep 2019.

[Iw13] Iwama, H.; Muramatsu, D.; Makihara, Y.; Yagi, Y.: Gait Verification System for Criminal Investigation. IPSJ Trans. on Computer Vision and Applications, 5:163–175, Oct. 2013.

[LL14] Lynnerup, Niels; Larsen, Peter Kastmand: Gait as evidence. IET Biometrics, 3(2):47–54, 6 2014.

[NTC05] Nixon, Mark S.; Tan, Tieniu N.; Chellappa, Rama: Human Identification Based on Gait. Int. Series on Biometrics. Springer-Verlag, Dec. 2005.

[Sh23] Shen, Chuanfu; Fan, Chao; Wu, Wei; Wang, Rui; Huang, George Q.; Yu, Shiqi: LidarGait: Benchmarking 3D Gait Recognition With Point Clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1054–1063, June 2023.

[Ta18] Takemura, Noriko; Makihara, Yasushi; Muramatsu, Daigo; Echigo, Tomio; Yagi, Yasushi: Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. IPSJ Transactions on Computer Vision and Applications, 10:1–14, 2 2018.

[WWP18] Wan, Changsheng; Wang, Li; Phoha, Vir V.: A Survey on Gait Recognition. ACM Comput. Surv., 51(5), aug 2018.

[Xu19] Xu, Chi; Makihara, Yasushi; Li, Xiang; Yagi, Yasushi; Lu, Jianfeng: Multimedia Tools and Applications, 78:26509–26536, 2019.

[YTT06] Yu, Shiqi; Tan, Daoliang; Tan, Tieniu: A Framework for Evaluating the Effect of View Angle, Clothing and Carrying Condition on Gait Recognition. In: 18th International Conference on Pattern Recognition (ICPR'06). volume 4, pp. 441–444, 2006.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# LVT Face Database: A benchmark database for visible and hidden face biometrics

Nelida Mirabet-Herranz[1], Jean-Luc Dugelay[2]

**Abstract:** Although the estimation of eHealth parameters from face visuals (images and videos) has grown as a major area of research in the past years, deep-learning-based models are still challenged by RGB lack of robustness, for instance with changing illumination conditions. As a means to overcome these limitations and to unlock new opportunities, thermal imagery has arisen as a favorable alternative to solidify different technologies such as heart rate estimation from faces. However, the reduced number of databases containing thermal imagery and the lack of health annotation of the subjects in them limits the exploration of this spectrum. Motivated by this, in this paper, we present our Label-EURECOM Visible and Thermal (LVT) Face Database for face biometrics. This database is the first that contains paired visible and thermal images and videos from 52 subjects with metadata of 22 soft biometrics and health parameters. Moreover, we establish the first study introducing the potential of thermal images for weight estimation from faces on our database.

**Keywords:** Face database, Visible spectrum, Thermal spectrum, eHealth, Weight estimation.

## 1 Introduction

Facial processing from visual content has gained a lot of attention in the past years since it allows for non-invasive contactless monitoring of a subject's health status, useful in numerous potential applications. Nowadays, there is a global trend to monitor eHealth parameters without the use of physical devices enabling their estimation in at-risk situations such as medical emergencies and road accidents besides at-home daily monitoring and telehealth. Automatic face recognition has consistently been one of the most active research areas of computer vision [MD18]. Beyond people identification and soft biometric prediction such as gender, age and ethnicity, a vast amount of health information belonging to a subject has been proved to be embedded in face visuals [RBC22]. The estimation of health indicators such as height, weight and Body Mass Index (BMI) from a single facial shot, has been explored in the literature by training a regression method based on the 50-layer ResNet-architecture [DBB18]. Past them, researches have extracted the called micro-signals from faces, information that has played important roles in media security and forensics [Wu20]. An established concept in the past fifteen years is derived from the fact that blood draws more light than the ambient tissues therefore subtle changes in blood volume can be captured by cameras based on the above-mentioned light absorption. This has allowed for remote photoplethysmography (rPPG). Researches have shown how a mobile phone camera has enough resolution to capture rPPG signal from faces leading to a successful Heart

[1] Dept. of Digital Security, EURECOM, 450 Route des Chappes, 06410 France, mirabet@eurecom.fr
[2] Dept. of Digital Security, EURECOM, 450 Route des Chappes, 06410 France, dugelay@eurecom.fr

Rate (HR) estimation [Ra16]. Following the same principle, recent works have successfully approximated the Blood Pressure (BP) of a subject thanks to the difference between the times a pulse wave reaches two different parts of the face [LWM20]. In up-to-date investigations, the ratio of oxygenated hemoglobin with respect to total hemoglobin (SpO2) has been computed from facial videos employing Convolutional Neural Networks (CNN) that consider the direct current and alternating current components extracted from the RGB signals of facial videos [AOI23].

Facial eHealth models traditionally based their estimations on images acquired in the visible spectrum. Despite those networks have reached a significant level of maturity with practical success, deep learning approaches based on data images in the visible spectrum are affected by compromising factors such as occlusion and illumination changes. Thermal imagery has proved itself as a powerful caption tool [MD18]. Computer vision researchers have affirmed it as superior to visible imaging in hard conditions such as the presence of smoke, dust and absence of light sources [ED22]. Thermal imagery operates by detecting electromagnetic radiation in the medium MWIR ($3-8\mu m$) and long LWIR ($8-15\mu m$) wave infrared spectrum [RMY17] where skin heat lays within. This capability enables thermal images to overcome the lack of illumination or some types of occlusions. However, works have highlighted how the thermal heat captured by thermal cameras can be affected by various factors such as ambient temperature or intense physical activity [MD18].

To enable the next step towards more accurate eHealth models and because we believe in the potential of thermal imagery, we introduce a new database with visuals collected using a paired thermal-visible camera and annotated with health traits from each subject. The main contributions of this work are the following: 1) We present our Label-EURECOM Visible and Thermal Face Database for face biometrics composed of 612 images and 416 videos from 52 different subjects and a compendium of 22 health metrics and soft biometrics annotated per person. 2) We propose the first study, up to the authors' knowledge, on weight estimation from facial thermal imagery.

The rest of the paper is organized as follows, Section 2 lists existing databases containing thermal visuals and some descriptors of them as well as motivates the use of thermal images for health-related applications. In Section 3, our LVT Face Database for face biometrics newly collected is presented in detail. Section 4 includes a brief state-of-art on weight estimation from facial images and the results of an up-to-date weight estimator when re-trained with our new thermal images. Finally, Section 5 summarizes and concludes with the future directions of our work. The LVT Face Database for face biometrics is publicly available upon request.

## 2 Potential of visible and thermal paired data

Existing biometric systems and facial eHealth applications, are based on databases acquired in the visible or, lately popular, Near InfraRed (NIR) spectrum. Particular studies have however focused on the thermal spectrum for applications such as cross-spectrum face recognition algorithms or HR estimation.

**Relevant thermal databases:** Interest in employing thermal face images has grown in the past years, nevertheless, this regard has been restricted mostly to tasks such as landmarks and face detection and Face Recognition (FR) [MD18, Ku22]. A relevant subset of FR is Cross-FR (CFR) discipline that aims to identify a person's image in the thermal spectrum from a gallery containing face images acquired in the visible spectrum [An21]. Only a few databases have been provided involving visuals acquired in thermal spectra and among them, the ones covering health-related metadata are few. In Tab. 1, we present an exhaustive selection of relevant databases that include visuals in the thermal spectrum and some key descriptors of them including their year of release, the number of subjects, images and videos present in the database and their initial intended purpose. One of the first datasets containing thermal visual data was presented in 2003 [KB03]. The data was acquired at the University of Notre Dame and contains images from 240 distinct subjects with four views with different lighting and facial expressions with the purpose of recognizing individuals. Beyond people recognition, Wang *et al.* establish a similar database for expression recognition containing both spontaneous and intended expressions of more than 100 subjects [Wa10] while Gault *et al.* recorded thermal videos from 32 subjects under three imaging scenarios and their paired rPPG signals for HR estimation [GF13]. In 2018 two new databases were acquired for FR with multiple illuminations, pose and occlusion variations [MD18] and including imagery from different modalities namely visible, thermal, near-infrared and a computerized facial sketch and 3D images of each volunteer's face [Pa18]. In the same year, Barbosa *et al.* collected thermal videos from 20 healthy subjects in two phases: phase A (frontal view acquisitions) and phase B (side view acquisitions) and the corresponding PPG and thoracic effort simultaneously recorded for HR and Respiratory Rate (RR) estimation [Ba18]. More recently, two large-scale visible and thermal datasets have been assembled. Abdrakhmanova *et al.* gathered a combination of thermal, visual, and audio data streams to support machine learning-based biometric applications [Ab21] and Poster *et al.* presented the largest collection of paired visible and thermal face images up to date. Variability in expression, pose, and eyewear were recorded [Po21]. Following, a thermal face dataset with annotated face bounding boxes and facial landmarks composed of 2556 images was introduced [Ku22].

| Year | Dataset | # of subjects | # of images | # of videos | Objective |
|---|---|---|---|---|---|
| 2003 | UND-X1 [KB03] | 241 | 4584 | - | FR |
| 2010 | NVIE [Wa10] | 215 | Not provided | Not provided | Expression recognition |
| 2013 | TH-HR [GF13] | 32 | - | 96 | HR |
| 2018 | VIS-TH [MD18] | 50 | 2100 | - | FR |
| 2018 | TUFTS [Pa18] | 113 | Over 10000 | 113 | FR |
| 2018 | TH-HR-RR [Ba18] | 20 | - | 40 | HR, RR |
| 2021 | Speaking faces [Ab21] | 142 | - | 45 hours | Biometric Authentication |
| 2021 | ARL-VTF [Po21] | 395 | 549712 | - | Cross-FR |
| 2022 | SF-TL54 [Ku22] | 142 | 2556 | - | Landmarks detection |
| 2023 | Ours: LVT | 52 | 612 | 416 | FR, Soft biometrics, e-health |

Tab. 1: Relevent face databases containing visuals in thermal spectra.

**Thermal data for eHealth:** Although the use of facial thermal imagery has traditionally focused on face recognition tasks, some researchers have intended for eHealth parameter estimation in the thermal spectrum showing the potential of this type of data. In 2017, Rai *et al.* suggested that thermal imaging systems have the prospective of providing details regarding physiological processes using skin temperature distributions due to processes such as blood perfusion. Indeed, cameras are often used to observe minute variations in temperature in the medical field in applications including the detection of malignant tumors [RMY17]. The assessment of eHealth parameters such as heart rate from face videos has been studied in depth in recent years. Up to the authors' knowledge, all methods need proper illumination difficult to achieve in uncontrolled environments. In 2018, Barbosa *et al.* presented a new method for remote HR monitoring based on periodic head movements caused by the cyclical ejection of blood flow from the heart to the head. This new algorithm was based on the use of thermal images as input data [Ba18]. Moreover, they proved possible the evaluation and measurement of a subject's RR by using temperature fluctuations under the nose during the respiratory cycle. Thermal imagery proved itself of high value to overcome illumination constraints since thermal images are light invariant. In the same line, other works continue investigating the future of heart rate and blood pressure extraction from thermal images through deep-learning approaches [NS21].

To the best of our knowledge, current literature focuses on HR, RR and BP from thermal face data. The estimation of other health traits such as SpO2 or weight from thermal images remains untouched by the community. The collection of a new database of visible face visual data and their thermal counterpart is motivated by the potential that thermal images and videos as input data have shown and by the limited number of publicly available databases containing this type of data and their associated health parameters annotation. Moreover, existing databases are limited to visual face information content and one or two parameters. We believe in the value that a database composed of more than 20 different soft biometric and health measures can add to the biometric and health research community.

## 3 Database description

In this section, we first introduce the recording setup of the database and the characteristics of the acquisition devices. We detail the data collection methodology as well as the database design and associated subjects' metadata.

**Acquisition material:** The visible and thermal face visual data was acquired with the dual sensor from the camera FLIR Duo R developed by FLIR Systems. The camera was designed for capturing simultaneously visible and thermal visuals by unmanned aerial vehicles. FLIR Duo R dual camera has been used in recent researches due to its suitability in data collection for different tasks such as face recognition and cross-spectrum applications [ED22, MD18]. The visible and thermal sensors of this camera are a CCD sensor with a pixel resolution of $1920\times1080$ and an uncooled VOx microbolometer with a pixel resolution of $640\times512$ respectively. Various devices were used for a health status assessment of the subjects. A contactless infrared thermometer with a precision of $\pm0.2°$ Celsius (C) between 34°C and 42.0°C and a precision of $\pm0.3$°C in the range of 42.1°C and 43.0°C

was used for computing the user's body temperature. For calculating the BP, an OMRON HEM-7155-E tensiometer was employed together with a LED finger oximeter for SpO2 measurement with a precision of $\pm 2\%$. For HR tracking, the subjects were asked to wear a Garmin Vivoactive®4 smartwatch that embeds an optical PPG sensor able to detect the heart rate by shining a green light through the subject's skin thus reflecting the red cells in the skin's blood vessels. For quantifying bodyweight related measures, we rely on the RENPHO®Body Fat Smart scale. When a subject steps on the device and after entering in the system their gender, age and height, the scale returns 13 metrics including weight and BMI.

**Visuals collection protocol:** Image and video acquisition were performed in an indoor environment where the ambient temperature was set to 25°C. In Fig. 1 we present the arrangements. The acquisition setup included a white wall acting as background, a chair at a fixed distance of 0.25 m from the camera which is placed at a height of 1 meter from the ground, and a two-point lighting kit placed to limit shadows allowing and easing segmentation of the subject from the background. Each volunteer participated in two separate acquisition sessions, with an average time interval of 6 weeks. Before the acquisition process, volunteers were asked to fill out and sign consent forms. The visual data includes 6 images per person (3 visible and their associated thermal pair) in each session with 3 different conditions, Neutral (N), Ambient light(A) and an occlusion in the form of eyeglasses (O) resulting in a total of 612 images. Fig. 2 illustrates example images of an individual from the database. In addition, four 60-second videos are recorded per subject in each session with N conditions. The first pair of videos (one in visible spectrum and its paired thermal) are taken after the subject has been resting for at least 5 minutes and the second pair follows moderate exercise in the form of climbing up stairs to increase their HR values making a total of 408 60s videos.


Fig. 1: Flir Duo R camera (left) and acquisition setup (right).

**Subjects' metadata:** Several metadata pieces of information were collected to describe the subject: gender, age and height. Other parameters were quantified to assess their health status: body temperature, HR, BP, SpO2, weight and BMI. In addition to weight and BMI, the smart scale provided other 11 variables: body fat and body water percentages, skeletal muscle, fat-free weight, muscle mass and bone mass, protein, subcutaneous and visceral fat, Basal Metabolic Rate (BMR) and metabolic age. Image and video filenames are constructed by indicating the visual data spectrum, subject id, session id (1 or 2) and in the case of the images the conditions at the time of acquisition (N, O or A).

Fig. 2: Example images obtained with Flir Duo R camera. The three variations are displayed in visible (upper row) and thermal (bottom row) spectra, from left to right: N, O and A.

**Summary:** The introduced database is devised as a compendium of images, videos, soft biometrics, and health parameters recorded from 52 different subjects in two sessions. It is composed of 612 and 416 face and shoulders images and 60-second videos respectively, corresponding to a total disk space of about 285 GB. The 52 recorded participants, 38 male and 14 female are from 13 different countries from 4 continents and their ages range between 22 and 51 years. An executive summary of the dataset is presented in Tab. 2.

| **Identities** | **Metadata** | | | |
|---|---|---|---|---|
| 52 subjects | **Soft biometrics** | **Health parameters** | | |
| 2 sessions | Gender | Body temperature | BMI | Body mass |
| **Visuals (Thermal and visible)** | Age | SpO2 | Body fat (%) | Bone mass |
| | Height | HR resting | Body water (%) | Proteins |
| 6 paired images in three conditions (N, O, A) | Weight | HR activity | Skeletal muscle | Subcutaneous fat |
| 1 paired 60s videos subject rested | **Biometric** | BP maximum | Fat-free weight | Visceral fat |
| 1 paired 60s videos after physical activity | ID | BP minimum | BMR | Metabolic age |

Tab. 2: Summary of the information contained in the LVT Face Database.

## 4 Preliminary assessment of the database

In this Section we present a preliminary evaluation of thermal data for eHealth parameters estimation to assess the applicability of the database. The suitability of thermal imagery for a subject's weight estimation from face images is tested.

**Weight estimation from face images:** Weight is a soft biometric trait and its estimation from a single facial shot has attracted interest in the research community in the latest years [MHMD23]. Besides being a soft biometric trait, weight is an indicator of a person's health condition, and unlike other biometric traits such as gender and height, body weight fluctuates during a person's life and needs to be periodically re-assessed. Remote estimation of this trait has been signaled of special interest in scenarios when a subject cannot be moved onto a scale due to different disabilities or in the case of road accidents. In such cases, estimating a person's weight from facial appearance allows for an

inexpensive and contactless measurement [MHMD23]. Although some researchers have intended to reduce the error presented by AI-based contactless weight models, existing methods still present several kilograms (kg) of error. Weight estimation models from face data are typically evaluated on the public dataset VIP_attribute consisting of 513 female and 513 male face images of different celebrities and their associated height, weight and BMI metadata [DBB18]. In 2018, Dantcheva *et al.* conducted for the first time a study on the possibility of estimating bodyweight from a subject's face by implementing a ResNet architecture with 50 layers [DBB18] and reported a Mean Absolute Error (MAE) of 8.15 kilograms (kg) of error and a Pearson's correlation coefficient ($\rho$) of $\rho = 0.77$. In 2020, Han *et al.* presented an auxiliary-task learning framework for weight estimation [HZS20] with gender and age as auxiliary traits obtaining in the same dataset a MAE of 7.20 kg. In 2023 Mirabet-Herranz *et al.* defined an optimal transfer learning protocol for a ResNet50 architecture and experimented with different influencing factors such as hair occlusions [MHMD23] achieving a MAE of 6.91 kg and a $\rho = 0.78$.

**Implementation details:** Weight estimation from face images has proved to be possible using deep learning structures known as Residual Neural Networks (ResNet) with 50 layers and a final regression layer [DBB18, MHMD23]. We selected likewise to those studies a ResNet50 structure and we carry out a two-step Transfer Learning (TL) protocol as illustrated in Fig. 3. From a largely trained model intended for age estimation from face images, we complete TL using the visible images in our LVT training set. In the second part, we continue with the pipeline by performing once more TL this time with the thermal images belonging to the LVT training set. Finally, each weight network is tested in the images of the same spectrum found in the LVT test set. Each weight model was re-trained during 10 epochs and the final regression layer during 10 more epochs. The first 20 layers in each TL step were fixed to be frozen. Adam optimizer was adopted, with a learning rate of 0.01 and Huber loss as selected in [MHMD23] with $\delta = 1$.

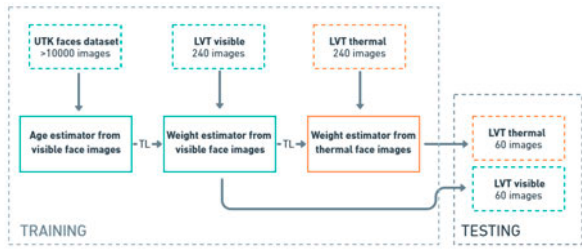



Fig. 3: Transfer learning protocol for weight estimation from visible and thermal images.

**Visible-thermal experimental results:** It is known in the research community that bone, muscle and body fat do not conduct equally temperature [Mo66]. Heat emission patterns can be used to characterize a person since they give information about the location of major blood vessels, skeleton thickness, amount of tissues, and muscle and fat amount [3]. Therefore we believe thermal imagery will access crucial information for weight estimation from faces neglecting skin tissues-related noise and the impact of certain occlusions namely hair. The weight distribution associated with the subjects present in our database

[3] https://biometrics.mainguet.org/types/face.htm#thermogram

has a maximum value of 116.2 kg, a minimum value of 52.3 kg, a Mean=73.54 kg and a STD=14.03 kg. We perform a subject-exclusive split of the training set (480 images from 40 subjects) and the testing set (120 images from 12 remaining subjects). Several metrics are reported in our experiments: the above-mentioned correlation coefficient $\rho$ and the MAE in kg, which are the most common units of measurement in weight estimation research; the root-mean-square error (RMSE) and the Percentage of Acceptable Predictions (PAP) used in [MHMD23] representing the percentage of the prediction whose error is smaller than 10% of the initial weight, i.e. a reasonable error in medical applications. In Tab. 3 the results of the weight network are presented. The metrics show that ResNet50 has a small advantage in the performance of weight estimation when re-trained using thermal data. Both the MAE and RMSE are lower for the thermal network at around 0.3kg. Moreover, the correlation coefficient between the predicted and original weight from the subjects is slightly higher for the thermal spectrum. This confirms the potential of thermal imagery for capturing hidden and more detailed information from human faces.

| Spectrum | MAE | RMSE | Correlation | PAP |
|---|---|---|---|---|
| Visible | 8.31 | 15.03 | 0.43 | 61.6% |
| Thermal | 7.98 | 14.73 | 0.49 | 61.6% |

Tab. 3: Comparison of weight estimation from faces between thermal and visible spectra images.

## 5 Conclusion

This paper presents the novel LVT Face Database for face biometrics. This database contains visuals from 52 subjects under different conditions, resulting in a total of 306 visible and 306 thermal images in addition to 204 visible and 204 thermal videos collected simultaneously using a paired camera (FLIR Duo R) allowing comparison or fusion of those different data types. The visuals acquired are associated with metadata belonging to the subjects both biometric- and health-related. To the best of our knowledge, this is the first database to provide visible-thermal face images and recordings with accompanying gender, age, body temperature, SpO2, BP, HR (resting and after physical activity), height, weight, BMI and 11 additional health metrics. We believe the extensive amount of parameters annotated by every subject will help unlock the potential of thermal data for assessing a person's health status. In addition, we provide preliminary experimental results of weight estimation from facial images using a baseline algorithm with ResNet50 architecture as a backbone, pre-trained with visible images. Results exhibit the potential of thermal data for contactless weight estimation. Based on this promising outcome, future work will focus on considering thermal imagery not only as an alternative to visible but also as a complement. The estimation of other parameters such as SpO2 or height from thermal depictions will be explored.

## Acknowledgment

This work has been partially supported by the European CHIST-ERA program via the French National Research Agency (ANR) within the XAIface project (grant agreement CHIST-ERA-19-XAI-011).

## References


[Ab21] Abdrakhmanova, Madina; Kuzdeuov, Askat; Jarju, Sheikh; Khassanov, Yerbolat; Lewis, Michael; Varol, Huseyin Atakan: Speakingfaces: A large-scale multimodal dataset of voice commands with visual and thermal video streams. Sensors, 21(10):3465, 2021.

[An21] Anghelone, David; Chen, Cunjian; Faure, Philippe; Ross, Arun; Dantcheva, Antitza: Explainable thermal to visible face recognition using latent-guided generative adversarial network. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). IEEE, pp. 1–8, 2021.

[AOI23] Akamatsu, Yusuke; Onishi, Yoshifumi; Imaoka, Hitoshi: Blood Oxygen Saturation Estimation from Facial Video Via DC and AC Components of Spatio-Temporal Map. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 1–5, 2023.

[Ba18] Barbosa Pereira, Carina; Czaplik, Michael; Blazek, Vladimir; Leonhardt, Steffen; Teichmann, Daniel: Monitoring of cardiorespiratory signals using thermal imaging: a pilot study on healthy human subjects. Sensors, 18(5):1541, 2018.

[DBB18] Dantcheva, Antitza; Bremond, Francois; Bilinski, Piotr: Show me your face and I will tell you your height, weight and body mass index. In: 2018 24th International Conference on Pattern Recognition (ICPR). IEEE, 2018.

[ED22] Eddine, Mohamed Jamel; Dugelay, Jean-Luc: GAIT3: An Event-based, Visible and Thermal Database for Gait Recognition. In: 2022 International Conference of the Biometrics Special Interest Group (BIOSIG). IEEE, pp. 1–5, 2022.

[GF13] Gault, Travis; Farag, Aly: A fully automatic method to extract the heart rate from thermal video. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 336–341, 2013.

[HZS20] Han, Dan; Zhang, Jie; Shan, Shiguang: Leveraging auxiliary tasks for height and weight estimation by multi task learning. In: 2020 IEEE International Joint Conference on Biometrics (IJCB). IEEE, pp. 1–7, 2020.

[KB03] Kevin, XCPJF; Bowyer, W: Visible-light and infrared face recognition. In: Workshop on Multimodal User Authentication. Citeseer, p. 48, 2003.

[Ku22] Kuzdeuov, Askat; Koishigarina, Darina; Aubakirova, Dana; Abushakimova, Saniya; Varol, Huseyin Atakan: SF-TL54: A Thermal Facial Landmark Dataset with Visual Pairs. In: 2022 IEEE/SICE International Symposium on System Integration (SII). IEEE, pp. 748–753, 2022.

[LWM20] Lu, Ye; Wang, Chaoqun; Meng, Max Q-H: Video-based contactless blood pressure estimation: A review. In: 2020 IEEE International Conference on Real-time Computing and Robotics (RCAR). IEEE, pp. 62–67, 2020.

[MD18] Mallat, Khawla; Dugelay, Jean-Luc: A benchmark database of visible and thermal paired face images across multiple variations. In: 2018 International Conference of the Biometrics Special Interest Group (BIOSIG). IEEE, pp. 1–5, 2018.

[MHMD23] Mirabet-Herranz, Nelida; Mallat, Khawla; Dugelay, Jean-Luc: New Insights on Weight Estimation from Face Images. In: 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG). IEEE, pp. 1–6, 2023.

[Mo66] Morley, MJ: Thermal conductivities of muscles, fats and bones. International Journal of Food Science & Technology, 1(4):303–311, 1966.

[NS21] Nair, Kavya S; Sarath, S: Illumination invariant non-invasive heart rate and blood pressure estimation from facial thermal images using deep learning. In: 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE, pp. 1–7, 2021.

[Pa18] Panetta, Karen; Wan, Qianwen; Agaian, Sos; Rajeev, Srijith; Kamath, Shreyas; Rajendran, Rahul; Rao, Shishir Paramathma; Kaszowska, Aleksandra; Taylor, Holly A; Samani, Arash et al.: A comprehensive database for benchmarking imaging systems. IEEE transactions on pattern analysis and machine intelligence, 42(3):509–520, 2018.

[Po21] Poster, Domenick; Thielke, Matthew; Nguyen, Robert; Rajaraman, Srinivasan; Di, Xing; Fondje, Cedric Nimpa; Patel, Vishal M; Short, Nathaniel J; Riggan, Benjamin S; Nasrabadi, Nasser M et al.: A large-scale, time-synchronized visible and thermal face dataset. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1559–1568, 2021.

[Ra16] Rahman, Hamidur; Ahmed, Mobyen Uddin; Begum, Shahina; Funk, Peter: Real time heart rate monitoring from facial RGB color video using webcam. 129. Linköping University Electronic Press, 2016.

[RBC22] Ross, Arun; Banerjee, Sudipta; Chowdhury, Anurag: Deducing health cues from biometric data. Computer Vision and Image Understanding, 221:103438, 2022.

[RMY17] Rai, Mritunjay; Maity, Tanmoy; Yadav, RK: Thermal imaging system and its real time applications: a survey. Journal of Engineering Technology, 6(2):290–303, 2017.

[Wa10] Wang, Shangfei; Liu, Zhilei; Lv, Siliang; Lv, Yanpeng; Wu, Guobing; Peng, Peng; Chen, Fei; Wang, Xufa: A natural visible and infrared facial expression database for expression recognition and emotion inference. IEEE Transactions on Multimedia, 12(7):682–691, 2010.

[Wu20] Wu, Min: Exploiting micro-signals for physiological forensics. In: Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security. pp. 1–1, 2020.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# Remote Cancelable Biometric System for Verification and Identification Applications

Hatef Otroshi Shahreza[1,2], Amina Bassit[3], Sébastien Marcel[1,4], Raymond Veldhuis[3,5]

**Abstract:** Cancelable biometric schemes protect the privacy of biometric templates by transforming them, with the help of a key, into an irreversible form that can be replaced if compromised. While these schemes provide more advantages in the user-specific key setting, their application with the user-specific key setting is limited in the identification scenario. Alternatively, the application-specific key setting can be used to employ cancelable biometric systems for the identification scenario. However, in an application-specific key setting, cancelable biometric schemes become static with respect to the protected template replacement; if a protected template or the key is compromised, then the replacement of all the protected templates stored within the same application is mandatory. In addition, experimental results show a degradation of performance for the application-specific key setting in cancelable biometric systems. In this paper, we consider a remote recognition protocol based on cancelable biometric schemes in the identification and verification scenarios so that trusted users can generate protected templates and send them to a server. The server can compare the protected query with the protected templates enrolled in the database for recognition. We investigate the user-specific key setting for cancelable biometric schemes for both verification and identification scenarios, which provides those systems with a dynamic replacement of compromised templates. In our experiments, we analyze different cancelable biometric schemes, including BioHashing, Multi-Layer Perceptron (MLP) Hashing, and Index-of-Maximum (IoM) Hashing. We evaluate their performances when applied within our proposed protocol for face recognition and speaker recognition on the IARPA Janus Benchmark C (IJB-C) and NIST-SRE04-16 datasets for user-specific key and application-specific key settings. The source code of all our experiments is publicly available to facilitate the reproducibility of our work.

**Keywords:** biometric template protection, cancelable biometric, face recognition, identification, speaker recognition, user-specific, verification.

## 1 Introduction

Biometric recognition systems became wildly deployed in authentication and identification solutions. However, in practice, the constant use of biometric data raises serious security and privacy concerns. In particular, it has been shown that the stored templates in the database of a biometric system can be used to reconstruct the underlying biometric data [OSM23b, Ma18a, OSM23c, OSKHM22a, OSM23a], which can lead to a crucial privacy threat for the enrolled users. Data regulations, such as EU General Data Protection Regulation (GDPR) [Re16], consider biometric data as sensitive information, which

[1] Idiap Research Institute, Martigny, Switzeralnd
[2] École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzeralnd
[3] University of Twente, Enschede, Netherlands
[4] Université de Lausanne (UNIL), Lausanne, Switzeralnd
[5] Norwegian University of Science and Technology, Gjøvik, Norway

must be protected. To address privacy-related issues in biometric systems, several biometric template protection (BTP) methods have been proposed in the literature. The ISO/IEC 24745 standard [ISO22] has also defined four requirements for each BTP scheme, including renewability, unlinkability, irreversibility, and performance preservation.

In general, BTP methods can be categorized into *cancelable biometrics* and *biometric cryptosystems*. In *cancelable biometric* schemes, a transformation function, dependent on a key, is used to generate protected templates, and the recognition is based on the comparison of protected templates [JLG04, OSKHM22b, RBB13, Ji17, Ot23]. In *biometric cryptosystems*, a key is either bound with (i.e., key binding schemes) or generated (i.e., key generation schemes) from the unprotected template, and then the recognition is based on correct generation or retrieval of the key [Ul04, Ra22, JW99, JS06].

In general, *cancelable biometric* schemes involve the use of a key in the process of generating protected templates. This key can either be *application-specific*, where the same key is used to protect all the templates within the same application, or *user-specific*, where a different key is used to protect the template of each user, even within the same application. However, in an application-specific key setting, if the key is compromised, then all the protected templates are affected. Moreover, a compromised template can affect the protection of the other protected templates within the same application, with an overwhelming probability the key can be recovered from that compromised template. Since the same key was used, then these require the replacement of all protected templates stored within the same application, which affects the dynamism of such cancelable systems. This limitation does not appear in the user-specific key setting because it only affects the compromised template and the compromised key of the same subject. This motivates us to investigate the user-specific key setting for *cancelable biometric* schemes specifically for the identification scenario.

In this paper, we focus on *cancelable biometric* methods and explore the application of user-specific key and application-specific key settings in these methods for identification and verification scenarios. While most works in the literature focus on the application of *cancelable biometrics* in the verification scenario, few works studied their application for the identification purposes [BG22, BCK09, Mu19, Os22]. In [BCK09], a fingerprint identification method is proposed in which each user has a sensor that has a symmetric key and is time-synchronized with the server. In [BG22], a format-preserving encryption method is used along with Bloom filters [Ra14], as a cancelable biometric, with an application-specific symmetric key in the identification scenario. In [Mu19, Os22], authors proposed indexing protected cancelable templates to accelerate the identification process. The main limitation of applying cancelable biometric systems for the identification scenario is that these systems are often employed in a centralized configuration, and thus the application of user-specific key setting in a centralized system is more suitable for verification, where each subject provides their own key and the system verifies the identity accordingly. Nevertheless, the user-specific key setting in a centralized system has limited application for identification in practice. Alternatively, the application-specific key setting can be used to employ cancelable biometric systems for both identification and verification scenarios. However, compared to user-specific key setting, application-specific key setting suffers

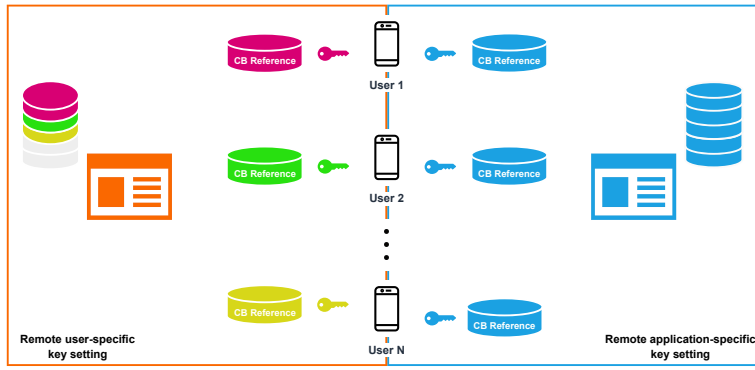



Figure 1: Enrollment in the Remote Cancelable Biometric System.

from security concerns in case the key or a template is compromised and also has inferior performance than the unprotected system.

In this paper, we present a remote recognition protocol, where trusted users can generate canceblable protected templates and send to the server. The server can compare the protected query with the templates enrolled in the database and return recognition result. In contrast to most cancelable biometric methods which are used for verification scenario in centralized systems, our remote protocol can be used for both identification and verification applications and can be used with both user-specific and application-specific key settings. In particular, our protocol enables application of user-specific key setting for identification scenario. In our experiments, we consider different *cancelable biometric* methods, including BioHashing [JLG04], Multi-Layer Perceptron (MLP) Hashing [OSKHM22b], and Index-of-Maximum (IoM) Hashing [Ji17] (i.e., Gaussian random projection-based hashing, shortly IoM-GRP). We evaluate the performance of each scheme in our proposed protocol for face recognition and speaker recognition in identification and verification scenarios on the IARPA Janus Benchmark C (IJB-C) [Ma18b] dataset (face recognition) and NIST-SRE04-16 [Sa17] dataset (speaker recognition) for user-specific key and application-specific key setups.

In the rest of the paper, we first present the protected remote biometric recognition protocol in Section 2. Next, we present our experiments in Section 3. Finally, the paper is concluded in Section 4.

## 2 Remote Cancelable Biometric System

In this section, we present our proposed protocol for a remote cancelable biometric system, which is illustrated in Figure 1 (enrollment) and Figure 2 (recognition) for both one-to-one (i.e., verification) and one-to-many (i.e., identification) comparison scenarios. We assume that each user is able to generate their own key that is safely kept with the user (e.g., as a token, or a seed stored at the user's device, etc.). This key is used to generate its protected reference during the enrollment phase (respectively registration phase) and its protected probe during the verification phase (respectively identification phase).

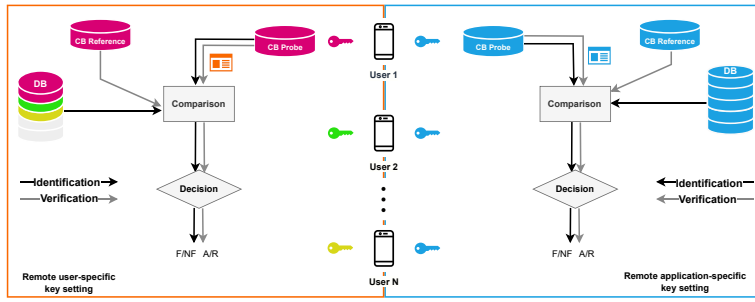



Figure 2: Recognition (identification/verification) in the Remote Cancelable Biometric System.

For the one-to-one comparison, the protected probe needs to be compared to the corresponding protected reference, and based on the comparison score a decision is made. For the one-to-many comparison, the protected probe needs to be compared to all protected references stored in the database, and based on the identification scenario (closed-set or open-set based) decision is made. In the case of closed-set identification, the rank of references is considered and the identity of the reference with the highest similarity is returned. In the open-set scenario, in addition to the value of the highest similarity is also compared to the threshold to avoid false identification.

In order to show the difference between the application-specific and user-specific scenarios, Figure 1 and Figure 2 present an overview of the system in both application-specific and user-specific key settings. In the application-specific key setting, we consider that for the same application, the users are sharing the same key that was distributed among the users during the setup phase. The risk of doing so is that this multiplies the chances of getting this key exposed. Therefore, for a remote biometric recognition scenario, it is safer to consider a user-specific key setting instead of an application-specific key setting in order to restrict the impact of the damage resulting from a leaked key.

## 3 Experiments

### 3.1 Experimental Setup

To evaluate the performance of the remote cancelable biometric system presented in Section 2, we consider face and speaker recognition in our experiments. For the face recognition system, we use ArcFace [De19] as our feature extractor and use the IARPA Janus Benchmark C (IJB-C) [Ma18b] dataset. The IJB-C dataset, which is one of the most challenging evaluation datasets in face recognition research, contains 31,334 images of 3,531 subjects. We use the *test4-G1* protocol in our experiments. For speaker recognition, we use ECAPA-TDNN [DTD20] as our feature extractor and use the NIST-SRE04-16 [Sa17] dataset. We use the *development* set of this dataset, which includes 1407 samples from 85 identities.

In our experiments, we consider different *cancelable biometric* methods, including BioHashing [JLG04], Multi-Layer Perceptron (MLP) Hashing [OSKHM22b], and Index-of-Maximum (IoM) Hashing [Ji17] (i.e., Gaussian random projection-based hashing, shortly IoM-GRP). We apply these schemes for face recognition and speaker recognition for both verification and identification scenarios. We should note that we do not evaluate the security aspect of this system (such as irreversibility and unlinkability) since the security of the mentioned *cancelable biometric* methods have been studied in the literature [JLG04, OSKHM22b, Ji17, OSSM23].

We use the Bob[3] toolbox [An12, An17] for implementation of the biometric pipeline in our experiments. To implement the *cancelable biometric* methods (i.e., BioHashing, MLP-Hashing, and IoM-GRP), we use the open-source implementation of these BTP schemes in Bob [OSM21, OSKHM21, OSKHM22b, Ot23]. The source code from our experiments is publicly available to facilitate the reproducibility of our results[4].

### 3.2 Analaysis

In order to evaluate the effect of the key with respect to the protected template generation, we compare the biometric performances of both application-specific key and user-specific key settings. We consider verification and identification (both open-set and closed-set) for the above scenarios in our experiments, and distinguish between the following experimental scenarios for verification (and respectively for identification):

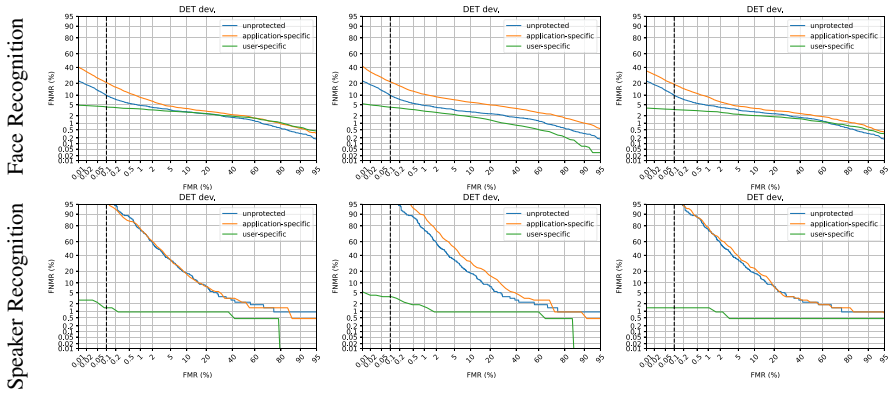
- **Unprotected scenario (baseline):** an unprotected probe $P_i$ is compared against an unprotected reference $R_j$ (respectively references $\{R_j\}_j$).
- **Application-specific key scenario:** a protected probe $P_i$ generated with the key $K$ is compared against a protected reference $R_j$ (respectively references $\{R_j\}_j$) generated with the same key $K$.
- **User-specific key scenario:** a protected probe generated with a key $K_i$ is compared against a protected reference (respectively references $\{R_j\}_j$) generated with its corresponding key $K_j$.

We consider verification and identification (both open-set and closed-set) for the above scenarios in our experiments.

#### 3.2.1 Verification Evaluation

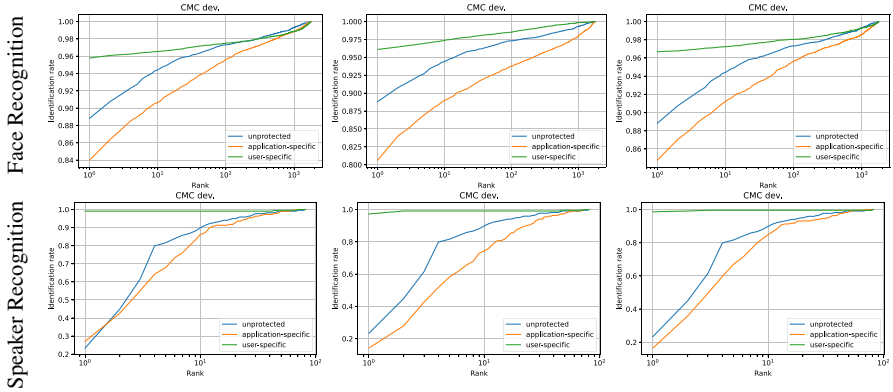
Figure 3 shows the Detection Error Tradeoff (DET) curves for evaluation of the remote cancelable biometric system using different BTP schemes for face and speaker recognition. As the results in this figure show the user-specific key achieves superior performance than the application-specific key and unprotected settings.

[3] Available at `https://www.idiap.ch/software/bob/`
[4] Source code: `https://gitlab.idiap.ch/bob/bob.paper.biosig2023_remote_cb`




(a) BioHash (b) MLP-Hash (c) IoM-GRP

Figure 3: DET curves of remote cancelable biometric system for face recognition (first row) and speaker recognition (second row) using (a) BioHashing, (b) MLP-Hashing, and (c) IoM-GRP schemes.

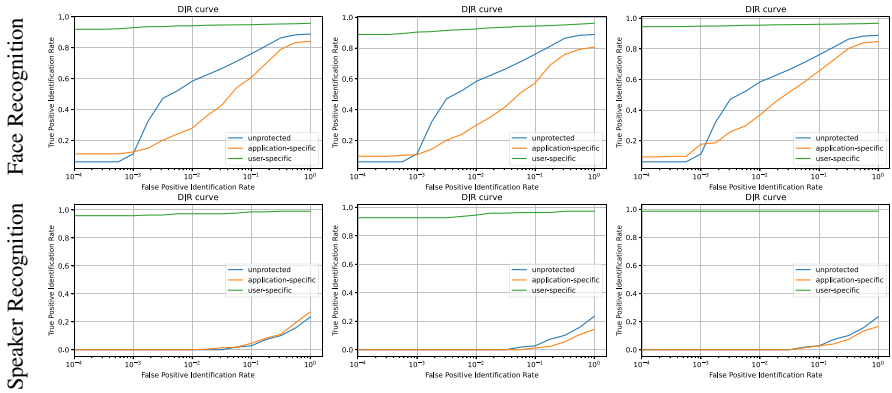



(a) BioHash (b) MLP-Hash (c) IoM-GRP

Figure 4: CMC curves (closed-set identification) of remote cancelable biometric system for face recognition (first row) and speaker recognition (second row) using (a) BioHashing, (b) MLP-Hashing, and (c) IoM-GRP schemes.

### 3.2.2 Identification Evaluation

Figure 4 and Figure 5 show the Cumulative Match Characteristics (CMC) plots (closed-set identification) and Detection and Identification Rate (DIR) plots (open-set identification) for face and speaker recognition in our remote cancelable biometric system using different BTP schemes. Similar to the verification scenario, these results also show that the user-specific key can lead to superior performance. We should highlight that as also discussed in Section 2, in application-specific key setup, the system is at risk that if the key is leaked all the templates need to be replaced with new protected templates. However, the use of




(a) BioHash (b) MLP-Hash (c) IoM-GRP

Figure 5: DIR curves (open-set identification) of remote cancelable biometric system for face recognition (first row) and speaker recognition (second row) using (a) BioHashing, (b) MLP-Hashing, and (c) IoM-GRP schemes.

a user-specific key can enable dynamic management of protected template storage. In the event that the key for one template is leaked, the revocation of that specific template is sufficient, preserving the protection of the remaining protected templates.

## 4 Conclusion

In this paper, we presented a remote cancelable biometric system and investigated its application for verification and identification (open-set or closed-set) applications. In the proposed protocol, trusted users can use a key to generate and send the protected templates to the server, and the server can use the protected template for comparison and decision making for recognition. We explored both user-specific and application-specific key scenarios in our remote cancelable biometric system. In contrast to the application-specific key setting, our experiments demonstrate that the user-specific key setting enhances biometric performance and mitigates the spread of damage caused by a compromised user's key. In addition to the application-specific key setting, our remote cancelable biometric system enables employing the user-specific key setting for verification and identification scenarios.

## Acknowledgment

This research is based upon work supported by the H2020 TReSPAsS-ETN Marie Skłodowska-Curie early training network (grant agreement 860813). This work was also supported by the PriMa project, which has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie (grant agreement 860315).

## 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: Proceedings of the 20th ACM Conference on Multimedia Systems (ACMMM). Oct. 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: ICML 2017 Reproducibility in Machine Learning Workshop. pp. 1–8, 2017.

[BCK09] Bringer, Julien; Chabanne, Hervé; Kindarji, Bruno: Anonymous identification with cancelable biometrics. In: 2009 Proceedings of 6th International Symposium on Image and Signal Processing and Analysis. IEEE, pp. 494–499, 2009.

[BG22] Bansal, Vidhi; Garg, Surabhi: A cancelable biometric identification scheme based on bloom filter and format-preserving encryption. Journal of King Saud University-Computer and Information Sciences, 34(8):5810–5821, 2022.

[De19] Deng, Jiankang; Guo, Jia; Niannan, Xue; Zafeiriou, Stefanos: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4690–4699, 2019.

[DTD20] Desplanques, Brecht; Thienpondt, Jenthe; Demuynck, Kris: ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In: Proc. of Interspeech 2020. pp. 3830–3834, 2020.

[ISO22] ISO/IEC 24745:2022(E) Information technologyy, cybersecurity and privacy protection – Biometric information protection, February 2022.

[Ji17] Jin, Zhe; Hwang, Jung Yeon; Lai, Yen-Lung; Kim, Soohyung; Teoh, Andrew Beng Jin: Ranking-based locality sensitive hashing-enabled cancelable biometrics: Index-of-max hashing. IEEE Transactions on Information Forensics and Security, 13(2):393–407, 2017.

[JLG04] Jin, Andrew Teoh Beng; Ling, David Ngo Chek; Goh, Alwyn: Biohashing: two factor authentication featuring fingerprint data and tokenised random number. Pattern Recognition, 37(11):2245–2255, 2004.

[JS06] Juels, Ari; Sudan, Madhu: A fuzzy vault scheme. Designs, Codes and Cryptography, 38(2):237–257, 2006.

[JW99] Juels, Ari; Wattenberg, Martin: A fuzzy commitment scheme. In: Proceedings of the 6th ACM Conference on Computer and Communications Security. pp. 28–36, 1999.

[Ma18a] 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.

[Ma18b] Maze, Brianna; Adams, Jocelyn; Duncan, James A; Kalka, Nathan; Miller, Tim; Otto, Charles; Jain, Anil K; Niggel, W Tyler; Anderson, Janet; Cheney, Jordan et al.: Iarpa janus benchmark-c: Face dataset and protocol. In: 2018 international conference on biometrics (ICB). IEEE, pp. 158–165, 2018.

[Mu19] Murakami, Takao; Fujita, Ryo; Ohki, Tetsushi; Kaga, Yosuke; Fujio, Masakazu; Takahashi, Kenta: Cancelable permutation-based indexing for secure and efficient biometric identification. IEEE Access, 7:45563–45582, 2019.

[Os22] Osorio-Roig, Dailé; Rathgeb, Christian; Shahreza, Hatef Otroshi; Busch, Christoph; Marcel, Sébastien: Indexing Protected Deep Face Templates by Frequent Binary Patterns. In: 2022 IEEE International Joint Conference on Biometrics (IJCB). IEEE, pp. 1–8, 2022.

[OSKHM21] Otroshi Shahreza, Hatef; Krivokuća Hahn, Vedrana; Marcel, Sébastien: On the Recognition Performance of BioHashing on state-of-the-art Face Recognition models. In: Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, pp. 1–6, 2021.

[OSKHM22a] Otroshi Shahreza, Hatef; Krivokuća Hahn, Vedrana; Marcel, Sébastien: Face Reconstruction from Deep Facial Embeddings using a Convolutional Neural Network. In: Proc. of the IEEE International Conference on Image Processing (ICIP). IEEE, pp. 1211–1215, 2022.

[OSKHM22b] Otroshi Shahreza, Hatef; Krivokuća Hahn, Vedrana; Marcel, Sébastien: MLP-Hash: Protecting Face Templates via Hashing of Randomized Multi-Layer Perceptron. arXiv preprint arXiv:2204.11054, 2022.

[OSM21] Otroshi Shahreza, Hatef; Marcel, Sébastien: Towards Protecting and Enhancing Vascular Biometric Recognition Methods via Biohashing and Deep Neural Networks. IEEE Transactions on Biometrics, Behavior, and Identity Science, 3(3):394–404, 2021.

[OSM23a] Otroshi Shahreza, Hatef; Marcel, Sébastien: Blackbox face reconstruction from deep facial embeddings using a different face recognition model. In: Proc. of the IEEE International Conference on Image Processing (ICIP). IEEE, 2023.

[OSM23b] Otroshi Shahreza, Hatef; Marcel, Sébastien: Comprehensive Vulnerability Evaluation of Face Recognition Systems to Template Inversion Attacks via 3D Face Reconstruction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[OSM23c] Otroshi Shahreza, Hatef; Marcel, Sébastien: Template Inversion Attack against Face Recognition Systems using 3D Face Reconstruction. In: Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2023.

[OSSM23] Otroshi Shahreza, Hatef; Shkel, Yanina Y; Marcel, Sébastien: Measuring Linkability of Protected Biometric Templates Using Maximal Leakage. IEEE Transactions on Information Forensics and Security, 2023.

[Ot23] Otroshi Shahreza, Hatef; Melzi, Pietro; Osorio-Roig, Dailé; Rathgeb, Christian; Busch, Christoph; Marcel, Sébastien; Tolosana, Ruben; Vera-Rodriguez, Ruben: Benchmarking of Cancelable Biometrics for Deep Templates. arXiv preprint arXiv:2302.13286, 2023.

[Ra14] Rathgeb, Christian; Breitinger, Frank; Busch, Christoph; Baier, Harald: On application of bloom filters to iris biometrics. IET Biometrics, 3(4):207–218, 2014.

[Ra22] Rathgeb, Christian; Merkle, Johannes; Scholz, Johanna; Tams, Benjamin; Nesterowicz, Vanessa: Deep face fuzzy vault: Implementation and performance. Computers & Security, 113:102539, 2022.

[RBB13] Rathgeb, Christian; Breitinger, Frank; Busch, Christoph: Alignment-free cancelable iris biometric templates based on adaptive bloom filters. In: Proceedings of the International Conference on Biometrics (ICB). IEEE, pp. 1–8, 2013.

[Re16] Regulation, General Data Protection: Regulation EU 2016/679 of the European Parliament and of the Council of 27 April 2016. Official Journal of the European Union, 2016.

[Sa17] Sadjadi, Seyed Omid; Kheyrkhah, Timothee; Tong, Audrey; Greenberg, Craig; Reynolds, Douglas; Singer, Elliot; Mason, Lisa; Hernandez-Cordero, Jaime: The 2016 NIST Speaker Recognition Evaluation. In: Proc. of Interspeech 2017. pp. 1353–1357, 2017.

[Ul04] Uludag, Umut; Pankanti, Sharath; Prabhakar, Salil; Jain, Anil K: Biometric cryptosystems: issues and challenges. Proceedings of the IEEE, 92(6):948–960, 2004.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# Facial image reconstruction and its influence to face recognition

Filip Pleško[1], Tomáš Goldmann[2], Kamil Malinka[3]

**Abstract:** This paper focuses on reconstructing damaged facial images using GAN neural networks. In addition, the effect of generating the missing part of the face on face recognition is investigated. The main objective of this work is to observe whether it is possible to increase the accuracy of face recognition by generating missing parts while maintaining a low false accept rate (FAR). A new model for generating the missing parts of a face has been proposed. For face-based recognition, state-of-the-art solutions from the DeepFace library and the QMagFace solution have been used.

**Keywords:** Face reconstruction, Face recognition, GAN, ArcFace, SFace, QMagFace

## 1 Introduction

Can reconstructing a corrupted facial image improve facial recognition accuracy? Image reconstruction is a challenging task, in which it is necessary to be sure that the drawn missing part blends well with the known environment and at the same time that the resulting image is not blurred. This paper focuses on the reconstruction of the face image, which adds even more difficulty to this task as the face contains many unique key features. Suppose a key feature such as an eye, nose, or mouth is covered. In that case, generating this feature in the correct location is necessary, as even a slight deviation can cause the resulting face to be deformed and easily recognizable as the generated one.

Current existing solutions try to solve this problem using different GAN architectures. This model architecture contains two other models competing against each other, playing the adversarial game in which they are trying to beat each other and, by doing so, improve the quality of the output. Initially, the generated images are blurry, and it is easy for the Discriminator to detect them. The Generator gradually improves and generates sharper images which are much more difficult to tell apart from real.

This paper further studies the effect of face reconstruction on face recognition tasks. We compare several state-of-the-art solutions for the face recognition task to study whether the reconstruction helps increase recognition precision while maintaining a low rate of false accept rate.

[1] Faculty of Information Technology, Department of Intelligent Systems, Bozetechova 2, Brno, Czech Republic, xplesk02@stud.fit.vutbr.cz
[2] Faculty of Information Technology, Department of Intelligent Systems, Bozetechova 2, Brno, Czech Republic, igoldmann@fit.vut.cz
[3] Faculty of Information Technology, Department of Intelligent Systems, Bozetechova 2, Brno, Czech Republic, malinka@fit.vut.cz

The main contributions of this paper may be stated as follows:

- We study how different layers of a neural network affect the reconstruction of a corrupted facial image.
- We perform experiments to evaluate the influence face reconstruction algorithm to face recognition performance.

## 2 Related works

A task such as facial reconstruction has been studied in several different papers. Each of them has developed a unique way of approaching the problems inherent in this task. In designing our solution, we came across approaches such as Generative face completion [Li17a], G-NST [ZHZ20] and DFNet [Ho19]. The Generative face completion is specific by using two discriminators and a semantic parsing network. One discriminator is used as a local loss for generating the missing parts and the other one is used as a global loss to check whether the generated parts fit into the whole image. The G-NST method uses an additional application of neural style transfer to achieve visual coherence. The method first performs image style clustering using a special model that recognizes different facial features. Then the style transfer is performed using VGG-16, which ensures visually pleasing results. DFNet is based on the well-known U-Net network, into which they designed a special fusion block that they connected to several decoder layers. This approach is designed to focus mainly on filling in the missing parts, as opposed to other solutions that try to generate the image as a whole.

In the last 10 years, face recognition possibilities have improved enormously. In 2014, an algorithm called DeepFace [Ta14] was introduced, which can be considered as the initiator of using a neural network approach to solve the face recognition problem. Consequently, FaceNet outperformed the algorithm. During this time, researchers are making efforts to obtain larger datasets of face images. Moreover, one potential way to improve the algorithms is by modifying the loss function. Modern loss functions for training neural networks in face recognition utilize the angle distribution of feature vectors, such as A-Softmax [Li17b], AM-Softmax [Wa18a], CosFace [Wa18b], ArcFace [De22], and SFace [Zh21]. In addition to the existing algorithms, a new approach utilizing feature distribution, similar to the previous algorithms, was published in 2021 [Me21]. This approach also takes into account the magnitude size of the feature vectors, known as MagFace. Furthermore, in [Te21], the algorithm was extended by adding the Quality Aware Metric as a comparison metric. Overall, for our experiments, we chose SFace and ArcFace as representatives that utilize feature distribution and Magnitude-Aware Loss function based on both angle distribution and magnitude size of the feature vector.

In our approach to corrupted image reconstruction, we have explored several different modifications of the neural network architecture. Our final model is different in that it uses strided convolutions for downsampling instead of pooling layers. We replaced fully connected layers with convolutions and added skip connections between the encoder and

decoder. All these modifications prove their ability to improve the final reconstruction results.

Nowadays, we focus on improving face recognition specifically for special cases of facial images, including damaged images, images with occlusions on the face, faces captured in difficult poses, and face images affected by varying lighting conditions. However, the image reconstruction algorithms are evaluated by image quality metrics that do not consider the ability to improve face recognition.

# 3 Face reconstruction

This section proposes our approach to reconstructing a corrupted facial image. First, we created a baseline model for image reconstruction. Then, we identified suitable modifications that could positively impact the credibility of the facial reconstruction. When designing the individual modifications, we were inspired by the paper from 2015 by Alec Radford, et al. [RMC16]. We individually investigated each modification, and the successful modifications were combined together. Finally, we applied those modifications to existing architectures and selected the best-performing one for face recognition experiments. At the end of the section, we introduced a dataset to simulate corrupted face images.

## 3.1 Architecture of proposed neural network

In our efforts to create a model that would best reconstruct the damaged areas in the image, we first created a base model to which we added various modifications and investigated how each affected the quality of the result. We assembled the final model from the modifications that helped the model improve the quality of the result.

The base model generator consisted of four encoder blocks and four decoder blocks, where each of the encoder blocks contained a convolutional layer followed by a MaxPooling layer. The encoder was ended by two dense layers, after which the decoder was connected. The decoder combined convolutions and transpose convolutions sequentially to produce the reconstructed image.

In the first modification, we tried to remove the fully connected layers from the generator model and replace them with convolutional layers. The network modified in this way has significantly fewer parameters, which allows more filters to be added to the convolutions. Adding more filters to each convolution aims to improve the model's ability to extract spatial features.

The second change focuses on replacing pooling layers with convolutional layers. If we properly set strides parameters we can retain the down-sampling functionality while preserving important information. The use of these so-called strided convolutions, as opposed to deterministic pooling layers, allows the network to learn its own spatial down-sampling.

In one of the modifications, we also tried to increase the number of convolutional layers in each encoder block to improve feature extraction. However, this modification decreased the final quality and was not used in the final model.

In the last modification of the model design, we observed the effect of adding skip connections to the generator on the image reconstruction quality. Direct skip connections between encoder and decoder layers can be used to preserve information about important details during data encoding and decoding. This allows the network to retain detailed information from higher resolutions.

Prior designs have aimed to improve the generator by presenting several architectural changes. These changes have been proposed as solutions to the challenges associated with the generator's performance. Some of the modifications were found to be suitable for face reconstruction. The following proposal investigates whether combining the individual modifications into a single model can improve the results even further. It also examines whether the combination of the individual modifications can work together. The combined model architecture is shown in Figure 1.

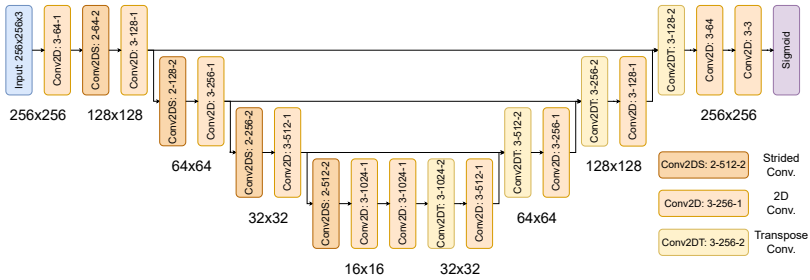



Fig. 1: Architecture of generator using all modifications together.

To find the best working architecture, we tried several different ones. We used models from the keras unet collection [Sh21], updated them with previously mentioned modifications, and tested their performance. The architecture with the best results will be selected for the face recognition experiments.

An important part of the GAN model is the discriminator. Discriminator has original and generated images on its input. Its task is to determine which image was generated and which is real. The harder it is for the discriminator to determine the difference between those two images, the better the generator results are. This feedback is used in the generator to improve the generation. For this task, we used a convolution neural network that extracts input features and classifies them into two classes. The architecture of this discriminator is shown in Figure 2.

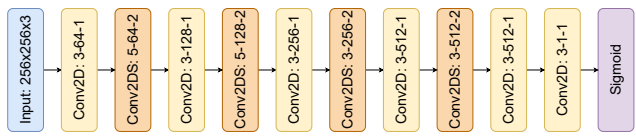



Fig. 2: Discriminator architecture

### 3.2 Training dataset

In this paper, we used CelebA [Li15] dataset for image reconstruction and face recognition tasks. This dataset does not contain damaged images, so some modification was necessary. Therefore, we created the working dataset that we called CelebA-C. This dataset was created by randomly drawing 30 lines of width $8-10$ px and length $10-20$ px filled with RGB Gaussian noise over the face in each image from the original CelebaA dataset. Figure 3 shows an example of corrupted images.

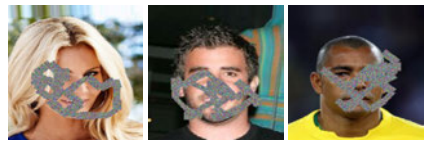


Fig. 3: Example of damaged images.

## 4 Experiments

First, we evaluated and compared the face reconstruction algorithm using PSNR and SSIM, see Section 4.1. Although metrics are commonly used to evaluate the performance of image reconstruction algorithms, these metrics do not consider the impacts on biometric recognition. Due to this drawback, we employ face recognition algorithms to analyze the influence on the accuracy of face recognition.

### 4.1 Performance metrics

Nowadays, the most commonly used metrics to determine the quality of a reconstructed or compressed image are PSNR (Peak Signal to Noise Ratio) (Eq. 1) and SSIM (Structural Similarity Index) (Eq. 2) [HZ10]. Both metrics are commonly used in image comparison to determine how much the modified image has changed from the original image. Calculating these metrics allows us to compare our solution to existing ones.

$$PSNR = 10 \times log_{10}\left(\frac{(2^n-1)^2}{MSE}\right), \tag{1}$$

where $MSE$ is the average squared difference between corresponding pixels of the original and processed images, and $n$ represents the number of bits per pixel in each image, most commonly 8.

$$SSIM(I,I') = \frac{(2\mu_I\mu_{I'}+C_1)(2\sigma_{II'}+C_2)}{(\mu_I^2+\mu_{I'}^2+C_1)(\sigma_I^2+\sigma_{I'}^2+C_2)}, \tag{2}$$

where $\mu_I$ and $\mu_{I'}$ are mean values of the images, $\sigma_{II'}$ stands for the covariance of the images and $\sigma_i^2$ and $\sigma_{I'}^2$ are differences between two images. $C_1 = (k_1L)^2$, $C_2 = (k_2L)^2$ are constants where $k_1 = 0.01$, $k_2 = 0.03$, $L = 255$ [WSB03].

### 4.2 Face reconstruction

We trained and tested the base architecture and all modifications individually. This allowed us to select modifications that improved the final quality and create the architecture that pushed the quality further. The comparison of individual modification to the base model is shown in Table 1.

Tab. 1: Comparison of individual modifications with the base model.

| Model | PSNR | SSIM |
|---|---|---|
| **Base model** | 22.641 | 0.710 |
| **No Dense layers** | 28.814 | 0.893 |
| **Strided convolutions** | 24.015 | 0.751 |
| **Skip connections** | 25.227 | 0.916 |

Using modifications that improved the generated image quality, we created a new model based on the base model but containing all those modifications. At the same time, we applied all those modifications to several architectures from the keras unet collection library and tested what architecture performed the best on image reconstruction. Of all the architectures we tested, U-net and V-net performed best, with V-net leading. Architecture performance comparison is shown in Table 2 with state-of-the-art solutions included as well. Examples of generated images by combined, U-net, and V-Net models are shown in Figure 4.


Fig. 4: Comparison of 3 different generators for generating damaged facial parts. We compare our combined model with tested modifications implemented into U-Net and V-Net architectures.

### 4.3 Face recognition

The second goal of the research is to evaluate the influence of face reconstruction on face recognition. For this purpose, we choose three different algorithms among which the QMagFace [Te21] algorithm is considered the most recent state-of-the-art algorithm for face reconstruction. In the case of ArcFace [De22] and SFace [Zh21], we used L2-distance as a distance metric to determine matches and non-matches. The embeddings generated by QMagFace are compared using the quality-aware score function.

Tab. 2: Comparison of performance of our model with existing solutions.

| Model | PSNR | SSIM |
|---|---|---|
| **Updated U-Net** | 33.736 | 0.968 |
| **Updated V-Net** | 34.326 | 0.972 |
| **Modification combination** | 33.659 | 0.969 |
| **Generative face completion** | 19.500 | 0.784 |
| **G-NST** | 29.655 | 0.937 |
| **DFNet** | 31.662 | 0.965 |

In Table 3, the accuracy is shown for the original CelebA dataset, CelebA-C, and for three datasets obtained by individual face reconstruction algorithms. It is obvious that face reconstruction based on the V-Net (#2) architecture provides the best accuracy for all selected face recognition algorithms.

Tab. 3: Summary of face recognition accuracy on the variants of CelebA dataset.

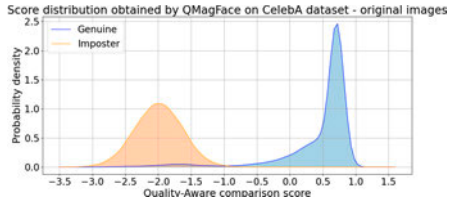
| | CelebA | CelebA-C | U-Net (#1) | V-Net (#2) | Proposed (#3) |
|---|---|---|---|---|---|
| SphereFace | 0.795 | 0.681 | 0.757 | 0.758 | 0.756 |
| ArcFace | 0.864 | 0.602 | 0.819 | 0.824 | 0.810 |
| QMagFace | 0.977 | 0.909 | 0.964 | 0.966 | 0.964 |

In addition, to better visualize the score distributions, the impostor and genuine density distributions were generated for the CelebA dataset, CelebA-C, and reconstructed images using the V-Net face reconstruction algorithm (#2). Due to the similarity between SFace and ArcFace, we introduced the score density distributions only for ArcFace and QMagFace, see Figure 5. As can be seen in Figure 5, the distribution of match pairs produced by the ArcFace algorithm after reconstruction resembles the distribution in the original dataset. However, in the case of the QMagFace algorithm, we can see the discrepancy between the distribution obtained from the original CelebA dataset and the distribution from reconstructed images. This may be due to the use of a quality-aware algorithm that takes into account the quality of the face images.

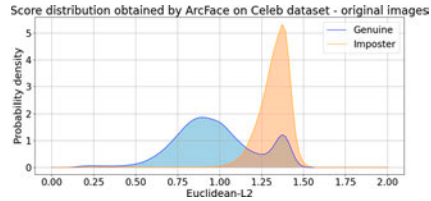
We found that face reconstruction has a positive effect on face recognition. It is evident that for all three recognition algorithms, the use of the reconstruction algorithm had a positive impact on accuracy.

## 5 Conclusion

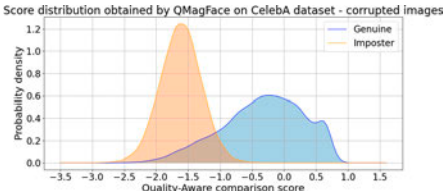
In this work, we have explored various modifications to the neural network architecture to ascertain their positive impact on the reconstruction of the corrupted face image. Based on the findings, we have developed and trained a novel model capable of high-quality reconstruction. When comparing the PSNR and SSIM metrics of our solution with the existing ones, we can see that our solution is as good as the others and even slightly bet-



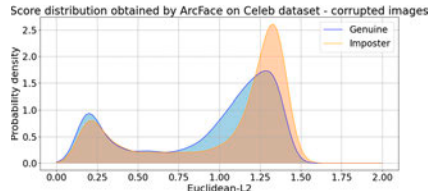

(a) CelebA dataset - QMagFace



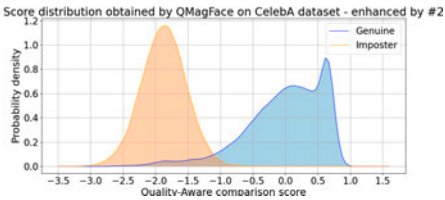

(b) CelebA dataset - ArcFace



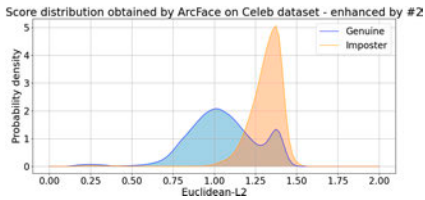

(c) CelebA dataset consists of corrupted images - QMagFace




(d) CelebA dataset consists of corrupted images - ArcFace




(e) Reconstructed CelebA images using V-Net (#2) - QMagFace




(f) Reconstructed CelebA images using V-Net (#2) - ArcFace

Fig. 5: Genuine and impostor score distributions obtained using QMagFace (a,c,e) and ArcFace (b,d,f).

ter. We then used this model to investigate how the reconstruction of a corrupted image affects face recognition. We have found from the results that the use of neural networks can significantly improve the ability of face recognition. The accuracy obtained by evaluating the matched pairs produced by QMagFace applied on CelebA is 97.7 % and for the CelebA-C is only 90.9 %. After reconstructing dataset images, the QMagFace accuracy is 96.4 %. Overall, the recognition accuracy using FaceQMag on the CelebA-C test dataset after image reconstruction was only 1.3 % lower than in the case of the original CelebA. A disadvantage of this solution is that the damage on the image must be masked with Gaussian noise.

## 6 Acknowledgment

This work was supported by the internal project of Brno University of Technology FIT-S-23-8151. Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

## References


[De22] Deng, Jiankang; Guo, Jia; Yang, Jing; Xue, Niannan; Kotsia, Irene; Zafeiriou, Stefanos: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10):5962–5979, oct 2022.

[Ho19] Hong, Xin; Xiong, Pengfei; Ji, Renhe; Fan, Haoqiang: , Deep Fusion Network for Image Completion, 2019.

[HZ10] Hore, Alain; Ziou, Djemel: Image quality metrics: PSNR vs. SSIM. In: 2010 20th international conference on pattern recognition. IEEE, pp. 2366–2369, 2010.

[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). December 2015.

[Li17a] Li, Yijun; Liu, Sifei; Yang, Jimei; Yang, Ming-Hsuan: Generative Face Completion. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5892–5900, 2017.

[Li17b] Liu, Weiyang; Wen, Yandong; Yu, Zhiding; Li, Ming; Raj, Bhiksha; Song, Le: Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 212–220, 2017.

[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. pp. 14225–14234, 2021.

[RMC16] Radford, Alec; Metz, Luke; Chintala, Soumith: , Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, 2016.

[Sh21] Sha, Yingkai: , Keras-unet-collection. `https://github.com/yingkaisha/keras-unet-collection`, 2021.

[Ta14] Taigman, Yaniv; Yang, Ming; Ranzato, Marc'Aurelio; Wolf, Lior: Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1701–1708, 2014.

[Te21] Terhörst, Philipp; Ihlefeld, Malte; Huber, Marco; Damer, Naser; Kirchbuchner, Florian; Raja, Kiran; Kuijper, Arjan: QMagFace: Simple and Accurate Quality-Aware Face Recognition. CoRR, abs/2111.13475, 2021.

[Wa18a] Wang, Feng; Cheng, Jian; Liu, Weiyang; Liu, Haijun: Additive margin softmax for face verification. IEEE Signal Processing Letters, 25(7):926–930, 2018.

[Wa18b] 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.

[WSB03] Wang, Z.; Simoncelli, Eero; Bovik, Alan: Multiscale structural similarity for image quality assessment. volume 2, pp. 1398 – 1402 Vol.2, 12 2003.

[Zh21] Zhong, Yaoyao; Deng, Weihong; Hu, Jiani; Zhao, Dongyue; Li, Xian; Wen, Dongchao: SFace: Sigmoid-constrained hypersphere loss for robust face recognition. IEEE Transactions on Image Processing, 30:2587–2598, 2021.

[ZHZ20] Zhao, Yanshun; Hu, Jinda; Zhang, Xindong: Face Restoration Based on GANs and NST. In: Proceedings of the 2020 5th International Conference on Mathematics and Artificial Intelligence. ICMAI 2020, Association for Computing Machinery, New York, NY, USA, p. 198–203, 2020.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# Human-centered evaluation of anomalous events detection in crowded environments

Giulia Orrù[1] Elia Porcedda[1] Simone Maurizio La Cava[1] Roberto Casula[1] Gian Luca Marcialis[1]

**Abstract:** Anomaly detection in crowd analysis refers to the ability to detect events and people's behaviours that deviate from normality. Anomaly detection techniques are developed to support human operators in various monitoring and investigation activities. So far, the anomaly detectors' performance evaluation derives from the rate of correctly classified individual frames, according to the labels given by the annotator. This evaluation does not make the system's performance appreciable, especially from a human operator viewpoint. In this paper, we propose a novel evaluation approach called "Trigger-Level evaluation" that is shown to be human-centered and closer to the user's perception of the system's performance. In particular, we define two new performance metrics to aid the evaluation of the usability of anomaly detectors in real-time.

**Keywords:** crowd, anomaly detection, human-centered, evaluation.

## 1 Introduction

Anomaly detection in crowd analysis is the process of identifying unusual or unexpected behaviours within a group of individuals [SSM17]. This analysis is widely utilized in security settings, where it can aid in identifying potential threats or incidents that require human intervention [Cr13]. The research community has proposed various methods for detecting irregularities, anomalies, or, in general, patterns that are not representative of expected behaviours in crowded environments. Among the techniques, it is possible to distinguish between methods relying on hand-crafted features and methods relying on deep learning features [AA22]. The textural and spatio-temporal features based on Gabor filters [Ha19] and Optical Flow [Zh15] and methods based on motion information, such as the speed of groups aggregation and disintegration [Or21], are examples of hand-crafted descriptors. However, designing an effective hand-crafted descriptor is not easy. As in other research fields, devoting the feature extraction step to deep classifiers raised the research community's attention. The most common approaches are based on ensemble [Si20], spatio-temporal CNNs [Zh16], and LSTM networks [SV22]. Consequently, data labelling has become increasingly important for machine learning algorithms to learn relevant information from annotated data. Despite the crucial role of labelling, there is no standard methodology nor an agreement on how to proceed [Wa22]. This is partially due to the multitude of tasks that can be included in crowd analysis, such as anomaly recognition and crowd density assessment [AA22]. However, even when the task is common, in addition to the annotator subjectivity, there may be differences in labelling approaches, leading to difficulty in comparing state-of-the-art (SOTA) methods [Be21]. In the detection of anomalous events in crowded environments, the predominant evaluation is at the

[1] University of Cagliari, DIEE, Piazza d'Armi, I-09123 Italy, {giulia.orru, simonem.lac, roberto.casula, marcialis}@unica.it, e.porcedda3@studenti.unica.it

frame-level, whereby the detection models flag each video frame as normal or anomalous, and the percentage of correctly classified frames is assessed [Lu20]. This evaluation does not consider the correct detection of the anomaly onset; therefore, it cannot assess the system's usefulness in real-time applications. In other words, we believe that a most appropriate, human-centered approach to evaluate the performance of crowd anomaly detection systems is necessary. We propose a novel one that allows the system's evaluation at the trigger level. This means that the proposed approach's purpose is to evaluate the proper detection of the *onset of the anomaly*. With *onset*, we refer to the time instant of the beginning of the anomalous sequence. In fact, in real-time applications, promptly detecting the start of the anomaly is essential to ensure effective intervention. From the point of view of the human operator, the system works correctly if it alerts him/her of an anomaly in a reasonable time window and does not raise false alarms. With this aim, the proposed approach led to the definition of two new performance metrics, which we assessed by experiments on two SOTA anomaly detection systems [Ko19, Or21]. Therefore, after overviewing the standard evaluation metrics (Section 2), we describe our human-centered approach leading to two new ones (Section 3). Experiments showing the pros and cons of our contributions are reported (Section 4), and the paper concludes accordingly (Section 5).

## 2 Common evaluation metrics for crowd anomaly detection

Detecting anomalies in crowded environments is commonly treated as a two-class classification problem [Ma10]. The Frame-Level (FL) evaluation criterion is the most commonly employed for video surveillance applications where a time frame sequence is available [RJV22]. A label for each video frame indicates whether or not it contains an anomaly (Figure 1). This criterion is useful for assessing the performance of these systems for offline analysis, as in calculating the duration of an anomalous event present in a video. This is achieved by providing the probability of the presence of an anomaly in a frame at a certain instant $t$ or in a time interval including it, then thresholding such a probability to decide whether the sample has to be considered normal or anomalous. It describes how to count *true positives* (TP), that is, anomalous frames correctly detected, and *false positives* (FP), that is, normal frames incorrectly classified as abnormal, at a given anomaly score threshold. Based on this count, it is possible to obtain various performance metrics such as accuracy, AUC, ROC curves, confusion matrices, F1-score, precision and recall [Je23].

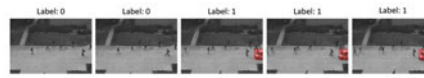



Figure 1: The FL evaluation criterion allows evaluating the duration of an anomaly: in the ground truth, all frames containing an anomaly are labelled with label 1 (positive sample).

However, these metrics allow for evaluation of the classifier used as an anomaly detector, but from a human operator's point of view, they may not be entirely meaningful. In fact, the operator's interest is to be promptly notified when the anomaly occurs. For this reason, the Frame-Level evaluation does not allow the assessment of the responsiveness of the anomaly detection system effectively and, thus, whether the system is able to warn a human operator in a timely fashion. Some SOTA works evaluate the state changes, i.e. the

onset and the end of the anomaly [Br18, SCS18]. However, the evaluation criterion is not standard: in [Br18], the authors evaluate the state change detection error on the total number of frames; in [SCS18], the false alarms percentage is instead calculated. Other temporal metrics for video anomaly detection have been defined in [DY22]. In particular, the Detection Delay, which describes the time between an anomaly and a subsequent alarm, the Alarm Precision, which evaluates the start number of anomalies detected correctly, i.e. in a time window following the actual onset, and the Average Precision Delay, a combination of the previous two have been proposed. These metrics constitute a first step for evaluating the real-time use of a crowd anomaly detector. However, they evaluate an alarm as correct only if subsequent to the ground truth, while as incorrect even if it is raised immediately before the occurrence of the trigger. Therefore, these performance metrics could be strongly influenced by the human operator's subjectivity in labelling the occurrence of the trigger since, if it is considered to be delayed with respect to the actual trigger, the alarm triggered by the system could be wrongly considered incorrect. Hence, they still do not take into account the subjectivity of the labelling, making it not human-centered.

## 3 Trigger-Level evaluation

While designing a real-time crowd anomaly detector, it is essential to determine if the system warns the human operator in time for an intervention. This determination can be made by assessing whether each video frame or batch of frames contains the start of the anomaly or not. Starting from the classification at the frame-level and exploiting the temporal continuous information of the video frames, it is, in fact, possible to obtain further characterizations of the analyzed scene, including the trigger of the anomaly. A sequence of frames classified as anomalous in the frame-level evaluation can be considered as a single anomalous sequence in the trigger-level (TL) evaluation. In this case, the initial frames of the sequence correspond to the trigger (Figure 2). From the human operator's point of view, knowing when to act is crucial. The duration of the anomalous event is much less critical, especially when the priority is to act fast. The TL evaluation measures the number of alarms successfully generated during an anomalous event onset versus the number of false alarms. Since this evaluation criterion is proposed to highlight the system's functioning in support of the human operator, we propose a parameter called "reaction window" $r$. The reaction window sets the time frame for which the human interprets two or more consecutive alarms as a single alarm. Consequently, we can evaluate the system's ability to successfully detect the anomaly trigger through the *Trigger-Level Detection Rate* (TLDR). In particular, it is possible to calculate the TLDR relative to a given reaction window $r$ as:

$$\text{TLDR}(r) = \frac{\text{\# detected anomaly onsets}}{\text{\# ground truth anomaly onsets}} = \frac{\sum_{i=1}^{N} do_i}{N} \quad \text{where} \quad do_i = \begin{cases} 0, & \text{if } \sum_{j=1}^{M} A_j \in [o_i - \frac{r}{2}, o_i + \frac{r}{2}] = 0 \\ 1, & \text{if } \sum_{j=1}^{M} A_j \in [o_i - \frac{r}{2}, o_i + \frac{r}{2}] \neq 0 \end{cases} \tag{1}$$

and $N$ is the number of anomaly onsets $o = o_1, ...o_N$ in the ground truth, while $M$ is the number of output alarms $A = A_1, ..., A_M$ of the detection system. $do_i$ is 1 only when the number of alarms $A$ that fall in the $o_i$ onset reaction window is equal to or greater than 1. It is important to point out that the sample of the onset of anomalous events can consist of a single frame or a batch of frames.

Another parameter of fundamental importance is the *False Trigger Detection Rate* (FTDR):

$$\text{FTDR(r)} = \frac{\text{\# false detected anomaly onsets}}{\text{duration of the sequence } (time)} = \frac{\sum_{j=1}^{M} FA_j}{time} \tag{2}$$

$$\text{where} \qquad FA_j = \begin{cases} 0, & \text{if } \sum_{i=1}^{N} A_j \notin [o_i - \frac{r}{2}, o_i + \frac{r}{2}] \neq 0 \\ 1, & \text{if } \sum_{i=1}^{N} A_j \notin [o_i - \frac{r}{2}, o_i + \frac{r}{2}] = 0 \;\&\; (\sum_{k=j-r}^{j-1} A_k = 0) \end{cases} \tag{3}$$

thus, $FA_j$ is a trigger that falls outside any reaction window associated with ground truth onsets $o$ and such that there are no triggers between the previous $r$ samples. In fact, two or more consecutive triggers in the same reaction window are considered a single false alarm.

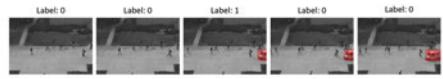



Figure 2: The proposed TL criterion allows using the detection system in a real-time context to intervene promptly: in the ground truth, only the frames relating to the anomaly onset are labelled with label 1 (positive sample).

The TL evaluation criterion, therefore, allows the evaluation of the usability of the detector in real time. A high TLDR indicates that the system can detect the onset of anomalies. A low FTDR indicates that the system does not trigger the human operator unnecessarily. Moreover, the TL evaluation criterion overcomes another limitation of the FL criterion given by comparing models that analyze an input of different sizes: if these take a different number of frames as input, then the number of analyzed samples is different, making the comparison unfair. The TL evaluation, instead, considering a single batch of frames as the beginning of the anomaly, allows for a more balanced comparison.

## 4 Experimental analysis

The models tested in this analysis are two: (i) a 3D-ShuffleNet network (abbreviated SNet) [Ko19], trained to perform human action recognition and fine-tuned on FL crowd anomaly detection. This model returns the probability that each input sequence of 16 frames contains an anomaly; (ii) an SVM (abbreviated GC_SVM) that classifies the features obtained by analysing the speed dynamics of group disruptions and aggregations [Or21]. This model returns the probability that each input sequence of 15 frames contains an anomaly.

Both models were trained in FL mode and we considered a batch of frames as anomalous whenever the probability predicted by the model is greater than 50%. Therefore, we added a module that converts the predicted FL labels into TL labels. Hence, the models can operate in both modes. In particular, we decided to consider as a trigger the first anomalous prediction of a series of at least five anomalous FL predictions. This conversion module can also be implemented with different approaches and can be incorporated into any crowd anomaly detection model that generates FL outputs.

To evaluate the proposed evaluation criteria, we employed the Motion Emotion dataset (MED), representing a fully controlled scenario, and the UFC-Crime dataset, representing an uncontrolled scenario. MED has 31 video sequences totalling around 44000 frames (30 frames per second) acquired with a fixed camera at a height overlooking individual paths.

The videos include normal and abnormal scenarios, with various crowd densities, labelled frame by frame as one of 5 classifications (panic, fight, congestion, obstacle, and neutral). For this work, starting from the TL labels produced in [Or21], we extracted the FL labelling. The UCF-Crime dataset contains 1900 surveillance videos containing normal events and 13 anomalies, including arrest, explosion, fight, and shooting. Due to the missing FL labels, we selected and labelled 240000 frames of 29 videos containing anomalous events more related to crowd analysis, namely fights and shootings.

Following [Or21], in this study, we considered non-overlapped groups of 20 frames, considering each of these batches of frames as an anomalous sample whenever one or more of the frames contains any event considered an anomaly, a normal sample otherwise. For the TL evaluation, we considered as triggers the first batch of frames of a sequence of consecutive batches classified as anomalous. In particular, we assessed the performance with both the FL and the TL criteria through a leave-one-out cross-evaluation. Thus, starting from N videos of each dataset, we excluded a single video at a time and tested the model trained on the N-1 videos with it for each repetition of the evaluation. We performed the experiments separately for the two datasets. For each repetition, we trained the model for 60 epochs with a learning rate of 0.01, reduced by a factor of 0.1 after epochs 30 and 45, employing batches of size equal to 128.

### 4.1 Results

In this Section, we point out the information we may obtain by Trigger-Level evaluation, comparing such information with the Frame-Level evaluation, and how they can be considered "complementary". The results showing the difference between the two evaluation criteria are summarized in Table 1.

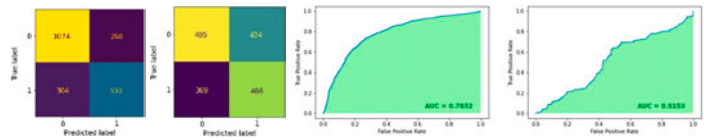
| | | *Frame Level* | | | | | *Trigger Level* | |
|---|---|---|---|---|---|---|---|---|
| **Dataset** | **Method** | **Accuracy** | **Precision** | **Recall** | **F1** | **EER** | **TLDR** | **FTDR** |
| **MED** | **SNet** | 73.89% | 59.91% | 62.65% | 61.25% | 28.12% | 85.19% | 0.45 |
| | **GC_SVM** | 57.27% | 53.35% | 56.79% | 55.02% | 54,86% | 77.78% | 1.01 |
| **UCF-C** | **SNet** | 65.41% | 50.00% | 49.00% | 51.00% | 26,99% | 53.00 % | 0.75 |
| | **GC_SVM** | 51.33% | 61.04% | 29.07% | 39.39% | 41.80% | 46.67% | 0.52 |

Table 1: FL and TL evaluations. The two criteria should be read as complementary.

The MED dataset simulates a controlled context with fixed-position cameras and sparse crowds. In the FL classification mode, SNet correctly classifies 73.89% of the batches, whereas the GC_SVM correctly classifies just 57.27% (Figure 3). The difference in performance is further highlighted by the ROC curves, representing the discrimination capability of the models at various thresholds, from which it is possible to observe higher overall performance with the SNet ($AUC = 0.78$) than with the GC_SVM ($AUC = 0.52$).

In the FL evaluation, the *recall* indicates the detectors' ability to recognise abnormal frames. For both models, the recall is less than 65%. A high percentage of anomalous frames are therefore not detected by the system: this results in a low precision in estimating the duration of the anomaly.

To evaluate whether the system is responsive, we should analyze the results with the TL evaluation. The TLDR, equal to about 85% for SNet and 78% for the GC_SVM, gives us a precise indication of the number of anomalous sequences correctly signalled to the

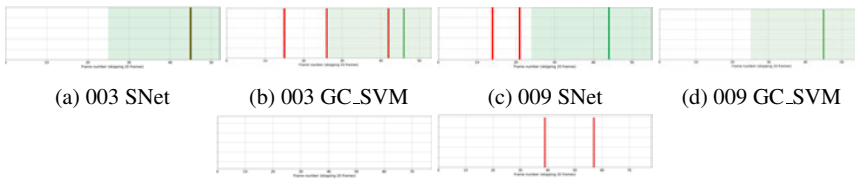



(a) (b) (c) (d)

Figure 3: Confusion matrix (CF) and ROC curves of the FL evaluation for the SNet (a-c) and the GC_SVM (b-d) approach on the MED dataset.

operator. The FLDR enables us to determine whether the detector can be utilised in real-time: we obtained a false alarm every two minutes for the SNet method and one per minute for the GC_SVM. This frequency of false alarms could represent a limit for many real-world applications since it could make the human operator lose trust in the detector.

We can inspect and discuss the reasons behind TL results through plots showing the ground truth and the alarms generated by the system of some test videos (Figures 4 and 5). The *green* areas of length equal to 40 batches related to the reaction window within which a trigger, produced by the system and marked with a *red* line, is considered correct. These areas are placed before and after the anomaly ground truth labels, marked with a *green* line. Among all the videos, three of them have been chosen which are significant due to the model behaviour (Figure 4): *Video 003*- In this panic situation, the SNet precisely detects the anomaly without any delay. Instead, the GC_SVM generates two false alarms, probably due to variations in the speed of disintegration of the groups that the system detects as anomalous. From the SNet FL evaluation, we obtained that 87.50% of the samples are correctly recognized for this video. The TL evaluation shows clearly and precisely that the SNet would be of fundamental importance in signalling this anomaly to a human operator.



(a) 003 SNet (b) 003 GC_SVM (c) 009 SNet (d) 009 GC_SVM

(e) 022 SNet (f) 022 GC_SVM

Figure 4: Trigger evaluation of video 003 (a-b), 009 (c-d) and 022 (e-f) of the MED dataset.

*Video 009*- This video contains only two groups of people who quickly disintegrate after a shot: both detectors do not work properly, the GC_SVM does not detect any anomaly, while the SNet detects the anomaly too early, generating two false alarms. Analyzing the FL classification for the SNet, we noticed that 90% of the frames are classified as anomalous and only 30.91% of the samples are correctly classified. This is because the network considers the overall behaviour of the crowd to be anomalous: indeed, such an arrangement is unusual and typical of specific contexts such as demonstrations or rallies. Unlike the FL evaluation, the TL evaluation allows the simulation of a real application in which

the operator would be warned immediately of the unusual behaviour of the crowd and would have shifted his/her attention to it.*Video 022*- This video depicts a normal situation with individuals walking down the street. The SNet correctly classifies the scene, while the GC_SVM leads to two false alarms. It is important to highlight that three out of four MED videos that do not contain anomalies are correctly classified by the SNet. The overall FTDR of 0.45 alarm/min is probably due to scenarios like the last one, in which the video, initially tagged as normal, exhibits unexpected activity, and the model tends to anticipate the detection. This aspect is of fundamental importance since many false alarms in normal situations make the detection system unreliable.

As shown in Table 1, the results obtained with the UFC-Crime dataset are worse than those obtained with the previous dataset. This is also confirmed by the confusion matrix for the FL classification (Figure 6). The ROC curves also confirm the performance deterioration, with an AUC value equal to 0.66 for the SNet and 0.44 for the GC_SVM. For this dataset, both evaluation criteria highlight the difficulty of anomaly detection. We have therefore selected, also in this case, some significant videos to better explain the network output and the consequent results (Figure 5): *Video 007*- In this representation of a sparse crowd relating to a fight in prison, the SNet detects any anomaly present but produces many false alarms. Analyzing the FL classification, we noticed that many false positives fall into the anomalous sequence broken up by single batches of frames, where the anomaly probability decreases. This highlights a limit of a TL evaluation: in long anomalous sequences, the normal fluctuations of the network probability are interpreted as new anomalous sequences. However, in FL mode, only 64.17% accuracy is achieved for this video, and, therefore, the two criteria agree in negatively evaluating the model's functioning. *Video 009*- This fight

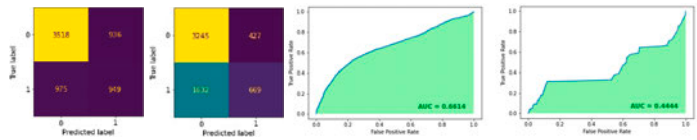
(a) 007 SNet (b) 007 GC_SVM (c) 009 SNet (d) 009 GC_SVM

Figure 5: Trigger evaluation of video 007 (a-b) and 009 (c-d) of the UFC-Crime dataset.

in the street is a case where the operator's sensibility strongly influences the labelling. The SNet does not detect the first anomaly as the approach of two people to a third could also be mistaken for an affection gesture and shows no abrupt or suspicious movements. When the fight begins, the model manages to identify the anomaly. Also, in this case, the final false alarm is related to the probability fluctuation during the anomaly. This behaviour occurs in several UCF-Crime videos, which present many long-lasting anomalies. In a real application context, it does not affect the use of the system as the operator has already been correctly alarmed by the initial trigger.

These described cases exemplify the functioning of the SNet on the UCF-Crime dataset. It is characterized by a correct anomaly classification but by fluctuations of the output during it. This leads to a malfunction in FL mode which is reflected in the TL mode.

These analyses highlighted the complementarity of the two evaluation approaches, namely FL and TL, since they provide different types of clues about the analyzed scenario, static and dynamic information, respectively. In particular, while the FL indicates whether a sin-



(a) (b) (c) (d)

Figure 6: Confusion matrix and ROC curves of the FL evaluation for the SNet (a-c) and the GC_SVM (b-d) approach on the UCF-Crime dataset.

gle frame (or batch of frames) is anomalous or not, the TL can determine the change of the state from a normal context to an anomalous one. However, among the two approaches, the TL evaluation is of fundamental utility for a human operator in interpreting the behaviour of a detector and, thus, for assessing its applicability in a video surveillance context.

## 5 Conclusions

Crowd anomaly detection is crucial in supporting the human operator in many video surveillance tasks. To our knowledge, no standard way is followed to evaluate the system's performance. The most common approach is to assess the rate of (un)correctly classified frames over the length of the video sequence. However, this does not allow the assessment of the responsiveness from the human-operator viewpoint. Therefore, in this work, we proposed a novel temporal evaluation criterion, called Trigger-Level, that aims to assess the ability of a crowd anomaly detection system to trigger an alarm to a human operator correctly. This evaluation focuses on the system's ability to detect the anomaly promptly, i.e. before it significantly impacts the crowd. In particular, two new performance evaluation metrics which allow a better analysis of the responsiveness of an anomaly detection system and its tendency to give false and late alarms were described. The analyses on two datasets and with two detection models demonstrated that the proposed labelling and evaluation approach allowed the assessment of whether the system is able to signal during operation when a human operator should intervene and to evaluate the number of false alarms. This analysis effectively highlights which detector is more suitable for real applications according to the requirements. Finally, we pointed out the complementarity of the frame-level and trigger-level evaluation criteria. Therefore, the two approaches can reveal different properties of the same system: the first provides valuable hints about the probability that the single frame represents an anomalous behaviour, while the second indicates when such an event starts and, thus, a human operator's intervention is required.

**Acknowledgments:** This work is supported by the Italian Ministry of Education, University and Research (MIUR) within the PRIN2017 "BullyBuster - A framework for bullying and cyberbullying action detection by computer vision and artificial intelligence methods and algorithms" (CUP: F74I19000370001). The project has been included in the Global Top 100 list of AI projects addressing the 17 UNSDGs (United Nations Strategic Development Goals) by the International Research Center for Artificial Intelligence under the auspices of UNESCO.

## References

[AA22] Aldayri, Amnah; Albattah, Waleed: Taxonomy of Anomaly Detection Techniques in Crowd Scenes. Sensors, 22(16):6080, 2022.

[Be21] Bendali-Braham, Mounir; Weber, Jonathan; Forestier, Germain; Idoumghar, Lhassane; Muller, Pierre-Alain: Recent trends in crowd analysis: A review. Machine Learning with Applications, 4:100023, 2021.

[Br18] Briassouli, Alexia: Unknown Crowd Event Detection from Phase-Based Statistics. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 1–6, 2018.

[Cr13] Cristani, Marco; Raghavendra, Ramachandra; Del Bue, Alessio; Murino, Vittorio: Human behavior analysis in video surveillance: A social signal processing perspective. Neurocomputing, 100:86–97, 2013.

[DY22] Doshi, Keval; Yilmaz, Yasin: Rethinking Video Anomaly Detection - A Continual Learning Approach. In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3036–3045, 2022.

[Ha19] Hao, Yu; Xu, Zhi-Jie; Liu, Ying; Wang, Jing; Fan, Jiu-Lun: Effective crowd anomaly detection through spatio-temporal texture analysis. International Journal of Automation and Computing, 16(1):27–39, 2019.

[Je23] Jebur, Sabah Abdulazeez; Hussein, Khalid A.; Hoomod, Haider Kadhim; Alzubaidi, Laith; Santamaría, José: Review on Deep Learning Approaches for Anomaly Event Detection in Video Surveillance. Electronics, 12(1), 2023.

[Ko19] Kopuklu, Okan; Kose, Neslihan; Gunduz, Ahmet; Rigoll, Gerhard: Resource efficient 3d convolutional neural networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0, 2019.

[Lu20] Luque Sánchez, Francisco; Hupont, Isabelle; Tabik, Siham; Herrera, Francisco: Revisiting crowd behaviour analysis through deep learning: Taxonomy, anomaly detection, crowd emotions, datasets, opportunities and prospects. Information Fusion, 64:318–335, 2020.

[Ma10] Mahadevan, Vijay; Li, Weixin; Bhalodia, Viral; Vasconcelos, Nuno: Anomaly detection in crowded scenes. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 1975–1981, 2010.

[Or21] Orrù, Giulia; Ghiani, Davide; Pintor, Maura; Marcialis, Gian Luca; Roli, Fabio: Detecting anomalies from video-sequences: a novel descriptor. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, pp. 4642–4649, 2021.

[RJV22] Ramachandra, Bharathkumar; Jones, Michael J.; Vatsavai, Ranga Raju: A Survey of Single-Scene Video Anomaly Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(5):2293–2312, 2022.

[SCS18] Sultani, Waqas; Chen, Chen; Shah, Mubarak: Real-World Anomaly Detection in Surveillance Videos. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6479–6488, 2018.

[Si20] Singh, Kuldeep; Rajora, Shantanu; Vishwakarma, Dinesh Kumar; Tripathi, Gaurav; Kumar, Sandeep; Walia, Gurjit Singh: Crowd anomaly detection using aggregation of ensembles of fine-tuned convnets. Neurocomputing, 371:188–198, 2020.

[SSM17] Swathi, HY; Shivakumar, G; Mohana, HS: Crowd behavior analysis: A survey. In: 2017 international conference on recent advances in electronics and communication technology (ICRAECT). IEEE, pp. 169–178, 2017.

[SV22] Sabih, Mohammad; Vishwakarma, Dinesh Kumar: A novel framework for detection of motion and appearance-based Anomaly using ensemble learning and LSTMs. Expert Systems with Applications, 192:116394, 2022.

[Wa22] Waqar, Sahar; Khan, Usman Ghani; Waseem, M Hamza; Qayyum, Samyan: The utility of datasets in crowd modelling and analysis: a survey. Multimedia Tools and Applications, pp. 1–32, 2022.

[Zh15] Zhao, Yu; Qiao, Yu; Yang, Jie; Kasabov, Nikola: Abnormal Activity Detection Using Spatio-Temporal Feature and Laplacian Sparse Representation. In (Arik, Sabri; Huang, Tingwen; Lai, Weng Kin; Liu, Qingshan, eds): Neural Information Processing. Springer International Publishing, Cham, pp. 410–418, 2015.

[Zh16] Zhou, Shifu; Shen, Wei; Zeng, Dan; Fang, Mei; Wei, Yuanwang; Zhang, Zhijiang: Spatial–temporal convolutional neural networks for anomaly detection and localization in crowded scenes. Signal Processing: Image Communication, 47:358–368, 2016.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# Automatic validation of ICAO compliance regarding head coverings: an inclusive approach concerning religious circumstances

Carla Guerra[1], João Marcos[2], Nuno Gonçalves[3]

**Abstract:** This paper contributes with a dataset and an algorithm that automatically verifies the compliance with the ICAO requirements related to the use of head coverings on facial images used on machine-readable travel documents. All the methods found in the literature ignore that some coverings might be accepted because of religious or cultural reasons, and basically only look for the presence of hats/caps. Our approach specifically includes the religious cases and distinguishes the head coverings that might be considered compliant. We built a dataset composed by facial images of 500 identities to accommodate these type of accessories. That data was used to fine-tune and train a classification model based on the YOLOv8 framework and we achieved state of the art results with an accuracy of 99.1% and EER of 5.7%.

**Keywords:** Facial Images, ICAO, ISO/IEC 19794-5, Head Covering Detection, Deep Learning

## 1 Introduction

Photographs used in identification documents must comply with certain requirements that guarantee standardization, in addition to allowing the person represented in the portrait to be properly identified through this image. Compliance with these requirements is based on quality metrics that measure, for example, the framing of the head in the photograph, the contrast with an homogeneous background, the restriction on the use of sunglasses or glasses whose lenses or frames partially or completely cover the eyes, among many others.

Two of the most relevant and extended public documents related to quality assessment in biometrics are the ISO/IEC 19794-5 standard [IS] and Doc 9030 [IC], created by the International Civil Aviation Organization (ICAO) based on that standard. These documents are actually a series of guidelines for the acquisition of high quality images, i.e., portrait-like images, for their inclusion in machine-readable travel documents like passports and ID cards. These guidelines are based on the typical impact that certain features like blur, occlusions, and resolution have in the quality of facial images and are intended to preserve the performance of Facial Recognition Systems (FRS). However, these reports do not specify the method to measure each of the features. In order to implement their recommended

[1] Institute of Systems and Robotics, University of Coimbra, Coimbra, Portugal, carla.guerra@isr.uc.pt
[2] Institute of Systems and Robotics, University of Coimbra, Coimbra, Portugal, joao.marcos@isr.uc.pt
[3] Institute of Systems and Robotics, University of Coimbra, Coimbra, Portugal and Portuguese Mint and Official Printing Office (INCM), Lisbon, Portugal, nunogon@deec.uc.pt

guidelines it becomes necessary to develop algorithms to automatically verify the compliance with the requirements. All of them can be verified using image processing techniques, combined with geometric measurements of elements detected in the photograph or with more sophisticated methods including deep learning.

In this work we focus on a specific requirement related to the use of head coverings in the photograph. Head coverings should not be accepted except in circumstances specifically approved by the Issuing State of the Machine Readable Travel Document (MRTD). Such circumstances may be religious, medical or cultural.

However, the work already found in the literature regarding this requirement ignores the case when veils, scarves or head coverings cannot be removed for religious reasons. Basically, it just looks for the presence of hats or caps.

Our main contribution in this paper is, thus, the proposal of an algorithm that automatically verifies the compliance with the requirements related to the use of head coverings, specifically considering religious cases.

To do so we built our own dataset for training and testing the algorithm, given the fact that none of the public dataset found on the literature considers the relevant particularities needed to verify ICAO compliance when in the presence of religious coverings.

We reached a very satisfactory performance, with an accuracy of 99.1% and an Equal Error Rate (EER) of 5.7%, which competes with the state of the art results.

## 2 Related Work

The University of Bologna's Biolab group played a significant role in popularizing methods adhering to the ISO/IEC 19794-5 standard. In 2009, they introduced the Biolab-ICAO framework [Ma09], which served as a benchmark tool for assessing the compliance of face images with ICAO requirements. Subsequently, in 2012, the benchmark underwent further refinement, leading to the presentation of an official database and testing protocol [Fe12]. Additionally, the authors proposed the BioLabSDK, the first documented method in the literature capable of evaluating 23 scene requirements.

Today, the Biolab-ICAO framework is used to evaluate algorithms via an online public competition called Face Image ISO Compliance Verification (FICV), hosted at the FVC-onGoing website [FV]. The FICV is considered the official evaluation tool for ISO/IEC 19794-5 standard and is used by most relevant works presented in the literature or commercial products. 23 scene requirements are evaluated individually in terms of EER. The EER is a standard metric to evaluate the performance of biometric systems and can be defined as the point where the False Acceptance Rate (FAR) and False Rejection Rate (FRR) curves intercept each other. Therefore, the EER represents the rate at which both acceptance and rejections errors are equal, i.e., FAR = FRR.

Notice that out of those 23, we focus particularly on the 'Hat/Cap' requirement which is a reduced version of what we called in this work 'Head Coverings' to accomodate many

more options, including religious cases that should be treated in a different way than just hat and caps.

To date, there are four published algorithms in the FVC-onGoing platform regarding the 'Hat/Cap' requirement: BioTest, BioPass Face, id3, and ICAONet [eSGB22]. Their results are summarized on Table 1.

| Algorithm | EER |
|---|---|
| ICAONet | 5.7% |
| id3 | 6.8% |
| BioPass Face | 9.8% |
| BioTest | 16.5% |

Tab. 1: Published results on the FVC-onGoing platform regarding the 'Hat/Cap' ICAO Requirement.

Three of the algorithms are own by private companies, therefore there is no detailed explanation about their methods. ICAO Net makes use of the significant advancements observed in deep learning over the past decade, which have notably improved accuracy compared to traditional hand-crafted methods. This progress has prompted researchers in the field of ICAO compliance verification to adopt deep learning techniques with remarkable success.

We point also the recent work by Hernandez-Ortega et. al [He22], who proposed the algorithm FaceQvec for evaluating the conformity of facial images with the same 23 algorithms defined by BioLab plus two regarding white-noise estimation and expression. The one regarding the head coverings in particular, also only considering Hats/Caps and no religious circumstances that might be considered compliant, looks for pixels with unnatural colour in the upper forehead region. However, the authors do not show results regarding that requirement in particular because of the lack on negative samples for the development and testing.

What we conclude is that there is some work already on the head coverings requirements but there is the need to extend these methods to consider more than just hats/caps. Also, facial images that can be used to train and test algorithms considering these particular concerns are lacking.

## 3 Data

To overcome the lack of available datasets to train and test algorithms to validate the compliance with ICAO requirements and, in particular, the 'Head Coverings' requirement, we built our own dataset. We collected facial images from people in controlled conditions, using many different accessories on the head, including religious options that could be considered compliant or not.

ICAO states that if head coverings are allowed, they shall be firm fitting and of a plain uniform colour with no pattern or no visible perforations and the region between hair lines, both forwards of the ears and chin including cheeks, mouth, eyes, and eyebrows shall be visible without any distortion or shadows [Wo18].

Our dataset is composed by 3500 images of 500 subjects gathered across volunteers from different ages, genders and origins. Table 2 shows the demographic distribution of identities.

| | Caucasian | | African | | Asian | |
|---|---|---|---|---|---|---|
| | Female | Male | Female | Male | Female | Male |
| Children/Teen [0,20] | 62 | 61 | 11 | 14 | 2 | 1 |
| Young Adult ]20-35] | 61 | 61 | 13 | 14 | 17 | 17 |
| Adult ]35-50] | 25 | 45 | 4 | 2 | 3 | 1 |
| Senior Adult ]50-65] | 16 | 26 | 1 | 0 | 0 | 1 |
| Senior ]65-inf] | 17 | 25 | 0 | 0 | 0 | 0 |

Tab. 2: Demographic distribution of identities in the built dataset.

Each volunteers takes 7 pictures:

- 2 with no head coverings at all;
- 3 with non-compliant head coverings (hat, caps, ribbons, etc);
- 1 with religious coverings that might be considered compliant;
- 1 with non-compliant religious coverings.

Samples of each picture taken can be seen in Figure 1.

Fig. 1: Samples of the pictures taken by each identity on the dataset. Top 3 are compliant, the bottom 4 are non-compliant.

In total, 1500 images are (potentially) compliant and 2000 images have one or more requirements that are non-compliant in terms of head coverings. We divided the dataset into training, validation and testing samples, randomly chosen.

## 4 Methods

In the field of object detection, the YOLO (You Only Look Once) network model [Re16] is well known for having the capability of detecting multiple objects in real-time. Therefore, we use YOLOv8 as our framework and we fine-tune the network parameters to achieve real-time performance and high accuracy on classifying facial images into ICAO compliant or non-compliant.

YOLOv8 was released on January 2023 by ultralytics and gives better results than its predecessor versions [To].

It has two parts: Head and Backbone. Backbone is responsible for generating feature pyramids after feature extraction. Head is responsible for identification and displaying bounding boxes along with objectness score [TCE23].

We trained our YOLOv8 model with the dataset specifically created for this purpose by us. The parameters chosen for training were:

- Model: yolov8s-cls.pt;
- Epochs: 10;
- Batch Size: 64;
- Image Size: 224 (pixels);
- Workers: 8;
- PreTrained: True;
- Optimizer: Adam;
- Initial Learning Rate: 0.001;
- Weight Decay: $5x10^{-}5$;
- Label Soothing Epsilon: 0.1;
- Model Layer Cutoff: None;
- Dropout (fraction): None.

Furthermore, YOLOv8 employs image augmentation techniques during training to enhance its performance. In each epoch, the model encounters slightly varied versions of the provided images. Notably, YOLOv8 utilizes mosaic augmentation, which involves

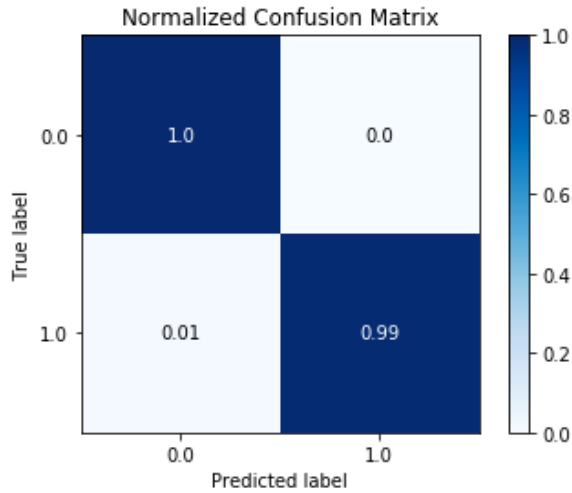



Fig. 2: Resulting confusion matrix. Class 0 stands for 'Non-compliant' and class 1 stand for 'Compliant'.

stitching together four training images to create a new composite image. This augmentation technique significantly contributes to the model's efficiency and learning capabilities [HZ20]. Compared to previous iterations, YOLOv8 demonstrates superior efficiency, thanks to its use of a larger feature map and a better optimized convolutional network [To]. For a deeper understanding of YOLOv8's functioning and detailed insights into its architecture, comprehensive information can be found in [Ro].

## 5 Results

The results obtained have shown that our method to automatically verify the compliance with the head coverings requirements can achieve very high accuracy levels (99.1%), failing only on 0.9% of the compliant samples - see Figure 2. The resulting loss curves during train and validation stages are shown in Figure 3. The tests were performed over a set of randomly chosen samples that represent 20% of dataset, making sure that all categories of images are balanced. The corresponding EER equals 5.7%, which is the same as the best performing algorithm already present in the literature, but now extended to be able to distinguish when a head covering might be considered compliant because of religious circumstances, which per se is an improvement.

## 6 Conclusions and Future Work

This work makes significant contributions in the form of a dataset and an algorithm aimed at automating the verification of compliance with ICAO requirements concercing the pres-

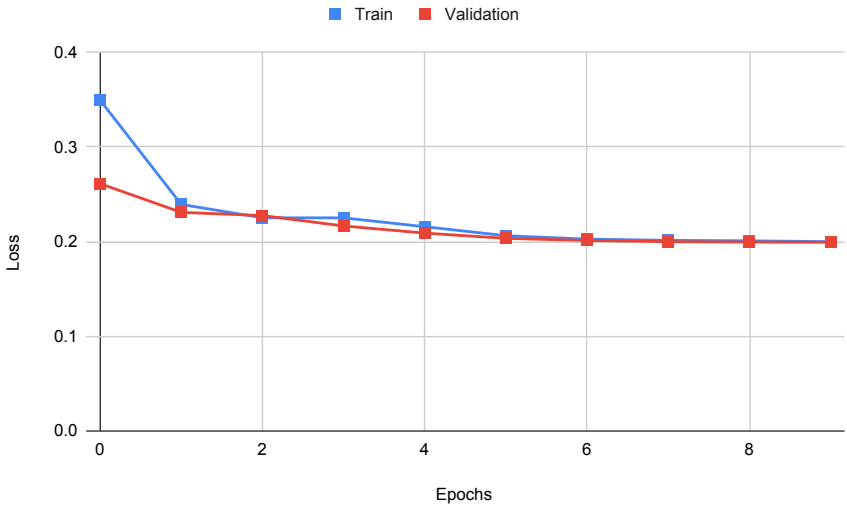



Fig. 3: Train and validation loss curves along the epochs.

ence of head coverings in facial images used in machine-readable travel documents. The existing methods fails to consider the acceptance of certain head coverings based on religious or cultural reasons, focusing primarily on the detection of hats or caps. In contrast, our approach specifically extends these considerations to include cases of religious coverings that can be accepted. To support our approach, we created a dataset consisting of 500 facial images representing diverse identities and accommodating the inclusion of more accessory types such as compliant and non-compliant religious coverings. Using this dataset, we fine-tuned and trained a classification model based on the YOLOv8 framework, resulting in a state-of-the-art performance with 99.1% of accuracy and an Equal Error Rate (EER) of 5.7%. These work highlights the lack of inclusion of religious factors when verifying compliance with head covering requirements, and demonstrates the efficacy of our approach in accurately identifying compliant head coverings. In the future we would like to consider also the case when there is a head covering that cannot be removed because of medical reasons, extending our dataset to include examples of it.

## 7 Acknowledgment

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 FACING2. This work has been supported by Fundação para a Ciência e a Tecnologia (FCT) under the project UIDB/00048/2020.

## References

[eSGB22] e Silva, Arnaldo Gualberto de Andrade; Gomes, Herman Martins; Batista, Leonardo Vidal: A collaborative deep multitask learning network for face image compliance to ISO/IEC 19794-5 standard. Expert Systems with Applications, 198:116756, 2022.

[Fe12] Ferrara, Matteo; Franco, Annalisa; Maio, Dario; Maltoni, Davide: Face image conformance to iso/icao standards in machine readable travel documents. IEEE Transactions on Information Forensics and Security, 7(4):1204–1213, 2012.

[FV] FVC-ongoing BioLab Benchmark. https://biolab.csr.unibo.it/fvcongoing/UI/Form BenchmarkAreas/BenchmarkAreaFICV.aspx9. Accessed: 2023-06-30.

[He22] Hernandez-Ortega, Javier; Fierrez, Julian; Gomez, Luis F; Morales, Aythami; Gonzalez-de Suso, Jose Luis; Zamora-Martinez, Francisco: FaceQvec: Vector quality assessment for face biometrics based on ISO compliance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 84–92, 2022.

[HZ20] Hao, Wang; Zhili, Song: Improved mosaic: Algorithms for more complex images. In: Journal of Physics: Conference Series. volume 1684. IOP Publishing, p. 012094, 2020.

[IC] ICAO: Doc 9303 - Machine Readable Travel Documents - Part 3: Specifications Common to all MRTDs. https://www.iso.org/standard/50867.html.

[IS] : ISO/IEC 19794-5:2011 Information technology — Biometric data interchange formats — Part 5: Face image data. https://www.iso.org/standard/50867.html.

[Ma09] Maltoni, Davide; Franco, Annalisa; Ferrara, Matteo; Maio, Dario; Nardelli, Antonio: Biolab-icao: A new benchmark to evaluate applications assessing face image compliance to iso/iec 19794-5 standard. In: 2009 16th IEEE International Conference on Image Processing (ICIP). IEEE, pp. 41–44, 2009.

[Re16] Redmon, Joseph; Divvala, Santosh; Girshick, Ross; Farhadi, Ali: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788, 2016.

[Ro] Roboflow: Whats New In YOLOv8. roboflow/whats-new-in-yolov8/9. Accessed: 2023-06-30.

[TCE23] Terven, Juan; Cordova-Esparza, Diana: A comprehensive review of YOLO: From YOLOv1 to YOLOv8 and beyond. arXiv preprint arXiv:2304.00501, 2023.

[To] Towards AI: YOLOv8 Is Here And It Gets Better. https://pub.towardsai.net/yolov8-is-here-and-it-gets- better54b12b87e3b9. Accessed: 2023-06-30.

[Wo18] Wolf, A: ICAO: Portrait Quality (Reference Facial Images for MRTD), Version 1.0. Standard. International Civil Aviation Organization, 2018.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# Generalizability and Application of the Skin Reflectance Estimate Based on Dichromatic Separation (SREDS)

Joseph Drahos, Richard Plesh, Keivan Bahmani, Mahesh Banavar, Stephanie Schuckers [1]

**Abstract:** Face recognition (FR) systems have become widely used and readily available in recent history. However, differential performance between certain demographics has been identified within popular FR models. Skin tone differences between demographics can be one of the factors contributing to the differential performance observed in face recognition models. Skin tone metrics provide an alternative to self-reported race labels when such labels are lacking or completely not available e.g. large-scale face recognition datasets. In this work, we provide a further analysis of the generalizability of the Skin Reflectance Estimate based on Dichromatic Separation (SREDS) against other skin tone metrics and provide a use case for substituting race labels for SREDS scores in a privacy-preserving learning solution. Our findings suggest that SREDS consistently creates a skin tone metric with lower variability within each subject and SREDS values can be utilized as an alternative to the self-reported race labels at minimal drop in performance. Finally, we provide a publicly available and open-source implementation of SREDS to help the research community. Available at https://github.com/JosephDrahos/SREDS

**Keywords:** Face Recognition, Privacy-Preserving, Feature Unlearning, Skin Reflectance.

## 1 Introduction

Face recognition systems are increasingly used as a form of biometric authentication for many new and existing systems. Research on the differential performance between demographics is an important topic of study to mitigate bias and ensure fairness [dFPM22, Ho22]. Modern facial recognition systems use deep learning pipelines to take an image of a person's face and create a unique template for that person. In such systems, the demographic information of a dataset is needed to assess or mitigate the differential performance of a particular face recognition algorithm. However, many of the large-scale datasets which have been aggregated from public images on the internet and used to train and benchmark face recognition networks lack self-reported race labels. Additionally, the large scale of such datasets makes it impractical and expensive to efficiently label demographics by human annotators. As a result, methods to automatically label a dataset can provide a valuable asset.

Our research focuses on the intersection of privacy preservation, bias mitigation, and skin tone metrics. We present our analysis of the Skin Reflectance Estimate based on Dichromatic Separation (SREDS) skin tone metric from [Ba21]. SREDS is a continuous skin

[1] Department of Electrical and Computer Engineering, 8 Clarkson Ave, Potsdam, NY,
{drahosj, pleshro, bahmank, mbanavar, sschucke} @clarkson.edu
This material is based upon work supported by the Center for Identification Technology Research and the National Science Foundation (NSF) under Grant No.1650503.

tone metric that can be used to automatically label skin tones on face datasets. Our goal is to evaluate SREDS' ability to label datasets compared to other skin tone metrics, assess the generalizability of SREDS on unseen data, and demonstrate an application of SREDS using a sensitive information removal approach when race labels are not available.

## 2 Background

### 2.1 Skin Tone Metrics

Previous methods formulated for generating a metric for subject skin tones to more accurately describe skin color are listed as follows: Fitzpatrick Skin Type (FST), Monk Skin Tone (MST) Scale, Individual Typology Angle (ITA), and Relative Skin Reflectance (RSR) [Fi88, Mo19, CCH91, Co19]. FST and its successor MST require a manual calculation from a survey of the subject, while ITA and RSR can be computed automatically via an algorithm. RSR was created to analyze skin tone for a specific dataset by fitting a Principal Component Analysis (PCA) model to the RGB space of the dataset. RSR is not resistant to changes in lighting and is specific to a particular dataset. The need for a skin tone metric that can be computed automatically and is more resistant to changes in lighting prompted the research that led to the Skin Reflectance Estimate based on a Dichromatic Separation (SREDS) [Ba21]. SREDS aims to decompose patches of skin into specular and diffuse components using the dichromatic reflectance model. A Kernal Principal Component Analysis (KPCA) is fit onto the diffuse components extracted from the dataset, resulting in a data-driven skin tone metric.

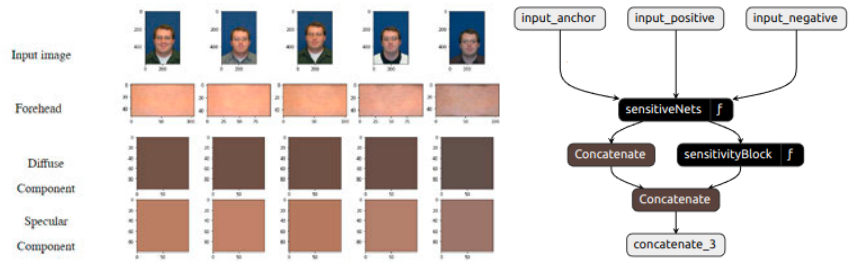
### 2.2 Bias Mitigation

The inclusion of demographic information in a dataset is to observe and attempt to eliminate the differential performance between demographic groups in FR models. Differing methods of bias mitigation have been attempted and documented at the feature, comparison, and post-comparison levels. A method of bias mitigation at the feature level is the triplet mining approach of [Se22] which used a triplet loss for discrimination-aware learning. Closely related triplets are mined based on race information to try and train a new representation that mitigates biased learning within the face embedding space of a pretrained model. At the comparison level, a learning classifier method reduces ethnic bias by introducing group and individual fairness to the decision process at the cost of matching performance [Te20c]. At the post-comparison level, an unsupervised method of score normalization has been presented to reduce bias between ethnic groups while increasing the performance of the system [Te20b].

### 2.3 Soft Biometric Privacy Preservation

Soft biometric information such as gender, race, age, etc. is stored within the templates created from FR systems and can be extracted without the user's consent [Te20a]. Meth-

ods of privacy preservation have been studied and introduced to protect users' sensitive information. The efforts in[OR15] produced a technique that morphed the input face with another face to mask the soft biometrics while maintaining matching performance. Another technique that added a perturbing element to the initial face image that would mask sensitive information while maintaining performance is [MR17]. Information removal networks attempt to remove sensitive information from the feature embedding space of the FR deep network. These methods require complex loss functions to maintain the performance accuracy of the network while also suppressing the racial information from the learned space, as performed in [Xu18]. A method that combines the methods from [Se22, Xu18] and was used within this research is [Mo20], which attempts to maintain the inter-identity distance using triplet loss and simultaneously unlearn[2] the facial features used to differentiate between demographic classes.




(a) Intra-subject variation between diffuse and specular components extracted using SREDS.

(b) SensitiveNets model implementation

Figure 1

## 3 Methodology

### 3.1 Skin Tone Metrics Assessment

The skin tone metrics outlined in section 2.1 will be used as a baseline to compare the previously developed methods to the performance of the SREDS measure. Individual typology angle (ITA) is a type of colorimetric analysis designed to measure acquired tanning [CCH91]. An RGB image is converted into CIE-Lab space [CF97], as follows: (1) the '$L$' component which quantifies luminance, (2) the '$a$' component - absence or presence of redness, and (3) the '$b$' component - yellowness. Using the '$L$' and '$b$' components, Pixel-wise ITA value, in degrees, can be estimated throughout an image as:

$$ITA = \frac{\arctan(L-50)}{b} * \frac{180}{\pi}. \quad (1)$$

To find suitable skin pixels in the image, a landmark extractor based on Dlib is used to detect the forehead, left cheek, and right cheek facial regions [Ki06]. For each facial region,

[2] The term unlearn will be used throughout the paper in the same context as introduced in the literature [Mo20].

ITA is computed over each pixel and smoothed using an averaging filter. The mode from each region's resulting values is averaged to result in a single skin tone estimate for a face.

Relative Skin Reflectance (RSR) is a process designed to relate the physical properties of the skin to the performance of facial recognition [Co19]. The pipeline works by removing the confounding effects of imaging artifacts on skin pixels and fitting a line in the direction of the greatest variance in the RGB color space using PCA. The resulting metric is related to the skin tone of each subject relative to the rest of the photos in the dataset. Assumptions include consistent lighting, the same acquisition camera, and constant background. As a further limitation, the metric only indicates where a subject lies regarding net skin reflectance relative to the other subjects in the dataset, rather than an absolute measure.

The process to compute SREDS begins by extracting patches of skin from the forehead, right, and left cheeks using Dlib landmarks of each face image. Using the dichromatic reflection model as a guide, Non-Negative Matrix Factorization (NNMF) is used to estimate the diffuse and specular components of the selected skin patches. KPCA is utilized on the extracted diffuse components to learn a skin tone gradient across the dataset. The averaged value of the first principal components of the extracted diffuse bases for a particular face defines that person's SREDS score. The KPCA model used for SREDS is data-driven, so the generalizability of the KPCA model onto unseen datasets is a point of interest within this study. A full description of the extraction of SREDS is found in [Ba21].

### 3.2 Datasets

For our experiments, we selected datasets that included demographic information of subjects across race, age, gender, orientation, and lighting. We utilized CMU Multi-PIE, MEDS-II, and Morph-II datasets [SBB03, Fo11, RT06]. Multi-PIE contains 750,000 sample images from 337 subjects images under 15 viewpoints with 19 illumination conditions. We selected three viewpoints (14 0, 05 1, 05 0) where full views of the face were captured for our testing, which reduced our sample images to 150,668 from 314 subjects. MEDS-II contains only 836 sample images from 425 subjects imaged in a controlled mugshot setting. Morph-II is a dataset from a longitudinal study that contains 55,063 sample images from 13,000 subjects within a controlled setting over 5 years. While MEDS-II and Morph-II datasets include uncontrolled illumination, Multipie includes controlled illumination samples. ITA, RSR, and SREDS scores were generated for all samples of each dataset.

### 3.3 Cross-Dataset Analysis

In prior work [Ba21], the intra-subject variance was used as a metric to describe the variance of a specific subject's skin tone score across multiple samples. The low intra-subject variance shows the metric can produce a consistent value of the same subject independent of external conditions. An example of intra-subject variation can be seen in figure 1a. We evaluate and compare the intra-subject variance across all of our evaluation datasets and compare it to other methods. In addition, we test the generalizability of the SREDS metric

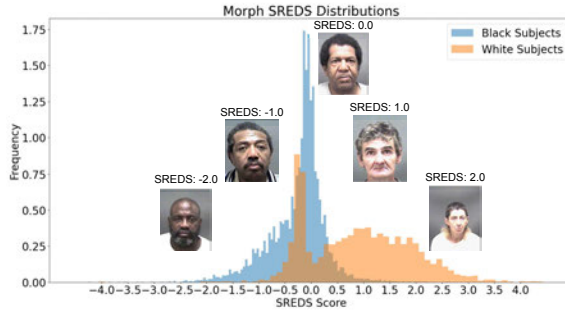



Figure 2: Distribution of SREDS Scores for Morph-II dataset separated by Race. Subjects with specific SREDS scores are shown for the range across the distribution.

by calibrating the skin tone gradient on one dataset and applying it to another, thereby testing its robustness to unseen datasets. The same experiment is run using the RSR PCA models for comparison. ITA does not have a training component and is only reported per dataset.

### 3.4 SREDS Agnostic Facial Recognition Model

To show the potential of SREDS for use in the replacement of race labels, we compared the outcome performance of SREDS versus ground-truth race labels when incorporated into SensitiveNets, a sensitive information removal network [Mo20]. SensitiveNets provides a novel privacy-preserving neural network feature representation to suppress the sensitive information of a learned space while maintaining the utility of the data. We reimplemented the sensitive removal network as our model for analysis of the suppression of race and skin tone. A diagram of our model is seen in figure 1b. SensitiveNets contains sensitive information removal dense layers added on top of a pre-trained face recognition backbone. Within our testing, we used a Resnet50 model pre-trained on VGGFace2 as the backbone, consistent with the cited literature [Mo20, Ca18, Xi]. The model's loss function requires a race classifier that acts as the sensitive information detector. The softmax probability from this detector describes the amount of racial information present within a subject's template and the goal of the loss function is to remove the sensitive race information and trend the classifier towards 50% accuracy. In our experiments, this classifier is either trained on race labels or SREDS scores binned into predetermined groups. The sensitive information removal $\varphi$ layers are then added and trained sequentially using an adversarial approach of triplet loss and an adversarial sensitivity regularizer loss which reduces the amount of sensitive race information from the embedding space. An in-depth look at the model and loss function can be found in the SensitiveNets literature [Mo20].

# 4 Experiment Results

Our experiments were performed to analyze how the consistency of SREDS performed relative to other skin tone metrics and the outcome of replacing race labels with SREDS-generated labels in a privacy preservation method.

## 4.1 Cross-Dataset Analysis Results

We performed the cross-dataset analysis of the two skin tone metrics described in Section 2 and SREDS across the three datasets listed in Section 3. We generated ITA, RSR, and SREDS for all subjects from the mentioned datasets. As part of background normalization, RSR assumed consistent lighting, the same acquisition camera, and a constant background. Only the Multi-PIE dataset meets all conditions. However, due to the lack of constant background in MEDS-II and MORPH-II, the background normalization step was bypassed for these datasets. ITA is a non-trainable method so we collected the ITA values from each subject of each dataset. To test SREDS consistency on unseen data we used the Kernal Principal Component Analysis (KPCA) fit to one dataset's diffuse components and used it to transform another dataset's diffuse components. The same process was recreated using the RSR PCA models on the same datasets' selected skin pixel values in order to compare these two methods.

| | Testing Dataset | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Morph-II | | | MEDS-II | | | Multi-Pie (Mugshot) | | |
| Training Dataset | SREDS | RSR | ITA | SREDS | RSR | ITA | SREDS | RSR | ITA |
| Morph-II | 0.419 | 0.539 | 0.645 | 0.681 | 0.493 | N/A | 0.157 | 0.468 | N/A |
| MEDS-II | 0.457 | 0.540 | N/A | 0.463 | 0.493 | **0.448** | 0.186 | 0.470 | N/A |
| Multi-Pie (Mugshot) | **0.399** | 0.538 | N/A | 0.674 | 0.493 | N/A | **0.138** | 0.304 | 0.401 |

Table 1: Cross dataset intra-subject variability analysis between SREDS, RSR, and ITA skin tone metrics. Bolded values are the lowest recorded intra-subject variability in that testing dataset. SREDS scores result in the least variable metric from Morph-II and Multi-Pie datasets and the second least variable metric in MEDS-II, behind ITA.

We computed the intra-subject variability of each dataset's skin tone metrics by calculating the standard deviation of each subject's individual skin tone measures and averaging across the dataset. The results of this analysis are seen in Table 1 and suggest that the learning-based algorithms (RSR and SREDS) perform better than ITA when evaluated on the dataset they are calibrated on. Viewing our cross-dataset results, we observe that in larger datasets (Morph, Multi-pie), SREDS outperforms both ITA and RSR even when calibrated on a different dataset, suggesting the generalizability of this approach.

## 4.2 Distribution of SREDS

To utilize SREDS by replacing race labels we needed a process to convert continuous SREDS scores into discrete labels. To understand the distribution of scores, the SREDS

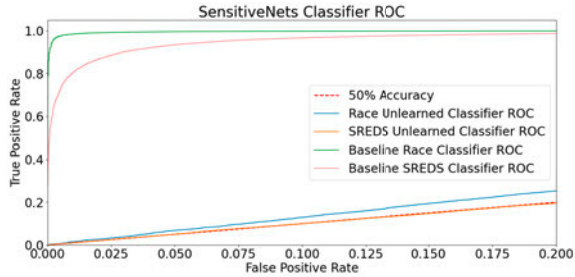



Figure 3: SensitiveNets Sensitive Information Classifier ROCs for both race labels and SREDS scores before and after training on Morph-II dataset. The goal of SensitiveNets training is for 50% classification accuracy. The unlearned classifier accuracy for both the RDM and SDM is nearly 50%, which shows SREDS scores and race labels perform similarly in this experiment.

scores across the Morph-II dataset were plotted within Figure 2. We split the dataset in half by the median SREDS score of -0.01 and binned the subjects into low and high SREDS scores to create a discrete labeling of the Morph-II dataset.

### 4.3 Comparison of Race Labels and SREDS in Sensitive Feature Unlearning

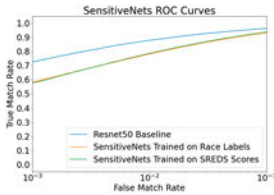
To see the effects of SREDS scores being used in place of self-reported race labels, we implemented two SensitiveNets models. One model is trained using the black and white subject race labels from the Morph-II dataset while the second model is trained using the binned SREDS value for the same subjects.

| Backbone | Classifier | Trained On | Tested On | ICA | FCA |
|---|---|---|---|---|---|
| Resnet50 | Race | Race Triplets | Morph | 0.985 | 0.47 |
| Resnet50 | SREDS | SREDS Triplets | Morph | 0.937 | 0.48 |

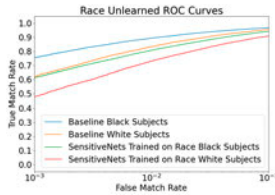
Table 2: Sensitive Information Removal Network Experiment Results
ICA: Initial Classification Accuracy, FCA: Final Classification Accuracy (Goal of sensitive information removal is for FCA to be 0.50)

The first model trained on race labels and the second model trained on SREDS scores will be referred to as the Race Unlearned Model (RUM) and the SREDS Unlearned Model (SUM) respectively. An outline of this testing plan is seen in Table 2 with the initial and final classification accuracy of the SensitiveNets classifiers. For both models, the sensitive information classifier ROCs were calculated and shown in Figure 3.

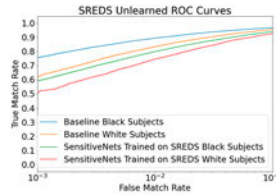
The two trained SensitiveNets models matching performances are compared to the baseline Resnet50 matching performance to evaluate the results of the training on matching performance in Figure 4. The feature unlearning experiments to preserve privacy show a similar drop in performance between training with race labels and training with SREDS




(a) ROC comparison for baseline (Resnet50), RUM, and SUM models




(b) ROC comparison between baseline and RUM model for White and Black subjects.




(c) ROC comparison between baseline and SUM model for White and Black subjects.

Figure 4: Comparison of biometric performance (matching) ROCs of baseline, RDM, and SDM, categorized by race labels on Morph-II dataset. Shows RDM and SDM suffer from a similar drop in matching performance when race or skin tone information is removed, respectively.

scores. The results suggest only a small (0.027) difference between the True Positive Rate (TPR) of RUM and SUM at $10^{-3}$ False Positive Rate (FPR).

## 5 Conclusions

The existing feature unlearning methods in FR rely on large-scale and expensive-to-collect demographically-labeled datasets. Within this study, we demonstrate the ability of SREDS to mitigate this reliance by automatically extracting consistent skin tone information from face images. We have shown that SREDS outperforms other available skin tone metrics in producing continuous and less-variable skin tone estimates while generalizing well to unseen data. We have presented an application of extracted SREDS scores in the absence of race labels in a feature unlearning method and shown that SREDS could be used as a replacement.

### 5.1 Limitations and Future Work

Limitations of this work include our analysis of only black and white subjects due to the under-representation of other races in our datasets. This led to us only using two SREDS bins when categorizing our datasets to match the binary race labels. We tested using only one face matcher within our privacy-preserving method and have not seen how different networks affect our results. A limitation of using skin tone as a way to label datasets is that skin tone does not encapsulate the entirety of a self-reported race label. Skin tone is one physical characteristic that makes up race and cannot be used as an exact replacement.

Future work planned includes further analysis of the mapping of SREDS to multi-race demographic information and its use in different downstream biometric tasks, recreating our experiments with addition face matches [De22], and attempting a bias mitigation solution using SREDS scores and evaluating using fairness metrics [dFPM22, Ho22] on an even larger scale dataset (BUPT-Globalface) [WZD21].

## References


[Ba21] Bahmani, Keivan; Plesh, Richard; Sahu, Chinmay; Banavar, Mahesh; Schuckers, Stephanie: SREDS: A dichromatic separation based measure of skin color. In: 2021 IEEE International Workshop on Biometrics and Forensics (IWBF). IEEE, pp. 1–6, 2021.

[Ca18] Cao, Qiong; Shen, Li; Xie, Weidi; Parkhi, Omkar M.; Zisserman, Andrew: , VGGFace2: A dataset for recognising faces across pose and age, 2018.

[CCH91] Chardon, A; Cretois, I; Hourseau, C: Skin colour typology and suntanning pathways. Int J Cosmet Sci, 13(4):191–208, August 1991.

[CF97] Connolly, C.; Fleiss, T.: A study of efficiency and accuracy in the transformation from RGB to CIELAB color space. IEEE Transactions on Image Processing, 6(7):1046–1048, 1997.

[Co19] Cook, Cynthia M.; Howard, John J.; Sirotin, Yevgeniy B.; Tipton, Jerry L.; Vemury, Arun R.: Demographic Effects in Facial Recognition and Their Dependence on Image Acquisition: An Evaluation of Eleven Commercial Systems. IEEE Transactions on Biometrics, Behavior, and Identity Science, 1(1):32–41, 2019.

[De22] Deng, Jiankang; Guo, Jia; Yang, Jing; Xue, Niannan; Kotsia, Irene; Zafeiriou, Stefanos: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10):5962–5979, oct 2022.

[dFPM22] de Freitas Pereira, Tiago; Marcel, Sébastien: Fairness in Biometrics: A Figure of Merit to Assess Biometric Verification Systems. IEEE Transactions on Biometrics, Behavior, and Identity Science, 4(1):19–29, 2022.

[Fi88] Fitzpatrick, T B: The validity and practicality of sun-reactive skin types I through VI. Arch Dermatol, 124(6):869–871, June 1988.

[Fo11] Founds, Andrew; Orlans, Nick; Genevieve, Whiddon; Watson, Craig: , NIST Special Databse 32 - Multiple Encounter Dataset II (MEDS-II), 07 2011.

[Ho22] Howard, John J.; Laird, Eli J.; Sirotin, Yevgeniy B.; Rubin, Rebecca E.; Tipton, Jerry L.; Vemury, Arun R.: , Evaluating Proposed Fairness Models for Face Recognition Algorithms, 2022.

[Ki06] King, Davis: , Dlib C++ library, 2006.

[Mo19] Monk, Ellis: , Monk Skin Tone Scale, 2019.

[Mo20] Morales, Aythami; Fierrez, Julian; Vera-Rodriguez, Ruben; Tolosana, Ruben: , SensitiveNets: Learning Agnostic Representations with Application to Face Images, 2020.

[MR17] Mirjalili, Vahid; Ross, Arun: Soft biometric privacy: Retaining biometric utility of face images while perturbing gender. In: 2017 IEEE International Joint Conference on Biometrics (IJCB). pp. 564–573, 2017.

[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. Springer International Publishing, Cham, pp. 682–696, 2015.

[RT06] Ricanek, K.; Tesafaye, T.: MORPH: a longitudinal image database of normal adult age-progression. In: 7th International Conference on Automatic Face and Gesture Recognition (FGR06). pp. 341–345, 2006.

[SBB03] Sim, T.; Baker, S.; Bsat, M.: The CMU pose, illumination, and expression database. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(12):1615–1618, 2003.

[Se22] Serna, Ignacio; Morales, Aythami; Fierrez, Julian; Obradovich, Nick: Sensitive loss: Improving accuracy and fairness of face representations with discrimination-aware deep learning. Artificial Intelligence, 305:103682, 2022.

[Te20a] Terhörst, Philipp; Fährmann, Daniel; Damer, Naser; Kirchbuchner, Florian; Kuijper, Arjan: , Beyond Identity: What Information Is Stored in Biometric Face Templates?, 2020.

[Te20b] Terhörst, Philipp; Kolf, Jan Niklas; Damer, Naser; Kirchbuchner, Florian; Kuijper, Arjan: , Post-Comparison Mitigation of Demographic Bias in Face Recognition Using Fair Score Normalization, 2020.

[Te20c] Terhörst, Philipp; Tran, Mai Ly; Damer, Naser; Kirchbuchner, Florian; Kuijper, Arjan: Comparison-Level Mitigation of Ethnic Bias in Face Recognition. In: 2020 8th International Workshop on Biometrics and Forensics (IWBF). pp. 1–6, 2020.

[WZD21] Wang, Mei; Zhang, Yaobin; Deng, Weihong: Meta Balanced Network for Fair Face Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2021.

[Xi] Xie, Weidi: , Weidixie/Keras-VGGFACE2-RESNET50.

[Xu18] Xu, Depeng; Yuan, Shuhan; Zhang, Lu; Wu, Xintao: , FairGAN: Fairness-aware Generative Adversarial Networks, 2018.

N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco, and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

# A Wrist-worn Diffuse Optical Tomography Biometric System

Satya Sai Siva Rama Krishna Akula[1], Sumanth Dasari[1], Keerti Bajaj[1], Bhuvan Chennoju[1], Tejaswi Dhandu[1], Mostafizur Rahman[1], Reza Derakhshani[1]

**Abstract:** We present a Diffuse Optical Tomography (DOT)based biometric system that uses interior anatomical information for better privacy and security instead of external traits such as face or fingerprint. The DOT system has a wearable form factor covering the lower forearm and the wrist, where anatomical structures in the optical path of the DOT optodes capture the unique internal patterns used for biometrics. Our DOT scanner is low-cost, using COTS near-infrared LEDs and sensors. Our design also incorporates wrist vein imaging as a secondary modality to supplement the DOT. This paper details the design of the DOT system and the ensuing machine-learning pipeline. We demonstrate the utility of the DOT as a stand-alone biometric modality and the efficacy of its fusion with wrist vein patterns. Our early experimental findings show promising results, using a pilot dataset to achieve an area under the receiver operating characteristic curve (ROC AUC) of 0.999138 and an equal error rate (EER) of 1.27% for the DOT modality. The AUC and EER were 0.999655 and 0.48% for the wrist vein imaging modality only and 0.99989 and 0.21% for the fusion of both modalities.

**Keywords:** Biometric authentication, Vein imaging, Diffuse optical tomography (DoT), Performance evaluation

## 1 Introduction

Biometric traits such as fingerprints, iris patterns, voice, and facial features are among the most popular, each with its own pros and cons. Different biometric modalities have also been combined to enhance recognition accuracy and robustness and to deter presentation attacks, giving rise to multimodal biometric systems [USJ20, Ga06]. This paper proposes a new multi-modal wrist-worn biometric system using Diffuse Optical Tomography (DOT) as the primary and vascular imaging as the secondary modality. The combination provides a simple unitary user experience while leveraging the near-infrared imaging of the deep structural elements of the forearm using a novel application of DOT in biometrics. While DOT [BCH16] is a known method and has been widely applied in various medical applications, to the best of our knowledge, we are the first to apply it to the wrist and lower forearm for biometric identification [Di05]. The addition of vascular arcades as a supplementary modality and the blending of the two using machine-learning techniques are among the other highlights of this work. Vascular patterns have long been known to exhibit distinct characteristics, even among identical twins [Kr20]. The introduction of forearm/wrist DOT, besides providing a new and powerful source of entropy, significantly

[1] School of Science and Engineering, University of Missouri, Kansas City

boosts the security of the system. The combination of vascular and DOT imaging as an internal biometric source is inherently harder to attain surreptitiously and harder to tamper with. Unlike the ubiquitous face biometrics, altering or replicating such internal structures raises the bar for potential attackers [Du08] [TY13].

To summarize our unique contributions, we demonstrate the feasibility of forearm DOT as a biometric modality and show the utility of simultaneously adding captured vein patterns from the same area to the mix through a pilot study. We also present our scanner design innovations, employing a multi-path continuous-wave DOT scanning using a near-infrared (NIR) sensor-illuminator mesh made out of affordable commercial off-the-shelf (COTS) components. Data pre-processing, fusion, and machine learning analytics are also presented as a part of the pipeline. The resulting hardware-software POC is capable of real-time, end-to-end enrollment and matching, providing a new secure biometric identification solution [HB09]. The rest of this paper is organized as follows: section 2 details the hardware setup, section 3 presents the methodology for hardware usage, section 4 presents data collection and evaluation, and Sections 5 and 6 outline the conclusion and acknowledgments.

## 2 DOT Wristband Design

The optical system employed in the DOT wristband design consists of NIR LEDs as the illumination source[Ch04] and NIR detector/sensor arrays. The LEDs and sensors are arranged in blocks for uniformity, as depicted in Figure 1. The selection of the 870 nm wavelength was based on the absorption coefficients for both oxygenated and deoxygenated hemoglobin[Ch14][Hi02].

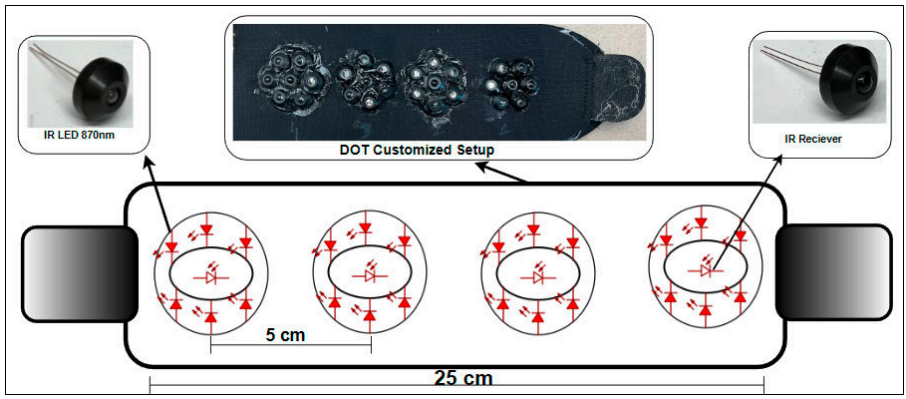



Fig. 1: DOT Wristband Setup

Several configurations of Infrared (IR) emitters and receivers were examined during the DOT design process. Experiments were conducted to explore distance ratios ranging from 1:1 to 1:6 between the IR receivers and emitters. Through these experiments, it was deduced that a 1:6 ratio with a 1.5cm gap between the components resulted in the best out-

comes. The final arrangement was determined to consist of one IR receiver and six IR emitters.

The setup shown in Figure 1 was finally selected. It consists of four units, namely S1, S2, S3, and S4. Each unit is comprised of a central IR sensor (100F5T-IR-JS-940NM) surrounded by six MTE8760N5 IR LEDs. The units were arranged linearly, with a 5 cm separation between the centers of adjacent units, as depicted in Figure 1. In the experimental setup, the control and communication of the sensors were facilitated by a Raspberry Pi (R Pi) device. The R Pi was the central controller and established a connection with an external host compute node. This arrangement enables the R Pi to exercise selective activation of pairs of blocks within the system.

In order to convert the analog signals received by the sensor into digital voltage values, an Adafruit ADS1115 Analog-to-Digital Converter (ADC) with a precision of 16 bits was employed. The utilized ADC can support sampling rates ranging from 8 to 860 samples per second and can be configured for 1 to 4 channels. It incorporates a programmable gain amplifier with a maximum gain of 16, which facilitates signal amplification. Communication with the ADC was established through I2C. The ADC can address up to 4 ADS1115 devices on a single 2-wire I2C bus, thereby allowing for a total of 16 single-ended inputs.

## 3 Methodology

This section describes the procedure for readings the DOT wristband signals. The DOT captures optical properties related to absorption and scattering, providing insights into tissue structure and function, with the former carrying the information of interest. The DOT method offers advantages in imaging deep tissue structures at shallow to medium depths. The positioning of Sensor-1 (S1), centrally on the palmar side of the hand near the nerve, with the wristband securely wrapped around the hand as shown in Figure 2(A), is the basis of the data collection apparatus. A 480-second data collection super-session was carried out, followed by the precise placement of the wristband 5 cm below the initial S1 location on the same hand to gather additional readings (in the future, the replication of the S1 optode array at multiple locations will obviate the need for this step). Given the proof of concept (POC) nature of this study[Yu05] [Hi02], the setup focuses on capturing data from two sensors (S1 and S2), despite the capability to capture data from four sensors. Consequently, two distinct files are obtained, with file-1 containing data from sensors S1 and S2 when the wristband is centered on the palmar side of the hand, and file-2 containing data from sensors S1 and S2 with the wristband positioned 5 cm below the previous hand location.

Our continuous-wave DOT specifically focuses on spatial patterns of absorption and scattering. The readings from the 16-bit ADC, Adafruit ADS1115, were converted to voltage values as follows:

$$Voltage = \frac{ADCReading * Full\ Scale\ Voltage}{(2^{16} * PGA)}$$

The frequency of anatomical DOT signals[HY16] are about zero given their permanence, however our sensors were capable of capturing higher-frequency biological signals [Xi19] [BS11], including heart rate (about 1 Hz), given that our ADC was operating at a sampling frequency of 128 Hz. The ADC device, configured with 2 channels and a programmable gain amplifier (PGA) setting of 2/3, facilitated the capture of signals within an absolute range of 0 to 43690. During the 480-second period ideally DOT should capture a total of

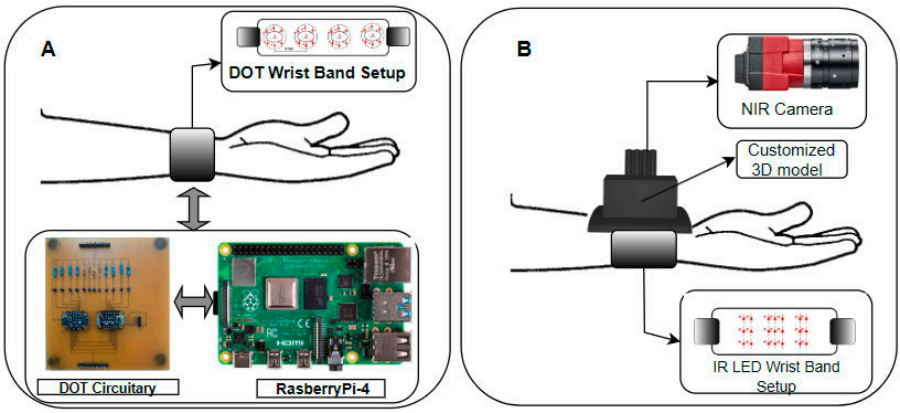



Fig. 2: (A) Setup to collect DOT signals and (B) Setup to collect IR vein images

30720 samples. However, factors such as the utilization of the Python spidev library and the limited processing speed of the Raspberry Pi as the master device introduced some overhead. As a result, approximately 25,000 readings, representing around 83.33 % of the original sensor data, were captured and saved. This data loss was considered negligible since the biological signals of interest (structural) exhibit significantly lower frequencies compared to the 128 Hz sampling rate. Section 4.2 details how this data was processed.

## 4 Data Collection and Evaluation

This section presents the methods used to evaluate the data collected with the experimental setup. The hardware captures DOT data and vein images. The data is then arranged as mated and unmated pairs and evaluated using a variety of metrics, including the false acceptance rate (FAR), the genuine acceptance rate (GAR), and the receiver operating characteristic (ROC) curve.

### 4.1 DOT and Vein Image Data Collection

This section shows samples of data that were collected. In Figure 3, images labeled A-1, A-2, A-3, and A-4, correspond to sensors S1, S2, S3, and S4, placed on different parts of the wrist. Images in 3A correspond to subject 1 and 3B correspond to subject 2. The placement

of the sensors is carefully considered to ensure accurate measurements, with Sensor S1 near the wrist providing potentially more precise readings than Sensor S4 positioned 5cm below the wrist on the dorsal side. Fig. 3 shows S1 - S4 readings for two subjects.

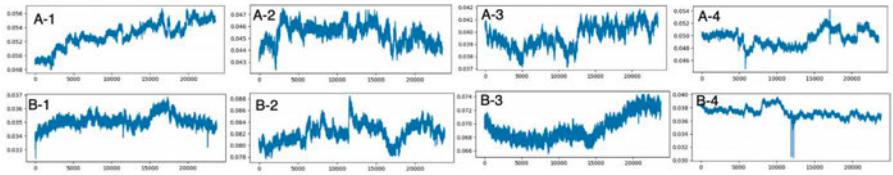



Fig. 3: DOT Readings A-x belongs to Subject 1, and B-x belongs to Subject 2, A-1, A-2, A-3, and A-4 belongs to sensors S1, S2, S3, and S4, respectively. S1 is placed on the palmar side near the wrist, S2 on the dorsal side near the wrist, S3 on the palmar side 5 cm below the wrist, and S4 on the dorsal side 5 cm below the wrist

The employed setup for IR imaging is depicted in Figure 2(B) and consists of a Customized IR camera hood and IR wristband. The IR Camera hood accommodates the NIR camera and is positioned 5 cm above the palmar side of the hand. The IR wristband comprises of three zones with a total of seven units interconnected in parallel, each unit containing 10 IR LEDs. These MTE8760N5 model IR LEDs operate at a wavelength of 870nm with a maximum power rating of 180mW. A regulated power supply maintains a consistent current of 50mA to the IR LEDs. When all three zones of the wristband are activated, IR light passes through the hand, making veins rich in deoxygenated hemoglobin visibly darker [Ji18] in the NIR image. The NIR camera used in this study is the Alvium 1800 U-501 NIR [Al], and the captured images are processed using the Vimba tool on a Windows PC. Adjustments to exposure and gain compensate for inter-subject variations and ensure consistent image quality, as shown in Figure 4. IR images labeled A and B demonstrate vein patterns under different configurations, including all LED zones, center zone only, sides only, and around the NIR camera.

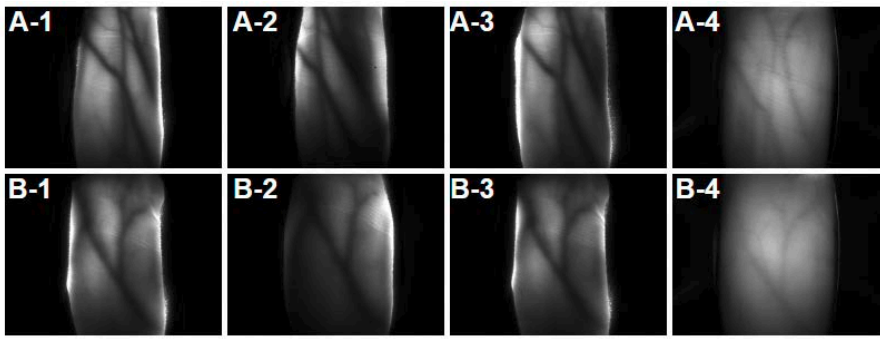



Fig. 4: IR images A-x belongs to Subject 1, and B-x belongs to subject 2, (A-1, B-1) - All LED zones on wristband turned on, (A-2, B-2) - Only Center zone LED's turned on, (A-3, B-3) Only LED's on either side turned on, (A-4, B-4) LED's around the NIR camera turned on.

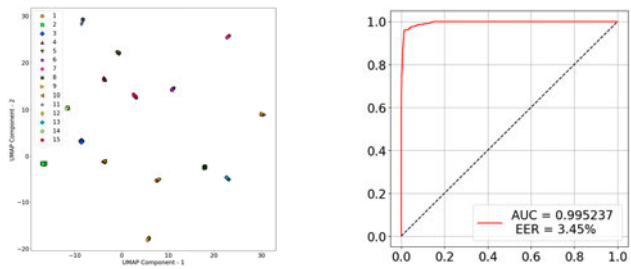
### 4.2 Data Analysis

This section delves into the computations performed to process the raw data. The discussion covers data preprocessing, analysis, and interpretation, providing a comprehensive account of the computational procedures adopted.

***Data Preprossessing:*** The data were acquired from 15 individuals with 9 trials per identity using the proposed DOT-vein scanner. The participants stayed still and breathed normally during the captures. Signals close to 1 Hz are primarily from the heartbeat. Thus Butterworth low pass filter with a 0.5Hz cut-off frequency was employed to filter out such non-structural DOT signals, followed by an outlier removal using the quantile techniques to ensure data quality. The DOT signal from each trial was segmented into multiple sub-signals and extracted the basic statistical features for each segment: the minimum, maximum, mean, and slope. This feature extraction reduced the time-dependent signals per channel to lower dimensions without losing the critical information that depends on the temporal signals. Table 1 shows the data preprocessing results. Feature extraction increased the feature space by approximately 12 folds, improving the subsequent comparisons and the machine learning algorithms.

| | Raw Data | | | Processed Data | | |
|---|---|---|---|---|---|---|
| Number of | Identities | Features | Trials | Identities | Features | Trials |
| | 15 | 4 | 9 | 15 | 48 | 9 |

Tab. 1: Summary of Raw and Processed Data




(a) UMAP for Visualizing DOT Data Separability

(b) ROC Curve for DOT Modality (all 9 trials)

Fig. 5: Visualizations of DOT data from 9 participants using tSNE and UMAP dimensionality reduction techniques.

***Data Separability:*** The visualization in Figure 5a illustrates the application of visualization dimensionality reduction techniques on the feature-extracted DOT Data with 12 features per channel for all individuals and reveals partial clustering of identities in UMAP [MHM20] spaces. This confirms the separation of identities by the DOT signal, but some overlaps exist. Thus other approaches, such as distance-based matchers over various higher-dimensional feature spaces and other machine learning-based learners, were used to classify the DOT data better.

***Similarity-based DOT Matching:*** This section presents the results of DOT verifications using cosine similarity as the similarity metric [Ib21] over 48 extracted features per identity. Figure 5b illustrates the Receiver Operating Characteristic (ROC) curve for 9 trials per individual, exhibiting an impressive AUC exceeding 0.99, indicating the accuracy of our DOT modality using a simple cosine similarity matcher, Whereas for the fusion 3 trials per each vein image were used to to be able to demonstrate the results with fusion.

### 4.3 Vein Image Matching

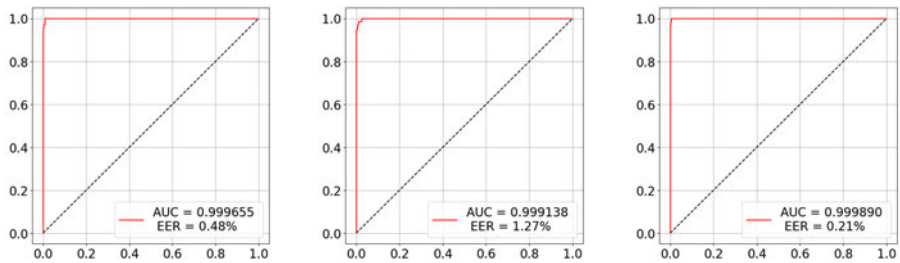
***Datasets Used:*** As we mentioned earlier, our experimental hardware can also capture wrist vein patterns using incident and IR illumination. To create the vein matchers to go with the DOT, we utilized the publicly available finger vein dataset, the SCUT FV Presentation Attack Database (SCUT FVD) [Qi18]. Permission was obtained to use the dataset. The SCUT FVD dataset consists of vein images from 100 participants, with six fingers per individual and six vein images per finger, resulting in a total of 600 unique fingers and 3600 captures. By incorporating this dataset in our experiments, we aimed to evaluate the effectiveness of our proposed method in a challenging cross-dataset scenario.

***Data Preprossessing and augmentation:*** The data pre-processing steps include random rotation up to 10 degrees), random horizontal flip, and color jittering. The latter involves making random modifications to the original image's brightness, contrast, saturation, and hue. This data augmentation step helps the model understand different representations of the data and increases the size of the pretraining dataset, enhancing the generalization capability of the model. A resizing step ensures that all input images have a uniform size of 224x224 pixels, matching the required input volume of our deep-learning models.

***Deep Feature Extraction:*** We fine-tuned a ResNet50 pre-trained on ImageNet using the SCUTFVD dataset of finger vein images. We used the cross-entropy loss function for fine-tuning. Features are extracted from the 'flatten' layer of the model. We obtained features of 2048 dimensionality. Since the proposed experiment focuses on vascular veins, this model is fine-tuned further using the data captured with our experimental setup (See Section 3.). We captured the forearm and wrist area vein images for 15 subjects, 6 images from each, for 90 images. This dataset was further divided into the training set and test set with 3 images per subject in each subset, i.e., 50% split for train and test. Only a subject-dependent test has been demonstrated as the dataset collected is insufficient to demonstrate subject-independent results. The model discussed earlier was fine-tuned using the training subset data for 10 epochs. The features were then matched using cosine similarity; the ROC curve presents the verification performance over the test set.

### 4.4 DOT-Vein Fusion

The scores from the vein image features and the DOT data were averaged to obtain a single fusion score for each trial per individual. Fig 6c represents the ROC curve of this




(a) Vein Image Modality (b) DOT Modality (3 trails only) (c) Fusion of both Modalities

Fig. 6: Receiver Operating Characteristic Curves for Biometric Data

fusion technique. Note that the results for vein image feature matching represent a subject-dependent protocol, where the identities (but not the images) used in training the models also appear for testing. But as the UTFVP dataset's experiment shows, with enough data, we can fine-tune the model to be able to extract quality features for subject-independent data as well.

## 5 Conclusion

Our pilot study shows that DOT of the forearm, as captured by our experimental wrist-worn scanner, has promise as a stand-alone biometric modality. We also showed that the secondary modality, wrist veins, can be fused with DOT to produce higher accuracy when compared to each modality by itself. The ROC AUC for the well-known vein image matching using deep features was 0.99965, yet a competitive 0. 99914 for the experimental DOT modality. The fusion of both modalities demonstrated an even better AUC of 0.99989. The vein image template matching technique exhibited a false acceptance rate of zero up to a genuine acceptance rate of 98.51%. Similarly, DOT data achieved a FAR of zero up to 74.81%. When both modalities were fused, the system achieved a zero FAR up to a GAR of 98.51%. This is a more desirable operating point, as operating at a low FAR is crucial for maintaining security and minimizing unauthorized access. We note that these results come from a very small pilot study, and thus a high observational variance must be considered. Our follow-up studies will be carried out using a larger data collection.

## 6 Acknowledgements

This work was sponsored by the Army Research Laboratory under Cooperative Agreement W911NF-21-2-0252. The views and conclusions contained are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes, notwithstanding any copyright notation herein. Dr. Derakhshani is also a consultant for Jumio.

## References

[Al] Alvium® 1800 USB Camera with High-Performance Sony Sensors: , Alvium 1800 U-501 NIR. Available online.

[BCH16] Barolet, Daniel; Christiaens, François; Hamblin, Michael R: Infrared and skin: Friend or foe. Journal of Photochemistry and Photobiology B: Biology, 155:78–85, 2016.

[BS11] Bagha, Sangeeta; Shaw, Laxmi: A real time analysis of PPG signal for measurement of SpO2 and pulse rate. International journal of computer applications, 36(11):45–50, 2011.

[Ch04] Christov, Ivaylo I: Real time electrocardiogram QRS detection using combined adaptive threshold. Biomedical engineering online, 3(1):1–9, 2004.

[Ch14] Chen, Chen; Tian, Fenghua; Liu, Hanli; Huang, Junzhou: Diffuse optical tomography enhanced by clustered sparsity for functional brain imaging. IEEE transactions on medical imaging, 33(12):2323–2331, 2014.

[Di05] Diamond, Solomon Gilbert; Huppert, Theodore J; Kolehmainen, Ville; Franceschini, Maria Angela; Kaipio, Jari P; Arridge, Simon R; Boas, David A: Physiological system identification with the Kalman filter in diffuse optical tomography. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005: 8th International Conference, Palm Springs, CA, USA, October 26-29, 2005, Proceedings, Part II 8. Springer, pp. 649–656, 2005.

[Du08] Ducros, Nicolas; da Silva, Anabela; Dinten, Jean-Marc; Peyrin, Francoise: Fluorescence diffuse optical tomography: A simulation-based study comparing time-resolved and continuous wave reconstructions performances. In: 2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro. pp. 388–391, 2008.

[Ga06] Galbally-Herrero, J; Fierrez-Aguilar, Julian; Rodriguez-Gonzalez, JD; Alonso-Fernandez, Fernando; Ortega-Garcia, Javier; Tapiador, Marino: On the vulnerability of fingerprint verification systems to fake fingerprints attacks. In: Proceedings 40th Annual 2006 International Carnahan Conference on Security Technology. IEEE, pp. 130–136, 2006.

[HB09] Hartung, Daniel; Busch, Christoph: Why vein recognition needs privacy protection. In: 2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing. IEEE, pp. 1090–1095, 2009.

[Hi02] Hielscher, AH; Bluestone, AY; Abdoulaev, GS; Klose, AD; Lasker, J; Stewart, M; Netz, U; Beuthan, dan J: Near-infrared diffuse optical tomography. Disease markers, 18(5-6):313–337, 2002.

[HY16] Hoshi, Yoko; Yamada, Yukio: Overview of diffuse optical tomography and its clinical applications. Journal of biomedical optics, 21(9):091312–091312, 2016.

[Ib21] Ibtehaz, Nabil; Chowdhury, Muhammad EH; Khandakar, Amith; Kiranyaz, Serkan; Rahman, M Sohel; Tahir, Anas; Qiblawey, Yazan; Rahman, Tawsifur: EDITH: ECG biometrics aided by deep learning for reliable individual authentication. IEEE Transactions on Emerging Topics in Computational Intelligence, 6(4):928–940, 2021.

[Ji18] Jiang, Huabei: Diffuse optical tomography: principles and applications. CRC press, 2018.

[Kr20] Krivokuca, Vedrana; Gomez-Barrero, Marta; Marcel, Sébastien; Rathgeb, Christian; Busch, Christoph: Towards Measuring the Amount of Discriminatory Information in Finger Vein Biometric Characteristics Using a Relative Entropy Estimator. Handbook of Vascular Biometrics, p. 507, 2020.

[MHM20] McInnes, Leland; Healy, John; Melville, James: , UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, 2020.

[Qi18] Qiu, Xinwei; Kang, Wenxiong; Tian, Senping; Jia, Wei; Huang, Zhixing: Finger Vein Presentation Attack Detection Using Total Variation Decomposition. IEEE Transactions on Information Forensics and Security, 13(2):465–477, 2018.

[TY13] Tilton, Catherine J; Young, Matthew: Standards for Biometric Data Protection. In: Security and Privacy in Biometrics, pp. 297–310. Springer, 2013.

[USJ20] Uliyan, Diaa M; Sadeghi, Somayeh; Jalab, Hamid A: Anti-spoofing method for fingerprint recognition using patch based deep learning machine. Engineering Science and Technology, an International Journal, 23(2):264–273, 2020.

[Xi19] Xiang, Jinxi; Dong, Yonggui; Xue, Xiaohui; Xiong, Hao: Electronics of a Wearable ECG With Level Crossing Sampling and Human Body Communication. IEEE Transactions on Biomedical Circuits and Systems, 13(1):68–79, 2019.

[Yu05] Yu, Guoqiang; Durduran, Turgut; Lech, Gwen; Zhou, Chao; Chance, Britton; Mohler III, Emile R; Yodh, Arjun G: Time-dependent blood flow and oxygenation in human skeletal muscles measured with noninvasive near-infrared diffuse optical spectroscopies. Journal of biomedical optics, 10(2):024027–024027, 2005.

## *GI-Edition Lecture Notes in Informatics*

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-314 Gesellschaft für Informatik e.V. (GI) (Hrsg.)
INFORMATIK 2021Computer Science & Sustainability
27. September– 01. Oktober 2021, Berlin

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
Online 8.–10. September 2021

P-317 M. Gandorfer, C. Hoffmann, N. El Benni, M. Cockburn, T. Anken, H. Floto (Hrsg.)
Informatik in der Land-, Forst- und Ernährungswirtschaft
Fokus: Künstliche Intelligenz in der Agrar- und Ernährungswirtschaft
Referate der 42. GIL-Jahrestagung
21.–22. Februar 2022 Agroscope, Tänikon, Ettenhausen, Schweiz

P-318 Andreas Helferich, Robert Henzel, Georg Herzwurm, Martin Mikusz (Hrsg.)
FACHTAGUNG SOFTWARE MANAGEMENT 2021
Fachtagung des GI-Fachausschusses Management der Anwendungsentwicklung und -wartung im Fachbereich Wirtschaftsinformatik (WI-MAW), Stuttgart, 2021

P-319 Zeynep Tuncer, Rüdiger Breitschwerdt, Helge Nuhn, Michael Fuchs, Vera Meister, Martin Wolf, Doris Weßels, Birte Malzahn (Hrsg.)
3. Wissenschaftsforum:
Digitale Transformation (WiFo21)
5. November 2021 Darmstadt, Germany

P-320 Lars Grunske, Janet Siegmund, Andreas Vogelsang (Hrsg.))
Software Engineering 2022
21.–25. Februar 2022, Berlin/Virtuell

P-321 Veronika Thurner, Barne Kleinen, Juliane Siegeris, Debora Weber-Wulff (Hrsg.)
Software Engineering im Unterricht der Hochschulen SEUH 2022
24.–25. Februar 2022, Berlin

P-322 Peter A. Henning, Michael Striewe, Matthias Wölfel (Hrsg.))
DELFI 2022 Die 20. Fachtagung Bildungstechnologien der Gesellschaft für Informatik e.V.
12.–14. September 2022, Karlsruhe

P-323 Christian Wressnegger, Delphine Reinhardt, Thomas Barber, Bernhard C. Witt, Daniel Arp, Zoltan Mann (Hrsg.)
Sicherheit 2022
Sicherheit, Schutz und Zuverlässigkeit
Beiträge der 11. Jahrestagung des Fachbereichs Sicherheit der Gesellschaft für Informatik e.V. (GI)
5.–8. April 2022, Karlsruhe

P-324 Matthias Riebisch,
Marina Tropmann-Frick (Hrsg.)
Modellierung 2022
Fachtagung vom 27. Juni - 01. July 2022,
Hamburg

P-325 Heiko Roßnagel,
Christian H. Schunck,
Sebastian Mödersheim (Hrsg.)
Open Identity Summit 2022
Fachtagung vom 07. - 08. July 2022,
Copenhagen

P-326 Daniel Demmler, Daniel Krupka, Hannes Federrath (Hrsg.)
INFORMATIK 2022
26.–30. September 2022
Hamburg

P-327 Masud Fazal-Baqaie, Oliver Linssen,
Alexander Volland, Enes Yigitbas,
Martin Engstler, Martin Bertram,
Axel Kalenborn (Hrsg.)
Projektmanagement und
Vorgehensmodelle 2022
Trier 2022

P-328 Volker Wohlgemuth, Stefan Naumann,
Hans-Knud Arndt, Grit Behrens,
Maximilian Höb (Editors)
Environmental Informatics 2022
26.–28. September 2022,
Hamburg, Germany

P-329 Arslan Brömme, Naser Damer,
Marta Gomez-Barrero, Kiran Raja,
Christian Rathgeb, Ana F. Sequeira,
Massimiliano Todisco, Andreas Uhl (Eds.)
BIOSIG 2022
14. - 16. September 2022,
International Conference

P-330 Informatik in der Land-, Forst- und Ernährungswirtschaft
Fokus: Resiliente Agri-Food-Systeme
Referate der 43. GIL-Jahrestagung
13.–14. Februar 2023Osnabrück

P-331 Birgitta König-Ries, Stefanie Scherzinger,
Wolfgang Lehner, Gottfried Vossen
(Hrsg.)
Datenbanksysteme für Business,
Technologie und Web (BTW 2023)
06.–10. März 2023, Dresden

P-332 Gregor Engels, Regina Hebig,
Matthias Tichy (Hrsg.)
Software Engineering 2023
20.–24. Februar 2023, Paderborn

P-333 Steffen Becker & Christian Gerth (Hrsg.)
SEUH 2023
23.–24. Februar 2023, Paderborn

P-334 Andreas Helferich, Dimitri Petrik,
Gero Strobel, Katharina Peine (Eds.)
1st International Conference on Software Product Management
Organized by „GI Fachgruppe Software Produktmanagement im Fachbereich Wirtschaftsinformatik (WI PrdM )“,
Frankfurt, 2023

P-335 Heiko Roßnagel, Christian H. Schunck,
Jochen Günther (Hrsg.)
Open Identity Summit 2023
15.–16. June 2023, Heilbronn

P-336 Lutz Hellmig, Martin Hennecke (Hrsg.)
Informatikunterricht zwischen
Aktualität und Zeitlosigkeit
20.-22. September 2023, Würzburg

P-338 René Röpke und Ulrik Schroeder (Hrsg.)
21. Fachtagung
Bildungstechnologien (DELFI)
11.-13. September 2023, Aachen

P-339 Naser Damer, Marta Gomez-Barrero,
Kiran Raja, Christian Rathgeb,
Ana F. Sequeira, Massimiliano Todisco,
Andreas Uhl (Eds.)
BIOSIG 2023
20.-22. September 2023, Darmstadt

P-340 Axel Kalenborn, Masud Fazal-Baqaie,
Oliver Linssen, Alexander Volland,
Enes Yigitbas, Martin Engstler,
Martin Bertram (Hrsg.)
Projektmanagement und
Vorgehensmodelle 2023
16. und 17. November 2023, Hagen

P-341 Gunnar Auth und Tim Pidun (Hrsg.)
6. Fachtagung Rechts- und
Verwaltungsinformatik (RVI 2023)
26.–27. Oktober 2023, Dresden

All volumes of Lecture Notes in Informatics can be found at
https://dl.gi.de/handle/20.500.12116/21.

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 informa-tion 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 978-3-88579-733-3

The proceedings of the BIOSIG 2023 include scientific contributions of the annual international conference of the Biometrics Special Interest Group (BIOSIG) of the Gesellschaft für Informatik (GI). The conference was held in Darmstadt on 20.-22. September 2023. 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.